

# **User Guide**

## **Data Science Workbench R-6.5**



## Contents

1.	Abc	ut This Guide	7
	1.1.	Document History	7
	1.2.	Overview	7
	1.3.	Target Audience	7
2.	Intr	oducing BDB Data Science Workbench	8
	2.1.	Introduction	8
	2.2.	Supported Web Browsers	8
3.	Get	ting Started with the Data Science Workbench	8
	3.1.	Accessing Data Science Workbench	8
4.	Ove	rview of the Data Science Workspace(s)	10
	4.1.	Tree-node Menu	10
	4.2.	Header Menu-Options	13
	4.3.	Tabbed Menu Strip - Options	19
5.	Dat	a Sources	25
	5.1.	CSV File	25
	5.2.	Data Service	28
	5.2.	1. Data Service with Conditions (Filters)	30
	5.3.	Cassandra Reader	36
	5.4.	Data Store Reader	38
	5.5.	Zip File	41
	5.6.	SFTP Reader	42
	5.7.	HDFS Reader	44
	5.8.	Excel File	45
	5.9.	Removing a Data Source from the Workspace	48
6.	Stat	istical Analysis	48
	6.1.	Hypothesis Testing	49
	6.2.	Correlation	57
7.	Dat	a Preparation	60
	7.1.	Data Type Definition	60
	7.2.	Filter	62
	7.2.	1. Column Filter	63



	7.2.2	2.	Row Filter	64
	7.3.	Miss	sing Value Replacement	66
	7.4.	Forr	nula	69
	7.5.	Nori	malization	70
	7.5.2	1.	Min-Max Normalization	70
	7.5.2	2.	Zero-Score	73
	7.5.3	3.	Decimal-Scaling	74
	7.6.	Sam	ple	77
	7.6.2	1.	Sampling Methods	77
	7.6.2	2.	Steps to Apply a Sampling Method	77
	7.6.3	3.	Result View for the Available Sampling Methods	78
	7.7.	Split	Data	82
	7.8.	Enco	oder	85
	7.9.	Outl	lier Detection	87
	7.9.2	1.	Interquartile Range	87
8.	Data	a Wri	ters	91
	8.1.	Data	a Store Writer	92
	8.2.	SFTF	P Writer	95
	8.3.	File	Writer	97
	8.3.2	1.	CSV Writer	97
	8.3.2	2.	JSON Writer	99
	8.3.3	3.	ZIP Writer	. 100
	8.4.	Data	abase Writer	. 102
	8.4.2	1.	Internal Data Writer	. 102
	8.4.2	2.	Cassandra Writer	.106
9.	Save	ed Wo	orkflows	.110
	9.1.	Ope	ning a Workflow	. 111
	9.2.	Dele	eting a Workflow	.111
	9.2.2	1.	Delete Connection in a Workflow	.112
	9.3.	Rena	aming a Workflow	.112
	9.4.	Auto	p-Save	.114
	9.5.	Shar	ring a Workflow	. 115
	9.6.	Publ	lish a Workflow as Service	.117
	9.7.	Pull	from VCS	.119



9.8. F	Push into VCS	
10. Sch	eduler	
10.1.	New Schedule	
10.1.1	Configuring General Tab	
10.1.2	2. Configuring Data Source	
10.1.3	8. Configuring a Data Writer	
10.1.4	. Scheduling a New job	
10.1.5	. Notification	
10.2.	Status	
10.2.1	Model Retraining in Scheduler	
11. Sav	ed Models	
11.1.1	Saving a Trained Model	
11.1.2	2. Importing a Model	
11.1.3	8. Reading a Saved Model	
12. Dee	ep Learning Workspace	
12.1.	Pre-Packaged Models	
12.2.	Working with Deep Learning Workspace	
12.2.1	Creating a New Model	
12.2.2	2. Data Preprocessing	
12.2.3	8. Running the NumPy Script(s)	
12.2.4	. Model Training	
12.2.5	. Model Data	
12.2.6	5. Tensor Board	
12.3.	Apply Model	
12.4.	Prediction using Trained Models	
13. R W	/orkspace	
13.1.	Algorithms	
13.1.1	. Clustering	
13.1.2	P. Forecasting	
13.1.3	8. Association	235
13.1.4	Regression Analysis	
13.1.5	. Classification	
13.1.6	5. Tree-Based Modeling	272
13.1.7	7. Anomaly Detection	
13.1.8	B. Dimensionality Reduction	



13.1.9.	Ensemble Models	
13.1.10	). Instance Based	
13.1.1	L. Regularization	
13.1.12	2. Structural Equation Modelling	
13.1.13	3. Automated ML	
13.1.14	ł. SVD	
13.1.1	5. WLS Relative Std	321
13.1.10	5. WLSR Input Weights	
13.1.1	7. BLR	324
13.1.18	3. Performance Metrics	325
13.2.	Apply Model	327
13.3.	Performance	
13.4.	Custom Scripts (R Scripts)	
13.4.1.	Creating a New Script	
13.4.2.	Saved Scripts	
14. Pyth	on Workspace	344
14.1.	Algorithms	
14.1.1.	Forecasting	
14.1.2.	Regression	351
14.1.3.	Classification	
14.1.4.	Tree-Based modeling	
14.1.5.	Anomaly Detection	
14.1.6.	Clustering	401
14.1.7.	Dimensionality Reduction	410
14.1.8.	Ensemble Models	416
14.1.9.	Instance-based Algorithm	433
14.1.10	). NLP	436
14.2.	Custom Scripts (Python Scripts)	453
14.2.1.	Creating a New Python Script	453
14.2.2.	Saved Python Scripts	457
14.3.	Jupyter Notebooks	464
15. Con	iguration	470
15.1.	Configuring Python Server	470
15.2.	Configuring R Server	472
15.3.	Configuring Spark Server	474



15.4	l.	Configuring Process Queue	477
16.	Libra	ary Management	478
17.	Sign	ing Out	480
17.1	L.	Forgot Password Option	481
17.2	2.	Force Login	482



## **1. About This Guide**

#### 1.1. Document History

The following table gives an overview of the most recent document updates:

Product Version	Date (Release Date)	Description	
Predictive Workbench 1.0	June 9 <sup>th</sup> , 2015	First Release of the document	
Predictive Workbench 2.0	Feb 18 <sup>th</sup> , 2016	Updated document	
Predictive Workbench 2.0	May 31 <sup>st</sup> , 2016	Modified document	
Predictive Workbench 2.5	November 9 <sup>th</sup> , 2016	Updated document	
Predictive Workbench 2.5.1	January 3 <sup>rd</sup> , 2017	Updated document	
Predictive Workbench 2.5.3	March 16 <sup>th</sup> , 2017	Updated document	
Predictive Workbench 3.0	August 31 <sup>st</sup> , 2017	Updated document	
Predictive Workbench 3.0	November 22 <sup>nd</sup> , 2017	Modified document	
Predictive Workbench 3.2	January 25 <sup>th,</sup> 2018	Updated document	
Predictive Workbench 3.5	April 15 <sup>th</sup> , 2018	Updated document	
Predictive Workbench 3.6	August 20 <sup>th</sup> , 2018	Updated document	
Predictive Workbench 3.7	October 10 <sup>th</sup> , 2018	Updated document	
Predictive Workbench 3.8	December 1 <sup>st</sup> , 2018	Updated document	
Predictive Workbench 4.0	December 31 <sup>st</sup> , 2018	Updated document	
Predictive Workbench 4.2	March 25 <sup>th</sup> , 2019	Updated document	
Predictive Workbench 4.3	April 24 <sup>th</sup> , 2019	Updated document	
Predictive Workbench 4.4	June 7 <sup>th</sup> , 2019	Updated document	
Data Science Workbench 4.5	August 5 <sup>th</sup> , 2019	Updated document	
Data Science Workbench 4.6	November 15 <sup>th</sup> , 2019	Updated document	
Data Science Workbench 5.0	February 17 <sup>th</sup> , 2020	Updated document	
Data Science Workbench 6.5	September 30 <sup>th</sup> , 2021	Updated document	

Note:

- a. The Predictive Workbench plugin is renamed as Data Science Workbench from R-4.5 onwards.
- b. The Spark ML and PySpaces are experimental workspaces so the detailed description of those workspaces is not included in the current document.

#### 1.2. Overview

This guide covers steps to:

- Access the BDB Data Science Workbench
- Server requirements and configuration details for the BDB Data Science Workbench
- Designer Part of the BDB Data Science Workbench
- Result or Analysis Part (Visualizing the analyzed data) of the BDB Data Science Workbench
- Creation and use of various Data Science Models

#### 1.3. Target Audience

This guide aims at business professionals, data analysts, data scientists, and statisticians who use BDB Data Science Workbench tool to conduct various experiments with data as in a Data Science Lab.



## 2. Introducing BDB Data Science Workbench

#### 2.1. Introduction

BDB Data Science Workbench provides the required environment for its users to create AI and ML models to empower their business insights. These Models can be used to envision the future outcomes of business processes based on past data. It is a user-friendly tool that shields users from mathematical complexity and offers an interactive graphical interface to provide a smooth, intuitive experience. It enables the users to discover hidden patterns in their data by Applying various statistical algorithms provided by the popular R statistical language, Spark ML, Python, and Deep Learning using Neural Network.

#### 2.2. Supported Web Browsers

The BDB Platform is a web browser-based application. The users can run the BDB Platform and its various plugins on the below-given versions of the browsers:

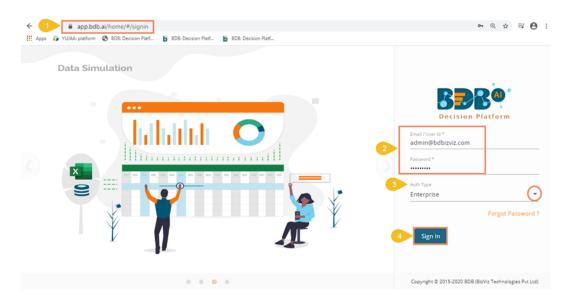
Mozilla Firefox/ Firefox ESR	Latest Version		
Microsoft Edge	Latest Version		
Apple Safari	10		
Google Chrome	Latest Version (recommended web browser)		

## **3. Getting Started with the Data Science Workbench**

#### **3.1.** Accessing Data Science Workbench

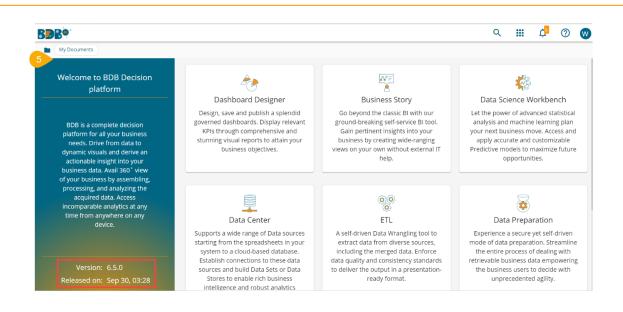
This section explains how to access the BDB Platform and the variety of plugins that it offers:

- i) Open BDB Enterprise Platform Link: https://app.bdb.ai
- ii) Enter your credentials.
- iii) Select an Auth Type from the drop-down menu.
- iv) Click the 'Sign In' option.



v) BDB Platform homepage opens.





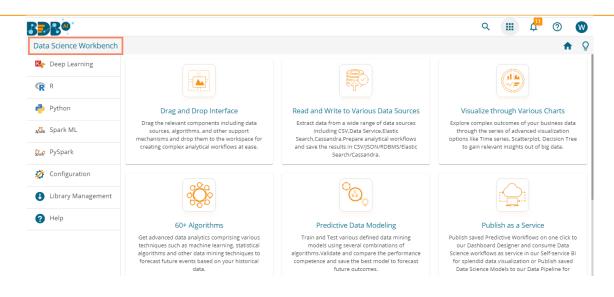
Note:

- a. The above screen opens only for those newly created users who have not yet created any document using the BDB Platform.
- b. If the user has created some documents previously, then the Platform homepage opens, displaying the '**My Documents**' page by default.
- vi) Click the 'Apps' 🛄 icon.
- vii) All the available plugin applications get displayed.
- viii) Select the 'Data Science Workbench' plugin.



- ix) The Data Science Workbench homepage opens.
- x) The major Data Science Workspaces get listed on this page.





This document aims to describe all the significant components and the related workflows in detail.

## 4. Overview of the Data Science Workspace(s)

This section describes all the options and icons provided on the landing page of the different Data Science Workspaces. The landing page of any selected Data Science Workflow contains the following Menus:

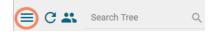
#### 4.1. Tree-node Menu

The Tree-node menu has all the available component connectors to run a data science execution. The components are provided in the hierarchical order via a tree structure menu. All the main categories are included as tree nodes, and sub-categories are committed as petals to the respective tree nodes. E.g. The following image displays the R Workspace landing page where '**Data Writer**' is the main category to which '**File Writer**' is committed as a subcategory and '**CSV Writer**,' and '**JSON Writer**' are displayed at the second level of the hierarchy.

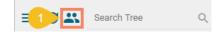
E C Search Tree Q
E Saved Workflows
> 📄 Data Source
▷ ∑ Statistical Analysis
> 🎲 Data Preparation
Algorithms
🧐 Apply Model
Reformance
4 🖀 Data Writer
🌸 Data Store Writer
File Writer
🔄 CSV Writer
🦾 Json Writer
Database Writer
Scustom Scripts
Scheduler
Saved Models



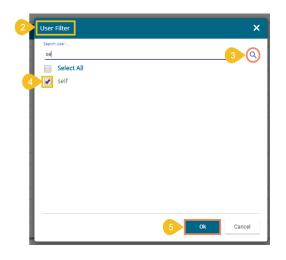
c. Click the '**Menu**'  $\equiv$  option Next to the '**Search**' box to collapse the tree structure menu from the homepage.



- d. The User Filter functionality is provided to restrict the display of the Workspace list to the other user of the same space.
  - 1) Click the 'User Filter' icon.



- 2) The User Filter window appears.
- 3) Search for the specific user.
- 4) Select the user(s) by a checkmark in the given box.
- 5) Click the '**Ok**' option.



- 6) The Workflow saved by the user gets displayed only to the selected user(s).
- e. Click the '**Search**' con to search across the entire tree-node menu.

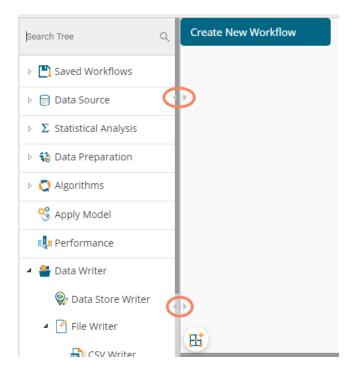
≡ Ca	csv	0
🔺 😝 Dai	ta Source	
-	CSV File	4
🔺 🏪 Dai	ta Writer	
⊿ 🎽	File Writer	
	🗐 CSV Writer	



f. Click on the icon to show or hide the gridlines on the workspace.

	Create New Workflow
1	
t	۶.
-	
-	
t	<i></i>
	E

g. The user can use these scrolling icons to increase or decrease horizontal space for the Tree Menu.



Note: This document is created focusing on each petal of the tree structure menu. All the available major and minor categories are described at length to understand a Predictive process.



#### 4.2. Header Menu-Options

- 1. **Run:** Click the '**Run**' icon to run the process and display the Result set view. This option can be applied to the data source, algorithms, and data preparation components.
- 2. **Refresh**: The **'Refresh'** C icon is provided the clear the cache memory and runs the component/ workflow.
- 3. **Create New Workflow:** Click the **'Create New Workflow**' **+** icon to clean the workspace removing the current component connectors.

The 'Create New Workflow' dialog box opens. Click the 'Yes' option to clean the workspace.

Create New workflow		×
Do you want to create new workflow ?		
	Yes	No

- 4. Clear Cache:
  - a. After using the '**Run**' option by default, the data gets cached in the server for the Next 10 minutes. For the latest Results, users need to rerun the workflow.
  - b. The user needs to click the 'Clear Cache' <sup>2</sup> option to remove the cached data before running the workflow (again).
  - c. If the user changes any component parameter which is to be applied to fetch the Result then, the 'Clear Cache'  $\mathcal{D}$  icon needs to be clicked.

If you get a message to clear cache to execute your process as shown in the following image:

	Component Co	nsole	Summary	Result	Visualization	Properties	<b>+</b> ↑	Ŧ
	10/07/2019 - 12:32:00	: CSV0 is s	started					
r	10/07/2019 - 12:32:01	: CSV0 is o	completed					
	10/07/2019 - 12:32:01	: Hypothe	sis Testing1 is s	tarted				
	10/07/2019 - 12:32:02 : Hypothesis Testing1 is completed							
	10/07/2019 - 12:39:32 execution. If you want	_				artial Cached is used	in	

Please follow the below-given steps to Clear Cache:

- i) Click the 'Clear Cache' 梦 icon from the header menu.
- ii) A message appears to confirm.
- iii) Click the '**Ok**' option.



Cache Confirm		×
Do you want to clear cache?		
	Ok	Cancel

iv) A message appears to confirm that the workspace-specific cache is cleaned. The below message appears for the R Workspace:



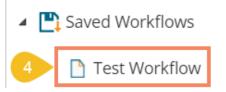
- 5. **Save:** Use the '**Save**' icon to save a created predictive workflow.
  - i) Create a workflow by connecting various configured components.
  - ii) Click the '**Save**' **b** icon from the landing page header menu.
  - iii) A new window appears to confirm the action.
    - a. Provide a Workflow Name.
    - b. Click the 'SAVE' option.

Save Workflow		×
Workflow Name		
Test Workflow		
	2 Save	Cancel
25	1.2	

iv) A success message appears.



v) The selected workflow gets saved to the list of Saved Workflows.



#### Copyright © 2015-21 BDB



- 6. Save As: Click the 'Save As' 🕍 icon to copy a data science workflow with the desired name.
  - i) Create a workflow by connecting various configured components.
  - ii) Click the 'Save As' 🕍 icon.
  - iii) A new window appears to confirm the task.
    - a. The Workflow Name contains the suffix '\_1' by default (If wished, users can also modify the name of workflow manually).
    - b. Click the 'Save' option.

Save Workflow		×
1 Workflow Name Test Workflow_1		
	2 Save	Cancel

- iv) A success message appears.
- v) The selected workflow gets saved by the new name in the 'Saved Workflows' list.

BBB@					
Data Science Workbench 🛚 👺	3		The workflow is s	aved.	
E C Search Tree Q	Component	Console	Summary	Result	V
Saved Workflows	Show 10 • er	ntries			
4 Test Workflow_1	SepalLength	Sep	alWidth	PetalLen	igth
Test Workflow	5.1	3.5		1.4	

- Parallel Processing: The user can enable or disable parallel processing by clicking the 'Parallel Processing' icon on the R landing page header. This option is only available for the R Workspace.
  - a. Click the 'Parallel Processing' 🔀 icon.



- b. The 'Parallel Processing Enabled' dialog box opens with a checkmark in the given box.
- c. Provide No. of CPU Cores in the given space.
- d. Click the 'Save' option.



2	Parall	el Processing Enabled		×
	0	Enable Parallel Processing 🗭	No. Of CPU Cores 3 2	
			4 Sav	ve

- e. The parallel processing gets enabled for the R Workspace.
- f. Click the 'Information' icon to get information about Parallel Processing.

			,	
Correla Parallel Processing	gDisabled			×
ws Enable Paralle	el Processing 🗖			
Parallel Processing will work only on :				Save
<ul><li>Data Type Definition</li><li>Filter</li></ul>				Save
<ul> <li>Formula</li> <li>R-Naive Bayes with validation</li> </ul>	Summary	Result	Visualization	Properties
valuation				

The Parallel Processing works only on Data Type Definition, Filter, Formula data Preparation components and R-Naïve Bayes (with Validation) algorithm.

Note: The user gets the Parallel Processing Disabled screen as given below:

ē	Parall	el Processing Disabled	×
	0	Enable Parallel Processing	
		Sa	ve

- 8. Version Control Panel: The user gets a dialog box to import Workflows, Custom Scripts, and Saved Models from Version Control Service (VCS) by clicking on the 'Version Control Panel' icon. This icon is available only for the Python Workspace.
  - a. Click the 'Version Control Panel' 💆 icon.





b. A dialog box opens displaying Workflows, Custom Scripts, and Saved Models categories to be imported.

e 2	Import entity from	vcs	×
	search		×
	Workflows		
3	Custom Scripts		
	Saved Models		

c. Select a Workflow/ Custom Script/ Saved Model and click the import icon.

Import entity from VCS	×
search	×
Workflows	
vcs_all Latest Version: v7	4
LTTS_Demo_1 Latest Version: v3	$\bigcirc$
vcs_workflow Latest Version: v3	$\bigcirc$
DP_DEmo Latest Version: v2	$\bigcirc$
Deploy_sanity Latest Version: v2	$\bigcirc$
Custom Scripts	
Saved Models	

- d. A success message appears.
- e. The selected entity (workflow/custom script/saved model) gets imported under the specific section.



BBB®				
Data Science Workbench    5   The selected entity is imported.				
≡ C 🚜 Search Tree Q	Create New Workflow			
◢ 🖺 Saved Workflows			1 -	
6 vcs_all	•••		·	
🎦 Zip_read_write_pdf	· · · · · · · · · · · · · · · · · · ·			

Note: If the user tries to import an existing workflow, script, or model then a warning message appears, and the selected entity does not get imported.

Data Science Workbench 🥐			Er	ror: Error	occurred	while sav	ing the in	nported e	entity data	a. Please t	ry again.	
E C 👪 Search Tree	Q	Create New Workflow										

- 9. Back: Click the 'Back' <i>icon to return on the Data Science homepage from any specific workspace.
- 10. **Full Screen/Full-Screen Exit:** Click the '**Full Screen**' C icon to display the selected Workspace on the full screen. The '**Full-Screen Exit**' icon appears to exit the full-screen view.

Data Science Workbench	A C + -→ B K + -→ Q
E C Search Tree C	Sample WF
Saved Workflows	CSV File Sample
Data Source	
👂 🎲 Data Preparation	
Pre Packaged Models	(H)
> ္ခ်ိန့် Models	Component Console Summary Result Visualization Properties 上 T
👂 蒬 Custom Scripts	
Model Training	20/8/2019 - 11:18:31 : Process added to Queue 20/08/2019 - 11:18:15 : CSV0 is started.
🤗 Apply Model	20/08/2019 - 11:18:19 : CSV0 is completed.
🖻 🏜 Data Writer	20/08/2019 - 11:18:19 : Sample1 is started.
	20/08/2019 - 11:18:20 : Samplet is completed. 20/08/2019 - 11:18:20 : Process Completed

Note: The user can also use the 'Esc' key to close the full-screen view.

11. Start Tour: Click the '**Start Tour**'  $\mathbf{Q}$  icon to begin the auto-guided tour for the selected workspace.



BBB@					<b>Д</b> <sup>11</sup> (?)	W
Data Science Workbench 🕵	Saved Workflows		► G +	- G 🖬	⊌ ← ::	Q
E C Search Tree Q	Access a list of all the saved and migrated NN workflows with the					
E Saved Workflows	credited options to view/edit, rename or delete along with the functionality	Sample	1			
Data Source	to deploy them to Designer.		¢			
👂 🎲 Data Preparation	« Prev Next » End tour					
Pre Packaged Models	Component Console Summ	ary Result N	Visualization	Properties	Ŧ	Ť

Note:

- a. Click the 'Next' option to proceed in the guided tour of the selected workspace.
- b. Click the '**Prev**' option to go back to the guided tour of the selected workspace.
- c. Click the 'End tour' option to end up the guided tour.

#### 4.3. Tabbed Menu Strip - Options

1. **Component:** The **'Component'** tab displays the required configuration fields for the dragged elements onto the workspace.

Component	Console	Summary	Result	Visualization	Properties	-	<u>L</u>	Ť
General	Configure (	CSV						
Properties	Select File		iris_1.csv					
Þ	Delimiter		ļ					
			s the format giv ould contain the o					
						Up	load	

Note: The component tab may display various sub-tabs as per the selected components onto the workspace.

- E.g., If the dragged data source is a CSV file, then the component tab displays General and Properties fields, while for a Cassandra Reader as a data source, the component tabs display General, Properties, and Column Selection.
- 2. **Console:** The **'Console'** tab displays the date and time for the entire process.
  - i) Click the **'Console'** option.
  - ii) The workflow process records (starting and ending time) get displayed:

	Component	sole Summa	ry Result	Visualization	Properties	<u>↓</u> 1	
	10/07/2019 - 12:32:00	: CSV0 is started					
	10/07/2019 - 12:32:01	: CSV0 is completed					
Þ	10/07/2019 - 12:32:01	: Hypothesis Testing	g1 is started				
	10/07/2019 - 12:32:02	: Hypothesis Testing	g1 is completed				

3. Summary: Click the 'Summary' tab to display the R and Spark Server overview of the process.



Component	Console Summary	Result	Visualization	Properties	<u>+</u> ⊥
s	ummary of the data				
Min. : 0 1st Qu.:1000 Median :1900 Mean :1838	Internal :18 Drive :15	Min. : 0.000 1st Qu.: 2.000 Median : 3.000 Mean : 3.969 3rd Qu.: 5.000	) Min. : 1.00		
(Other) :50 offered_ctc	previou BDB Fresher Cognizant Technology solut Accenture Solutions Pvt. I TCS CGI Information Systems	: 18 1 tions: 12 M .td : 8 M : 7 3 : 5 M :139 previous_ctc	lin. : 1.00 Ast Qu.: 56.75 Median :112.50 Mean :112.50 Ord Qu.:168.25 Max. :224.00 team		

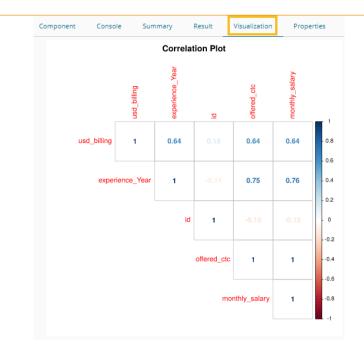
4. **Result:** Click the **'Result'** tab to display a Result list view based on the selected execution.

Component	Console	Summary		Visualization	Properties					<u>+</u> ⊥
						Search: Java+				
candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	previous_ctc	team	expyrsper_ctc	monthly_salary	cur_monthly_payment
3	Java+UI	Accenture Solutions Pvt. Ltd	з	1024000	18-07-1980	650000	BU 11	256000	85333	85333
8	Java+UI	HCL Technologies	8	845000	20-05-2018	650000	BU 11	281667	70417	0
127	Java+UI	UST global	127	900000	17-07-2017	600000	BU 11	333333	75000	62500
130	Java+UI	CGI Information Systems	130	750000	21-08-2017	0	BU 11	0	62500	0
131	Java+UI	Mphasis Ltd	131	750000	17-07-2017	450000	BU 11	277778	62500	54167
155	Java+UI	NTT Data	155	750000	21-08-2017	550000	BU 11	375000	62500	0
157	Java+UI	Navriti Technologies	157	550000	21-08-2017	400000	BU 11	275000	45833	45833
205	Java+UI	BDB	205	924000	01-12-2016	792000	BU 10	264000	77000	67500
206	Java+UI	BDB	206	864000	01-12-2016	702000	BU 10	172800	72000	52000
207	Java+UI	BDB	207	907200	01-12-2016	777600	BU 10	259200	75600	67500
						Previous	1 N	lext		

#### Note:

- a. The '**Result**' tab gets displayed for the given data only after data is configured and the '**Run**' option has been selected. Up to 50000 cells can be displayed in the Result view.
- b. The user can search for specific data using the 'Result' tab.
- 5. **Visualization:** Click the **'Visualization'** tab to display a graphical representation of the Result data. E.g., The following image displays a Correlation in the chosen data via the **'Correlation Plot'** chart.





6. **Properties:** Click the '**Properties**' tab to display properties for the current workflow on the Workspace.

Component	Console	Summary	Resu	lt	Visualization	Properties
Created By				Will		
Created At				2019	-08-20 11:06:57 +	0530
Last Modified B	у			Will		
Last Modified A	t			2019	-08-20 11:06:57 +	0530
Version				4.5.0		

7. **Data Insight**: Click the '**DataInsight**' tab from the Python workspace to display a detailed profiling report for the uploaded/processed data. The report opens displaying Overview, Variables, Correlations, Missing Values, and Sample sections.

Overview: It displays an overview of the uploaded dataset.

Compo	nent C	onsole	Summary	Result	Visualization	Properties	DataInsig	nt		+	Ŧ
Profili	ing Report					Overview	Variables	Correlations	Missing values	Sample	Î
•	Ove	rview									
	Data	set info	D			Variables	types				
	Number	of variables		21		Numeric			7		
	Number	of observatio	ns	224		Categorical			10		
	Missing	cells		148 (3.1%)		Boolean			0		
	Duplicate	rows		0 (0.0%)		Date			1		
	Total size	in memory		36.9 KIB		URL			0		
	Average	record size i	n memory	168.6 B		Text (Unique)			0		
						Rejected			3		
						Unsupported			0		-
										1	



Co	omponent	Console	Summary	Result	Visualization	Properties	DataInsight				*	Ŧ
P	Profiling Repo	ort					Overview	Variables	Correlations	Missing values	Sample	•
	Varia	bles										
	candidate Numeric	e_id		Distinct count Unique (%) Missing (%) Missing (n) Infinite (%) Infinite (n)	224 100.0% 0.0% 0 0.0% 0	Mean Minimum Maximum Zeros (%)	112.5 1 224 0.0%		o <sub>5</sub> 6	<sup>40</sup>	e details	
- A	cur_mont Numeric	hly_payment		Distinct count Unique (%) Missing (%) Missing (n) Infinite (%) Infinite (n)	73 32.6% 0.0% 0 0.0% 0	Mean Minimum Maximum Zeros (%)	55244.13839 0 700000 29.9%			ogge	e details	•
											ł	ŧ

#### Variables: All the variables get displayed via a column chart.

**Correlations:** It displays the correlational values via the Correlation Plot chart.



Missing Values: It displays the missing values through the column chart.

	Overview	Variables	Correlations	Missing values	Sample
Missing values					
Count Matrix Heatmap Dendrogram					
et at					
	224 179				
	134 89				
	44				
0.0 de de la de de la d	D				



F	irst rows	3							
	candidate_id	cur_monthly_payment	current_status	designation	expected_joining_date	experience_Year	expyrsper_ctc	gender	id
0	1	125000	Transferred	QA Manager	2018-07-02	15	120000	Male	1
1	2	125000	Resigned	QA Architect	2018-01-12	10	150000	Male	2
2	3	85333	Terminated	Senior Software Engineer	1980-07-18	4	258000	Male	3
3	4	52000	Transferred	QA Engineer	2018-03-18	5	130000	Female	4
4	5	43333	Transferred	QA Engineer	1972-04-15	3	208000	Male	5
5	6	0	Declined	Senior Software Engineer	2018-05-20	4	233333	Male	6
6	7	0	Absconded	AWS Consultant	2018-08-10	3	216667	Male	7
7	8	0	Declined	Senior Software Engineer	2018-05-20	3	281667	Male	8
8	9	0	Declined	QA Engineer	2017-02-20	2	260000	Male	9
	10	0	Declined	Business Analyst	2017-02-08	2	325000	Male	10

Note: The **'Download Report'** icon gets provided to download the entire DataInsight report.

8. **Status:** Click the '**Status**' tab to view the live job status of a running Spark job.

Search Tree Q	Co	omponent C	Console nove all jobs	Summary Resu	lt Visualization	n Prope	rties	Status		÷ -
🔄 Saved Workflows								Sea	rch:	
- 🎲 Data Preparation		Workflow Name	Run by	Start time	End Time	Status	View Log	Live job status	Summary	Actions
- 😋 Algorithms - 😳 Algorithms		save_feb5	user1A	5/Feb/2019- 14:55:50	5/Feb/2019- 14:55:53	success	۲	۲	۲	<b>■</b> 🛱
- a Performance		save_feb5	user1A	5/Feb/2019- 14:52:21	5/Feb/2019- 14:52:25	success	۲	٩	۲	<b>■ <sup>^</sup></b>
🆀 Data Writer 🕵 Custom Scripts		wtfinal	user1A	21/Nov/2018- 12:3:5	21/Nov/2018- 12:3:18	failed	۲	۲	۲	<b>■ <sup>†</sup></b>
are share	•••	sqlTr_P	user1A	21/Nov/2018- 11:46:58	21/Nov/2018- 11:47:4	success	۲	۲	۲	<b>■ <sup>†</sup></b>
🔛 Saved Models		sparkSpilitdata	user1A	11/Oct/2018- 12:39:38	NA	in progress	۲	۲	۲	<b>•</b>

Note: The Status tab appears when the user needs to check the live job status of a running job inside the Spark Workspace. The '**Status**' tab does not appear for other workspaces.

 Center-Top-Bottom icons: These icons have been provided on the tabbed Menu Strip to customize the workspace and view space as per the user requirement. The Default view of the Data Science Workspace canvass is as shown below:



Data Science Workbench	⊠ç•					•	G	+	Ð	•	¥.	←	53	Q
E C Search Tree Q	Create New Work	kflow												
Saved Workflows														
Data Source	F >													
👂 🎲 Data Preparation														
▶ 🎲 Pre Packaged Models														
> နွင့် Models														
Scustom Scripts														
Model Training														
强 Apply Model 🛛	(• )													
👂 🐣 Data Writer	Ħ													
	Component	Console	Summary	Result	Visualization	Properties							+	Ť

a. Click the **'Center'** icon to get equal space for the workspace and process view space.

Data Science Workbench	α <sub>¢</sub>		G   +	<i>•</i>	<b>⊠</b> ←	53	Q
≡ C ♣ Search Tree Q	Test Workflow						
Saved Workflows	Missing.						
Data Source							
🕨 🎲 Data Preparation	CSV						
Pre Packaged Models							
⊳ နွို Models	Component Console Summary Result Visualization Propertie	es				<u>+</u>	Ť
▶ 🕵 Custom Scripts	Created By Will						
Model Training	Created At 2019-08-20 11:06:57 +0530						
얥 Apply Model	Last Modified By Will						
🕨 🚢 Data Writer	Last Modified At 2019-08-20 11:06:57 +0530						
	Version 4.5.0						

b. Click the '**Top**' icon to maximize view space and minimize the workspace on the Predictive landing page.

Data Science Workbench	30					•	G	+	Ð		÷	Ω.	Q
≡ C ♣ Search Tree Q	Component	Console	Summary	Result	Visualization	Properties						+	Ŧ
▶ 🖺 Saved Workflows	General	Config	ure CSV										
🕨 😝 Data Source	Properties	Select F	lle	sam	pledata.csv								
👂 🎲 Data Preparation		Please	ensure csv file fol	llows the form	at given below:								
Pre Packaged Models					n the column headers. ce. It should be a single	word or two words	concate	nated	by underso	ore(_).			
⊳್ಳಿಸ್ಥೆ Models					any special characters.	-							
▶ 🕵 Custom Scripts			CSV header shoul		numerals. It should be le data type.	with at least one alpi	habet.						
Model Training		6) The	CSV header shoul	d not use single	or double quotes, dot,	brackets, and hyphe	ın.						
😚 Apply Model 🦂	Þ												
👂 🚢 Data Writer												Uploa	ad



Note: Click the '**Bottom**' icon to maximize the workspace as displayed in the default view of the Predictive Workspace landing page.

### **5. Data Sources**

Acquiring data from a data source is the initial step to move ahead in the Data Science Workbench. The 'Data Source' tree-node offers the following types of data source connectors:

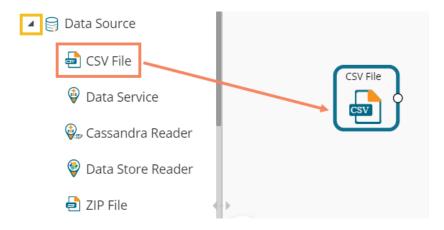
- a. CSV File
- b. Data Service
- c. Cassandra Reader
- d. Data Store Reader
- e. Zip File
- f. SFTP Reader
- g. HDFS Reader

The present section aims at describing the steps to get data from all the above-mentioned data sources.

Note: The Data Source list may differ based on the Workspace. The configuration steps as displayed below for a specific data source connector remains the same across the workspaces.

#### 5.1. CSV File

- i) Select and drag the 'CSV File' component onto the workspace.
- ii) Click the 'CSV File' component.



- iii) Configure the following fields for a data source:
  - a. Select File: Browse a CSV file.
  - b. Delimiter: Mention the delimiter used in the CSV file (it is a comma).
- iv) Click the 'Upload' option.



Component	Console	Summary	Result	Visualization	Properties		<del>+</del>	<u>+</u>
General	Configu	re CSV						
Properties	Select File		iris (1	).csv				
	Delimiter		,					
		sure csv file follo		t given below: the column headers.				
	2) The C	SV headers should	not have space	e. It should be a single	e word or two words concatenated by underscore(_	_).		
				ny special characters. E iumerals. It should be v	E.g- %, #, \$,@,*, etc. with at least one alphabet.			
	5) All rov	ws in a column sho	uld have singl	e data type.				
>	6) The C	SV header should i	not use single	or double quotes, dot,	brackets, and hyphen.			
						-	Upload	d

v) The user should get a success message, as highlighted in the image given below:

Component	Console	Summary	Result	Visualization	Properties	*	<u>+</u>
General	Configu	re CSV					
Properties	Select File		iris (1)	.CSV			
	Delimiter						
Þ	1) First r 2) The C 3) The C 4) The C 5) All rov	SV headers should SV headers should SV header should r vs in a column sho	hould contain not have space not contain an not have just nu uld have single	the column headers. e. It should be a single t y special characters. E. umerals. It should be w	th at least one alphabet.		
				Uplo	aded.		

- vi) Click the '**Run**' **>** or '**Refresh**' <sup>C</sup> icon.
- vii) Users will be redirected to the 'Console' tab to display the progress of the process.a. It first displays that the process has been initiated.

Component	Console	Summary	Result	Visualization	Properties	Ŧ	Ť
10/7/2019 - 15:	36:19 : Process	Initiated					

b. The completion of the process is marked with a green checkmark on the dragged component.



Create New Workflow						
) E	CSV	File				
Component Con	sole Summary	Result	Visualization	Properties	Ţ	Ť
10/07/2019 - 15:36:22	: CSV0 is started					
10/07/2019 - 15:36:23	: CSV0 is completed					

- viii) After the Console process gets completed, the uploaded data appears under the 'Result' tab.
- ix) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged data source component on the workspace.
  - **b.** Click the 'Result' tab.

			<u></u>	·	
now 10 🔻	entries			Search:	
Number	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
B	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

#### • Rules to be followed while uploading a CSV File

- 1. The first row provided in the CSV file should contain the column headers.
- 2. The second row of the CSV file should contain the data under all the headers without any 'null' or 'NA.'
- 3. CSV headers should not have space. It should be a single word or two words concatenated by an underscore (\_).
- 4. CSV headers should not contain any special characters. E.g. %, #, \$, @,\*, etc.
- 5. CSV headers should not contain single or double quotes, dots, brackets, and high-fen.
- 6. CSV headers should not contain merely numbers. Numerals should be used with at least one alphabet.
- 7. CSV header should not exceed 50 characters.
- 8. All rows in a column should have the same data type.

#### Note:

a. The supported file types are the '.csv' and '.tsv'

#### Copyright © 2015-21 BDB



- b. All the supported data sources get the 'General' tab to configure the following information for any tree-node component:
  - i. Component Name: A predefined name of the component is displayed in this field
  - ii. Alias: A predefined component name appears with the number to provide a record of its sequence in the workflow.
  - iii. Description (it is an optional field)(E.g. the following image displays the 'General' tab for a CSV data source.)

Click the	'Upload'	option	after	providing	the	required	details
CHER THE	opioau	option	ancer	providing	uic	required	actans

Component	Console	Summary	Result	Visualizatio	on	Properties	<u>+</u>	Ť
General	Basic							
Properties	Component N	ame	CSV					
	Alias		CSV0					
r	Description		Optional					
							Uploa	ad

#### 5.2. Data Service

- i) Select and drag the 'Data Service' component onto the workspace.
- ii) Click the **'Data Service'** component.



- iii) The **'Properties'** fields open for the Data Service data source connector under the **'Components'** tab.
- iv) Configure the 'Data Service Properties':
  - a. Select Data Connector: Select a data source from the drop-down menu
  - b. Select Data Service: Select a query service from the drop-down menu
  - c. Fields:
    - The following tables get displayed:
    - i. Column Header
    - ii. Data Type
  - d. Click the 'Next' option.



Component	Console	Summary	Result	Visualiza	ation	Properties	+ †	<u>+</u>	
General	Data Service	Properties							
Properties	Select Data Co	nnector	Connector_hiring	-	Ð				
Batch Query	Select Data Ser	rvice	Progress_data	•					
	Fields								
	Colur	nn Header		Data	a type				
	sourc	ce		strir	ng				
	num			long	5				
	ct			float	t				
>									
							N	lext	

- e. The Batch Query tab gets displayed.
- f. Click the 'Apply' option from the Batch Query page.

Component	Console	Summary	Result	Visualization	Properties	+ +	<u>+</u>
General	Column S	Selection		0			
Properties	Sele	ct Columns For	Select	•			
Batch Query	Bato	h Query					
	1				Generate		
•							
,							_
						Appl	у

g. A success message appears if the Apply is successful.

Component	Console	Summary	R	esult	Visualiza	tion	Properties	<u>↓</u> ↑	<u>+</u>
General	Column S	election				0			
Properties	Selec	t Columns For		Select		•			
Batch Query	Batc	h Query							
							Generate		
•									
				Арр	ly Successfu	IL			

- h. Click the 'Run' or 'Refresh' icon to start the Console process.
- i. The completion of the Console process gets marked with the Green checkmark on the data source connector.



			► C
Create New Workflow			
) E	Data Se		
Component Con	sole Summary	Result	Visuali
	: Data Service0 is starte : Data Service0 is compl		

j. The uploaded data from the data service data source can be seen under the 'Result' tab.

Component Console	Summary Result	VisualizationProperties $\frac{1}{7}$
Show 10 • entries		Search:
source	num	ct
Referral	63	28
CareerNet	58	26
BDB	34	15
Internal	18	8
Showing 1 to 4 of 4 entries		Previous 1 Next

Note: The Batch Query tab appears only for the R Workspace. The Properties tab for the Data Service connector in the Python Workspace appears with the Apply option if the Data Service does not contain any filter.

Data Science Workbench 🜪				▶ (	C   + 😔 🖬	<b>₹</b> +	0 Q
E C Search Tree	٩	Component	Console Summ	nary Result	Visualization	Properties	<u>+</u> <u>↓</u>
🔺 😝 Data Source		General	Data Service Propert	ies			
🖨 CSV File		Properties	Select Data	Test_Connector_Mys	ql 👻		
🏺 Data Service			Connector				
💡 Data Store Reader			Select Data Service	Mysql1	•		
🛃 ZIP File			Fields				
彈 SFTP Reader			Column Header		Data type string		
▹ ∑ Statistical Analysis		Þ	salary		float		
👂 🎲 Data Preparation							_
Algorithms							Apply

#### 5.2.1. Data Service with Conditions (Filters)



The Conditions tab appears for the Data Service that has filters. The Data Science Workbench supports Text, LOV, and Batch Query control types to configure the Conditions tab. The section aims to explain them in detail.

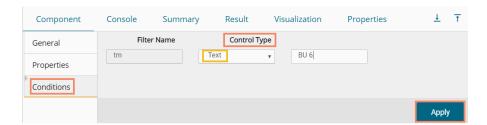
#### 5.2.1.1. Text Control Type

The filter value needs to be configured manually for the '**Text**' control type option. The user can enter multiple filter values separated by a comma.

- i) Drag a Data Service data source connector to the workspace canvas.
- ii) Choose a Data Connector and Data Service using the 'Properties' tab.
- iii) Click the 'Next' option.

Component	Console	Summar	y Result	Visual	ization	Properties	<del>*</del>	<u>+</u>
General	Data Service	Properties	5					
Properties	Select Data Co	nnector	HiringData DB	•	Ð			
	Select Data Ser	vice	team_det	•				
	Fields							
	Colur	nn Header		Data	a type			
	emp	name		stri	ng			
	gend	er		stri	ng			
	sourc	e		stri	ng			
>	refer	ral_of		stri	ng			
							Nex	t

- iv) The Conditions tab opens.
- v) Select the '**Text**' as a Control Type option.
- vi) Manually enter the Filter value.
- vii) Click the 'Apply' option.



- viii) Click the 'Run' or 'Refresh' icon to begin the Console process.
- ix) The completion of the process gets marked by a green checkmark on the top of the component.

				► C
Create New Worl	cflow			
,	- ·	Data Se.		
Component	Console	Summary	Result	Visuali
		Service0 is started Service0 is complet	ted	



Component	Console	Sum	mary Resu	ılt Visualiza	tion	Properties					+	-
Show 10 v entries Search:												
emp_name	gender	source	referral_of	designation	team	previous_organisation	skills	expected_joining_date	experience	monthly_salary	usd_billing	
Emp 145	Male	Referral	Ahamad	QA Engineer	BU 6	Omni globe Information Technology PVT. LTD	Selenium	2017-07-10	2	54167	1750	5
Emp 147	Female	Referral	Ahamad	Senior QA Engineer	BU 6	Test Mile Software Testing Pvt Ltd	Selenium	2017-07-24	3	58333	2000	-
Emp 148	Male	Referral	Ahamad	QA Engineer	BU 6	Test Mile Software Testing Pvt Ltd	Selenium	2017-07-24	3	50000	1750	
Emp 160	Female	Referral	Dhandapani	Lead Software Engineer	BU 6	Oracle	Selenium	2017-09-28	12	208333	4000	
Emp 163	Female	Referral	Ahamad	Senior QA Engineer	BU 6	Athenahealth	Selenium	2017-08-09	4	91667	2300	
Emp 167	Male	Referral	Tania	Senior QA Engineer	BU 6	Support.com	Selenium	2017-09-01	3	71667	2000	

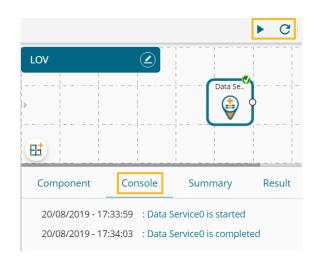
x) The filtered data for team BU 6 gets uploaded from the selected data service.

#### 5.2.1.2. LOV Control Type

- i) The Conditions tab opens.
- ii) Select 'LOV' as a Control Type option.
- iii) Select another Data Connector and Data Service from the lookup.
- iv) Select filter value
- v) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General		Filter Name		ol Type			
Properties	ID		LOV	Ŧ			
Conditions	Select Data	Connector	Mysql_pr	•			
	Select Data	Service	Distinct_filter	•			
	Select Filter	r(s)	2 checked	•			
•			<ul><li>✓ 15</li><li>✓ 60</li></ul>				
						Арр	ly

vi) Run the component to get data.



vii) The filtered data for the provided values gets uploaded.



ihow 10 🔻 entries		_		
Checking_account_status	Loan_Duration	Credit_History	Purpose_of_the_loan	Credit_Amount
>= 200 DM	15	existing credits paid back duly till now	business	2687
no checking account	60	existing credits paid back duly till now	car (new)	6527
<0 DM	15	no credits taken/all credits paid back duly	car (new)	950
no checking account	15	delay in paying off In the past	fumiture/equipment	960
no checking account	15	existing credits paid back duly till now	radio/television	3568
no checking account	15	critical account/other credits existing (not at this bank)	car (used)	3368
0 ←<200 DM	15	no credits taken/all credits paid back duly	car (new)	1778
no checking account	15	existing credits paid back duly till now	radio/television	1386
no checking account	15	delay in paying off in the past	radio/television	1478
no checking account	15	existing credits paid back duly till now	fumiture/equipment	2708

#### 5.2.1.3. Batch Query

- i) Drag and drop a Data Service connector to the Workspace canvas.
- ii) Configure the Properties tab.
- iii) Click the **'Next'** option.

Component	Console	Summary	Result	Visualization	Propert	ties	<del>*</del>	<u>+</u>
General	Data S	ervice Propertie	S					
Properties	Select D	ata Connector	bat	ch_query	-	G		
	Select D	ata Service	iris	batch	•			
	Fields							
		Column Header			Data ty	ype		
		row1			long			
		row2			double	le		
		row3			double	le		
		row4			double	le		
		row5			double	le		
>		rowб			string	F 9		
							Nex	ĸt



iv) If the selected data does not contain a filter, then while clicking the '**Next**' option, the Batch Query tab appears.

Component	Console	Summary	Result	Visualization	Properties	+ †	<u>+</u>
General	Column	Selection		0			
Properties	Sele	ect Columns For	Select	•			
Batch Query	Bato	ch Query					
				(	Generate		
						Арр	ly
Or							

The 'Conditions' tab opens (if the selected data service contains filter values). Select the 'Batch Query' option as the Control Type. Select the 'Next' option.

Component	Console	Summary	Result	Visualization	Properties	<u>1</u> T
General		Filter Name		Control Type		
Properties	catego	ory	Bate	ch Query	•	
Conditions						
						Next

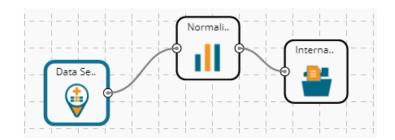
The 'Batch Query' tab appears.

- v) Select a column using the '**Columns**' drop-down menu.
- vi) Click the 'Generate' option to generate a batch query.
- vii) Click the 'Apply' option after configuring the '**Conditions**' tab.

Component	Console	Summary	Result	Visualization	Properties	÷ 1
General	Col	umn Mapping			0	
Properties		Filter Names		Columns		
Conditions		category		row6	•	
		SELECT DIST	INCT row6 FR	OM iric in	2 Generate	
Batch Query		SELECT DIST	INCT IOWOT N			
						3 Apply
						3 Apply



- viii) The 'Apply Successful' message appears.
- ix) To see batch-wise completion of the process under the 'Console' tab connect the Data Service component to a data writer. E.g., the following image displays the normalization and internal data writer connectors connected to the Data Service component.



- x) Configure the component and run the workflow.
- xi) Open the batch-wise completion of the process that can be seen under the 'Console' tab.

	Component	Console	Summary	Result	Visualizatior	n Properties		
	10/07/2019 - 14:52	2:13 : Batch	Process started					
Þ	10/07/2019 - 14:52	2:14 : Data	Service0 is started	l for setosa				
	10/07/2019 - 14:52:15 : Data Service0 is started for versicolor							
	10/07/2019 - 14:52	2:16 : Data	Service0 is started	l for virginica				
	10/07/2019 - 14:52	2:22 : Num	per of Rows fetche	ed: 50 for virgir	nica			
	10/07/2019 - 14:52	2:22 : Num	per of Rows fetche	ed: 50 for setos	a			
	10/07/2019 - 14:52	2:22 : Num	per of Rows fetche	ed: 50 for versi	color			
	10/07/2019 - 14:52	2:22 : Data	Service0 is comple	eted for virginio	ta			
	10/07/2019 - 14:52	2:22 : Data	Service0 is comple	eted for setosa				
	10/07/2019 - 14:52	2:22 : Data	Service0 is comple	eted for versico	olor			
Þ	10/07/2019 - 14:52	2:22 : Norm	alization1 is starte	ed for virginica				
	10/07/2019 - 14:52	2:22 : Norm	alization1 is starte	ed for setosa				
	10/07/2019 - 14:52	2:22 : Norm	alization1 is starte	ed for versicolo	or			
	10/07/2019 - 14:52	2:24 : Norm	alization1 is comp	oleted for virgin	nica			

#### Note:

- a. The Result tab displays no data in the case of the Batch Query option in the R workspace.
- b. The Batch Query option is available only for the **R** and **Python** Workspaces.
- c. The user can develop a data service via the Data Management module of the BDB Platform.
- d. **'Fields'** option under the '**Properties**' tab appears only after selecting the appropriate query service.
- e. LOV service provided under the '**Conditions**' tab can contain only one column, in case of more than one column, a warning message appears.
- f. The user can configure the following information for a data service data source via the 'General' tab:
  - i. Alias Name
  - ii. Description (it is an optional field)



#### • Rules to be Followed while Creating a Data Service

- 1. The data service header should not have space. It should be a single word or two words concatenated by an underscore (\_).
- 2. The data service header should not contain any special characters. E.g. %, #, \$, @,\*, etc.
- 3. The data service header should not contain single or double quotes, dots, brackets, and high-fen.
- 4. The data service header should not contain merely numbers. Numerals should be used with at least one alphabet.
- 5. The data service header should not exceed 50 characters.

#### 5.3. Cassandra Reader

The Cassandra Reader data source connector is provided for R and Spark ML workspaces.

- i) Select and drag the 'Cassandra Reader' component onto the workspace.
- ii) Click on the dragged 'Cassandra Reader' component.



- iii) Users will be redirected to the 'Properties' tab of the component.
- iv) Configure the required properties:
  - a. Select Data Connector: Select a data connector using the drop-down menu
  - b. Host Name: Data connector specific hostname will be displayed
  - c. Port Number: Port number will be displayed
  - d. User Name: Username gets displayed
  - e. Password: Enter the password
  - f. Cluster Name: Enter a cluster name
  - g. Select Key Space: Select a keyspace from the drop-down menu
  - h. Select Table: Select a table from the drop-down menu
  - i. Limit No. of the row to fetch: Select an option using the drop-down menu. Two options are provided, as shown below:
    - 1. Select all Rows
    - 2. Limit By



- j. Max. No. of Rows to be fetched: Enter a number to decide maximum fetched rows. (This option appears only if the 'Limit By' option has been selected using the 'Limit by Row' field. The default value for this field is 1000).
- v) Click the 'Next' option.

Component	Console	Summary	Result	Visualization	Properties	Status	*	<u>1</u>
General	Data Serv	ice Properties						
Properties	Select Data	Connector	Cassandra	-				
Column Selection	Host Name							
	Port Numbe	er.						
	Username		cassandra					
	Password							
	Cluster Nam	ne	Cluster					
	Select Key S	pace	ра	-				
	Select Table		Prod_table_ne	w1 -				
	Limit No: of	rows to fetch	Limit by	٣				
>	Max no: of r	ows to be	100					
	fetched							
							Next	

- vi) Users will be redirected to the 'Column Selection' tab.
- vii) Select the required columns from the list.
- viii) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	Status	+	<u>+</u>
General	Meta Data							
Properties	Headers		Туре	s	pecify			
	UID		TIMEUUID					
Column Selection	Attrition		TEXT					
	Skills_Group	bing	TEXT					
	current_sta	tus	TEXT					
	designation		TEXT					
	experience		FLOAT					
	gender		TEXT					
	monthly_sa	lary	FLOAT					
	source		TEXT					
	team		TEXT					
•								
							Appl	у

ix) Run the component process for fetching data clearing the Cache.



	►	C
Cassand.		

x) The 'Console' tab opens to display the progress of the process. The completion of the Console process is marked through the green checkmark on the top of the component.

Create New Workflow		-111 1 1 1 1 1 1	1     -     -     1     -     1     -     1     -       1     1     1     1     1     1     1				
							1
		<u> </u>	-1 = -1 = -1 = -1 = -1 = -1 = -1 = -1 =				
iiiiiiiiiiiii		Cassand					
		[ 🛛 🖓 J					
Component Console	Summary	Result	Visualization	Properties	Status	Ť	Ť
10/7/2019 - 16:23:47 : 10/07/	/2019 - 16:23:37 : ca	ssandra0 is st	arted				
10/7/2019 - 16:23:47 : <b>10/07</b> /	2019 - 16:23:43 : ca	ssandra0 is co	mpleted				

- xi) After the Console process gets completed, users can view the Result data using the 'Result' tab.
- xii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged data source component on the workspace.
  - **b.** Click the 'Result' tab.

Component	Console Sum	nmary Result	Visualization Properti	es Status				<u>+</u> ↑
Show 10	▼ entries						Search:	
Attrition	Skills_Grouping	current_status	designation	experience	gender	monthly_salary	source	team
No	QA	joined	senior qa engineer	4	female	42392.9	referral	bu 2 qa
No	UI and Java Developer	joined	associate software engineer	1.3	male	30896	drive	bu 1 engineering
No	ВІ	joined	senior software engineer	4.3	male	33582.5	portal	bu 1 ps
No	UI and Java Developer	joined	sr.ui developer	3.3	male	38783.56	agency	bu 1 ps
No	DEVOPS	joined	senior software engineer	3.4	male	41471.15	referral	bu 2 engineering
No	QA	absconded	qa engineer	0	female	18581.32	drive	bu 1 qa
No	UI and Java Developer	joined	senior software engineer	3.5	female	45610.9	referral	bu 1 engineering
Yes	UI and Java Developer	resigned	senior software engineer	3.11	male	33230	portal	bu 2 ps
No	ВІ	joined	associate software engineer	1	male	28261.4	drive	bu 2 ps
Yes	BI	resigned	senior software engineer	4	male	32308.33	portal	bu 2 ps

## 5.4. Data Store Reader

i) Select and drag the 'Data Store Reader' component onto the workspace.



#### ii) Click the 'Data Store Reader' component.



- iii) Users will be redirected to the 'Properties' tab of the component.
- iv) Configure the required properties:
  - a. Select Data Store: Select a data store using the drop-down menu.
  - b. Limit No. of Documents to Fetch: Select an option using the drop-down menu. Two options are provided, as shown below:
    - 1. Fetch all Documents
    - 2. Limit By
  - Max. No. of Documents to be Fetched: Enter a number to decide maximum fetched documents (This option appears when the 'Limit By' option has been selected using the 'Limit No. of Documents to Fetch' field. Users can select any positive integer value).
- v) Click the 'Next' option.

Component	Console Summary	Result Visi	ualization Pi	roperties	<u>+</u> ↑	<u>+</u>
General	Data Store Reader Prop	erties				
Properties	Select Data Store	Hiring_Data	•	Ð		
Column Filter	Limit No. of Documents	Limit By	٣			
	to Fetch					
	Max No. of Documents	9	\$			
	to be Fetched					
	Fields					
	Column Header	Field Definitio	n Data	a type		
>	name	dimension	strir	ng		
r	gender	dimension	strir	ng		
	source	dimension	strir	ıg		
					Next	

- vi) The 'Column Filter' tab opens.
- vii) Select the required columns from the drop-down list.
- viii) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Column Sele	ection					
Properties	8 checked		-				
Column Filter							
						Apply	

- ix) Run the component by clearing the previous cache to get Data.
- x) The 'Console' tab opens to display the progress of the process. The completion of the Console Process is marked by a green checkmark on the top of the dragged Datastore component.



- xi) The user can view the Result data using the 'Result' tab.
- xii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged data source component on the workspace.
  - b. Click the 'Result' tab.

ow 10 •	entries				S	earch:	
name	gender	source	designation	team	skills	id	salary
Emp ID 112	male	portal	senior software engineer	bu 1 engineering	Pl/Sql developer	45	38585.33
Emp ID 112	male	portal	senior software engineer	bu 1 engineering	Pl/Sql developer	1560	38585.33
Emp ID 112	male	portal	senior software engineer	bu 1 engineering	Pl/Sql developer	3401	38585.33
Emp ID 112	male	portal	senior software engineer	bu 1 engineering	Pl/Sql developer	3894	38585.33
Emp ID 112	male	portal	senior software engineer	bu 1 engineering	Pl/Sql developer	5734	48231.67
Emp ID 112	male	portal	senior software engineer	bu 1 engineering	Pl/Sql developer	7110	48231.67
Emp ID 112	male	portal	senior software engineer	bu 1 engineering	Pl/Sql developer	8357	48231.67
Emp ID 112	male	portal	senior software engineer	bu 1 engineering	Pl/Sql developer	9111	48231.67
Emp ID 113	male	portal	senior software engineer	bu 2 engineering	Java	46	25718.58

Note: Empty values present in any row of the numeric column get replaced with zero (0) while reading data from a data store reader.



# 5.5. Zip File

- i) Select and drag the 'Zip File' component onto the workspace.
- ii) Click the **'ZIP File'** component.



- iii) The Properties tab opens for the Zip File.
- iv) Browse a Zip file.
- v) Click the **'Extract'** option.

Component	Console	Summary	Result	Visualization	Properties	± Ť
General	Upload ZIP					
Properties	Select File		Browse		This is a required field	
þ.						Extract

vi) After extracting data, the following message appears.

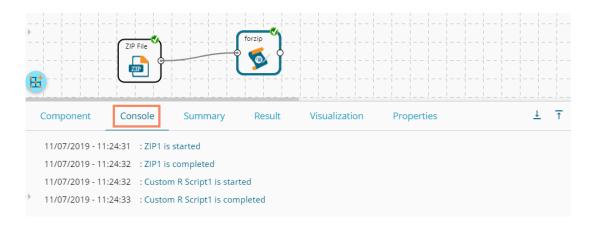
Component	Console	Summary	Result	Visualization	Properties		<u>+</u>	Ť
General	Uploa	d ZIP						
Properties	Select F	ile	merg	ge.zip				
<u>&gt;</u>	The file	has been extracted	l in a secure loci	ation and can be acces	sed by using variable ZIP_FILE_L	OCATION 👁		
				Extract	ed Successfully			

- vii) Connect the dragged ZIP file to a script component to read the extracted data from the ZIP file.
- viii) Click the 'Apply' option for the Custom Group tab of the script component.

	ZIP File		forzip						
Component	Console	Summary	Result	Visualizatio	n Pro	perties		<u>+</u>	Ť
General	Dynam	nic Fields							
Custom Group									
•									
								Apply	



- ix) After getting the 'Apply Successful' message, click the 'Run' or 'Refresh' icon to get started with the process.
- x) The progress of the process appears under the 'Console' tab, and the completion of the process gets marked by the green checkmarks on the top of the dragged components.



xi) After the successful completion of the Console process, open the 'Result' tab to view the Result data.

Component Console	Summary	Result	Visualization	Properties	5	+	<u>+</u>
Show 10 • entries			Searc	h:			
country	democracy	y_score					
USA	19						
Albania	16						
Turkey	16						
China	5						
Sudan	10						
Showing 1 to 5 of 5 entries				Previous	1	Ne	xt

Note:

- a. In the R workspace, the ZIP file can have files with the following extensions-.csv, .xlsx, and .json
- b. The ZIP file will have the following properties:
  - 1) Extensions supported for ZIP will be ".zip",".tar",".rar",".7z", "tar.gz".
  - 2) The ZIP file data source should only get connected to the Custom Scripts. If connected to any other component, an error should occur, saying, "Cannot be connected. Connect to Custom Scripts".
  - 3) After uploading a ZIP file, the contents of the ZIP file get shown in UI after decompressing it.
  - 4) Within the script, the files in the zip can be accessed with the drive location.

#### 5.6. SFTP Reader

The SFTP reader is provided to handle enormous data for the Python Workspace. The SFTP reader can read data from any file extension using a relevant Script.

i) Select and drag the 'SFTP Reader' component onto the workspace.



- ii) Click the 'SFTP Reader' component.
- iii) The Properties tab opens for the SFTP data source connector.
- iv) Configure the required details:
  - a. Host address
  - b. Port number
  - c. Username
  - d. Password
  - e. Remote Path
  - f. Click the 'Connect' option. It should return a notification that it is successfully connected.
  - g. The user can select a file with a double click from the available options. The selected file path gets mentioned.
- v) Click the 'Apply' option.

Component	Console Summary	Result	Visualization	Properties	<del>*</del>	<u>+</u>
General	Configure SFTP					
Properties	Host			0		
	Port					
	User name	ftpuser				
	Password					
	Remote path	/home/ftpuse	r			
2	Selected File Path : /home/	ftpuser/	Connect	Successfully Connected		
					Арр	ly

- vi) A success message should appear, stating that the data source has been applied.
- vii) Run the component process to get data.
- viii) Completion of the 'Console' process gets marked by a green checkmark on the top of the dragged SFTP reader component.



			► C
SFTP DS			
Component	Console	Summary	Result
21/8/2019 - 12:57	7:35 : Process	added to Queue	
21/08/2019 - 12:5	55:24 : SFTP R	Reader0 is started.	
> 21/08/2019 - 12:5	57:23 : SFTP R	Reader0 is complet	ted.
21/08/2019 - 12:5	57:23 : Proces	ss Completed	

ix) The fetched data appears under the 'Result' tab.

Componer	nt Conso	ole Summa	ry Result	Visualizat	tion Propertie	25				<u>+</u> ⊥
Show 10	ow 10 • entries Search:									
play_id	game_id	home_team	away_team	posteam	posteam_type	defteam	side_of_field	yardline_100	game_date	quar
46	2009091000	PIT	TEN	PIT	home	TEN	TEN	30.0	2009-09-10	900
68	2009091000	PIT	TEN	PIT	home	TEN	PIT	58.0	2009-09-10	893
92	2009091000	PIT	TEN	PIT	home	TEN	PIT	53.0	2009-09-10	856

## 5.7. HDFS Reader

The HDFS Reader is provided for the PySpark workspace. The HDFS reader loads distributed data (in batches) and supports only CSV extension.

- i) Select and drag the 'HDFS Reader' component onto the workspace.
- ii) Click the **'HDFS Reader'** component.

🔺 🎯 Data Source	e	HDFS Re
🔊 HDFS R	eader	

iii) The Properties tab opens for the HDFS data source connector.



- iv) Configure the required fields.
- v) Click the 'Connect' option. It should return a success message that is successfully connected.
- vi) The user can select a file with a double click from the available options. The selected file path gets mentioned.
- vii) Click the '**Apply**' option.
- viii) Run the workflow after getting the success message.
- ix) The progress of the process appears under the '**Console**' tab. The completion of the process gets marked by a green checkmark on the top of the dragged HDFS reader component.

				►	C + 5
ŀ	HDFS_DS				
•			HD	FS Re.	
E	Ħ			_	
	Component	Console	Summary	Result	Visualization
	21/8/2019 - 15:49	9:33 : Process	added to Queue		
	21/08/2019 - 15:4	49:17 : HDFS I	Reader0 is started	ł.	
•	21/08/2019 - 15:4	49:22 : HDFS I	Reader0 is comple	eted.	
	21/08/2019 - 15:4	49:22 : Proces	s Completed		

x) The fetched data appears under the 'Result' tab.

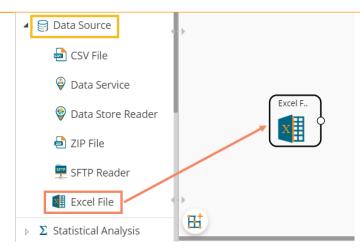
Component	Console	Summar	ry Res	ult Visualiz	ation	Properties		$\frac{4}{7}$ $\frac{1}{2}$	
Show 10 • entries Search:									
candidate_id	name	gender	source	designation	team	previous_organisation	skills	expected_joining_dat	
11	Renuka	Female	Orgspire	Senior QA Engineer	BU 6	Thomson Reuters	Selenium	2017-02-05 18:30:00	
12	Arunjunai Sathis R	Male	Orgspire	QA Engineer	BU 6	Accenture Solutions Pvt. Ltd	Selenium	2017-03-23 18:30:00	
13	Jalavathi Batchu	Female	CareerNet	Senior QA Engineer	BU 6	Harman Connected Services	Selenium	2017-03-07 18:30:00	
14	lanaki	Female	Referral	OA Architect	BU 6	Support.com	Selenium	2017-02-12 18:30:00	

## 5.8. Excel File

The Excel File reader is provided in the **Python Workspace** to handle minutes to large data from your spreadsheets and make it analytics-ready.

- i) Select and drag the 'CSV File' component onto the workspace.
- ii) Click the **'Excel File'** component.





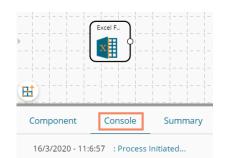
- iii) Configure the following fields for a data source:
  - a. Select File: Browse an Excel file.
  - b. Sheet Name: Provide the sheet name.
- iv) Click the **'Upload'** option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>⊥</u> <u>⊺</u>
General	Config	ure Excel					
Properties	Select Fi	le	Hiring	g Data.xlsx			
•	Sheet N	ame	Hiring	Data			
							Upload

v) The user should get a success message, as highlighted in the image given below:

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General	Configure	Excel						
Properties	Select File		Hiring Data.xl	lsx				
•	Sheet Name	2	0					
				Uploaded.				

- vi) Click the '**Run**' **>** or '**Refresh**' <sup>C</sup> icon.
- vii) The users will be redirected to the '**Console**' tab to display the progress of the process. a. It first displays that the process has been initiated.





- Excel
   Summary

   16/3/2020 11:7:2
   : Process added to Queue

   16/03/2020 11:06:49
   : Excel0 is started.

   16/03/2020 11:06:50
   : Excel0 is completed.

   16/03/2020 11:06:50
   : Process Completed.
- b. The completion of the process is marked with a green checkmark on the dragged component.

- viii) After the Console process gets completed, the uploaded data appears under the 'Result' tab.
- ix) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged data source component on the workspace.
  - **b.** Click the '**Result**' tab.

10W 10 ¥	entries											Search:	
usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	previous_ctc	team	expyrsper_ctc	month
4000	Male	Indeed	15	1	Management, Selenium	Athenahealth	1	1800000	2018-07-02 00:00:00	2000000	BU 6	120000	150000
4000	Male	Orgspire	10	2	Selenium	Support.com	2	1500000	2018-01-12 00:00:00	2000000	BU 6	150000	125000
2600	Male	Orgspire	4	з	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000	1980-07-18 00:00:00	650000	BU 11	256000	85333
2300	Female	Referral	5	4	Selenium	Inventateq	4	650000	2018-03-18 00:00:00	580000	BU 6	130000	54167
1750	Male	Referral	3	5	Selenium	Tekinspy	5	520000	1972-04-15 00:00:00	500000	BU 6	208000	43333
D	Male	BM5 Innolabs	4	6	java	CGI Information Systems	6	980000	2018-05-20 00:00:00	730000	BU 7	233333	81667
D	Male	Orgspire	3	7	AWS	Cognizant Technology solutions	7	650000	2018-06-10 00:00:00	510000	BU 7	216667	54167
D	Male	BMS Innolabs	3	8	Java+UI	HCL Technologies	8	845000	2018-05-20 00:00:00	650000	BU 11	281667	70417
2000	Male	Referral	2	9	Selenium	Support.com	9	520000	2017-02-20 00:00:00	500000	BU 6	260000	43333
D	Male	SkillRecruit	2	10	XLS, Report	Altisource	10	650000	2017-02-06 00:00:00	380000	BU 11	325000	54167

#### • Rules to be followed while uploading a CSV File

- 9. The first row provided in the Excel file should contain the column headers.
- 10. The second row of the Excel file should contain the data under all the headers without any 'null' or 'NA.'
- 11. Excel headers should not have space. It should be a single word or two words concatenated by an underscore (\_).
- 12. Excel headers should not contain any special characters. E.g. %, #, \$, @,\*, etc.
- 13. Excel headers should not contain single or double quotes, dots, brackets, and high-fen.
- 14. Excel headers should not contain merely numbers. Numerals should be used with at least one alphabet.
- 15. Excel header should not exceed 50 characters.
- 16. All rows in a column should have the same data type.

Note:



a. The Excel File component supports the .xlsx file type.

## 5.9. Removing a Data Source from the Workspace

- i) Right-click on the data source connector (in the workspace).
- ii) A context menu appears.
- iii) Click the 'Delete' option.
- iv) The selected Data Source component gets removed from the workspace.

OR

Click the 'Create New Workflow' icon to remove the connector(s) from the workspace.

•	୯ 🕂 ୫ 🖬
	Create New Workflow
Data St	
Delete	

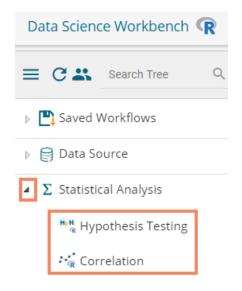
Note: The same set of steps applies to remove all types of data source connectors.

# **6. Statistical Analysis**

Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. Given a hypothesis about a population, for which the user wishes to draw inferences, statistical inference consists of two things, first selecting a statistical model of the process that generates the data and second deducing propositions from the model.

The R workspace provides two Statistical Analysis options as described below:

- 1) Hypothesis Testing
- 2) Correlation





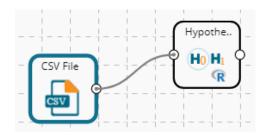
The Python Workspace provides Hypothesis Testing as a Statistical Analytics option.

Data Science Workbench 🥐						
E C Search Tree						
Saved Workflows						
🕨 😝 Data Source						
Σ Statistical Analysis						
<sup>8</sup> 쁞 Hypothesis Testing						
👂 🎲 Data Preparation						

## 6.1. Hypothesis Testing

A statistical hypothesis test is a method of statistical inference. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets.

i) Drag the Hypothesis Testing component to the workspace and connect it to a configured data source.



- ii) Click the Hypothesis Testing component to open the configuration fields.
- iii) The user needs to configure various properties fields based on the Hypothesis Testing component. The following are some possibilities of the various Properties fields when a specific **method** has been selected to perform Statistical Analysis:
  - a. One Sample t-test: The one-sample t-test compares the mean of sample data to a known value. For example, one may want to know how sample means get compared to the population means. For this, one should run a one-sample t-test.
    - i. Statistical Analysis:
      - 1. **Method**: Select an option from the drop-down menu. Other properties fields get displayed based on the selection of the Method option. (In this case, the selected method is 'One Sample t-test')
    - ii. Column Selection:
      - 1. Input Columns: Select any one column from the drop-down menu (it lists only Numeric Column)
    - iii. Other Parameters
      - 1. Mean Value: Pass any integer/ decimal value. The default value for this field is 0.
      - 2. Alternative Hypothesis: select any one option from the drop-down menu (provided choices for this field are- Two-Sided, Greater, Lesser)



3. **Confidence Interval**: the textbox takes a single number between 0 and 1 (the default value for this field is 0.95)

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Statistical Ana	lysis					
Properties	Method		One Sample t-test	•			
	Column Select	tion					
	Input Columns		SepalLength	- 0			
	Other Parame	eters					
	Mean Value		0	0			
	Alternative Hypot	thesis	Two Sided	- 0			
•	Confidence Inter	val	0.95	0			
						Apply	

- **b.** Two Sample t-test: A two-sample t-test is used to test the difference between two population means. A typical application is to determine whether the means are equal.
  - i. Statistical Analysis:
    - 1. **Method**: Select an option from the drop-down menu. Other properties fields get displayed based on the selection of the Method option. (In this case, the selected method is 'Two Sample t-test')
  - ii. Column Selection:
    - 1. First Distribution: Select any one column from the drop-down menu (it lists only Numeric and Factor Columns)
    - 2. Second Distribution: Select any one column from the drop-down menu (it lists only Numeric and Factor Columns)

#### iii. Other Parameters

- 1. Mean Value: Pass any integer/ decimal value. The default value for this field is 0.
- 2. Alternative Hypothesis: select any one option from the drop-down menu (provided choices for this field are- Two-Sided, Greater, Lesser)
- 3. Confidence Interval: the textbox takes a single number between 0 and 1 (the default value for this field is 0.95)
- 4. Paired t-Test: It has two values: True and False (The default value is False)



Component	Console Summary	Result	Visualization	Properties	÷ <u>1</u>
General	Statistical Analysis				
Properties	Method	Two Sample t	-test 🗸		
	Column Selection				
	First Distribution	Number	•	0	
	Second Distribution	SepalLength	•	0	
	Other Parameters				
	Mean Value	0		0	
	Alternative Hypothesis	Two Sided	•	0	
<i></i>	Confidence Interval	0.95		0	
	Paired t-Test	False	•	0	
					Apply
					тери

#### c. Chi-Square Test:

A Chi-Square Test is used to determine whether there is a significant association between the two variables.

- i. Statistical Analysis
  - 1. Method: Select an option from the drop-down menu. Other properties fields get displayed based on the selection of the Method option. (In this case, the selected method is 'Chi-Square Test')

#### ii. Column Selection:

- 1. First Distribution: Select any one column from the drop-down menu (it lists only Numeric and Factor Columns)
- 2. Second Distribution: Select any one column from the drop-down menu (it lists only Numeric and Factor Columns)

#### iii. Other Parameters

- 1. Simulate p-Value: It has two values: True and False (The default value for this field is True)
- 2. Number of Replicates: It takes positive integers (The default value for this field is 2000)
- 3. Continuity Correction: It has two values- True and False (The default value for this field is True)

Component	Console	Summary	/ Result	Visualization	Properties	+	<u>+</u>
General	Statistical An	alysis					
Properties	Method		Chi-Square Test	•			
	Column Sele	ction					
	First Distributio	n	Number	•	0		
	Second Distribu	ition	Select	•	0		
	Other Param	leters					
	Simulate p- valu	ie	True	•	0		
	Number of Rep	licates	2000		0		
>	Continuity Corr	ection	True	•	0		
							_
						Apply	



**d. One-Way ANOVA**: There are many situations where the user may want to compare the mean between multiple groups. The ANOVA test can tell if the groups have similar performances. One-way ANOVA takes one target variable and one independent variable at a time.

#### i. Statistical Analysis

- 1. **Method**: Select an option from the drop-down menu. Other properties fields get displayed based on the selection of the Method option. (In this case, the selected method is 'One Way ANOVA')
- ii. Column Selection:
  - 1. **Target Variable**: Select any one column from the drop-down menu (it lists only Numeric Columns)
  - 2. Independent Variables: Select any one column from the drop-down menu (it lists only Numeric and Factor Columns)

#### iii. Other Parameters

1. Contrasts: Select an option from the given choices to display a list of contrast items that can be used for some variables in the model. (the provided options are contr. treatment, contr. poly, contr. sum, contr. Helmert)

Component	Console	Summary	Result	Visualizatio	on Properties	*	<u>+</u>
General	Statistical An	alysis					
Properties	Method	One-V	Vay ANOVA	•			
	Column Sele	ction					
	Target Variable	Numb	er	- 0	•		
	Independent	Sepal	Length	-	•		
	Variables						
	Other Param	eters					
	Contrasts	None	Selected	•			
						Apply	,

- e. Two-Way ANOVA: There are many situations where the user might want to compare the mean between multiple groups. The ANOVA test can tell if the groups have similar performances. Two-way ANOVA takes one target variable and multiple independent columns at a time.
  - i. Statistical Analysis
    - 1. **Method**: Select an option from the drop-down menu. Other properties fields get displayed based on the selection of the Method option. (In this case, the selected method is 'Two-Way ANOVA')
  - ii. Column Selection:
    - 1. **Target Variable**: Select any one column from the drop-down menu (it lists only Numeric Columns)
    - 2. Independent Variables: Select any one column from the drop-down menu (it lists only Numeric and Factor Columns)
  - iii. Other Parameters
    - 1. **Contrasts:** Select an option from the given choices to display a list of contrast items that can be used for some variables in the model. (the provided options are contr. treatment, contr. poly, contr. sum, contr. Helmert)



Component	Console Si	ummary	Result	Visuali	zation	Properties	+	Ŧ
General	Statistical Analys	is						
Properties	Method	Two-W	ay ANOVA	•				
	Column Selectior	ı						
	Target Variable	Numbe	r	•	0			
	Independent	1 check	ked	•	0			
	Variables							
	Other Parameter	S						
•	Contrasts	contr.p	oly	•				
							Apply	

- **f. Z-test**: Z-test is a statistical test where normal distribution is applied and is used for dealing with problems relating to large samples when  $n \ge 30$ .
  - i. Statistical Analysis:
    - 1. **Method**: Select an option from the drop-down menu. Other properties fields get displayed based on the selection of the Method option. (In this case, the selected method is 'Z-test')
  - ii. Column Selection:
    - 1. **Input Columns**: Select any one column from the drop-down menu (it lists only Numeric Column)
  - iii. Other Parameters
    - 1. Mean Value: Pass any integer/ decimal value. The default value for this field is 0.
    - 2. Alternative Hypothesis: select any one option from the drop-down menu (provided choices for this field are- Two-Sided, Greater, Lesser)
    - 3. **Confidence Interval**: The textbox takes a single number between 0 and 1 (the default value for this field is 0.95)

Component	Console	Summary	Result	Visualization	Properties	<del>*</del>	<u>+</u>
General	Statistical Anal	ysis					
Properties	Method	Z-test		•			
	Column Select	ion					
	Input Columns	Numb	er	- 0			
	Other Parame	ters					
	Mean Value	0		0			
	Alternative	Two S	ided	- 0			
•	Hypothesis						
	Confidence Interv	al 0.95		0			
						Apply	

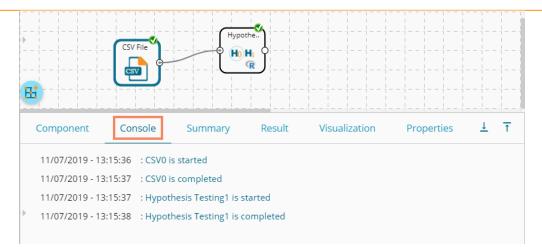


- **g.** Fisher Test: Fisher's exact test is a statistical test used to determine if there are non-random associations between two categorical variables.
  - i. Statistical Analysis
    - 1. **Method**: Select an option from the drop-down menu. (In this case, the selected method is **'Fisher Test'**)
  - ii. Column Selection:
    - 1. **First Distribution**: Select any one column from the drop-down menu (it lists only Numeric and Factor Columns)
    - 2. **Second Distribution**: Select any one column from the drop-down menu (it lists only Numeric and Factor Columns)
  - iii. Other Parameters
    - 1. Simulate p-Value: It has two values: True and False (The default value for this field is True)
    - 2. Number of Replicates: It takes positive integers (The default value for this field is 2000)
    - **3.** Alternative Hypothesis: select any one option from the drop-down menu (provided choices for this field are- Two-Sided, Greater, Lesser)
    - 4. **Confidence Interval**: The textbox takes a single number between 0 and 1 (the default value for this field is 0.95)

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Statistical An	alysis					
Properties	Method		Fisher Test	-			
	Column Sele	ction					
	First Distributio	n	Number	-	9		
	Second Distribu	ution	SepalLength	-	9		
	Other Param	neters					
	Simulate p- valu	Je	True	-	9		
	Number of Rep	licates	2000		9		
•	Alternative Hyp	othesis	Two Sided	-	9		
	Confidence Inte	erval	0.95		3		
						_	_
						Apply	

- iv) After a successful configuration, runs the workflow.
- v) The 'Console' tab opens, displaying the progress of the process.
- vi) The success of the process gets indicated through the green marks on the components.





#### vii) Click the 'Result' tab to see the Result view of the data.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> <u>⊥</u>
Show 10	• entries				Search:	
Number	SepalLength	SepalW	/idth	PetalLength	PetalWidth	Species
1	5.1	3.5		1.4	0.2	setosa
2	4.9	3		1.4	0.2	setosa
3	4.7	3.2		1.3	0.2	setosa
4	4.6	3.1		1.5	0.2	setosa
5	5	3.6		1.4	0.2	setosa
6	5.4	3.9		1.7	0.4	setosa
7	4.6	3.4		1.4	0.3	setosa
8	5	3.4		1.5	0.2	setosa
9	4.4	2.9		1.4	0.2	setosa
10	4.9	3.1		1.5	0.1	setosa
Showing 1 to 10	0 of 150 entries			Previous 1	2 3 4 5	15 Next

#### viii) Click the 'Summary' tab to see the summary of the Hypothesis Test.

C	omponent	Console	Summary	Result	Visualization	Properties	<del>*</del>	<u> </u>
,		Summary of the	Model					
	One	Sample t-test						
		df = 149, p-valu						
		onfidence interv	e mean is not eq /al:	ual to 0				
	sample estim mean of x 75.5	ates:						
		End of Summary						
·		End of Summary						



Note:

- a. Other properties fields get displayed based on the selection of the 'Method' option.
- b. The Hypothesis Testing provided under the Python Workspace contains the same steps of configuration, but the Other Parameters fields vary as per the selected testing method. Please find all the Other Parameters variations provided below based on a specific testing method. Click the 'Information' <sup>(1)</sup> icon to get the details of these fields.
  - i. One Sample t-test

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> ↑	Ŧ
General	Statistical Ar	nalysis					
Properties	Method		One Sample t-test	•			
	Column Sele	ection					
	Input Columns		Number	-	0		
	Other Paran	neters					
	Population Mea	an	0		0		
	Axis		None		0		
•	Dealing With M	lissing	Propagate	•	0		
	Value						
						Apply	y

#### ii. Two Sample t-test

Component	Console	Summary	Result	Visualization	Properties		<del>*</del>	Ŧ
General	Statistical Ar	nalysis						
Properties	Method	C	Two Sample t-test	•				
	Column Sele	ection						
	First Distributio	on	Number	•	0			
	Second Distrib	ution	SepalLength	-	0			
	Other Paran	neters						
	Axis		None		0			
•	Equal Variance		True	•	0			
	Dealing With M	lissing	Propagate	•	0			
						A	pply	

iii. Chi-Square Test



Component	Console	Summary	Result	Visualization	Properties	4	-	<u>+</u>
General	Statistical Ar	nalysis						
Properties	Method	[	Chi-Square Test	•				
	Column Sele	ection						
	First Distributio	on	Number	-	0			
	Second Distrib	ution	Select	-	0			
	Other Paran	neters						
	Delta Degrees	of	0		0			
•	Freedom							
r	Axis		0		0			
						A	oply	

#### iv. Z-test

Component	Console	Summary	Result	Visualization	Properties	÷	•	<u>+</u>
General	Statistical A	nalysis						
Properties	Method		Z-test	-				
	Column Sele	ection						
	First Distributio	on	Number	-	0			
	Second Distrib	ution	Select	-	0			
	Other Parar	neters						
	Mean Value		0		0			
•	Alternative Hyp	oothesis	Two Sided	-	0			
	Delta Degrees	of	0		0			
						А	pply	

Way

There are no Other Parameters fields provided for the methods One-Way ANOVA and Two-

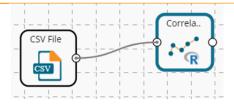
ANOVA.

## 6.2. Correlation

Correlation is a statistical inference method that measures the degree to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

- i) Drag the Correlation component to the workspace and connect it to a configured data source.
- ii) Click the Correlation component to open the configuration fields.





- iii) Configure the following properties fields for the Correlation component:
  - a. Input Columns: Select any two columns using the drop-down menu
  - **b.** Method: Select a method using the drop-down menu. The available methods are:
    - i. Pearson
    - ii. Kendall
    - iii. Spearman
  - **c. Missing Value Method:** Select the required option using the drop-down menu. The available methods to Apply the Missing Value are:
    - i. Everything
    - ii. All. obs
    - iii. Complete. obs
    - iv. Na.or. complete
    - v. Pairwise.complete.obs

v. runvv	sc.compicte.	005				
Component	Console Summa	ary Result	Visualization	Properties	+++++++++++++++++++++++++++++++++++++++	<u>+</u>
General	Column Selection					
Properties	Input columns	5 checked	- 0			
	Method	1 checked	-			
	Missing value method	1 checked	•			
Þ						
					Apply	1
					трру	

- a. Click the 'Apply' option.
- iv) Run the workflow.
- v) The progress of the process gets displayed in the 'Console' tab.



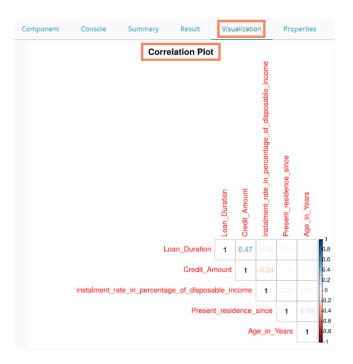
- vi) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged correlation component onto the workspace.
  - b. Click the 'Result' tab.



now 10 🔻 entries				Search:	
category	Loan_Duration	Credit_Amount	instalment_rate_in_percentage_of_disposable_income	Present_residence_since	Age_in_Years
Loan_Duration	1	0.465738245237381	0.0935215165673161	0.0348946077169088	-0.025185706702483
Credit_Amount	0.465738245237381	1	-0.238537324761332	0.0181460663030051	0.0173077340771623
instalment_rate_in_percentage_of_disposable_income	0.0935215165673161	-0.238537324761332	1	0.0410097613966184	0.055433135422757
Present_residence_since	0.0348946077169088	0.0181460663030051	0.0410097613966184	1	0.185288601533654
Age_in_Years	-0.0251857067024839	0.0173077340771623	0.0554331354227578	0.185288601533654	1

Note: The selected dataset has more columns than displayed in the below-given Result view.

- vii) Click the 'Visualization' tab.
- viii) The probable values of the selected columns get displayed via the Correlation Plot.



ix) Click the 'Summary' tab to view the model summary.



	Component Console	Summary	Result	Visualization	Properties
	Summary of th	e model			
	Columns used in the algori	thm			
*	Loan_Duration (i Credit_Amount (i instalment_rate_in Present_residence_ Age_in_Vears (i	nteger) _percentage_of_di since (integer)	sposable_inco	me (integer)	
	Loan_Duration Credit Min. :-0.02519 Min. 1st Qu.: 0.03489 1st Qu Median : 0.09352 Median Mean : 0.31379 Mean	- :-0.23854 .: 0.01731 : 0.01815 : 0.25253			
	3rd Qu.: 0.46574 3rd Qu Max. : 1.00000 Max. instalment_rate_in_percen	: 1.00000	e income Pres	ent residence since	
	Min. :-0.23854 1st Qu.: 0.04101 Median : 0.05543		Min. 1st	 :0.01815 Qu.:0.03489 an :0.04101	
	Mean : 0.19029 3rd Qu.: 0.09352		Mean	0.25587 Qu.:0.18529	
	Max. : 1.00000 Age_in_Years Min. :-0.02519		Max.	:1.00000	
4	1st Qu.: 0.01731 Median : 0.05543 Mean : 0.24657 3rd Qu.: 0.18529 Max. : 1.00000				
	End of Summar	у			

Note: The displayed Result, Visualization, and Summary tabs are based on the selection of the Kendall method. The user may have a slight variation based on another selection.

# 7. Data Preparation

Components provided under the **Data Preparation** tree-node help in preparing the raw data from the data source and make it suitable for analysis. They organize data to gain accurate Results out of it. The list of the Data Preparation components may vary based on the different Workspace, but the configuration steps remain the same. This section aims at listing all the available Data Preparation components collectively.

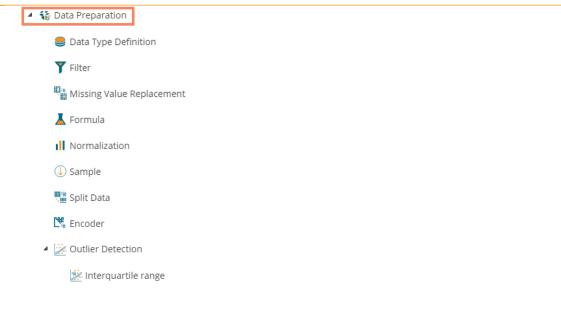
Note: The Data Preparation list may vary based on various Data Science Workspaces, but the configuration process remains the same for all.

## 7.1. Data Type Definition

The Data Type Definition option can be used to change the name, data type of the data source column. This component helps users to prepare data and make it suitable for further analysis.

- i) Navigate to the landing page of any Data Science Workspace.
- ii) Click the 'Data Preparation' tree node.
- iii) Various data preparation options get displayed (The below-given list displays the Data Preparation options provided under the R Workspace since it includes all the available Data Preparation components).





- iv) Drag the 'Data Type Definition' component and connect it to a configured data source onto the workspace.
- v) Click the 'Data Type Definition' component (in the workspace).

	Data Ty	ו
CSV File		0

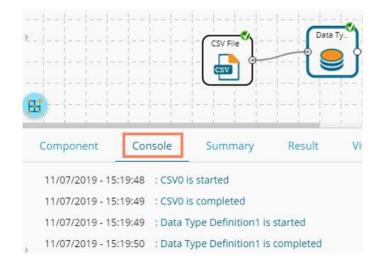
- vi) The 'Properties' tab opens.
- vii) Configure the following 'Data Type Mapping' details:
  - a. Column Name: Select a column name that you want to change
  - b. Alias Name: Enter an alias name for the required source column
  - c. Primary Data Type: Select a primary data type column that you want to change
  - d. **Date Format:** Select a date format that you want to display (the Date format is optional for date Data Type)
  - e. 'Add' option :: Click on this icon to add one more row of the 'Data Type Mapping' fields
- viii) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Data Type M	apping					
Properties	Colun usd_t id	pill 🔻	AliasName USDBilling Employeee	PrimaryDataType Integer  String	-	+	
						Appl	y

ix) Run the workflow by clearing the previous Cache.



x) Open the 'Console' tab to see the progress of the process. The completion of the Console process gets marked by the green checkmarks on the top of the dragged components.



- xi) After the Console process gets completed, users can view the Result data using the 'Result' tab
- xii) Follow the below-given steps to display the Result view:
  - a. Click the dragged Data Type Definition component in the workspace.
  - **b.** Click the 'Result' tab.
- xiii) The user can see the given column names on the selected columns in the displayed **Result** data.

Component	Console	e Sumn	nary Result	Visualization	Properties			<u>+</u> ⊥
Show 10	entries						Search:	
USDBilling	gender	source	experience_Year	candidate_id	skills	previous_organisation	EmployeeeID	offered_ctc
4000	Male	Indeed	15	1	Management, Selenium	Athenahealth	1	1800000
4000	Male	Orgspire	10	2	Selenium	Support.com	2	1500000
2600	Male	Orgspire	4	3	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000
2300	Female	Referral	5	4	Selenium	Inventateq	4	650000
1750	Male	Referral	3	5	Selenium	Tekinspy	5	520000
0	Male	BMS Innolabs	4	6	java	CGI Information Systems	6	980000
0	Male	Orgspire	3	7	AWS	Cognizant Technology solutions	7	650000
0	Male	BMS Innolabs	3	8	Java+UI	HCL Technologies	8	845000
2000	Male	Referral	2	9	Selenium	Support.com	9	520000
0	Male	SkillRecruit	2	10	XLS, Report	Altisource	10	650000
Showing 1 to 10	of 224 entries	s				Previous 1 2	3 4 5	23 Next

### 7.2. Filter

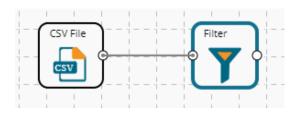
This data preparation component is used to filter the data by column or row.

- i) Select and Drag the 'Filter' component onto the workspace.
- ii) Connect the 'Filter' component to a configured data source component.

#### Copyright © 2015-21 BDB

#### www.bdb.ai





iii) Configure the filter component as described below:

# 7.2.1. Column Filter

- i) Select a column from the 'Selected Columns' context menu.
- ii) Click the **'Apply'** option to configure the data.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Column Filter	]					
Row Filter	Selected Columns	; 3 (	checked	•			
Column Filter			/ team				
		e	xpyrsper_ctc				
		n	nonthly_salary				
		c	ur_monthly_pay	ment			
			/ name				
		c	urrent_status				
			<ul> <li>designation</li> </ul>				
						Appl	y

- iii) Run the workflow by clearing the previous cache.
- iv) The 'Console' tab opens to display the progress of the process. The completion of the Console process gets marked by green checkmarks on the top of the dragged components.



Ē				Fi	Iter
	Component	Console	Summary	Result	Visualization
•	11/07/2019 - 15:4 11/07/2019 - 15:4 11/07/2019 - 15:4 11/07/2019 - 15:4	0:43 : CSV0 is 0:43 : Filter1	s completed is started		

- v) After the Console process gets completed, users can view the Result data using the 'Result' tab.
- vi) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component in the workspace.
  - **b.** Click the 'Result' tab.
- vii) The filtered data gets displayed via the 'Result' tab.

Component	Console Summary	ResultVisualizationProperties####
Show 10 •	entries	Search:
team	name	designation
BU 6	Ahsan R	QA Manager
BU 6	Rajive Raveendra Pai	QA Architect
BU 11	Amit Kumar Soni	Senior Software Engineer
BU 6	Ritu	QA Engineer
BU 6	Vedprakash	QA Engineer
BU 7	Vedprakash	Senior Software Engineer
BU 7	Animesh Srivastava	AWS Consultant
BU 11	Vikram Bharti	Senior Software Engineer
BU 6	Sudharshan Reddy	QA Engineer
BU 11	Ajish.T.Thomas	Business Analyst
Showing 1 to 10 c	of 224 entries	Previous 1 2 3 4 5 23 Next

## 7.2.2. Row Filter

- i) Drag the Filter Component to the workspace and connect it to a configured data source.
- ii) Click the **'Filter'** component.
- iii) The 'Column Filter' tab gets displayed (by default).

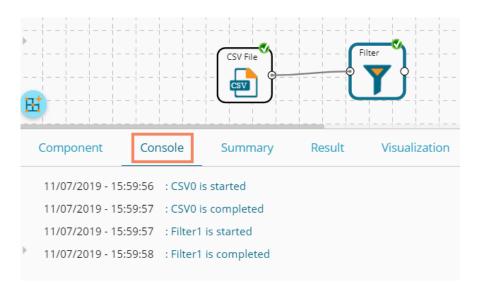
#### www.bdb.ai



- iv) Select a column using the context menu.
- v) Select the 'Row Filter' tab from the 'Component' menu list.
- vi) Configure the required fields:
  - a. Double click on the components from **Columns, Operators,** and **Functions** in the sequence as shown in the image below
  - b. A formula gets entered in the given box (E.g., in this case, the entered formula is [id]>SELECT(200)
  - c. Click the 'Apply' option.

General       Row Filter         Row Filter       [id]>SUM(200)         Column Filter       Id         Skills       SUBSTRING         skills       STRLEN         id       STRLEN         mame       MAX         designation       AVERAGE         SUM       Conditional functions         IFELSECONDITION       Divide	Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
skills     SUBSTRING     Equal to       id     STRLEN     Not Equal to       team     Mathematical functions     Greater than       name     MIN     Less than       designation     AVERAGE     Less than or equal to       SUM     Conditional functions     Multiply       IFELSECONDITION     Divide	Row Filter	Row Filte				0		
Apply	Þ	[	skills id team name		SUBSTRING STRLEN Mathematical functions MAX MIN AVERAGE SUM Conditional functions	Equal to Not Equal to Greater than Greater than or equal to Less than Less than or equal to Multiply	Apply	

- vii) Run the workflow by clearing the previous cache.
- viii) The 'Console' tab opens to display the progress of the process. The completion of the Console process is marked by the green tick marks on the top of the dragged components.



- ix) After the Console process gets completed, users can view the Result data using the 'Result' tab
- x) Follow the below-given steps to display the Result view:

#### www.bdb.ai



- **a.** Click the dragged data preparation component on the workspace.
- b. Click the 'Result' tab.
- xi) The filtered data, as per the applied formula, gets displayed under the 'Result' tab.

Com	iponent Console	5	Summary	Result	Visualization	Properties		+	<u>+</u>
Show	10 • entries					Search	:		
skil	lls	id	team		designation		name		
Java,	, Big Data	201	BU 10		Software Developer		Ranjana		
Java,	, Big Data	202	BU 10		Sr Big Data Developer		Saquib		
Java,	, Big Data	203	BU 10		Sr Big Data Developer		Mayur		
Java,	, Big Data	204	BU 10		Big Data Developer		Ishana		
Java	+UI	205	BU 10		Sr Software Developer		Arnav		
Java	+UI	206	BU 10		Sr Software Developer		Kanakpriya		
Java	+UI	207	BU 10		Sr Software Developer		Vijay		
Java		208	BU 10		Sr Software Developer		Arghya		
Java		209	BU 10		Sr Software Developer		Anamika		
ios e	Dev, Java	210	BU 10		iOS Developer		Gurdeep		
Showi	ing 1 to 10 of 24 entries					Previous	1 2 3	Nex	αt

#### Note:

- a. The expression should retain Boolean output.
- b. Users can not use Data manipulation functions.
- c. The Row Filter functionality provided under the Spark workspace takes the specific column name in between the @ symbols.

E.g., @cat@ as displayed below.

General	Row Filter
Row Filter	cat=@cat@
Condition Filter	
	Columns
	Number
	PetalLength PetalWidth
	SepalLength
	SepalWidth
	cat

## 7.3. Missing Value Replacement

Users can replace the missing data in the specified variable with the determined value. The user is



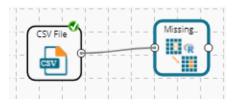
provided with a list of options that can be considered for replacement.

i) Drag a data source on the workspace, configure it, run it, and check the data using the 'Result' tab. (in this

Search	1:									
oining_date	previous_ctc	team	expyrsper_ctc	monthly_salary	cur_monthly_payment	name	current_status	designation	referral_of	joining_stat
	2000000	BU 6	120000	150000	125000	Ahsan R	Transferred	QA Manager		Joined
	2000000	BU 6	150000	125000	125000	Rajive Raveendra Pai	Resigned	QA Architect		Joined
	650000	BU 11	256000	85333	85333	Amit Kumar Soni	Terminated	Senior Software Engineer	Ritu	Joined
	580000	BU 6	130000	54167	52000	Ritu	Transferred	QA Engineer	Ahamad	Joined
	500000	BU 6	208000	43333	43333	Vedprakash	Transferred	QA Engineer	Ahamad	Joined
	730000	BU 7	233333	81667	0	Vedprakash	Declined	Senior Software Engineer		Declined
	510000	BU 7	216667	54167	0	Animesh Srivastava	Absconded	AWS Consultant		Absconded
	650000	BU 11	281667	70417	0	Vikram Bharti	Declined	Senior Software Engineer		Declined
	500000	BU 6	260000	43333	0	Sudharshan Reddy	Declined	QA Engineer	Tania	Declined
	380000	BU 11	325000	54167	0	Ajish.T.Thomas	Declined	Business Analyst		Declined

case, the selected input data is displayed in the following image)

- ii) Select and drag the 'Missing Value Replacement' component onto the workspace.
- iii) Connect the 'Missing Value Replacement' component to a configured data source.
- iv) Use the Right-click on the 'Missing Value Replacement' component to configure.

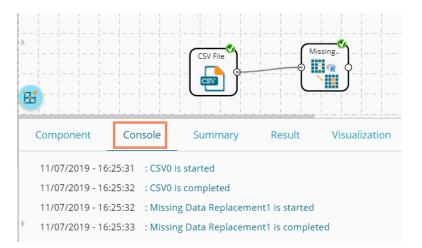


- v) Choose the replacement value by configuring the following fields:
  - a. Column Name: Select a column using the drop-down that contains some missing values.
  - b. **Replacement Options**: Select a replacement option using the drop-down menu. The following replacement options are provided under this field:
    - 1. Mean
    - 2. Median
    - 3. Mode
    - 4. Maximum
    - 5. Minimum
    - 6. Remove Entire Row
    - 7. Remove Entire Column
    - 8. Custom Replacement
- vi) Click the 'Apply' option.



Component	Console Summa	y Result Visual	ization Properties	± T
General	Replacement Values	]		
Properties	Column Name referral_r 🔻	Replacement Options Custom Replaceme	<b>6</b> +	
				Apply

- vii) Run the workflow by clearing the previous cache.
- viii) The user can be redirected to the 'Console' tab to display the progress of the process.



- ix) After the Console process gets completed, the user can view the Result data using the 'Result' tab.
- x) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged data preparation component on the workspace.
  - **b.** Click the 'Result' tab.
- xi) The missing values in the selected column get replaced with the selected custom replacement value.

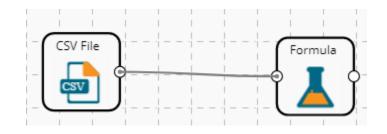
Com	ponent Console	Summary	Result	alization Pro	operties	÷.
lary	cur_monthly_payment	name	current_status	designation	referral_of	joining_status
	125000	Ahsan R	Transferred	QA Manager	no record found	Joined
	125000	Rajive Raveendra Pai	Resigned	QA Architect	no record found	Joined
	85333	Amit Kumar Soni	Terminated	Senior Software Engineer	Ritu	Joined
	52000	Ritu	Transferred	QA Engineer	Ahamad	Joined
	43333	Vedprakash	Transferred	QA Engineer	Ahamad	Joined
	0	Vedprakash	Declined	Senior Software Engineer	no record found	Declined
	0	Animesh	Absconded	AWS Consultant	no record	Absconded



# 7.4. Formula

The user can create a calculated column using **'Formula.'** A formula can be formed by using available columns, functions, and operators.

- i) Select and drag the 'Formula' component onto the workspace.
- ii) Connect the 'Formula' component to a configured data source.
- iii) Click on the **'Formula'** component.

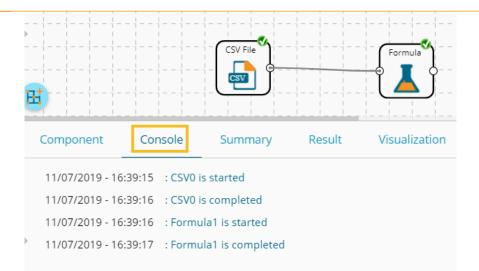


- iv) Configure the required component fields to Apply a formula:
  - a. 'Columns,' 'Functions,' and 'Operators': Double click on these lists enter a formula in the given box.
  - b. Formula Name: Enter a formula name in the given field.
  - c. Click 'Apply' to configure the formula.

Component	Console	Summary	Result	Visualiz	ation	Properties		<del>*</del>	<u>+</u>
General	F	ormula Name	Formula			0			
Formula	-		[usd_billing]+[n	nonthly_salar	ry]				
		C	Columns offered_ctc expected_join previous_ctc	ing_date	MONTH	ETWEEN HSBETWEEN BETWEEN	Operators Multiply Divide Add		
Þ			team expyrsper_ctc monthly_salar	_		ME MBEROFMO MBEROFWE	Subtract Power of Exponential		
								Apply	/

- v) Run the workflow by clearing the previous cache.
- vi) The 'Console' tab opens displaying the progress of the process. The completion of the Console process gets marked by the green checkmarks on the top of the dragged components.





- vii) After the Console process gets completed, the user can view the Result data using the 'Result' tab.
- viii) Follow the below-given steps to display the Result view:
  - a. Click the dragged data preparation component on the workspace.
  - **b.** Click the 'Result' tab.
- ix) A new column containing the data based on the inserted formula gets added to the Result data. (E.g., the '**Formula**' column as displayed below.)

	Component Console	e Summary	y Result	Visualization	Properties		<u>+</u> ⊥
•	cur_monthly_payment	name	current_status	designation	referral_of	joining_status	Formula
	125000	Ahsan R	Transferred	QA Manager		Joined	154000
	125000	Rajive Raveendra Pai	Resigned	QA Architect		Joined	129000
	85333	Amit Kumar Soni	Terminated	Senior Software Engineer	Ritu	Joined	87933
	52000	Ritu	Transferred	QA Engineer	Ahamad	Joined	56467
	43333	Vedprakash	Transferred	QA Engineer	Ahamad	Joined	45083
Þ	0	Vedprakash	Declined	Senior Software Engineer		Declined	81667
	0	Animesh Srivastava	Absconded	AWS Consultant		Absconded	54167

## 7.5. Normalization

This component controls the relevant data. It attempts to convert the available data from a larger range to a smaller range. It can be done over numerical columns.

## 7.5.1. Min-Max Normalization

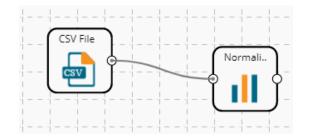
It implements a linear transformation of the original data values and sets a new range for all the data values to fit in. The user can fix the New Maximum and New Minimum Value for the data from the new field. Consequently, each value "v" from the original interval gets mapped into value "new\_v" following the below-given formula:

www.bdb.ai



$$new_v = \frac{v - min_x}{max_x - min_x} \cdot (new_max_x - new_min_x) + new_min_x$$

- i) Select and drag **the 'Normalization'** component onto the Workspace.
- ii) Connect the **'Normalization'** component to a configured data source.
- iii) Click the 'Normalization' component.



iv) Configure the following component fields:

#### Properties

- a. Column Selection
  - i. **Select a Column**: Select a column using the drop-down menu (Only the numerical column gets selected)
- b. Behavior
  - i. Normalization Type: Select 'Min-Max' normalization type from the drop-down menu
  - ii. New Maximum: Set a new maximum value (the Default value for this field is 1)
  - iii. **New Minimum:** Set a new minimum value (the Default value for the New Minimum field is 0)
- v) Click the 'Apply' option.

Component	Console Summa	ary Result	Visualization	Properties	+	Ť
General	Column Selection					
Properties	Select Columns	1 checked	<b>.</b> 0			
	Behavior					
	Normalization Type	Min-Max	•			
	New Maximum	100				
	New Minimum	0				
•						
					Apply	/

vi) Run the workflow by clearing the previous cache.



vii) The 'Console' tab opens displaying the progress of the process. The completion of the Console process gets marked by the green checkmarks.

				Þ	• 0	+	Ð
(	Create New Workflow	N					
Ē	Ħ	CSV FI	le Norr	nali.			
	Component	Console	Summary	Result	Vis	sualizat	ion
	10/09/2019 - 12:05:	10 : CSV0 i	s started				
	10/09/2019 - 12:05:	12 : CSV0 i	s completed				
•	10/09/2019 - 12:05:	12 : Norma	alization1 is start	ed			
	10/09/2019 - 12:05:	14 : Norma	alization1 is com	oleted			

- viii) After the Console process gets completed, the user can view the Result data using the 'Result' tab.
- ix) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged Formula component in the workspace.
  - **b.** Click the 'Result' tab.

how 10 • ent	ries								Search:	
usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	pr
72.727272727272727	Male	Indeed	15	1	Management, Selenium	Athenahealth	1	1800000	02-07-2018	200
72.727272727272727	Male	Orgspire	10	2	Selenium	Support.com	2	1500000	12-01-2018	200
47.2727272727273	Male	Orgspire	4	3	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000	18-07-1980	650
41.8181818181818	Female	Referral	5	4	Selenium	Inventateq	4	650000	18-03-2018	580
31.8181818181818	Male	Referral	3	5	Selenium	Tekinspy	5	520000	15-04-1972	500
0	Male	BMS Innolabs	4	6	Java	CGI Information Systems	6	980000	20-05-2018	73(
0	Male	Orgspire	3	7	AWS	Cognizant Technology solutions	7	650000	10-06-2018	51(
0	Male	BMS Innolabs	3	8	Java+UI	HCL Technologies	8	845000	20-05-2018	65
36.3636363636364	Male	Referral	2	9	Selenium	Support.com	9	520000	20-02-2017	50
0	Male	SkillRecruit	2	10	XLS, Report	Altisource	10	650000	06-02-2017	38



### 7.5.2. Zero-Score

This normalization is known as **Zero Mean Normalization**, which is calculated on the **mean** and **standard deviation** for each attribute. It determines whether a specific value is above or below average. It also signifies the exact proportion of the variance from the fixed limit of aver3age. After Applying **'Zero-Score'** normalization, each feature has a mean value of zero (0). The unit of each value is the number of (estimated) standard deviations away from the (estimated) mean. Zero score normalization may be sensitive to small values of **'**<sup> $\sigma_x$ </sup> ' new value the **'new\_v'** can be found by using the following expression:

$$new_v = \frac{v - \mu_x}{\sigma_x}$$

- i) Select and drag the 'Normalization' component onto the Workspace
- ii) Connect the 'Normalization' component to a configured data source
- iii) Click the 'Normalization' Component
- iv) Configure the required component fields:

#### **Properties**

- a. Column Selection
  - i. **Select a Column**: Select a column using the drop-down menu (Only the numerical column gets selected)
- b. Behavior
  - i. **Normalization Type:** Select '**Zero-Score**' normalization type from the drop-down menu
- v) Click the 'Apply' option.

Component	Console	Summary	Result	Visuali	zation	Properties	+ †	<u>+</u>
General	Column Select	ion						
Properties	Select Columns	1 c	hecked	•	0			
	Behavior							
	Normalization Typ	Zei	ro-Score	•	0			
•								
							App	ly

- vi) Run the workflow by clearing the previous cache.
- vii) The user gets redirected to the 'Console' tab to display the progress of the process. The completion of the Console process is marked by the green checkmarks on the top of the dragged components.



		G
	Create New Workflow	
•	CSV File Normali	
	Component Console Summary Result	Ņ
	10/09/2019 - 13:26:43 : CSV0 is started	
L	10/09/2019 - 13:26:45 : CSV0 is completed	
$(\cdot)$	10/09/2019 - 13:26:45 : Normalization1 is started	
	10/09/2019 - 13:26:47 : Normalization1 is completed	

- viii) After the Console process gets completed, the user can view the Result data using the 'Result' tab.
- ix) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component in the workspace.
  - **b.** Click the 'Result' tab.

how 10 • ent	tries								Search:	
usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	F
72.727272727272727	Male	Indeed	15	1	Management, Selenium	Athenahealth	1	1800000	02-07-2018	20
72.727272727272727	Male	Orgspire	10	2	Selenium	Support.com	2	1500000	12-01-2018	20
47.2727272727273	Male	Orgspire	4	3	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000	18-07-1980	6
41.8181818181818	Female	Referral	5	4	Selenium	Inventateq	4	650000	18-03-2018	5
31.8181818181818	Male	Referral	3	5	Selenium	Tekinspy	5	520000	15-04-1972	5
0	Male	BMS Innolabs	4	6	Java	CGI Information Systems	6	980000	20-05-2018	7
0	Male	Orgspire	3	7	AWS	Cognizant Technology solutions	7	650000	10-06-2018	5
0	Male	BMS Innolabs	3	8	Java+UI	HCL Technologies	8	845000	20-05-2018	6
36.3636363636364	Male	Referral	2	9	Selenium	Support.com	9	520000	20-02-2017	5
0	Male	SkillRecruit	2	10	XLS, Report	Altisource	10	650000	06-02-2017	3

### 7.5.3. Decimal-Scaling

The decimal point of the value of each element is moved by its maximum absolute value. A modified value **'new\_v'** can be obtained using the following formula:



$$new_v = \frac{v}{10^c}$$

Note: In the decimal-scaling expression, 'c' is the smallest integer so that  $max(new_v) < 1$ .

- i) Select and drag the 'Normalization' component onto the Workspace.
- ii) Connect the 'Normalization' component to a configured data source.
- iii) Click the **'Normalization'** Component.
- iv) Configure the required component fields:

#### Properties

- a. Column Selection
  - i. **Select a Column**: Select a column using the drop-down menu (Only the numerical column gets selected).
- b. Behavior
  - i. Normalization Type: Select 'Decimal Scaling' normalization type from the drop-down menu.
- v) Click 'Apply' to configure the fields:

Component	Console	Summary	Result	Visualization	Properties	<del>+</del> †	Ŧ
General	Column Se	lection					
Properties	Select Colum	ns	1 checked	•	0		
	Behavior						
	Normalization	n Type	Decimal Scaling	•	0		
•							
						Apply	/

- vi) Run the workflow by clearing the previous cache.
- vii) The 'Console' tab opens displaying the progress of the process. The completion of the Console process gets marked by the green checkmarks on the top of the dragged components.



Normalization			► C
		Norma	
Component	Console	Summary	Result
	5:59 : CSV0 5:59 : Norm		

- viii) After the Console process gets completed, users can view the Result data using the 'Result' tab.
- ix) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged data preparation component on the workspace.
  - **b.** Click the 'Result' tab.

Compone	ent Console	Sumn	nary Result	Visualization	Properties					+
Show 10	• entries								Search:	
usd_billi	ng gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	prev
0.4	Male	Indeed	15	1	Management, Selenium	Athenahealth	1	1800000	02-07-2018	20000
0.4	Male	Orgspire	10	2	Selenium	Support.com	2	1500000	12-01-2018	20000
0.26	Male	Orgspire	4	3	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000	18-07-1980	65000
0.23	Female	Referral	5	4	Selenium	Inventateq	4	650000	18-03-2018	58000
0.175	Male	Referral	3	5	Selenium	TekInspy	5	520000	15-04-1972	50000
0	Male	BMS Innolabs	4	6	Java	CGI Information Systems	6	980000	20-05-2018	73000
0	Male	Orgspire	3	7	AWS	Cognizant Technology solutions	7	650000	10-06-2018	51000
0	Male	BMS Innolabs	3	8	Java+UI	HCL Technologies	8	845000	20-05-2018	65000
0.2	Male	Referral	2	9	Selenium	Support.com	9	520000	20-02-2017	50000
0	Male	SkillRecruit	2	10	XLS, Report	Altisource	10	650000	06-02-2017	38000

Note:

- a. Normalization displays columns containing only numerical data.
- b. 'New Maximum Value' must be higher than 'New Minimum Value.



# 7.6. Sample

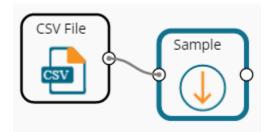
This component can be used to select a subsection of data from a large dataset. The sample component supports the following sample types:

# 7.6.1. Sampling Methods

- 1. **First N:** It selects the first N records from the data source. E.g., If the chosen value for "N" is 10, then it will select the first ten records from the data.
- 2. Last N: It selects the last N records from the data source. E.g., If the chosen value for "N" is 5, then it will select the last five records from the data.
- **3. Every Nth:** It selects every Nth record from the data source, wherein "N" indicates an interval. E.g., If N=3, then the 3<sup>rd</sup>, 6<sup>th</sup>, and 9<sup>th</sup> records get selected from the data.
- 4. **Simple Random:** It selects records randomly as per the value of "N" or the percentage mentioned for "N" from the data source. E.g., If the selected value for "N" is four then, it selects randomly any four records from the data source. If the selected value for "N" is 4% then, it selects 4% of records from the data source.
- 5. **Systematic Random:** It selects data based on the bucket size. E.g., If the chosen value for the bucket is two then, it selects 1<sup>st</sup>, 3<sup>rd</sup>, 5<sup>th</sup> records or 2<sup>nd</sup>, 4<sup>th</sup>, 6<sup>th</sup>records from the data source.

# 7.6.2. Steps to Apply a Sampling Method

- i) Select and drag the 'Sample' component onto the workspace.
- ii) Connect the 'Sample' component to a configured data source.
- iii) Click the 'Sample' component.



iv) Configure the required component fields:

#### **Properties**

- a. Sampling Information
  - i. Sampling Type: Select an option from the drop-down menu
  - ii. Limit Rows by Select an option from the drop-down menu. This field will offer two options, as described below:
    - Numbers of Rows: By selecting this option, it will display a new field 'Number of Rows.'
    - 2. **Percentage of Rows**: By selecting this option, it will display the new field '**Percentage of Rows**.'
- b. Sample Size Limit
  - i. **Maximum Rows**: The maximum number of rows that can be viewed in the 'Result' tab (It is an optional field)
- v) Click the 'Apply' option.



Component	Console Summ	ary Result	Visua	alization	Properties	+ +	<u>+</u>
General	Sampling Informatio	n					
Properties	Sampling Type	First N	•				
	Limit Rows by	Number of Rows	•	]			
	Number of Rows	5					
	Sample Size Limit						
	Maximum Rows	9					
•							
						Appl	у

- vi) Run the workflow by clearing the previous cache.
- vii) The 'Console' tab opens displaying the progress of the process. The completion of the process gets marked by the green checkmarks on the top of the dragged components.

			► C
	Create New Workflow		
•		CSV File	Imple
	Ŧ		_
	Component	e Summary	Result
	11/09/2019 - 11:55:42 : C	SV0 is started	
	11/09/2019 - 11:55:43 : C	SV0 is completed	
•••	11/09/2019 - 11:55:43 : Sa	ample1 is started	
	11/09/2019 - 11:55:45 : Sa	ample1 is completed	

- viii) After the Console process gets completed, open the 'Result' tab to view Result data.
- ix) While accessing the 'Result' tab, the user gets the Result view based on the selected Sampling Type.

# 7.6.3. Result View for the Available Sampling Methods



### 1. First N (Where 'N' is 1 number of the row)

Component	Console Sur	nmary	Result	Visua	lization	Properties	+	Ť
General	Sampling Informa	tion						
Properties	Sampling Type	First N	٧	•				
	Limit Rows by	Numb	er of Rows	•				
	Number of Rows	5						
	Sample Size Limit							
	Maximum Rows	9						
•								
							_	
							Appl	у

how 10 •	entries								Search:	
usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	previou
4000	Male	Indeed	15	1	Management, Selenium	Athenahealth	1	1800000	02-07-2018	2000000
4000	Male	Orgspire	10	2	Selenium	Support.com	2	1500000	12-01-2018	2000000
2600	Male	Orgspire	4	3	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000	18-07-1980	650000
2300	Female	Referral	5	4	Selenium	Inventateq	4	650000	18-03-2018	580000
1750	Male	Referral	3	5	Selenium	Tekinspy	5	520000	15-04-1972	500000

2. Last N ('N' is 10% and maximum rows are 7)

Component	Console	Summary	Result	Visualizatio	on Properties	<u>+</u> <u>+</u>
General	Sampling In	formation				
Properties	Sampling Type	L	ast N	•		
	Limit Rows by	P	Percentage of Rows	•		
	Percentage of	Rows	10		0	
	Sample Size	Limit				
	Maximum Row	/S	7			
•						
						Apply



	Component	Console	Sum	mary Result	Visualization	Prope	rties						÷ 1
	Show 10 🔻	entries									Search:		
•	usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	previous_ctc	team	expyrs
	3025	Male	BDB	5	202	Java, Big Data	EDB	202	1382400	01-12-2016	1123200	BU 10	276480
	2625	Male	BDB	4	203	Java, Big Data	BDB	203	1041600	01-12-2016	892800	BU 10	297600
	1500	Female	BDB	2	204	Java, Big Data	BDB	204	480000	01-12-2016	480000	BU 10	240000
	2625	Male	BDB	4	205	Java+UI	BDB	205	924000	01-12-2016	792000	BU 10	264000
	3025	Female	BDB	5	206	Java+UI	BDB	206	864000	01-12-2016	702000	BU 10	172800
÷	2625	Male	BDB	4	207	Java+UI	BDB	207	907200	01-12-2016	777600	BU 10	259200
	2225	Male	BDB	4	208	Java	BDB	208	748800	01-12-2016	662400	BU 10	213943
	Showing 1 to 7 o	f 7 entries									Previ	ious 1	Next

### 3. Every Nth (Interval is 3, and the maximum rows are 7)

Component	Console	Summary	e Result	Visualiz	ation	Properties		<u>↓</u> †	Ŧ
General	Sampling Int	formation							
Properties	Sampling Type	[	Every Nth	•					
	Step Size		1						
	Sample Size	Limit							
	Maximum Row	s	7						
• >									
							A	\pply	

Component	Console	Sumr	mary Result	Visualization	Properties					*
how 10 •	entries								Search:	
usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	previo
4000	Male	Indeed	15	1	Management, Selenium	Athenahealth	1	1800000	02-07-2018	2000000
4000	Male	Orgspire	10	2	Selenium	Support.com	2	1500000	12-01-2018	2000000
2600	Male	Orgspire	4	3	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000	18-07-1980	650000
2300	Female	Referral	5	4	Selenium	Inventateq	4	650000	18-03-2018	580000
1750	Male	Referral	3	5	Selenium	Tekinspy	5	520000	15-04-1972	500000
1750	Male	BMS Innolabs	4	6	Java	CGI Information Systems	6	980000	20-05-2018	730000
2300	Male	Orgspire	3	7	AWS	Cognizant Technology solutions	7	650000	10-06-2018	510000

4. Simple Random (the 'Maximum Rows' are 7). The randomly selected seven rows will be displayed.



Component	Console Summa	ry Result	Visualization	Properties	<u>↓</u> ↑	⊥
General	Sampling Information					
Properties	Sampling Type	Simple Random	v			
	Limit Rows by	Percentage of Rows	T			
	Percentage of Rows	10	0			
	Sample Size Limit					
	Maximum Rows	7				
•						
					Apply	/

how 10 🔻	entries									Search:		
usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	previous_ctc	team	exp
1750	Male	CareerNet	2	17	Selenium	Aspire Infinite Solutions And	17	460000	20-05-2018	350000	BU 6	2300
2300	Male	BMS Innolabs	4	29	Selenium	Test Mile Software Testing Pvt	29	1050000	03-04-2017	700000	BU 6	2625
2200	Male	BMS Innolabs	3	31	Java	Aptean India Pvt Ltd	31	725000	15-05-2017	525000	BU 7	2416
0	Male	Referral	3	35	Selenium	Genpact	35	750000	15-05-2017	650000	BU 6	2272
3600	Male	CareerNet	7	38	Selenium	Wipro Technologies	38	1500000	15-05-2017	1150000	BU 8	2027
2200	Male	CareerNet	4	40	AngularJS	ConnectM Technology	40	840000	11-04-2017	600000	BU 1	2333
0	Male	CareerNet	5	48	Java	Oracle	48	1300000	15-05-2017	830000	BU 7	2600

# 5. Systematic Random (Bucket Size is 10).

Component	Console	Summary	/ Result	Visualization	Properties	+	<u>+</u>
General	Sampling Inf	formation					
Properties	Sampling Type		Systematic Random	•			
	Bucket Size		10				
	Sample Size	Limit					
	Maximum Row	s	7				
•							
						Apply	y



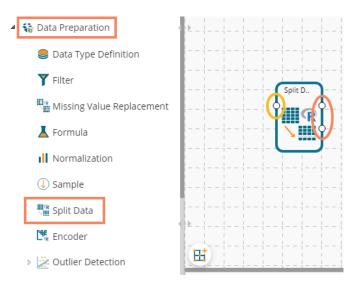
how 10 •	entries									Search:	
usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joining_date	previous_ctc	team
2600	Male	Orgspire	4	3	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000	18-07-1980	650000	BU 11
	Female	CareerNet	4	13	Selenium	Harman Connected Services	13	850000	08-03-2017	600000	BU 6
0	Male	CareerNet	3	23	Java	NTT Data	23	770000	17-04-2017	450000	BU 7
0	Male	Emuser	6	33	DotNet	CitiusTech Healthcare Technolo	33	1050000	15-05-2017	775000	BU 4
4000	Male	Referral	20	43	Java, Management	Trigent	43	2100000	31-03-2017	2750000	BU 7
2200	Male	CareerNet	3	53	DotNet	HP	53	950000	05-06-2017	700000	BU 4
4600	Male	Referral	16	63	Selenium, Management	TEK Systems Global Services	63	2800000	28-04-2017	0	BU 8

# 7.7. Split Data

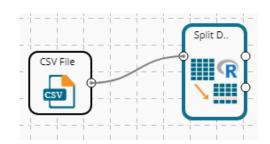
The Split Data component is used to split a dataset into training and testing per percentage and method. Once the most suitable model is decided from the trained data, users can pass test data to validate the model.

Split Data appears as a leaf node under the Data Preparation Tree node (the current description displays the Split Data component provided under the R Workspace).

The Split Data consists of two connector nodes: An upper node for the **training data set** and a lower node for the **testing data set**.



i) Select the '**Split Data**' component and connect it with a valid data source.





- ii) Click the 'Split Data' component in the workspace.
- iii) The user gets directed to the Properties fields provided under the 'Components' tab
- iv) The user can choose the size of the first partition:
  - a. Relative (train): Enter a value to decide the ratio of train data out of the dataset (Type: Decimal, Range: 0-1 and the sum of train and test data should be 1)
  - b. Relative (test): Enter a value to decide the ratio of train data out of the dataset (Type: Decimal, Range: 0-1 and the sum of train and test data should be 1)

Note: If the user does not want to configure the Advanced tab then the 'Apply' option provided for the '**Properties**' tab must be clicked, otherwise click the 'Apply' option provided for the Advanced tab.

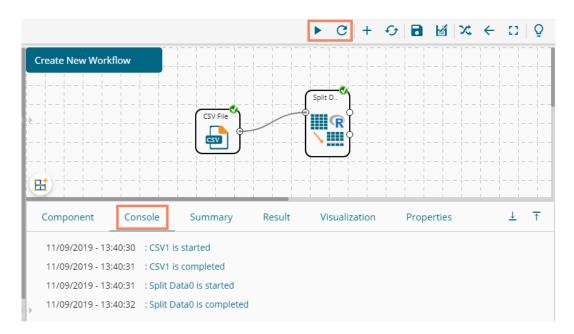
Component	Console	Summary	Result	Visualiza	tion	Properties	<u>+</u>	. ↑
General	Choose Split	Ratio						
Properties	Train Split Ratio	0	.7		0			
Advanced	Test Split Ratio	0	.30					
							Ар	ply

- v) The user can configure the sampling type using the Advanced fields if needed.
  - a. Sampling Type: Select any one option from the drop-down menu
    - i. Linear Sampling
    - ii. Shuffled Sampling
    - iii. Stratified Sampling
- vi) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	+	Ŧ
General	Sampling Ty	pe					
Properties	Sampling Type		near Sampling	•			
Advanced			Search				
			<ul> <li>Linear Sam</li> </ul>	pling			
			Shuffled Sampl	ing			
			Stratified Samp	ling			
•		_					
						Apply	y



- vii) Run the workflow after clearing the cache.
- viii) The 'Console' tab opens displaying the progress of the process. The completion of the Console process gets marked by the green checkmarks on the top of the dragged components.



- ix) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component in the workspace.
  - **b.** Click the 'Result' tab.

The Result tab displays two data sets separated by a sub-tab. As shown in the below-given images:

i. Select the 'Split 1' tab to see one set of data (the training dataset)

Split 1 Split 2	ies		Search:	
sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa



	Console Summary	Result Visua	lization Propertie	25 🛟
Split 1   Split 2     Show   10 <ul> <li>entri</li> </ul>	es		Search:	
sepal_length	sepal_width	petal_length	petal_width	species
7.6	3	6.6	2.1	virginica
4.9	2.5	4.5	1.7	virginica
7.3	2.9	6.3	1.8	virginica
6.7	2.5	5.8	1.8	virginica
7.2	3.6	6.1	2.5	virginica
6.5	3.2	5.1	2	virginica
6.4	2.7	5.3	1.9	virginica
6.8	3	5.5	2.1	virginica
5.7	2.5	5	2	virginica
5.8	2.8	5.1	2.4	virginica

#### ii. Select the 'Split 2' tab to see another set of data (the testing dataset)

Note:

- a. The current document covers steps to deal with a CSV File dataset for all the R Data Preparation components. Similar steps can be followed for a Data Service data set.
- b. The Data Preparation list may vary based on different workspaces, but the configuration process remains the same. All the unique Data Preparation components are explained under this section.

#### 7.8. Encoder

Encoding operation determines the existence of a string value in a selected column within each row in a worksheet. It converts categorical values in a worksheet to numeric values (only zero and one) required by machine learning algorithms.

i) Drag the Encoder component and connect it with a configured data source.



- ii) Click the Encoder component to configure the Properties tab:
  - a. Column Selection
    - i. Feature: Select a column using the drop-down option. All the string value columns get listed.
  - b. Output Data Behaviour
    - i. Remove First Dummy: Select an option from the drop-down menu (out of True/False)
    - ii. Remove Most Frequent Dummy: Select an option from the drop-down menu (out of True/False)
    - iii. Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	<del>+</del>	<u>+</u>
General	Column Sel	lection					
Properties	Features		1 checke	d 👻			
	Output Dat	a Behaviour					
	Remove First	Dummy	False	-			
	Remove Most Frequent Dun		False	•			
•							
						Apply	/

- iii) Run the workflow.
- iv) The Console tab opens displaying the process. The completion of the Console process gets marked by the green marks on the top of the dragged components.

				► C
	Freate New Workflow			
E	Ħ	CSV File		coder F
	Component	nsole	Summary	Result
	11/09/2019 - 16:14:59			
	11/09/2019 - 16:15:00		1.1	
•	11/09/2019 - 16:15:00	: Encode	er1 is started	
	11/09/2019 - 16:15:01	: Encode	er1 is completed	

- v) Open the Result tab to see the processed data.
  - a. Click the Encoder component.
  - b. Click the 'Result' tab to open the Result view.(The data of the selected column gets displayed by the 0 and 1 numbers)



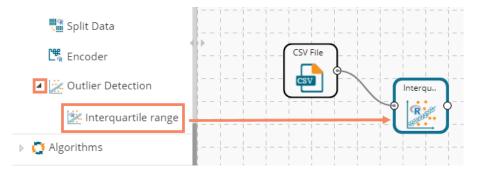
how 10 🔻	entries					Search:	
sepal_length	sepal_width	petal_length	petal_width	species	species_setosa	species_versicolor	species_virginic
5.1	3.5	1.4	0.2	setosa	1	0	0
4.9	3	1.4	0.2	setosa	1	0	0
4.7	3.2	1.3	0.2	setosa	1	0	0
4.6	3.1	1.5	0.2	setosa	1	0	0
5	3.6	1.4	0.2	setosa	1	0	0
5.4	3.9	1.7	0.4	setosa	1	0	0
4.6	3.4	1.4	0.3	setosa	1	0	0
5	3.4	1.5	0.2	setosa	1	0	0
4.4	2.9	1.4	0.2	setosa	1	0	0
4.9	3.1	1.5	0.1	setosa	1	0	0

# 7.9. Outlier Detection

This component is used to discover patterns in the data set that do not follow the expected behavior. It lists the outlying values based on the statistical distribution between the first and third quartiles. Interquartile Range has been provided as a sub-algorithm type.

### 7.9.1. Interquartile Range

i) Drag the Interquartile Range component to the workspace and connect it to a configured data source.



- ii) Configure the following fields in the 'Properties' tab:
  - a. Output Information
    - i. Output Mode: Select a mode of display for output data.
      - 1. **Show Outlier**: Select this option to add a Boolean column to the input data identifying whether the Resultant value is an outlier.
      - 2. Remove Outlier: Select this option to remove outlying values from the input data.
  - b. Column Selection
    - i. Feature: Select an input column that can be used to perform the analysis.
  - c. Behavior
    - i. **Fence Coefficient:** Enter the permissible deviation limit for values from the Interquartile Range (The default value for this field is 1.5)



#### d. New Column Information

i. New Column Name: Enter a name for the new column containing the predicted values (This column appears only when 'Show Outliers' is selected as an Output Mode).

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> ↑	<u>+</u>
General	Output Infor	mation					
Properties	Output Mode		Show Outliers	•			
Advanced	Column Sele	ction					
	Feature		Wind_speed	•	0		
	Behavior						
	Fence Coefficie	nt	1.5		0		
	New Columr	n Informatio	n				
	New Column N	ame	OutliersDetected1		0		
						Apply	

Properties fields with the 'Remove Outliers' option selected to display Output Information.

Component	Console	Summary	Result	Visualization	Properties	<del>+</del>	Ŧ
General	Output Infor	rmation					
Properties	Output Mode		Remove Outliers	•			
Advanced	Column Sele	ection					
	Feature		Wind_speed	•	0		
	Behavior						
	Fence Coefficie	nt	1.5		0		
•							
						Apply	y

Note: If the user does not need to configure the '**Advanced'** tab, then the 'Apply' option must be clicked from the Properties tab.

- iii) Click the 'Advanced' tab and configure if required:
  - a. Input Data Handling
    - i. **Missing Values**: Select a method to deal with missing values from the drop-down menu.
      - 1. **Ignore**: Select this option to skip the records containing missing values in the columns.
      - 2. **Stop**: Select this option to stop the application of the algorithm if a value is missing in any column.
- iv) Click the **'Apply'** option.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Input Data I	Handling					
Properties	Missing values	;	Ignore	•			
Advanced			Search				
			✓ Ignore				
			Stop				
•				_			
						Analy	
						Apply	

- v) Run the workflow after clearing the cache.
- vi) The 'Console' tab opens, displaying the process. The completion of the Console process gets marked by the green checkmarks on the top of the dragged components.



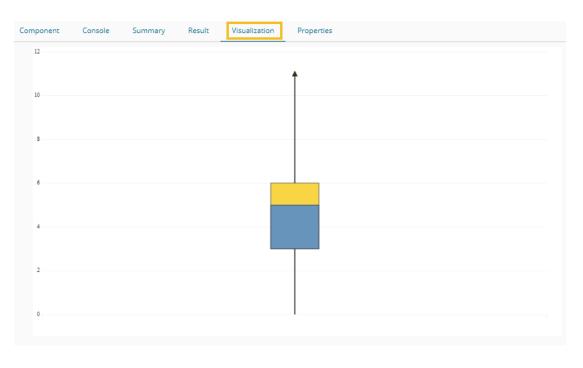
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged Outlier component.
  - **b.** Click the 'Result' tab.

A new column '**OutliersDetected1'** displays in the Result data (If '**Show Outliers**' option has been selected).



ow 10 🔻 en	tries						Search:	
Wind_speed	Humidity	Temperature_Sandburg	Temperature_ElMonte	Inversion_base_height	Pressure_gradient	Inversion_temperature	Visibility	OutliersDetect
8	20			5000	-15	30.56	200	FALSE
6		38			-14		300	FALSE
4	28	40		2693	-25	47.66	250	FALSE
3	37	45		590	-24	55.04	100	FALSE
3	51	54	45.32	1450	25	57.02	60	FALSE
4	69	35	49.64	1568	15	53.78	60	FALSE
6	19	45	46.4	2631	-33	54.14	100	FALSE
3	25	55	52.7	554	-28	64.76	250	FALSE
3	73	41	48.02	2083	23	52.52	120	FALSE
3	59	44		2654	-2	48.38	120	FALSE

- viii) Click the 'Visualization' tab.
- ix) The Result data is displayed via the Box Plot chart.



OR

The outliers column is removed from the Result data (If the '**Remove Outliers**' option has been selected).



ow 10	▼ entries								Search:	
f_week	ozone_reading	pressure_height	Wind_speed	Humidity	Temperature_Sandburg	Temperature_ElMonte	Inversion_base_height	Pressure_gradient	Inversion_temperature	Visibili
	3.01	5480	8	20			5000	-15	30.56	200
	3.2	5660	6		38			-14		300
	2.7	5710	4	28	40		2693	-25	47.66	250
	5.18	5700	3	37	45		590	-24	55.04	100
	5.34	5760	3	51	54	45.32	1450	25	57.02	60
	5.77	5720	4	69	35	49.64	1568	15	53.78	60
	3.69	5790	6	19	45	46.4	2631	-33	54.14	100
	3.89	5790	3	25	55	52.7	554	-28	64.76	250
	5.76	5700	3	73	41	48.02	2083	23	52.52	120
	6.94	5700	3	59	44		2654	-2	48.38	120



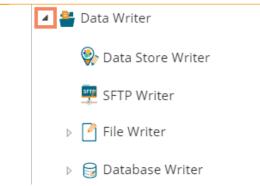


# 8. Data Writers

Data Writers are provided to store the Results of the Data Science Workspace in flat files or databases for further in-depth analysis. The Data Science Workspace contains the following types of Data Writers across the various Workspaces.

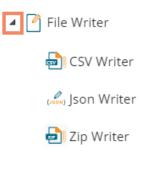
- 1. Data Store Writer
- 2. SFTP Writer (only available for the Python Workspace at present)
- 3. File Writer
- 4. Database Writer





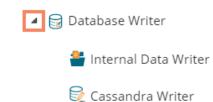
The File writer has the following categories:

- 1. CSV Writer
- 2. JSON Writer
- 3. Zip Writer (only for the Python Workspace)



The Database Writer has the following categories:

- 1. Internal Data Writer
- 2. Cassandra Writer



Find the step by step description for each data writer given below:

### 8.1. Data Store Writer

The elastic Search Writer component is listed under the Data Writer Tree node. The Data Store Writer allows the user to write the processed data onto the Elastic Search server, which makes it more distributed.

i) Drag the Data Store Writer component to the workspace and connect it with a configured data source or any valid combination of a data source with other given components. (In this case, there is a combination of CSV file with a Missing Value Replacement and Encoder components to bring the input data to the Data Store writer)





- ii) Click on the connected Data Store Writer component.
- iii) The component tab for the data writer opens.
- iv) Configure the required component properties.
  - i. Select Data Store: Select a datastore from the drop-down menu or select the 'Create New Data Store' option from the drop-down menu
  - Select Operation Type: This field appears by choosing an existing Data Store. Select an option from the drop-down menu (Overwrite/Append/Upsert).
     OR

**Data Store Name**: This field appears by choosing the 'Create New Data Store' option. The user can define a name for the data store.

- iii. **Select Insertion Type**: Select an insertion type from the drop-down menu (Full Insertion/Batch Insertion)
- iv. The user gets all the Dimensions, Measures, and Time fields from the selected data source.
- v. They can define hierarchy by dragging the required Dimensions using the 'Drill Definition' box.
- v) Click the 'Apply' option.

Component	Console Summary Result Visualization Properties	<u>+</u> <u>↓</u>
General	Data Store Writer Properties	
Properties	Select Data Store Create New Data Store	
	Data Store Name Data Store Writer	
	Select Insertion Type Full Insertion	
	Dimensions Hierarchy Definition +	
	previous_organisation	
	referral_of	
	designation	
	joining_status	
	E E	
	skills O	
		Apply

Note: If the selected insertion type is '**Batch Insertion**,' the Properties configuration displays the '**Next**' option.



Component Co	nsole Summary Result	Visualization Properties	+	<u>+</u>
General Properties	Data Store Writer Properties Select Data Store	Data Store Writer		
Advanced	Select Operation Type Select Insertion Type	Overwrite           Batch Insertion		
>	Dimensions previous_organisation referral_of designation joining_status skills team	Hierarchy Definition + Drill Definition - 1 × team × designation × name × Control Control C		
			Nex	t

The user gets redirected to the '**Advanced**' fields to configure the Batch Query Properties. Select and then click the 'Apply' option as displayed in the following image:

Component	Console	Summary	Result	Visualizati	ion	Properties		*	<u>+</u>
General	Batch (	Query Properties							
Properties		Select Dimension fo	or Batch Query		Select Query	ed Dimension for Batch	0		
Advanced		previous_organisat	ion		team				
		referral_of							
		designation							
		joining_status							
		source		+					
		name							
		current_status							
		gender							
>		skills					J		
								Apply	1

The user can move only one dimension at a time from the list of 'Select Dimension for Batch Query' value for the batch query.

- vi) Run the workflow after getting the success message.
- vii) Users will get the process status under the 'Console' tab. The completion of the process is marked with green checkmarks on the components.





viii) The data will be saved in the desired format to the selected Data Store Writer after the Console process gets completed.

Note:

a. The user also gets the 'General' fields for the Data Store Writer component, but they need not configure it.

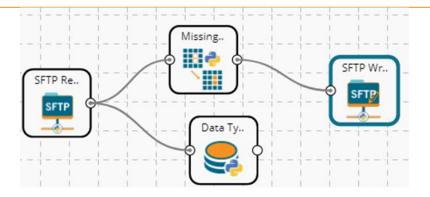
Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Basic						
Properties	Component N	Name	Data Store Writer				
Advanced	Alias		Data Store Writer	2			
P	Description		Optional				
						Next	

### 8.2. SFTP Writer

The SFTP Writer is available under the Python Workspace to write the processed data securely.

i) Drag and drop the SFTP writer to the workspace and connect it to the configured combination of the data source and other relevant components to create a workflow.



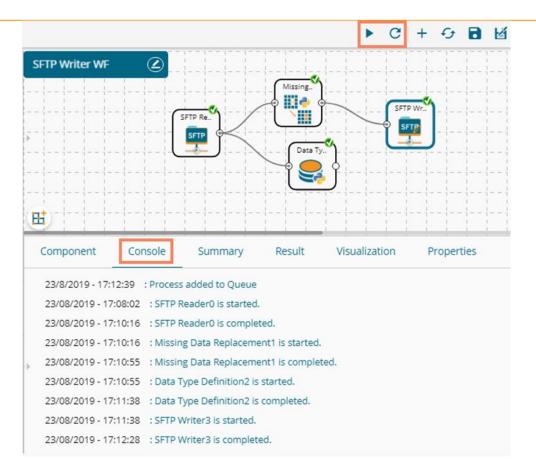


- ii) Click on the writer to get the configuration fields.
- iii) Fill in the required details to configure the properties of the SFTP Writer.
  - **a.** Host address
  - b. Port Number
  - c. Username
  - d. Password
  - e. Remote path
  - f. Data to Write
  - g. Name
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	+	Ŧ
General	Configure S	FTP Writer					
Properties	Host						
	Port						
	User name		ftpuser				
	Password						
	Remote path		Second Second				
	Data to Write		2 checked	•			
	Name		dfmdlkjldfg		0		
						Apply	

- v) Run the workflow after getting the success message.
- vi) The stepwise completion of the process gets displayed in the 'Console' tab. The completion of the Console process is marked by the green checkmarks on the top of the components.





vii) The processed data gets written at the configured SFTP file/location through the SFTP writer.

#### 8.3. File Writer

The user can write output data to flat files like CSV, TEXT, and DAT files using the File Writer.

#### 8.3.1. CSV Writer

i) Drag and drop the **CSV Writer** component and connect it to a configured workflow to get the input.

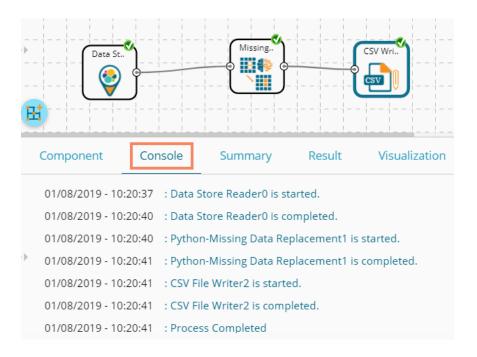




- ii) Click on the CSV Writer component to access component properties.
- iii) Enter the 'File Name' in the displayed field.
- iv) Click the 'Apply' option.

, Ħ	Data St.		Missing.	CSV Wri	)	
Component	Console	Summary	Result	Visualization	Properties	<u>↓</u> T
General	File Name					
Properties	Name		CSV Writer			
						Apply

- i) Run the workflow after getting the success message.
- v) The process status gets displayed under the 'Console' tab, and green checkmarks get displayed a the top of the dragged components indicating completion of the process.





- vi) The data gets written in the CSV File.
- vii) Click the 'CSV Writer' component.
- viii) A pop-up message appears with a link to download the CSV file.

N	Download File	×
-	Click Here to download CSV Writer.csv	

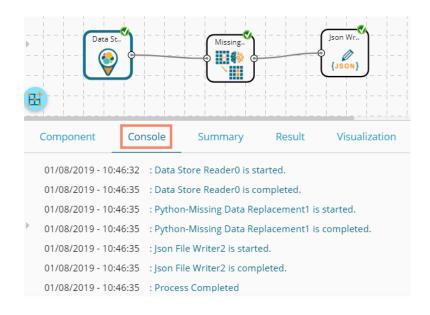
ix) Click the link to download the CSV file.

### 8.3.2. JSON Writer

- ii) Drag and drop the **'JsonWriter'** component to the workspace and connect it to the configured workflow to get input.
- iii) Click on the 'JsonWriter' component to access component properties.
- iv) Enter 'File Name' in the displayed space.
- v) Click the 'Apply' option.

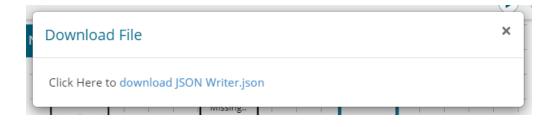
Component	Console	Summary	Result	Visualization	Properties	± T
General	File Name					
Properties	Name		JSON Writer			
*						
						Apply

- vi) Run the workflow after getting the success message.
- vii) The process status gets displayed under the 'Console' tab, and the completion of the process gets marked by green checkmarks on the dragged components.





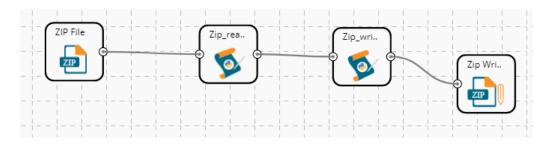
- viii) A pop-up message appears with a link to download the **JSON** file.
- ix) Click the link to download the JSON file.



### 8.3.3. ZIP Writer

This data writer helps the user to write the processed data into a Zip file.

Drag the 'Zip Writer' from the Data Writer tree-node and connect it to a configured data source and other relevant components to create a workflow.
 (The Zip writer requires a relevant script to write the data in the specified Zip file.)



- ii) Click the dragged Zip Writer component to get the configuration fields.
- iii) Provide the file name to configure the Zip Writer properties.

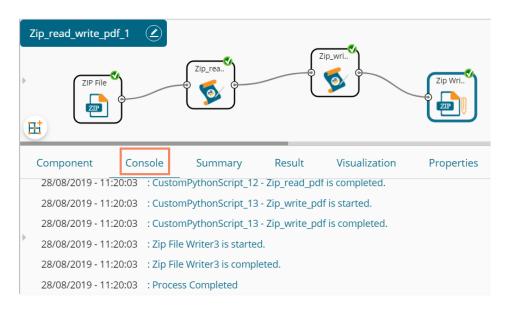
Component	Console	Summary	Result	Visualization	Properties	⊥ ⊺
General	File Name					
Properties	Name		Zip_pdf_prod			
•						
						Apply

- iv) Run the workflow after getting a success message.
- v) The Console process displays the process step by step.



	Com	ponent	t C	Console	Su	mmary	y Result	Visualization	Properties
	28/8	/2019 -	- 11:20:18	B : Proce	ess addec	l to Qu	eue		
Þ	28/0	8/2019	9 - 11:20:0	)3 : ZIPC	) is starte	d.			
	28/0	8/2019	9 - 11:20:0	)3 : ZIPC	) is comp	leted.			
	28/0	8/2019	9 - 11:20:0	)3 : Cus	tomPytho	onScrip	ot_12 - Zip_read_p	df is started.	
	28/0	8/2019	- 11:20:0	3 : 1	Number	sepal_l	ength sepal_wid	th petal_length petal_	width species
	0	1	5.1	3.5	1.4	0.2	setosa		
	1	2	4.9	3.0	1.4	0.2	setosa		
	2	3	4.7	3.2	1.3	0.2	setosa		
	3	4	4.6	3.1	1.5	0.2	setosa		
	4	5	5.0	3.6	1.4	0.2	setosa		
	5	6	5.4	3.9	1.7	0.4	setosa		
	6	7	4.6	3.4	1.4	0.3	setosa		
	7	8	5.0	3.4	1.5	0.2	setosa		
	8	9	4.4	2.9	1.4	0.2	setosa		
ľ	9	10	4.9	3.1	1.5	0.1	setosa		
	10	11	5.4	3.7	1.5	0.2	setosa		
	11	12	4.8	3.4	1.6	0.2	setosa		
	12	13	4.8	3.0	1.4	0.1	setosa		
	13	14	4.3	3.0	1.1	0.1	setosa		

vi) The completion of the success process gets indicated by a green checkmark on the top of all components in the selected workflows.



- vii) The processed data gets written in a Zip file through the Zip Writer.
- viii) After the process gets completed, click on the Zip writer component from the workflow.
- ix) The 'Download File' dialog box appears to download the Zip file.
- x) The user can download the Zip file by clicking on the link mentioned in the dialog box.

	-	$\mathbf{\circ}$	
Download File		×	
Click Here to download Zip_pdf_prod.zip	_		

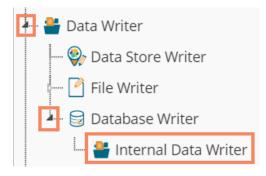


### 8.4. Database Writer

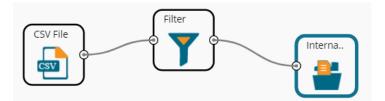
### 8.4.1. Internal Data Writer

The user can store data in databases like MySQL, MSSQL, and Oracle by Internal Data writer.

- i) Click the **'Data Writer'** tree node option.
- ii) Select the 'Database Writer' option.
- iii) Select and drag the 'Internal Data Writer' component to the workspace.



iv) Drag and Connect the 'Internal Data Writer' component to a configured data source and other related components to create a workflow.



- v) Click the 'Internal Data Writer' component to access the Component properties. The user gets different 'Properties' fields based on the selected table operation as described below:
   a. Selecting the 'Create a New Table' option as the 'Table Operation':
  - i. **Data Connector Name**: All the available data connectors in particular user id get listed. Select a data connector from the drop-down menu.
  - ii. **Type**: This field is preselected based on the selected data Connector.
  - iii. **Number of Rows in a batch**: Enter a number to limit the entries of rows for one batch
  - iv. Database Name: Select a database name from the drop-down menu
  - v. **Password:** Enter the database password
  - vi. Table Name: Select the 'Create New Table' option from the list
  - vii. Table Operation: Select an option from the drop-down menu
  - viii. Create a New Table: It is an optional field. It appears when the user selects the 'Create New Table' option from the 'Table Name' drop-down menu.
  - ix. **Auto Increment:** Select an option to enable or disable the auto-increment. By enabling this option, a new column gets added to the dataset, and the same column gets selected as the primary key by default.
  - x. Auto Increment Label: Enter a name for the auto-increment label



- xi. **Column Selected from a model**: Select columns that are needed to be written into the selected database.
- vi) Click the '**Next**' option.

Component	Console Summary	Result Visualization	Properties	÷ .
General	Internal Data Writer Pro	perties		
Properties	Data Source Name	predictive_prod -	G	
Schema Viewer	Туре	mysql		
	Number of Rows in a	1000	0	
	batch			
	Database Name	predictive_analysis -		
	Password			
	Table Name	Create New Table 🗸		
	Table Operation	Upsert -		
	Create New Table	RCNRWF	9	
•	Auto Increment	Disable -		
	Column selected from	7 checked 🗸		
	model			
				Next

- vii) The user gets the 'Schema Viewer' tab to select the primary keys.
- viii) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	⊥ ⊺
General	Internal Dat	a Writer Prop	perties			
Properties	Select Primary	Keys	Select	-		
Schema Viewer						
						Apply

Note: The selected Auto Increment Label appears as the selected Primary Keys by default, if the '**Auto Increment**' option is enabled.

- 1. Enable the 'Auto Increment' option from the 'Properties' tab.
- 2. Click the 'Next' option.



Component	Console Summary	Result Visualization	Properties	+	Ŧ
General	Internal Data Writer Prope	erties			
Properties	Data Source Name	predictive_prod -	Ð		
Schema Viewer	Туре	mysql			
	Number of Rows in a batch	1000	0		
	Database Name	predictive_analysis -			
	Password				
	Table Name	Create New Table			
	Table Operation	Append to Table -			
	Create New Table	Internal_DW_new	0		
	Auto Increment	Enable -			
•	Auto Increment Label	AIL	•		
	Column selected from	7 checked -			
	model				_
				Next	i

- 3. The Schema Viewer tab opens.
- 4. The configured Auto Increment Label gets selected as a Primary Key by default.
- 5. Click the 'Apply' option to save the configuration.

Component	Console	Summary	Result	Visualization	Properties	<del>*</del>	Ţ
General	Internal Data	a Writer Proj	perties				
Properties	Select Primary I	Keys 1	checked	•			
Schema Viewer			previous_organisa	ation			
			id				
			team				
			monthly_salary				
			name				
•			designation				
			🖌 AIL				
						Apply	/

#### b. Selecting an Existing Table as the 'Table Operation':

- i. Data Connector Name: Select a data connector from the drop-down menu
- ii. Type: Displays a type based on the selected data connector
- iii. Number of Rows in a batch: Enter a number to limit the entries of rows for one batch
- iv. Database Name: Select a database name from the drop-down menu
- v. Password: Enter the database password
- vi. Table Name: Select an existing table name from the drop-down menu
- vii. **Table Operation**: Select an option using the drop-down menu. The following are the provided choices:



- 1. Append Table
- 2. Overwrite Table
- viii. **Column Selected from a model**: Select columns that are needed to be written into the selected database.

Component	Console Summary	Result Visualiza	tion Properties	+	<u>+</u>
General	Internal Data Writer Pro	operties			
Properties	Data Source Name	predictive_prod	• <i>f</i>		
Schema Viewer	Туре	mysql			
	Number of Rows in a	1000	0		
	batch				
	Database Name	predictive_analysis	•		
	Password				
	Table Name	InternalDW_new	•		
	Table Operation	Overwrite Table	•		
	Column selected from	7 checked	•		
	model				

ix. **Details of the Selected table**: Displays column headers from the selected table. ix) Click the **'Next'** option.

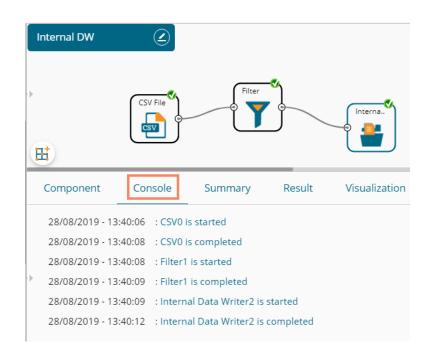
The internal data writer can extract new or changed records while loading data from the MySQL database. The Schema View tab has been added to the internal database writer to extract data using the delta data load type.

x) The Schema Viewer tab opens displaying the selected Primary Keys (in this case, no Primary Keys is selected).

Component	Console	Summary	Result	Visualization	Properties	± ∓		
General	Internal Da	ata Writer Prope	erties					
Properties		Selected Primary Keys						
Schema Viewer		No prin	nary keys Fo	r this table				
r								
						Apply		



- xi) Click the 'Apply' option.
- xii) Run the Workflow after getting the success message.
- xiii) The progress of the process gets displayed in the 'Console' tab, and the completion of the process gets marked by the green tick marks on the dragged components.

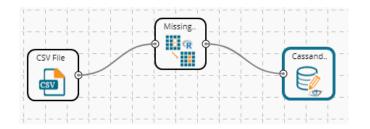


xiv) The processed data gets saved in the selected database.

### 8.4.2. Cassandra Writer

Cassandra Writer can be used to store the data science executions.

- i) Open the 'Database Writer' tree node.
- ii) Select and drag the 'Cassandra Writer' component to the workspace.
- iii) Connect the Cassandra Writer to a configured data source or relevant components to create a Workflow.



- iv) Click the 'Cassandra Writer' component to access it.
- v) Configure the following Properties details:
  - a. Selecting Create New Table as a Table option
    - i. Select Data Connector: Select a data connector using the drop-down menu
    - ii. **Host Name**: Based on the chosen data connector a hostname gets displayed (the user cannot edit this field)
    - iii. Port Name: The server port number gets displayed (the user cannot edit this field)



- iv. **Username**: The username of the selected connection appears by default. (the user cannot edit this field)
- v. Password: the database password
- vi. No. of rows in a batch: Enter a number to limit the entries of rows for one batch
- vii. Select Key Space: Select a keyspace using the drop-down menu
- viii. **Replication Factor**: The replication factor mentioned in the selected '**Key Space**' get displayed (the user cannot edit this field)
- ix. Select Table: Select the 'Create a New Table' option from the drop-down list
- x. Select Columns: Select the columns that you want to write
- xi. Consistency: Select an option from the drop-down list
- xii. New Table: Provide a name for the newly created table
- xiii. New time UUID column name: Enter a UUID column name
- xiv. Click the **'Next'** option.

Component	Console Summary	Result Visualization	Properties	<u>+</u> ⊥
General	Data Service Properties			
Properties	Select Data Connector	Cassandra_06	•	
Key Specification	Host name	cassandra-0.cassandra-prod		
	Port Number	9042		
	Username	cassandra		
	Password			
	No: of rows in a batch	100	0	
	Select Key Space	pa	•	
	Replication Factor	2		
	Select Table	Create new table	•	
	Select columns	7 checked	•	
	Consistency	TWO	•	
•	New table	Cassandra_Writer_New		
	New time uuid column	uuid		
	name			
				Next

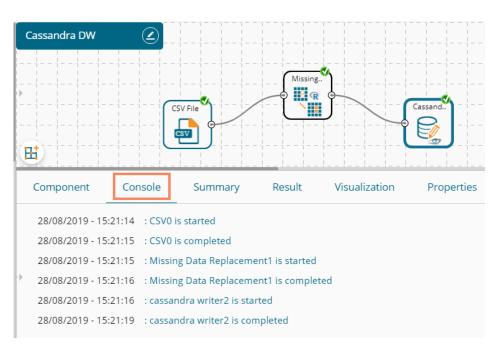
- vi) The 'Key Specification' tab opens.
- vii) Configure the following information:
  - a. **Headers**: All the columns from the data set get listed.
  - b. **Partition Key (Name)**: The Partition Key determines which node stores the data. It is responsible for data distribution across the nodes.
    - The UUID Column name gets displayed under the 'Partition Key' window.
    - The user can select and move any column from 'Header' (Select Column) to 'Partition Key' space.
    - The sequence of the columns listed under Partition Key can be arranged by using the '**Up**' or '**Down**' options.
  - c. **Clustering Key**: The Clustering Key is a storage engine process that sorts data within the partition. It determines per-partition clustering.
    - The items listed under the Clustering Key box can be arranged by using the '**Up**' or '**Down**' options.
    - The user can select any column from 'Headers' (Select Column) to the 'Clustering Key' space.



#### viii) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties		+	<u>+</u>
General	He	aders			Partition Key			
Properties		team			Name uuid			
Key Specification		skills		>	uulu			Up
		name		<				Down
		current_status						
		designation						
		joining_status			Clustering Key			
		referral_of			Name	Order		
				>				Up
				<				Down
ľ								
							Арр	oly

- ix) Run the workflow after getting the success message.
- x) The step-by-step process gets displayed under the Console tab. The completion of the process gets marked by the green checkmarks.



- Note: The user gets some defined consistency levels while defining the KeySpace, which can be overridden based on the selected replica nodes. The user gets the following options for the Consistency field:
  - One
  - Two
  - Three
  - Quarum



#### b. Selecting an Existing Table as Table Operation

Configure the following **Properties** details:

- i. Select Data Connector: Select a data connector from the drop-down menu
- ii. Host Name: Enter database server details (from where the user wants to fetch data)
- iii. Port Name: The server port number
- iv. **Username**: Username of the selected connection appears by default (Users cannot edit this field)
- v. **Password**: the database password
- vi. No. of rows in a batch: Enter a number to limit the entries of rows for one batch
- vii. Select Key Space: Select a keyspace using the drop-down menu
- viii. **Replication Factor**: Replication factor in the selected '**Key Space**' gets displayed (Users cannot edit this field)
- ix. Select Table: Select a table from the drop-down menu
- x. **Choose Columns**: Select columns from the drop-down menu that users want to be written in the data writer.
- xi. Consistency: Select an option using the drop-down menu
- xii. **Settings**: Select an option using the drop-down menu The following choices are provided:
  - 1. Append Table (to select an existing table the selected settings option should be Append)
  - 2. Overwrite Table

Component	Console	Summary	Result	Visualization	Properties	÷ 1
General	Data Service	Properties				
Properties	Select Data Con	nector	Cassandra_06	-		
Key Specification	Host name		cassandra-0.cass	andra-prod		
	Port Number		9042			
	Username		cassandra			
	Password		•••••			
	No: of rows in a	batch	100		0	
	Select Key Spac	e	ра	-		
	Replication Fact	or	2			
	Select Table		Cassandra_Writer	_new 👻		
•	Select columns		7 checked	•		
·	Consistency		ONE	-		
	Settings		Append	*		

- xiii. The list of column headers existing in the table gets displayed once the user selects an existing table.
- xiv. Click the 'Apply' option.



#### OR

Configure the Key Specification settings and click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties		+	<u>+</u>
General	He	eaders		[	Partition Key			
Properties		skills			Name team			
Key Specification		name current_status designation		~	com			Up Down
		joining_status						
		referral_of			Clustering Key			
•				~	Name	Order		Up Down
							Ар	ply

- xi) Run the workflow after getting the success message.
- xii) The process status gets displayed under the '**Console**' tab. The completion of the Console process gets marked by the green checkmarks on the dragged components.
- xiii) The data gets saved in the selected Cassandra Writer.

## 9. Saved Workflows

The user can save a workflow by clicking the '**Save**' con provided on the workspace menu row. All the saved Workflows the selected Workspace gets listed under the '**Saved Workflow**' tree node. This section explains various options assigned to a saved workflow.

- i) Navigate to any Data Science Workspace (in this case, the Python Workspace has been selected).
- ii) Click the 'Saved Workflow' tree node.
- iii) Select a saved workflow from the list and use a right-click on it.
- iv) A context menu opens with various options (As shown below):



Dat	Data Science Workbench						
=	C Search Tree	٩					
<b>2</b> ⊡Ľ	Saved Workflows	-					
3	🗋 jupyterWorkflow						
	Open	-					
	Delete	-					
	Rename	-					
4	Auto Save	or E					
	Share	or E					
	Publish as Service	-					
	Pull from VCS	-					
	Push into VCS						

## 9.1. Opening a Workflow

- i) Select a workflow from the list of **Saved Workflows** and use a right-click on it.
- ii) Select the '**Open**' option from the context menu.
- iii) The selected workflow gets displayed in the right pane of the screen.

≡ C Search Tree	jupyterWorkflo	w 🖉								$ \frac{1}{1} = -\frac{1}{1} $ $ \frac{1}{1} = -\frac{1}{1} $ $ \frac{1}{1} = -\frac{1}{1} $
<ul> <li>Saved Workflows</li> <li>jupyterWorkflow</li> </ul>			CSV File		Jupyter.					$\frac{1}{1} = -\frac{1}{1}$ $\frac{1}{1} = -\frac{1}{1}$ $\frac{1}{1} = -\frac{1}{1}$ $\frac{1}{1} = -\frac{1}{1}$
Open Delete										
Rename Auto Save	or Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u></u>	Ŧ	T
Share Publish as Service Pull from VCS	or E									
Push into VCS										

**Note:** The workflow name gets displayed on the left side of the workspace menu row while opening a workflow.

## 9.2. Deleting a Workflow

- i) Select a workflow from the list of **Saved Workflows** and use a right-click on it.
- ii) Select the 'Delete' option from the context menu.



Saved Workflows						
	) jupyterWorkflow					
	Open					
2	Delete					
	Rename					
	Auto Save					
	Share					
	Publish as Service					
	Pull from VCS					
	Push into VCS					

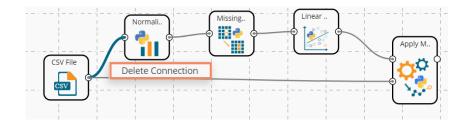
- iii) A dialog box appears to confirm the deletion.
- iv) Click the '**Ok**' option.

Delete work flow		×
Do you want to delete selected workflow ?		
	4 Ok	Cancel

v) The selected workflow gets removed from the list.

#### 9.2.1. Delete Connection in a Workflow

A Right-click on the inter-node connection displays the '**Delete Connection**' option in the workflow. Click the '**Delete Connection**' option to delete a connection.



### 9.3. Renaming a Workflow

- i) Select a workflow from the list of **Saved Workflows** and use a right-click on it.
- ii) Select the 'Rename' option from the context menu.



4 🖺 S	Saved Workflows
	🗋 jupyterWorkflow
	Open
	Delete
2	Rename
	Auto Save
	Share
	Publish as Service
	Pull from VCS
	Push into VCS

or

Open a Saved Workflow.

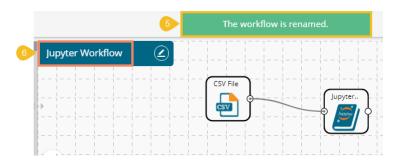
Click the '**Rename**' icon provided next to the workflow name.



- iii) The Rename Workflow window opens.
- iv) Enter a new/modified name for the workflow.
- v) Click the 'Yes' option.

3 Rename Wo	orkflow			×
4 Workflow nar Jupyter Wor				
		5	Yes	No

- vi) A success message appears.
- vii) The workflow gets renamed.





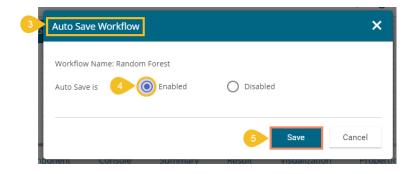
### 9.4. Auto-Save

The workflow gets auto-saved by enabling this option for a saved workflow.

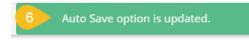
- i) Select a workflow from the list of Saved Workflows and use right-click on it.
- ii) Select the 'Auto Save' option from the context menu.

4 🖺	Saved Workflows	
0	🗋 Random Forest	
	Open	
	Delete	
	Rename	
2	Auto Save	
	Share	
	Publish as Service	i
	Pull from VCS	
	Push into VCS	

- iii) The 'Auto Save Workflow' window opens.
- iv) Select the 'Enabled' option by using the checkbox.
- v) Click the 'Save' option.



vi) A message appears to inform the user that the Auto-Save option is updated.



- vii) Open the Workflow.
- viii) Edit some Component information.
- ix) Click the 'Apply' option.



≡ C 👪 Search Tree Q	Random Forest						
A 🖺 Saved Workflov 7							
🗋 Random Forest	CSV File	Random					
Open Delete							
Rename							
Auto Save 8	Component Co	onsole Summary Res	ult Visualization	Properties	DataInsight	-	±⊤
Share							
Publish as Service	General	Output Information					
Publish as Service Pull from VCS	General Properties	Output Information Algorithm Type	Classification	•			l
			Classification	•			
Pull from VCS	Properties	Algorithm Type					

- x) A message confirms that the edited information has been applied.
- xi) Another message on the top appears to inform the user that the Workflow is auto-saved.

BBB®				۹ 🏢	🕂 🕐 🖤
Data Science Workbench	10 The workflow is	auto-saved.	▶ C +	€ B 🛛	$\mathfrak{F} \   \ \leftarrow \ \ \boxdot \   \ \check{O}$
= C 🔹 Search Tree Q Compone	ent Console Summary	Result Visualization	Properties	DataInsight	<u>+</u> <u>↓</u>
General	Output Information				
Random Forest	Algorithm Type	Classification	•		
Normalizer Advanced	Show Probability	True	•		
Normalization_Min Max	Column Selection				
Naive Bayes	Features	5 checked	• 0		
MVRcheck	Target Variable	gender	• 0		
Multi-Linear Regression_Iri	New Column Information				
Missing Value Replacement	Predicted Column Name	PredictedValues1	0		
🗅 Max Absolute scaler	Probability Column Name	Probability1	0		
Logistic Regression_Perforr		11 Apply Su	iccessful		

#### 9.5. Sharing a Workflow

The user can share a saved workflow with other users and groups through this option.

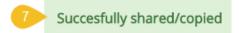
The following options are available to share a selected workflow:

- 1. **Share With**: This option allows the user to share a file with the selected users or user groups. Any changes made to the file get transferred to all the users with whom the file has been shared.
  - i) Select a workflow from the list of **Saved Workflows** and use right-click on it.
  - ii) Select the 'Share' option from the context menu.
  - iii) The 'Share With' option gets displayed (by default)
  - iv) Select either 'Group' or 'Users'
    - a. By selecting a group, all group members inside the group get listed. You can exclude the users by not selecting them from the group.
    - b. The users can also get excluded by not selecting a username from the list when the 'Users' option has been selected.
  - v) Select a specific group or user from the list by putting a checkmark in the given box.
  - vi) Click the 'Apply' option.



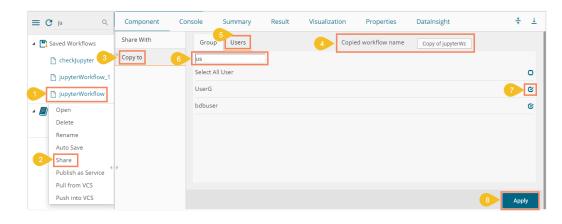
≡	C ju Q	Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u> <u>↓</u>
- E	Saved Workflows	Share With	4 Group	Users					
	CheckJupyter	Copy to	Search						
	🚹 jupyterWorkflow_1		Select A	ll Group					0
	🗋 jupyterWorkflow		Admin F	tole					5 🕑
	Open		Viewer	Role					<b>∨</b> 0
	Delete		PAPerm	issions					<b>∨</b> 0
	Rename		Data sci	entist					<b>v</b> 0
2	Auto Save Share		permiss	ions					<b>∨</b> 0
	Publish as Service	• >	PAGrou	p					<b>∨</b> 0
	Pull from VCS								
	Push into VCS								6 Apply

vii) A success message appears.



viii) The selected workflow gets shared with the chosen user(s)/group(s).

- 2. **Copy To**: This option creates a copy and shares the copy with the selected users and user groups. Any change to the original file after sharing does not display for the users that received the shared file via the '**Copy To**' method.
  - i) Select a workflow from the list of Saved Workflows and use right-click on it.
  - ii) Select the 'Share' option from the context menu.
  - iii) Select the 'Copy To' option.
  - iv) The Workflow name gets displayed with the 'copy of' prefix.
  - v) Select either 'Group' or 'Users'
    - a. By selecting a group, all group members inside the group get listed. The users can be excluded by not selecting them from the group.
  - b. The user can also get excluded by not selecting a username from the list when the 'Users' option has been selected.
  - vi) Use search space to search for a specific user.
  - vii) Select a specific group or user from the list by putting a checkmark in the given box.
  - viii) Click the '**Apply**' option.



ix) A success message appears.



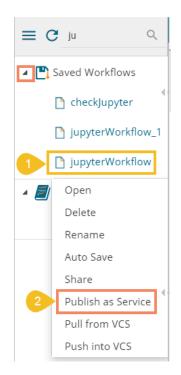


x) The copied workflow gets shared with the chosen users/groups.

#### 9.6. Publish a Workflow as Service

The Data Science workflow can be deployed to the BDB Dashboard Designer as a service.

- i) Select a Workflow from the list of **Saved Workflows** and use a right-click on it.
- ii) Select the '**Publish as Service**' option from the context menu.



- iii) A success message appears to assure that the workflow has been published.
- iv) The published workflows get marked by a checkmark (as displayed below).

Data Science Workbench	n 🥐	3	The workflow is published.	
E C ju Q	Create New Workflow		7 - 7 - 7 - 7 - 7 - 7 - 7 - 7 - 7 - 7 -	
Saved Workflows				
🎦 checkJupyter	A B A A A A A A A A A A A A A A A A			
🎦 jupyterWorkflow_1				
4 jupyterWorkflow			i i	

- v) Navigate to the Dashboard Designer homepage.
- vi) Click the 'New' option.
- vii) Click the 'Dashboard' option.



Simple Drag and Drop UI Drag any chart or grid from the Component Panel and drop to your designer canvas. It is incredibly intuitive and smooth functionality for designing dashboard.	Highly Interactive and Drill Down Each dashboard can have multiple widgets and filters which can be altered in a click to get data analytics at the speed of thought. SDK methods help to achieve the various functional	Mobile, Tablet and Desktop Views Customizable layouts to view the interactive dashbeards from phones, tablets, and desktops. Making the data insights accessible all the time for	
Drag any chart or grid from the Component Panel and drop to your designer canvas. It is incredibly intuitive and smooth functionality for designing	Each dashboard can have multiple widgets and filters which can be altered in a click to get data analytics at the speed of thought. SDK methods	Mobile, Tablet and Desktop Views Customizable layouts to view the interactive dashboards from phones, tablets, and desktops. Making the data insights accessible all the time for	
Drag any chart or grid from the Component Panel and drop to your designer canvas. It is incredibly intuitive and smooth functionality for designing	Each dashboard can have multiple widgets and filters which can be altered in a click to get data analytics at the speed of thought. SDK methods	Customizable layouts to view the interactive dashboards from phones, tablets, and desktops. Making the data insights accessible all the time for	
Drag any chart or grid from the Component Panel and drop to your designer canvas. It is incredibly intuitive and smooth functionality for designing	Each dashboard can have multiple widgets and filters which can be altered in a click to get data analytics at the speed of thought. SDK methods	Customizable layouts to view the interactive dashboards from phones, tablets, and desktops. Making the data insights accessible all the time for	
intuitive and smooth functionality for designing	analytics at the speed of thought. SDK methods	Making the data insights accessible all the time for	
	requirements based on the customer needs.		
60+ Components and Themes	Advanced Visualizations	Import Custom Component	
Check out our exclusive range of visualization components to select the most suitable for your data. You can govern the overlook and feel of the	Explore complex business data in the most simplified way using our advanced visualization components. Avail a series of interactive charting	Incorporate any third-party visualizations to your dashboard, including D3 Charts, DataTables etc.	
	Check out our exclusive range of visualization components to select the most suitable for your	Check out our exclusive range of visualization components to select the most suitable for your data. You can govern the overlook and feel of the components. Avail a series of interactive charting	

- viii) The Dashboard canvas opens.
- ix) Click the 'Data Connectors' icon  $\ge$  to display all the available data connectors.

BBB®	A Hor, 8 Untitled x +	
		<b>9</b>
		e
		t≡
		8
		Q
		_

- x) Click the '**Create New Connection**' option + provided Next to the '**Data Science Models'** option on the Data Connector page.
- xi) A new connection gets created and added below.
- xii) The connection-specific details get displayed on the right.
- xiii) Select the deployed Data Science workflow as a data source via the drop-down menu.
- xiv) After selecting the Data Science Workflow the FIELD SET tab displays the available fields.

BBB@		Home Untitled 🗙	+			
CSV	- <mark>12</mark> +	Name	Connection	-1		() × (
Excel	Excel		/s jupyterWo	rkflow		G -
Data Service	+	Load at Start	Yes	No		
Data Science Service 10	+	Timely Refresh	Yes	No		1
Connection-1	Î	Refresh Interval	5		Minute(s	i)
Data Store	+	FIELD SET CAL	CULATED FIELDS	CONDITION		
Data Sheet	14	diameter				
Web Socket	+	height				
Merged Data	+	length				



xv) Once the data connection is established the selected predictive workflow can be used as a data source to the Dashboard Designer.

Note:

- b. If a deployed Predictive Workflow has a summary, it can be viewed using the Dashboard Designer tool.
- c. If the model included in the selected saved NN Workflow contains NumPy script, then after the successful deployment of that workflow still users cannot create a dashboard based on it.
- d. The dashboards created based on the deployed Python workflows also support Bokeh charts.

#### 9.7. Pull from VCS

The option helps to pull the workflow from the Version Controlling Service.

- i) Select a workflow from the Saved Workflow list.
- ii) Click the 'Pull from VCS' option.
- iii) A window opens like below:
  - a) The branch name for pull comes pre-written.
  - b) The details of the existing version get displayed from where the user can select the desired version using the radio button.
  - c) Click the '**Pull**' option.

Pull from Version	Controlling System		×
Branchname for Pull *			
VCS_branch			
Version	Commit Date	Committed by	
O v1	22 Aug, 2019 12:07:27 GMT	team-pa@bdb.ai	
O v2	22 Aug, 2019 12:10:01 GMT	team-pa@bdb.ai	
Commit Message			
Warning: Deployed v	workflow will undeploy after Pull op	eration.	
		Pull Cano	el

d) A success message appears to indicate that the selected entity has been pulled from the VCS.



#### 9.8. Push into VCS

The option helps to push the workflow into the Version Controlling Service.



- i) Select a workflow from the Saved Workflow list.
- ii) Click the 'Pull from VCS' option.
- iii) A window opens like below:
  - a) The branch name for push comes pre-written.
  - b) Provide Commit message (it is mandatory)
  - c) Click the 'Push' option.

Push into Version Controlling System	×
Branchname for Push *	
VCS_branch	
Commit Message (required) *	
vcstest_push	
Warning: Please save data if selected entity has been updated to prevent any data lose. In case if updated version exists on configured repository, data will be updated.	
Push	Cancel

d) A success message appears to indicate that the selected entity has been pushed into the VCS.



Note: At present, the Pull from VCS and Push into VCS options are available only for the Python workflows.

## **10. Scheduler**

The Scheduler component helps to schedule the Data Science workflows as per the requirement.

#### 10.1. New Schedule

This section explains the steps to schedule a new job. Scheduling a new job is a continuous step by step process as described below:

- i) Navigate to the Predictive homepage.
- ii) Click the 'Scheduler' tree node.
- iii) Two options get displayed:
  - a. New Scheduler
  - b. Status
- iv) Select the 'New Schedule' option from the menu.



💶 i Scheduler								
	New Schedule							
	討 Status							

v) The 'General' tab opens.

## 10.1.1. Configuring General Tab

- i) The 'General' tab opens (by default) by clicking the New Schedule.
- ii) Fill in the required information:
  - a. Model Name: Select a model name using the drop-down menu.
  - b. Job Name: Enter a job name.
  - c. **Description**: Describe the job (optional field).
  - d. Use Existing Data Connector: Use radio buttons to select an option.
    - i. Select '**Yes**' to use an existing data connector.
    - ii. Select 'No' for not using an existing data connector.

(Only Data service and Data Store data connectors can be allowed to use an existing data connector option.)

- e. Use Existing Datawriter: Use radio buttons to select an option.
  - i. Select 'Yes' to use an existing data writer.
  - ii. Select 'No' for not using an existing data writer.
- iii) Click the 'Next' option.

Component	Console	Summary	Result	Visualiza	tion	Properties	+	<u>+</u>
General	Basic							
Data Source	Workflow Name		Scheduler WF	•				
Data Writer	Job Name		Sample Schdeule	: Job				
Schedule	Description		Optional					
Notification	Select Server for		Sample R Server					
	Scheduling							
	Use Existing Data	C	) Yes	No				
	Connector							
	Use Existing Data	writer	) Yes	No				
							Next	

iv) The 'Data Source' tab opens.

## 10.1.2. Configuring Data Source

Provide the required information to configure a data source:

- i) The 'General' fields to configure the data source appears by default.
- ii) The user can fill in the required fields:



- a. Component Name: A default name provided for the component.
- b. Alias Name: The user can enter a name for the component.
- c. Description: Users can describe the component (optional).
- iii) Click the 'Next' option.

Component	Console	Summary	/ Result	Visualizati	on	Properties	+	<u>+</u>
General	General	Properties	Conditions					
Data Source	Mapping							
Data Writer	Basic							
Schedule	Component Name Alias		Data Service					
Notification			Data Service					
	Description		Optional					
•								
							Ne	xt

- iv) The user gets redirected to the 'Properties' fields.
- v) Configure the following fields (to configure a new data source):
  - a. Select Data Connector: Select a data connector from the drop-down menu
  - b. Select Data Service: Select a data service from the drop-down menu
  - c. Based on the selected data service the below-given columns get displayed
    - i. Column Header
    - ii. Data Type
- vi) Click the 'Next' option.

Component	Console	Summary	Result	Visualizatio	on Properties	<u>+</u> ↑	<u>+</u>
General	General	Properties	Conditions				
Data Source	Mapping						
Data Writer							
Schedule	Select Data	Hirir	igData DB				
Notification	Connector						
	Select Data Ser	vice team	team_det 👻				
•	Column Header emp_name gender source referral_of	r Data t string string string string					
						Next	i -



- vii) The '**Conditions**' tab opens (If conditions are available, else the user gets redirected to the 'Mapping' page).
- viii) Configure the required 'Conditions' fields.
- ix) Click the 'Next' option.

Component	Console	Summary	Result	Visualizat	ion P	roperties	<u>+</u> ↑	Ţ
General	General	Properties	Conditions					
Data Source			Mapping					
Data Writer	Filter Name	Cont	trol Type					
Schedule	tm	Te	ext v	BU10				
Notification								
•								
							Next	

- x) The user gets redirected to the 'Mapping' tab.
- xi) Configure the column header information from the data service that is used for the selected model columns.
- xii) Click the 'Next' option.

Component	Console Summary	Result	Visualization	Properties	<del>*</del>	<u>+</u>
General	General Properties	Conditions				
Data Source	Mapping					
Data Writer	Column selected from model	C	olumn Header from dat	a service		
Schedule						
Notification	emp_name		emp_name •			
	gender		gender 🔻			
	source		source 🔻			
	referral_of		referral_of •			
•	designation		designation 🔹			
	team		team 🔻			
					Next	

xiii) The 'Data Writer' tab opens.

Note: The user can skip this step if the existing data connector is used. The user needs to configure the data source.



## 10.1.3. Configuring a Data Writer

The Data Writer fields are reliant on the selected data writer types. The scheduler is provided with two kinds of data writers: 1. Data Writer, and 2. Data Store Writer.

Component	Console	Summary	Result	Visualization
General	Data Write	er		
, Data Source	Data Writer	Туре	Data Writer	•
Data Writer	Data Source	Name	Search	
Schedule	Number of F	_		
Notification	batch	1	🖌 Data V	/riter
	Database Na	ame 2	Data Store	Writer
	Password			
	Table Name			
•	Table Opera	tion		- 1
	Column Sele	cted	13 checked	•

#### 1. Data Writer

- i) Fill in the required details to configure a database writer.
- ii) Click the 'Next' option.

Component	Console Summary	Result Visualization	Properties 🗧 🛓	_
General	Data Writer			
Data Source	Data Writer Type	Data Writer	•	
Data Writer	Data Source Name	HiringData DB	•	
Schedule	Туре	mysql		
Notification	Number of Rows in a batch	1000	0	
	Database Name	BDB_Hiring_Data	•	
	Password			
	Table Name	Create New Table	•	
	Table Operation	Overwrite Table	•	
	Create New Table	InternalDW	0	
•	Auto Increment	Enable	•	
	Auto Increment Label	AIL	0	
	Column Selected	8 checked	•	
			Next	

iii) The 'Schedule' tab opens.



#### 2. Data Store Writer

Users can directly use the predictive workflows to create Business Stories if the workflows are written using the Elastic Search Writer.

- i) Select 'Data Store Writer' as a Data Writer Type to schedule a Predictive workflow.
- ii) The Data Store Writer Properties appears.
- iii) Drag and drop the required dimensions to define a hierarchical drill.
- iv) Click the '**Next**' option.

Component	Console Summary	Result	Visualization	Properties	+	Ŧ
General	Data Writer					
Data Source	Data Writer Type	a Store Writer	<b>•</b>			
Data Writer				_		
Schedule	Data Store Writer Pro	perties	Hierarchy Defini	tion	+	
Notification	Dimensions		Drill Definition	designation 🗙	×	
	emp_name	€ L	emp_nam			
	gender	Ö	- inp_nam			
	source	O L				
	referral_of	0 O				
• •	designation	Q L				
	*****	11				
					Ne	xt

v) The 'Schedule' tab opens.

**Note**: The user can skip this step if the existing data writer has been marked to use.

#### 10.1.4. Scheduling a New job

The user can select a time to schedule a new job using this section. The refresh interval option appears as per the selected scheduling time.

- i) **Start Date**: Select a start date and time for the scheduled job (It should be higher than **the Current System Date and Time)**
- ii) Select a Job Refresh Interval option:

E.g., When the selected time range is '**Hourly**,' the selected interval option can be as described below:

Every\_hour: Selecting this option refreshes the scheduled job after every selected interval. OR

At: Selecting this option refreshes the scheduled job at the selected hour.

- iii) **Start Time:** Select a start time higher than the current system time.
- iv) **End Date**: Select an end date and time for the scheduled job (It should be higher than the Start date and the Current System Date and Time).
- v) Run Now: Select this option to run the scheduled job on Applying.
- vi) Click the 'Next' option.
- vii) The 'Notification' tab opens.



## 10.1.4.1. Job Refresh Intervals Details

- Hourly: By selecting this option, the user can schedule the job on an hourly basis.
  - Select a specific hour by using the below-given options: Every\_hour: Selecting this option refreshes the scheduled job after the selected hourly interval.

OR

At: Selecting this option refreshes the scheduled job at the selected hour.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Hourly	Daily Weekly	Monthly	Yearly			
Data Source	Custom Ci	on Expression					
Data Writer Schedule Notification	Start Date Every	1h	g 27 2019 00:00 our(s)				
	<ul> <li>At</li> <li>End Date</li> </ul>	12 <b>v</b> 0	)0 ▼ Ig 29 2019 00:00	<b>i</b>			
÷	€ Run Now	ſ					
						Next	

• **Daily:** By selecting this option, the user can schedule the job daily.

1. Select a specific day by using the below-given options:

**Every\_ Days**: the scheduled job gets refreshed after every selected number of days. E.g., if 2 is selected then; the scheduled job gets refreshed every alternate day at the set time.

#### OR

**Every Week Day**: the scheduled job gets refreshed daily till the end date.

1. Select the Start time.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Hourly Da	aily Weekly	Monthly	Yearly			
, Data Source	Custom Cron	Expression					
Data Writer	Start Date	Tuo Au	g 27 2019 00:00:	-			
Schedule	Start Date	Tue Au	g 27 2019 00.00.0	i			
Notification	Every	1 D	ays				
	O Every Week	Day					
	Start Time	12 • 0	0 •				
	End Date	Thu Au	ıg 29 2019 00:00:1	i			
•	Run Now						
						Next	

• Weekly: By selecting this option, the user can schedule the job every week. Select a day or days of the week when the scheduled job can be refreshed.

Component	Console Su	immary	Result	Visualization	Properties	+	Ŧ
General	Hourly Daily	Weekly	Monthly	Yearly			
Data Source	Custom Cron Ex	pression					
Data Writer	Start Data	Ortiture	04.0010.00.00.0				
Schedule	Start Date	Sat Aug	24 2019 00:00:0	i			
Notification	🕑 Monday 🗆 Tue 🗆 Saturday 🔲 Su		dnesday 🗆 Tl	hursday 🔲 Friday			
	Start Time	12 • 0	0 🔻				
	End Date	Fri Aug	21 2020 01:00:0	i			
	Run Now						
Þ							
						Next	

• **Monthly**: By selecting this option, users can schedule the job every month. This time the range can be used to set schedule refresh for more than a month. Select a specific day of the month by using the below-given options: E.g., Set monthly refresh interval (E.g., the first day of every month)

#### OR

Set a specific day after the desired monthly interval (the first Monday of the every month)



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Hourly	Daily Weekly	Monthly	Yearly			
, Data Source	Custom Cro	on Expression					
Data Writer Schedule Notification	Start Date		g 24 2019 00:00:0 month(s)	i			
Notification	○ The First	<ul> <li>Monday</li> </ul>	• of every 1	month(s)			
	Start Time	12 • 0	0 🔻				
	End Date	Fri Aug	21 2020 01:00:0	i			
•							
						Next	

• Yearly: By selecting this option, users can schedule the job every year. This time range is provided for jobs that run for more than one year.

Select a specific day of the month by using the below-given options:

Set a date for any month (E.g., The 1<sup>st</sup> January of every year till it approaches the end date) Or

Select a day of any month (E.g. The 1<sup>st</sup> Monday of January every year until it approaches the end date)

Component	Console	Summary	Result	Visualization	Properties	+	Ŧ
General	Hourly	Daily Weekly	Monthly	Yearly			
Data Source	Custom Cro	on Expression					
Data Writer Schedule Notification	Start Date • Every • The First	January	24 2019 00:00: • ▼ 1 ▼ of Janu				
÷	Start Time End Date <i>ඔ</i> Run Now	12 V 00	21 2020 01:00:0	=			
						Next	

• **Custom Cron Expression:** The user can schedule a more flexible and customizable schedule runs by using the 'Custom Cron Expression' option. The scheduled workflow can be more specific with the custom cron expression that supports timing up to minutes and seconds. Users need to enter a valid Cron Expression in the given field.



Component	Console Sumn	nary Result	Visualization	Properties	+	Ţ
General	Hourly Daily	Weekly Monthly	Yearly			
Data Source	Custom Cron Expres	ssion				
Data Writer	Start Date	Tue Aug 27 2019 00:00:(	_			
Schedule	Start Date	Tue Aug 27 2019 00:00:0	<b></b>			
Notification	Cron Expression	0 0 12 * *				
	End Date	Thu Aug 29 2019 00:00:1	<b></b>			
•						
					Next	

**Note**: By selecting the 'Use Existing Data Connector' and 'Use Existing Data Writer' options the 'Schedule' tab gets displayed immediately after the 'General' tab.

#### 10.1.5. Notification

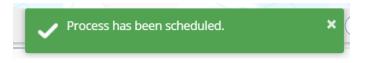
The '**Notification**' tab opens to configure the email settings to get a notification.

- i) Configure the below-given fields:
  - a. Enable Email Notification: Use a checkmark in the box to enable email
  - b. Email Address: Enable this option by using checkmarks in the box
  - c. **Send Mail when Server is not running**: Users can checkmark in the box to enable this option. By enabling this option, the user gets an email when the R server is not running.
  - d. Send Mail when Process is Completed Successfully: Users can put a checkmark in the box to enable this option. By enabling this option, the user gets mail after the process is completed.
  - e. Send Mail when the Process is a Failure: Users can checkmark in the box to enable this option. By enabling this option, the user gets an email when the process fails.
- ii) Click the 'Apply' option.



Component	Console Summ	ary Result	Visualization	Properties	<u>+</u> ↑	<u>+</u>
General	Email Notification					
, Data Source	Enable Email					
Data Writer	Notification					
Schedule	Email Address	bdbuser@bdb.ai				
Notification	Send Mails When					
	Server is not Running					
	Send mail when					
	process is completed					
	successfully					
•	Send mails when the					
	process is a failure					
					Apply	/

iii) A success message appears.



iv) The scheduled job/ process gets added to a list provided under the 'Status' tab.

C Refresh								Search		
Task Name	Frequency	Start Date	End Date	Next Run	Status	Scheduled By	Workflow Name	Data Source	Logs	Actions
Sample Schdeule Job	customCronExpression	27/Aug/2019- 0:0:0	29/Aug/2019- 0:0:0	NA	Stopped	Will	Scheduler WF	team_det	View Logs	×.
Sample Schdeule Job	Hourly	27/Aug/2019- 0:0:0	29/Aug/2019- 0:0:0	27/Aug/2019- 0:0:0	Active	Will	Scheduler WF	team_det	View Logs	X.

Note:

- a. The PDF summary gets sent through email for the scheduled workflows.
- b. Multiple email addresses can be entered into a comma-separated value.
- c. At present, Spark Workflows are not supported by Scheduler.

### 10.2. Status

This section displays detailed information for all the scheduled jobs.

- i) Click the 'Scheduler' tree node.
- ii) Select the 'Status' option.



🔺 皾 Scheduler	
New Sch	nedule
👼 Status	

- iii) The Component tab opens with a list of all the scheduled jobs.
- iv) Click the 'View Logs' icon.

C Refresh								Search:		
Task Name	Frequency	Start Date	End Date	Next Run	Status	Scheduled By	Workflow Name	Data Source	Logs	Actions
Sample Schdeule Job	customCronExpression	27/Aug/2019- 0:0:0	29/Aug/2019- 0:0:0	NA	Stopped	Will	Scheduler WF	team_det	View Logs	×.
Sample Schdeule Job	Hourly	24/Aug/2019- 0:0:0	29/Aug/2019- 0:0:0	24/Aug/2019- 0:0:0	Active	Will	Scheduler WF	team_det	View Logs	.≼ ₽
job 23 aug	Hourly	23/Aug/2019- 14:0:0	23/Aug/2019- 16:0:0	23/Aug/2019- 14:0:0	Active	Will	Scheduler WF	team_det	View Logs	X.

- v) The logs of the selected workflow get displayed under the '**Component**' tab.
- vi) Click the 'Refresh Log' option to refresh the logs.

	Component	onsole	Summary	Result	Visualization	Properties	÷	<u> </u>
	23/Aug/2019 - 01:19:4	43	Data Service0 is st	arted.			C Refrest	ı Log
ľ	23/Aug/2019 - 01:19:4	15	Number of Rows f	etched: 8				
	23/Aug/2019 - 01:19:4	45	Data Service0 is co	ompleted				
	23/Aug/2019 - 01:19:4	15	Filter2 is started.					
	23/Aug/2019 - 01:19:4	16	Filter2 is complete	d				
	23/Aug/2019 - 01:19:4	16	Data Store Writer	is started.				
	23/Aug/2019 - 01:19:5	55	Data Store Writer	is completed				

#### **Related Actions for a Scheduled Job:**

Options	Name	Description
1	Edit	To edit/update the scheduled job details
	Stop	To stop the scheduled job
×	Remove	To remove the scheduled job from the list
	Start	To start the scheduled job

Note:

- a. The 'Edit' option allows the user to update/ edit all the tabs for the selected job.
- b. The user can click the '**Start**' button to restart the scheduler for a scheduled job until it reaches the end date.

#### www.bdb.ai



c. The user can enable 'Edit' and 'Remove' actions only after stopping the Scheduled job.

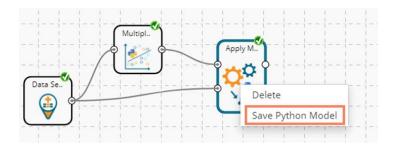
### 10.2.1. Model Retraining in Scheduler

The users can monitor the model retraining steps through the scheduler.

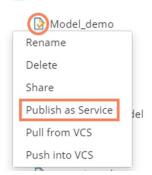
- i) Create a Workflow or select a workflow with an Apply Model component.
- ii) Run the workflow.



iii) Save the Model.

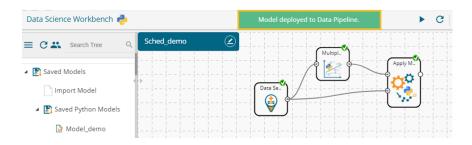


- iv) Navigate to the Saved model.
- v) Select the '**Publish as Service**' option to deploy the model to the Data Pipeline.
  - ▲ 🛐 Saved Python Models





vi) A success message appears to confirm the deployment.



vii) Navigate to the Scheduler and select the same workflow using the 'General' tab.

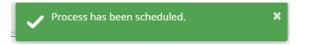
viii) Select the saved model from the 'Model to Retrain' drop-down menu.

E C 🎝 Search Tree	Component	Console	Summary	Result	Visualization	Properties	DataInsight	+	<u>+</u>
Performance	General	Basic							
🕨 🐣 Data Writer	Data Source	Workflow Na	me	Sched_demo	•				
🕨 💆 Custom Scripts	Schedule	Job Name		sched_mar					
🛛 🗾 Jupyter Notebooks	Notification	Description		Optional					
<ul> <li>▲ Scheduler</li> <li>New Schedule</li> <li>➡ Status</li> <li>▲ Saved Models</li> </ul>	•	Use Existing Connector Model To Ret	Data	Yes     Model_demo	No Vo				
								Nex	t

ix) Configure the required steps to schedule the workflow.

Data Science Workbench 🥐				•	G + G	B 🖌 🤞	← 📾 :   🤉
E 🖸 🚢 Search Tree 🔍 Q	Component	Console Summary	Result	Visualization	Properties	DataInsight	÷ 1
💵 Performance	General	Email Notification					
🚔 Data Writer	Data Source	Enable Email Notification	×				
🕵 Custom Scripts	Schedule	Email Address	pa@bdb.ai				
🗾 Jupyter Notebooks	Notification	Send Mails When Server is	×.				
📸 Scheduler		not Running					
New Schedule		Send mail when process is					
📆 Status		completed successfully					
🚯 Saved Models		Send mails when the	۲				
		process is a failure					

x) A message appears to inform that the process has been scheduled.



- xi) The users get redirected to the Status option displaying all the scheduled processes.
- xii) Click the 'View Log' option.



									Search:	
Task Name	Frequency	Start Date	End Date	Next Run	Status	Scheduled By	Workflow Name	Data Source	Logs	Actions
onAdmin_Sched	Hourly	16/Jan/2020-16:0:0	16/Jan/2020-18:0:0	NA	Stopped	anaghakn	scheduler_nonadmin	Burnedforest_Forecast	View Logs	/ = ×
nodelRetrain	Hourly	21/Jan/2020-16:0:0	21/Jan/2020-19:0:0	NA	Stopped	admin	Correlation_Model_save	lris_dataset	View Logs	/ = ×
ched_model	Hourly	28/Jan/2020-12:0:0	28/Jan/2020-16:0:0	NA	Stopped	anaghakn	sched_model	lris_Nov19	View Logs	/ = ×
chedd1	Hourly	29/Jan/2020-16:0:0	29/Jan/2020-22:0:0	NA	Stopped	anaghakn	schedd	lris_Nov19	View Logs	/ = ×
chedule_Feb5	customCronExpression	5/Feb/2020-17:0:0	5/Feb/2020-18:0:0	NA	Stopped	ShyamPd	Schedule_Feb5	German_credit_card_data	View Logs	/ = ×
rain_blubrich_01	Daily	5/Feb/2020-19:0:0	13/Feb/2020-19:0:0	NA	Stopped	admin	blubirch_train	input_blubirch_train	View Logs	/ = ×
est_Blurich	Daily	6/Feb/2020-13:0:0	14/Feb/2020-14:0:0	NA	Stopped	admin	Blubirch_infer	input_blubirch_train	View Logs	/ = ×
est_infer2	Daily	6/Feb/2020-21:0:0	12/Feb/2020-12:0:0	NA	Stopped	admin	Blubirch_infer	input_blubirch_train	View Logs	/ = ×
est_infer3	Daily	7/Feb/2020-14:0:0	13/Feb/2020-5:0:0	NA	Stopped	admin	Blubirch_infer_350	input_blubirch_train	View Logs	2 = X
ched_mar	Hourly	20/Mar/2020-16:0:0	20/Mar/2020-23:0:0	20/Mar/2020-16:0:0	Active	admin	Sched_demo	iris_filter	View Logs	/ = ×

xiii) The stepwise logs get displayed confirming the model retaining and upload to the Data Pipeline.

Componer	t Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u> <u>⊥</u>
20/Mar/20	0 - 03:06:55	Data	Service0 is st	arted.			C Refresh Log
> 20/Mar/20	0 - 03:06:55	Data	Service0 is co	mpleted.			
20/Mar/20	0 - 03:06:55	Multi	ple Linear Re	gression1 is started.			
20/Mar/20	0 - 03:06:55	Multi	ple Linear Re	gression1 is complete	ed.		
20/Mar/20	0 - 03:06:55	Appl	/ Model2 is st	arted.			
20/Mar/20	0 - 03:06:55	Mode	el Retrain is st	arted for Model_dem	0		
20/Mar/20	0 - 03:06:55	Appl	/ Model2 is co	mpleted.			
20/Mar/20	0 - 03:06:55	Mode	el Retrain is c	ompleted for Model_d	lemo		
20/Mar/20	0 - 03:06:55	Mod	el Retrain is co	ompleted and update	d into PipeLine.		

## **11. Saved Models**

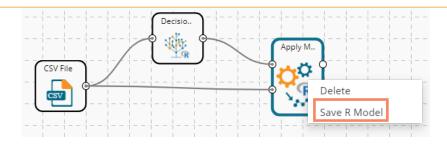
The user can save a trained model through the Apply Model component. The user can either split the dataset into training and testing, create a model with training data, and Apply the testing data. Another approach is to save the model and Apply the model over a new test data set.

The user can save a model after successful execution. The saved R models get listed under the 'Saved R Model' tree node. Users can select a saved R model from the list and use it to create a new workflow.

#### 11.1.1. Saving a Trained Model

- i) Create a Workflow with Apply Model or Open a saved workflow that contains an Apply Model.
- ii) Use right-click on the 'Apply Model' component.
- iii) A context menu opens.
- iv) Select the 'Save R Model' option (The 'Save' option for Python and Spark, which gets displayed as 'Save Python Model' and 'Save Spark Model' based on the selected workbench).

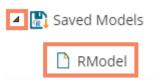




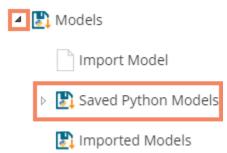
- v) The **Save R Model** window opens (The heading of the Save R Model gets changed as '**Save Python Model**' and '**Save Spark Model**' based on the selected workbench).
- vi) Enter the model name by which you wish to save the model.
- vii) Click the '**Ok**' option.

Save R Model		×
Save Model Name RModel		
	Ok	Cancel

- Note: The 'Save Model Name' is a mandatory field. The user cannot give in-between space for two words. The first character of the model name should be an alphabet and must be mentioned in a capital case.
- viii) The selected model gets saved in the 'Saved Models' list.



Note: The heading for Python Workbench is '**Models**' as it includes Imported Models together with **Saved Python Models**.



## 11.1.2. Importing a Model



This component lets a user import any localized model in the Python workbench to use it directly in the BDB platform.

- xiv) Navigate to the Models tree-node from the Python Workspace.
- xv) Click the '**Import Model**' component tree node.

⊳ 🦿	🖇 Custom Scripts
Þ	🔋 Scheduler
4	🕽 Models
	Import Model
	Saved Python Models
	🛐 Imported Models

- xvi) The General tab for the Import model opens.
- xvii) The user can edit the Component Name.
- xviii) Click the 'Next' option.

Component	Console	Summary	Result	Visualization	Properties	<u>⊥</u> <u>⊺</u>
General	Basic					
Properties	Component I	Name	Flower_Classific	ation_Model		
	Component	Туре	Custom Python	Model		
Þ	Component (	Description	Optional	4		
						Next

- xix) The Properties tab opens.
- xx) Upload the model file.
  - a. Select a file from the system.
  - b. Click the 'Upload' option.
  - c. A success message appears to convey that the selected file has been uploaded.



		The file is uploa	ided.		► C + O E	N R + C   (
Component	Console	Summary	Result	Visualization	Properties	÷
General	Upload	d Custom Mode	l Files			
Properties	Step 1	Sele	ct File	iris.zip	Upload	
	Step 2	Ente	r Python 8	<ul> <li>- Include Instr</li> <li>- Include Instr</li> <li>- CUSTOM_INF</li> <li>- Include Instr</li> <li>- CUSTOM_INF</li> <li>- Include Instr</li> <li>- CUSTOM_OU</li> <li>- Include Instr</li> <li>- CUSTOM_MOD</li> <li>- Piease Preps</li> <li>- example Data</li> <li>- Example Codd</li> <li>- from sklearn.et</li> </ul>	Write Python Code uction(s) to Acquire Model from "CU uction(s) to Prepare Input Data for N UT_DATA" Variable uction(s) to Assign Output Result of TPUT_DATA" Variable uction(s) to Assign Summary of Mo DEL_SUMMARY 'Variable are Model Input Data as per the Mod Cleaning is require for Model e: xternals import joblib blib load(CUSTOM_MODEL_DIR + '1	Addel using Model into del (if any) into iel requirement, for

- xxi) Enter a Python Code (Script)
- xxii) Click the 'Apply' option.

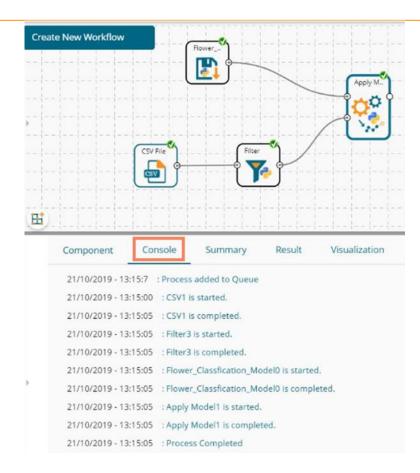
Upload Custom Model	ct File r Python	predicted = regresse a CUSTOM_OUTPUT_	oad(CUSTOM_MODEL_DIR + 'iris.sav') or.predict(CUSTOM_INPUT_DATA)		
Step Enter	r Python	1 from sklearn.extern 2 regressor = joblib.k 3 predicted = regress 4 CUSTOM_OUTPUT_	als import joblib oad(CUSTOM_MODEL_DIR + 'iris.sav) or predict(CUSTOM_INPUT_DATA)		
		2 regressor = joblib.lo 3 predicted = regressor 4 CUSTOM_OUTPUT_	oad(CUSTOM_MODEL_DIR + 'iris.sav') or.predict(CUSTOM_INPUT_DATA)		
		6 CUSTOM_MODEL_S	UNIA - publicarraine() DATA(PredictedValues") = predicted UMMARY = 'Custom Model Summary ()' m Model execution finished()		
			10 11 12 13 13 14 16	10 11 12 13 14 16	10 11 12 13 14 16

xxiii) A success message appears to convey that the custom model has been imported.

The Custom Model is imported.

- xxiv) Click the Run or Refresh icon to run the model.
- xxv) Stepwise completion of the process can be seen under the 'Console' tab. The green checkmarks at the top of the dragged components mark the completion of the console process.





xxvi) Click the '**Result**' tab to get the processed data (To open the Result tab first click on the Apply Model component, then click the 'Result' tab).



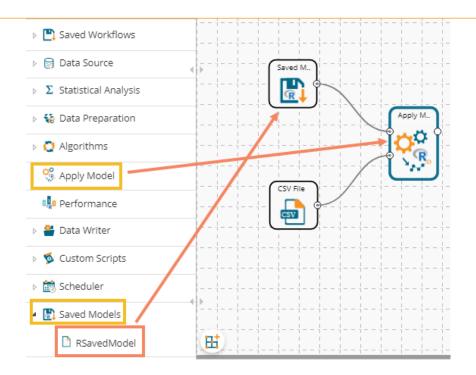
E C Search Tree Q	Creat	e New Workflow				
E Saved Workflows				Flower		Apply M.
👂 😝 Data Source				[]]		<b>ૻૣ૽૽ૢૢ૽૾૿૿ૣ</b>
▷ ∑ Statistical Analysis	Þ				$\langle /$	Ŭ.
👂 🎲 Data Preparation			CSV File	Filter		
Algorithms			l 💼 📬		)	
😤 Apply Model						
Performance	E.					
🖻 🚢 Data Writer		Component	Console	Summary	Result	Visualization
👂 🕵 Custom Scripts		Show 10 T	entries			-
▷ 📸 Scheduler		PredictedValue	25			
🔺 🛃 Models		setosa				
Import Model		setosa				
Saved Python Models	• >	setosa				
🖬 🔛 Imported Models	ľ	setosa				
Flower_Classfication_Model		setosa setosa				
		setosa				

## 11.1.3. Reading a Saved Model

The user can drag a saved model to the workspace and reuse the model for test data. A saved model can be connected to only Apply Model and new test data source to create a workflow.

- i) Select and drag a saved R model component onto the workspace.
- ii) Connect the dragged saved model component and a configured data source to an Apply Model component. Pass the Saved model data in the training node and data source's data in the testing node of the dragged Apply Model component (As shown in the following image).





- iii) Click on the dragged Saved Model component.
  - a. The 'Summary' tab opens by default displaying the model summary.
    - i. Click the 'Apply' option for the saved model component.

Component	Console	Summary	Result	Visualizat	tion P	roperties	+	Ŧ
General	Summary	r						
Summary	rpart(for weigh 0, Variable weight_ 193 weight_sl	***** Summary of ***** Summary of rmula = rings ~ d ht_viscera + weig ction = na.rpart, minsplit = 10, cp Importance _shell weight_w 18.176 17313 hucked he 36.468 1512	<pre>stage 1 ~~ he model - liameter + ht_shell - method = 0 = 0.005, hole .580 ight</pre>	height + weight + length, data = "anova", contro usesurrogate = diameter	= RProcessbd ol = rpart.c 1)) length w	188420a4ee44c21929816dd7626a control(, reight_viscera	92cf_11	L_
>	~~~~~~~	End of Summa www End of stage	1 summary	~~~~~~~~~~~			Apply	

b. Click the '**General**' tab to display the Basic information of the concerned Saved Model component.

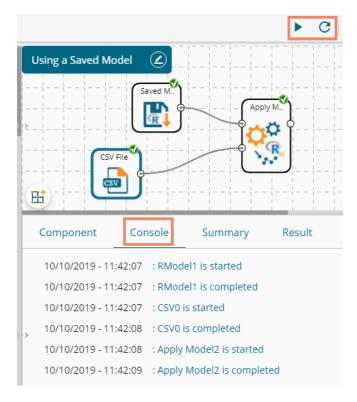


Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Basic						
Summary	Component Na	me R S	Saved Model				
	Alias	RN	1odel1				
7	Description	Op	tional	li			
						Apply	

iv) Click the 'Apply' option provided in the Apply Model component.

Component	Console	Summary	Result	Visualization	Properties	Ŧ	Ť
General	Basic						
	Component Nar	ne R A	opply Model				
•	Alias	Ap	ply Model2				
	Description	Op	tional				
				//			
						Apply	/

- v) Run the workflow after getting the success message.
- vi) The 'Console' tab opens displaying the progress of the process.





# vii) After the process gets completed under the Console tab, click the '**Result**' tab to see the processed data.

	COMPONE	NT CONSOLE	SUMMARY	RESULT	SUALIZATION PF	OPERTIES			$\left(\frac{\pm}{\uparrow}\right)$	
1	Show 10	▼ entries							Search:	
Þ	Month	Day_of_month	Day_of_week	ozone_reading	pressure_height	Wind_speed	Humidity	Temperature_Sandburg	Temperature_ElMonte	h
	1	1	4	3.01	5480	8	20			50
	1	2	5	3.2	5660	6		38		
	1	3	6	2.7	5710	4	28	40		26
	1	4	7	5.18	5700	3	37	45		59
	1	5	1	5.34	5760	3	51	54	45.32	14
	1	6	2	5.77	5720	4	69	35	49.64	15
	1	7	3	3.69	5790	6	19	45	46.4	26
	1	8	4	3.89	5790	3	25	55	52.7	55
Þ	1	9	5	5.76	5700	3	73	41	48.02	20
	1	10	6	6.94	5700	3	59	44		26
	Showing 1 to	o 10 of 358 entries						Previous 1 2	3 4 5 36 Ne	ext

#### Note:

- a. A mandatory condition to run the workflow with a 'Saved R Model' component is that the column headers and data type of the test data source should match with the selected saved model. Otherwise, an error notification of validation failure appears while running the workflow.
- b. The user can connect a data writer to the '**Apply Model**' component in a workflow containing a saved model.

#### 11.1.3.1. Renaming a Saved Model

- i) Select a model from the 'Saved Models' list.
- ii) Use a right-click on the selected saved model component.
- iii) A context menu opens.
- iv) Select the 'Rename' option.

🔺 💽 Saved Models					
	🗋 RModel				
	Rename				
	Delete				
	Publish as Service				

- v) A pop-up window appears to rename the model.
- vi) Enter a new 'Model Title' or modify the existing model title in the given field (if desired)
- vii) Click the 'Yes' option.



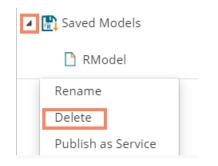
Rename Model		×
Model Title *Workflows used by this model will not work after rename		
RSavedModel		
	Yes	No

viii) The selected Saved Model gets renamed. A success message appears to notify for the same action.

Data Science Workbench 🧟				The model is renamed.				
Search Tree	Q		Compo	onent	Console	Summary	R	
∑ Statistical Analysis		T.	Show	10 • ent	tries			
🛛 🎲 Data Preparation		P	sex	length	diameter	height	wei	
O Algorithms			Μ	0.455	0.365	0.095	0.514	
🤗 Apply Model			Μ	0.35	0.265	0.09	0.225	
-		-11	F	0.53	0.42	0.135	0.677	
📲 Performance			Μ	0.44	0.365	0.125	0.516	
🐣 Data Writer			1	0.33	0.255	0.08	0.205	
蓫 Custom Scripts			T	0.425	0.3	0.095	0.351	
📸 Scheduler		-11	F	0.53	0.415	0.15	0.777	
			F	0.545	0.425	0.125	0.768	
🚯 Saved Models			Μ	0.475	0.37	0.125	0.509	
RSavedModel			F	0.55	0.44	0.15	0.894	

## 11.1.3.2. Deleting a Model

- i) Select a model from the 'Saved Models' list.
- ii) Right-click on the selected model.
- iii) A context menu opens.
- iv) Select the 'Delete' option.





- v) A new window opens, asking for confirmation for the deletion.
- vi) Click the '**Ok**' option.
- vii) The selected saved model gets removed from the **Saved Models** list.

Note: After renaming or deleting a Saved R Model, workflows used by the same model don't work.

#### 11.1.3.3. Sharing a Python Model

The user can share a saved model with other users or user groups. There are two options to share a selected model:

- 1. **Share With**: This option allows the user to share a file with the selected users or user groups. Any changes made to the file are transferred to all the users with whom the file has been shared.
  - i) Use right-click on a model from the list of **Saved Models** (In this case, a Python saved model is selected from the Python Workspace).
  - ii) Select the 'Share' option from the context menu.

🔺 🛃 Models					
	Import Model				
4	🛃 Saved Python Models				
1	PythonSavedModel				
	Rename				
	Delete				
2	Share				
	Publish as Service				
	Pull from VCS				
	Push into VCS				

- iii) The 'Share With' option gets displayed by default.
- iv) Select either the 'Group' or 'Users' option.
  - a. By selecting a group, all group members inside the group get listed. Users can be excluded by not selecting them from the group.
  - b. Users can be excluded by not selecting a username from the list when the '**User**' option has been selected.
- v) Select a specific group or user from the list by using checkmarks in the box.
- vi) Click the '**Apply**' option.



	Component	Console	Summary	Result	Visualization	Properties	± ∓
3	Share With	Group	Users				
	Copy to	Search					
		Select All Gro	oup				0
	Þ	Admin Role					5 🕑
		Viewer Role					<b>∀</b> 𝔄
							6 Apply

- vii) The saved model gets shared with the selected group of users.
- 2. **Copy To**: This option creates a copy and shares the copy with the selected users and user groups. Any changes to the original file after sharing will not show up for the users that received the shared file via the '**Copy To**' method.
  - i) Use right-click on a model from the list of the **Saved Models** (In this case, a Python saved model is selected from the Python Workspace).
  - ii) Select the 'Share' option from the context menu.

<b>a</b> 📳 1	Models
	Import Model
4	🛃 Saved Python Models
1	PythonSavedModel
	Rename
	Delete
2	Share
	Publish as Service
	Pull from VCS
_	Push into VCS

- iii) Select the 'Copy To' option.
- iv) The copied model name gets displayed.
- v) Select either the 'Group' or 'Users' option with a click.
  - a. By selecting a group, all group members inside the group get listed. Users can be excluded by not selecting them from the group.
  - b. Users can be excluded by not selecting a username from the list when the 'Users' option has been selected.
- vi) Select a specific group or user from the list by using checkmarks in the box.
- vii) Click the '**Apply**' option.



	Component	Console	Summary	Result	Visualization	Properties	<u>∔</u>
	Share With	4 Copied w	orkflow	opyOfPythonSave	dMo		
3	Copy to	name 5 Group	Users				
		archnewsp	ace				0
	Þ	Administra	tor				6 🕑
							7 Apply

viii) A copy of the model gets shared with the selected user or group.

## 11.1.3.4. Publishing a Saved Model as Service

The user can publish the saved Data Science models to the Data Pipeline module using this option. The user can access the published Data Science model using the ML model runner component to use them in a pipeline workflow.

- i) Select a model from the Saved Models list.
- ii) Open the context menu provided for the selected saved model.
- iii) Select the 'Publish as Service' option for the selected model.

4	💦 Saved Models
	🎦 RModel
	Rename
	Delete
	Publish as Service

iv) A success message appears to notify the user that the selected saved model is deployed to the Data Pipeline plugin.



v) Navigate to the Data Pipeline plugin using the '**Apps**' menu.





- vi) Open the 'Settings' page.
- vii) The published saved model gets added to the Data Science Models list.

Kafka Configuration	Da	ata Science Models				Search Script	٩
Dataprep Scripts		Model Name	Model Type	Created On		RSavedModel	
Data Science Models		FRAUD_AIINFINITY	Python Model	Tuesday, Oct	ober 1, 2019	Linked pipelines	
Logger		Iron_Ore	Python Model	Thursday, Oo			
		RSavedModel	R Model	Thursday, O	tober 10,		
				2019			
	Dataprep Scripts Data Science Models	Dataprep Scripts Data Science Models	Dataprep Scripts Model Name Data Science Models FRAUD_AIINFINITY Logger Iron_Ore	Data Science Models         Model Name         Model Type           Data Science Models         FRAUD_AIINFINITY         Python Model           Logger         Iron_Ore         Python Model	Data prep Scripts         Model Name         Model Type         Created On           Data Science Models         FRAUD_AIINFINITY         Python Model         Tuesday. Oct           Logger         Iron_Ore         Python Model         Thursday. Oct	Dataprep Scripts         Model Name         Model Type         Created On           Data Science Models         FRAUD_AIINFINITY         Python Model         Tuesday, October 1, 2019           Logger         Iron_Ore         Python Model         Thursday, October 10, 2019           Process         Pothon Model         Thursday, October 10, 2019	Data Science Models         Model Name         Model Type         Created On         RSavedModel           Data Science Models         FRAUD_AIINFINITY         Python Model         Tuesday, October 1, 2019         Inixed pipelines           Logger         Iron_Ore         Python Model         Thursday, October 10, 2019         EssuelModel         EssuelModel

- viii) Access the R Model runner component from the Component Pallet.
- ix) Open the Meta Information tab and scroll down the provided drop-down list.
- x) The published saved model from the Data science Workbench appears in the drop-down list.

Search Component	R Model
System Custom	Basic Information Meta Information
Reader 🔹	
Writer 🔹	popo01
Transformation 🔹	Insurance_Forecast
ML	Fraud_detection
	save_credit_data
Spark 🧲	HR_Model
÷	RSavedModel

## 11.1.3.5. Pull from VCS

The option helps to pull models from the Version Controlling Service.

- i) Select a model from the Saved Workflow list.
- ii) Click the 'Pull from VCS' option.
- iii) A window opens like below:
  - a) The branch name for pull comes pre-written.
  - b) The details of the existing version get displayed from where the user can select the desired version using the radio button.
  - c) Click the '**Pull**' option.



Pull from Version	Controlling System	×
Branchname for Pull *		
VCS_branch		
Version	Commit Date	Committed by
(O) v1	17 Oct, 2019 06:10:50 GMT	SUPPORT
Commit Message		
VCS_push		
Warning: Deployed	workflow will undeploy after Pull o	peration.
		Pull Cancel

d) A success message appears to indicate that the selected entity has been pulled from the VCS.

		Data is pulled.	
--	--	-----------------	--

## 11.1.3.6. Push into VCS

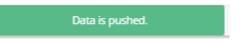
The option helps to push the workflow into the Version Controlling Service.

- i) Select a workflow from the Saved Workflow list.
- ii) Click the 'Pull from VCS' option.
- iii) A window opens like below:
  - a) The branch name for push comes pre-written.
  - b) Provide Commit message (it is mandatory)
  - c) Click the 'Push' option.

Branchname for Push *	
VCS_branch	
Commit Message (required) *	
VCS_push	
Warning: Please save data if selected entity has been updated to prevent any data lose. In case if updated version exists on configured repository, data will be updated.	
	Commit Message (required) * <u>VCS_push</u> Warning: Please save data if selected entity has been updated to prevent any data lose. In case if updated version exists on configured repository, data will be updated.



d) A success message appears to indicate that the selected entity has been pushed into the VCS.

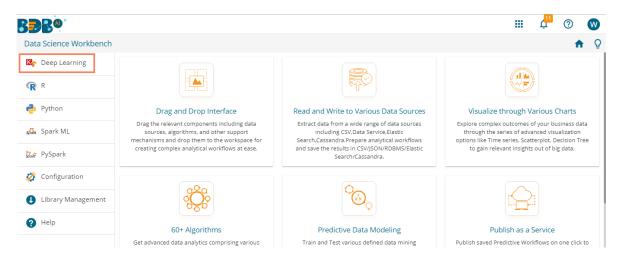


#### Note:

- a. At present, the **Pull from VCS** and **Push into VCS** options are available only for the Python Workspace.
- b. Data Science Models can get deployed multiple times to the Data Pipeline module and get marked to identify the deployed models.

# 12. Deep Learning Workspace

The user can select the Deep Learning Workspace from the Data Science landing page to access the Neural Network Environment under the Data Science Workbench.



The user gets redirected to the following screen by selecting the NN Workspace:



CR Search Tree Q Create New Workflow				
Saved Workflows   Data Source   Data Preparation   Pre Packaged Models   Models   Sustom Scripts   Model Training	ata Science Workbench	⊠ç-	The Neural Network Workbench is loaded.	- ci   9
Saved Workflows   Data Source   Data Preparation   Pre Packaged Models   Models   Source   Models   Model Training   Apply Model	C 🎝 Search Tree Q			
The preparation	🎦 Saved Workflows			
Pre Packaged Models & Models S Custom Scripts % Model Training % Apply Model	Data Source			
Models	🎲 Data Preparation	)		
S Custom Scripts	-			
Model Training       Apply Model				
🥦 Apply Model				
	•			

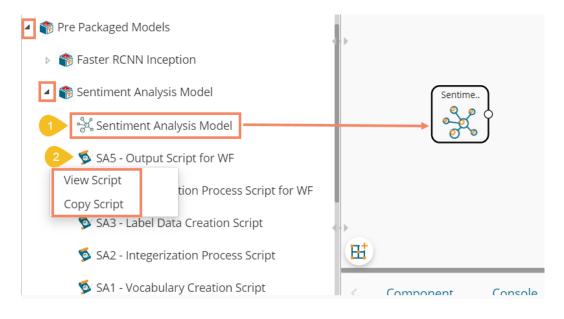
Note:

- a. Neural Network Space is applicable only for Python Environment.
- b. Keras (as a High-level API) is supported by the Tensorflow Backend.
- c. Tensorboard is attached for the Live Visual Tracking of Model during Training.
- d. Model Creation using Python Script is supported.
- e. A pre-trained Model of Sentiment Analysis is Provided along with its feature scripts.

The Component Tree-node menu displays various components with their sub-components to be used in the NN workspace as per requirement.

## **12.1.** Pre-Packaged Models

The component tree-node provided on the NN Workspace contains one node as Pre-Packaged Models which contains the Pre-trained Sentiment Analysis Model and its feature scripts.



• The user can use the Pre-trained Model in a Workflow.

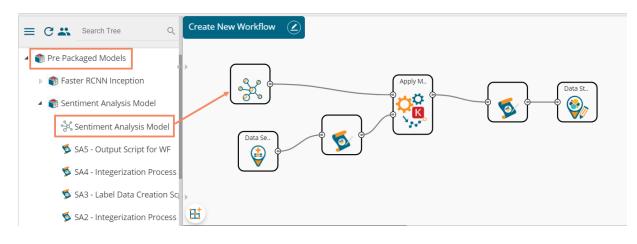
#### Copyright © 2015-21 BDB



- These Scripts can be used directly in Workbench Area using drag-n-drop Functionality.
- The user can Copy the Script, Modify the Code, and then use them as per their need.
- The user must use the 'NN Apply Model' that applies the selected NN-Model over input data to get predicted Results.
- Along with these Pre-trained Models and Scripts, you get support files for training this model (these can be viewed in the 'Supporting Files' tabs of View Model). These supporting files users can access using the SHARED\_PATH variable in the scripts.

Note: The featured scripts are provided with a Pre-trained Sentiment Analysis Model. If the users wish to modify the scripts OR refer to these scripts for other user-defined models, then they must be modified as per their requirements and need to avoid error(s) & incorrect calculation.

The following image displays a workflow created by using a pre-trained model:



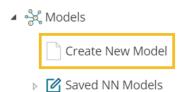
## 12.2. Working with Deep Learning Workspace

This section explains the general steps for Training a Neural Network Model. The entire process can be described in the below-mentioned parts:

## 12.2.1. Creating a New Model

The user needs to start the process from the creation of a new model.

i) Click on the 'Create New Model' option from the Models tree node.



- ii) A Dialog Box opens.
- iii) Provide a name for the Deep Learning model.
- iv) Click the 'Save' option.



Create New Neural Network Model		×
Neural Network Model Name NN_TestModel		
	Save	Cancel

Note:

- a. The user can use a maximum of 20 characters to provide a name for the newly created Model
- b. No other Special Character(s) except Underscore (\_) is allowed
- c. Model Name cannot begin with Space/Numeric Digit or Underscore
- d. Model Name should be unique
- v) A success message appears to assure that the new model has been created.
- vi) The new model gets listed under the 'Saved NN Models' tree node.

Data Science Workbench 🛚	•		The mo	del is created.	
= C 🚜 Search Tree Q	Create New Workflow				$\frac{1}{1} = -\frac{1}{1} =$
Saved Workflows			-		
😝 Data Source		 			
🎲 Data Preparation		 			
鹡 Pre Packaged Models					
နှင့် Models			 		
Create New Model				+ +	
🔺 🗹 Saved NN Models					
NN_TestModel		 			

#### vii) Use a right-click on the model and select the 'View Model' option.

Models
 Create New Model
 Saved NN Models
 Saved NN\_TestModel
 View Model
 Delete Model

#### Copyright © 2015-21 BDB



- viii) The component details open for the selected model, as shown in the following image:
   The user can view only the General tab displaying the Basic information about the newly created NN Model.
  - a. General: The Basic Details regarding the NN model are displayed in this tab.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Basic						
Supporting Files	Component	Name	Neural Netwo	rk Model			
Summary	Alias		NN_TestMod	el			
Model Script	Description		Description				
Model Status					\$		
Tensor Board							

Note: The remaining tabs do not display any information until the model gets trained.

# 12.2.2. Data Preprocessing

## 12.2.2.1. Creating a NumPy Script

This section describes data preprocessing from creating NumPy files to have the required data in a binary format that a Model Script can use for training or prediction purposes. In this section, the user must pre-process the data that is required for a model to get trained; we call this process '**Data Preprocessing**' or NumPy-fication.

Here, the user creates NumPy files; these files have the information of data in a binary format that can be fed into the model during/after training.

Use the '**Custom Script**' tree-node to create a new script inside the NN Workspace. The workflow for creating a new script is like the Python Workspace. The user can also choose an option to create a Utility Module Script.

- i) Select the 'Create New Script' option using the 'Custom Python Script' tree node.
- ii) The Component tab displaying the General tab opens.
- iii) Provide the Basic component information:
  - 1. Provide a Component Name.
  - 2. The Component Type comes pre-filled.
  - 3. Provide relevant Descriptions about the component.
  - 4. Select a script type by using the radio button.
  - 5. Click the 'Next' option.



	Component	Console	Summary	Result	Visualization	Properties	+ †	<u>+</u>
	Basic							
	Component Name		SA_1					
	Component Type		Algorithms					
	Description		Optional					
	Script Type		Custom Pytho Script	n O	Utility Module Script			
•								
	General	Script	Settings				Next	

- iv) The 'Script' tab opens.
- v) Insert script syntax in the Script Editor space.
- vi) Click the 'Validate' option. It should get the success message to move ahead.
- vii) Click the 'Next' option.

*
8
-

- viii) The 'Settings' tab opens.
- ix) Select a Script Type using the checkbox.

#### 1. Normal Python Script

If the selected script type is **Normal Python Script**, then the Primary Function Details gets displayed immediately after the Script Type to be configured:

- ii. Select a Primary Function Name from the drop-down list
- iii. Select an Input Data Frame option from the drop-down list
- iv. Provide a name for the Output Data Frame
- v. Provide the Summary Variable Name (if the View Summary option is enabled)
- vi. Enable 'Show Visualization' and 'Show Summary' options by enabling them in the boxes.



Component	Console	Summary	Result	Visualization	Properties	<u>+</u> ↑	<u>+</u>
Script Type							
	Normal Python Scri	pt 🔿 Mod	lel Object File Scri	pt			
Primary Func	tion Details						
	ary Function Ir	nput DataFrame	Output DataFrame	Summary Variable Nam	e		
create	_labels <b>v</b> df	- <b>-</b>	Output Data F	Summary			
Short Short	w Visualization			ary			

#### 2. Model Object File Script

If it is a Model Object File Script (i.e., NumPy File Creation), then the user needs to provide the following details to configure the Primary Function details:

- i. Select any one NN Model using the drop-down list, which can be associated with an Output NumPy Filename.
- ii. The Output File Name appears in the given box.
- iii. Describe the NumPy File.
- iv. Configure the Primary function details
  - a) Select a Primary Function Name using the drop-down list
  - b) Select an Input Data Frame using the drop-down list
  - c) Provide an Output name for the NumPy

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> <u>+</u>			
Script Type	lormal Python Scrip	it 🔘 Moo	del Object File S	Script					
Select	any NN Model	NN_Te	stModel	-					
Outpu	Output File Name			labeled_data					
Nump	y File Description	Option	al						
Primary Funct	ion Details								
	ry Function Inp	out DataFrame	Output Num	тру					
create_	labels <b>v</b> df	¥	Output Data	aFra					

Note: The user needs to create a model object File Script to get it listed along with the model.

- x) Configure the Function Parameters by providing relevant Property Display Name and defining the Control Type:
- xi) Click the 'Apply' option.



Compon	ent Conso	ole Su	ummary	Result	Visua	alization	Properties	5		<b>↓</b> ↑	<u>+</u>
	create_labels 🔻	df	•	Output Dat	aFra						
Function	Parameter De	finition									
	Function Parame	ters	Property [	)isplay Name		Control Type					
	column_header		column_	header		Column se	lector(Single)	Ŧ	\$		
Γ	no_of_classes		no_of_cl	asses		TextBox		v	•		
General	Scri	pt	Settings				Prev	vious	Ap	ply	

- xii) A Success message appears to confirm the creation of a Python script.
- xiii) The newly created NumPy script gets added to the model folder.

E Search Tree Q	Component	Console	Summary Result	Visualization	Properties		+	<u>+</u>
Pre Packaged Models	create_labe	els 🔻 df	▼ Output D	DataFra				
⊿ 🔩 Models								
Create New Model								
Saved NN Models	Function Parame	Parameters	Property Display Nan	ne Control Ty	pe			
NN_TestModel	column_l	neader	column_header	Column	selector(Single) 🔻	•		
<b>5</b> SA_1	no_of_cla	asses	no_of_classes	TextBox	· · · · ·	•		
👂 🕵 Custom Scripts								
Model Training								
🔓 Apply Model	General	Script	Settings	The	Python Script is save	ed.		

# 12.2.2.2. Copying a Pre-Packaged Script

The user can copy the existing scripts and use them if he wants to use the pre-packaged script instead of creating a new NumPy Script.

- i) Navigate to the pre-packaged Scripts option.
- ii) Select a pre-packaged script and click on it avail of the options.
- iii) Click the 'Copy Script' option.

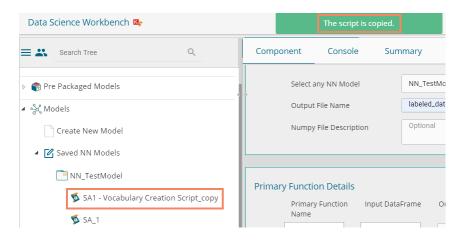


🕘 🌍 Pre Packaged Models
▷ 🌍 Faster RCNN Inception
4 🌍 Sentiment Analysis Model
ိုင်္ဂီ Sentiment Analysis Model
蓫 SA5 - Output Script for WF
蓫 SA4 - Integerization Process Script for WF
View Script Creation Script
Copy Script
SA1 - Vocabulary Creation Script

- iv) The Copy Pre-Packaged Script dialog box opens.
- v) Select the NN Model to Copy the script using the drop-down option.
- vi) Provide a name that you wish to display for the copied pre-package script.
- vii) Click the 'Yes' option.

ipt		×
NN_TestModel		•
Script_copy		
	Yes	No
	ipt NN_TestModel s Script_copy	NN_TestModel s Script_copy

- viii) A success message appears.
- ix) The copied script gets listed below the model.





Data Science Workbench	12 <sub>6</sub>	The script is copied.				
≡ C ♣ Search Tree Q	Create New Workflow					
<ul> <li>Image: Second Sec</li></ul>						
▷ 🎁 Pre Packaged Models						
🔺 - 🔆 Models						
Create New Model						
🔺 📝 Saved NN Models						
NN_TestModel						
🚿 SA3						
🕵 SA2						
💆 SA1	<b>H</b>					

x) The copied script for the 'NN\_TestModel' is as displayed below:

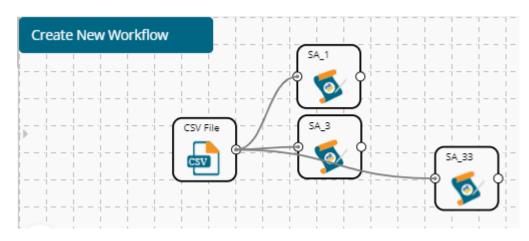
#### Note:

- a. Output for NumPy Script must be a NumPy array. The created NumPy script can be used with any Data-Source, and as the workflow gets completed, the NumPy file gets created and stored for future use with the selected NN Model.
- b. To access a NumPy file from the selected model use, FAKE\_PATH+ '/<filename>.'
- c. To access the shared NumPy file from the Pre-packaged models provided use, SHARED\_PATH+ '/<filename>.'
- d. The user can also add multiple files/scripts and click the 'Apply' option to enable them for the saved model.

## 12.2.3. Running the NumPy Script(s)

The user needs to run the script(s) created or copied to the selected model.

i) Connect the script component(s) to a data source.





- ii) Configure the required fields (The fields for all the script components and data source should be configured)
  - a) Data Source
    - i. Browse a file
    - ii. Click the 'Upload' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Configure	CSV					
Properties	Select File	te	estNN666.csv				
	Please ensu	e csv file follows t	the format giv	en below:			
		in the CSV file shou beaders should not			rd or two words concate	nated by	
						Uplo	ad

- b) Script 1 (SA\_1)
  - i. Configure the Custom Group options
    - 1. Utility Module Script Form
      - a. Select Utility (Package(s) to import (optional)
      - b. Select Utility Script(s) to import (optional)
    - 2. Dynamic Fields
      - a. Select Column Header from the drop-down.
    - 3. Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> ⊥
General	Utility Mod	ule Script Form				
Custom Group	Sele	ect Utility Package(s)	Select	•	(optional)	
	to li	nport				
	Sele	ect Utility Script(s) to	Select		(optional)	
	Imp	ort				
	* So	ript(s) listed under a	ny Utility Pack	age will be		
	ava	ilable in Package itse	lf			
	Dynamic Fi	elds				
•	Select	Column Header	text		•	
						Apply

- c) Script 2 (SA\_2)
  - i. Configure the Custom Group options
    - 1. Utility Module Script Form
      - a. Select Utility (Package(s) to import (optional)
      - b. Select Utility Script(s) to import (optional)
    - 2. Dynamic Fields
      - a. Select Column Header from the drop-down.
      - b. Provide Maximum Sequence Length



#### 3. Click the 'Apply' option.

Component	Console	Summary	Result	Visualization		Properties	+	<u>+</u>
General	Utility Mod	ule Script Form						
Custom Group	Sele	ct Utility Package(s)	Select		•	(optional)		
	to Ir	nport						
	Sele	ct Utility Script(s) to	Select		•	(optional)		
	Imp	ort						
	* Sc	ript(s) listed under a	iny Utility Pack	age will be				
	avai	lable in Package itse	lf					
	Dynamic Fi	elds						
	Select	Column Header	text		•			
	Maxin	num Sequence	140					
	Length	n						
							Apply	,

- d) Script 3 (SA\_3)
  - i. Configure the Custom Group options
    - 1. Utility Module Script Form
      - a. Select Utility (Package(s) to import (optional)
      - b. Select Utility Script(s) to import (optional)
    - 2. Dynamic Fields
      - a. Select Column Header from the drop-down.
      - b. No. of Output Classes.
    - 3. Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	1	Properties	+	Ŧ
General	Utility Mod	lule Script Form						
Custom Group	Sele	ect Utility Package(s)	Selec	t	•	(optional)		
	to l	mport						
	Sele	ect Utility Script(s) to	Selec	t	•	(optional)		
	Imp	port						
	* 50	cript(s) listed under a	any Utility Pa	ckage will be				
	ava	ilable in Package its	elf					
	Dynamic F	ields						
•	Select	t Column Header	Senti	ments	•			
	No. o	f Output Classes	3					
							Арр	ly

- iii) Run the workflow.
- iv) The completion of the process is marked with green checkmarks on the top of the dragged components.



					► C
Create New Workflow	SA1	) - (5A2 )			
		5 A2			
E =		SA3	<b>)</b>		
			<b>,</b>		
			$\begin{array}{cccccccccccccccccccccccccccccccccccc$		
Component C	onsole Summ	ary Result	Visualization	Properties	
	onsole Summ		Visualization	Properties	
	: Process added to Q		Visualization	Properties	
5/9/2019 - 16:53:33 05/09/2019 - 16:52:3	: Process added to Q	ueue	Visualization	Properties	
5/9/2019 - 16:53:33 05/09/2019 - 16:52:3 05/09/2019 - 16:53:1	: Process added to Q 3 : CSV0 is started.	ueue d.		Properties	
5/9/2019 - 16:53:33 05/09/2019 - 16:52:3 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1	: Process added to Q 3 : CSV0 is started. 7 : CSV0 is complete	ueue d. ript_13 - SA1 is starte	d.	Properties	
5/9/2019 - 16:53:33 05/09/2019 - 16:52:3 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1	: Process added to Q 3 : CSV0 is started. 7 : CSV0 is complete 7 : CustomPythonSc	ueue d. ript_13 - SA1 is starte ript_13 - SA1 is comp	d. leted.	Properties	
5/9/2019 - 16:53:33 05/09/2019 - 16:52:3 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1	: Process added to Q 3 : CSV0 is started. 7 : CSV0 is complete 7 : CustomPythonSc 7 : CustomPythonSc 7 : CustomPythonSc	ueue d. ript_13 - SA1 is starte ript_13 - SA1 is comp	d. leted. d.	Properties	
5/9/2019 - 16:53:33 05/09/2019 - 16:52:3 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1	: Process added to Q 3 : CSV0 is started. 7 : CSV0 is complete 7 : CustomPythonSc 7 : CustomPythonSc 7 : CustomPythonSc 7 : CustomPythonSc	ueue d. ript_13 - SA1 is starte ript_13 - SA1 is comp ript_12 - SA2 is starte	d. leted. d. leted.	Properties	
5/9/2019 - 16:53:33 05/09/2019 - 16:52:3 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1 05/09/2019 - 16:53:1	: Process added to Q 3 : CSV0 is started. 7 : CSV0 is complete 7 : CustomPythonSc 7 : CustomPythonSc 7 : CustomPythonSc 7 : CustomPythonSc 7 : CustomPythonSc	ueue d. ript_13 - SA1 is starte ript_13 - SA1 is comp ript_12 - SA2 is starte ript_12 - SA2 is comp	d. leted. d. leted. d.	Properties	

v) The script files are listed under the Supporting File tab for the selected model.

😑 C 🚢 Search Tree	Q Component	Console	Summary	Result	Visualization	Properties	*
🕨 🎲 Data Preparation	General	Suppor	rting Files				
🕨 🎁 Pre Packaged Models	Supporting Files		Name of File		Shape of File	File Description	
🔺 ်င္တိ Models	Summary		vocabulary.npy		(2216)	This file consist the vocab in dictionary format. Eg: {word: word_id}	
Create New Model	Model Script		integrized_logit_d	ata.npy	(159, 140)	This file consist of Logit_data (learning data) in format of Numpy Arrays	
Delete Model	Model Status		labeled_data.npy		(159, 3)	This file consist numpy array for Training Labeled_data for No.of Output Classes as	
NN_TestModel	Tensor Board					provided by user	
💆 SA3							
🕵 SA2							
💆 SA1	ľ						

# 12.2.4. Model Training

This part of the document describes the steps involved in the model training. The entire process of the Model, training involves '**Model Structure**' and '**Model Training**' sections. The user can create a Neural Network Model structure based on his/her problem statement. The user gets three options to form a structure for the selected model:

- i) User Interface
- ii) Script Editor
- iii) Model Configuration

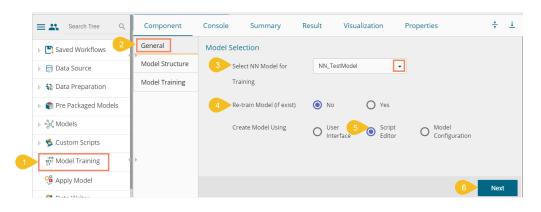
This section describes steps to create a Keras Model Structure using the preprocessed file details. The created model can then be used for training purposes.

i) Click the 'Model Training' tree node.

www.bdb.ai

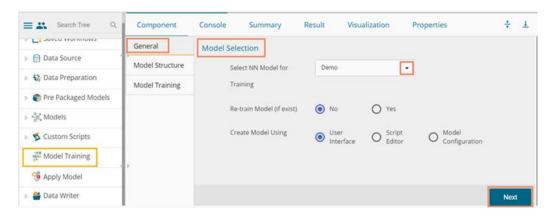


- ii) Configure the Model Selection fields provided under the 'General' tab:
  - a. Select the NN Model for Training: All Created Neural Network Models list here. The user needs to select a Model for which it needs the training.
  - b. Re-train Model (if exist): Opt for this option if the selected model is already created and required to re-train the existing model
  - c. Create Model Using: Select a medium through which the model structure can be created
    - i. User Interface
    - ii. Script Editor
    - iii. Model Configuration
  - d. Click the 'Next' option to proceed.



## 12.2.4.1. Create Model using User Interface

- i) Select the 'User Interface' as a model creation option.
- ii) Click the 'Next' option.



iii) The user gets another page to create the model by drag and drop of the various layers.

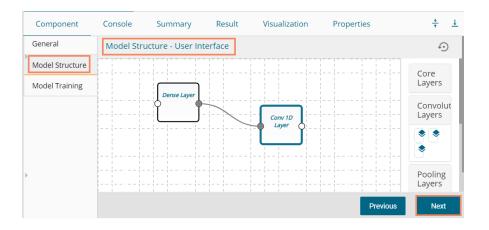


Component	Console	Summary	Result	Visualization	Properties	<u>+</u> <u>↓</u>
General	Model Stru	ucture - User Int	erface			Ð
Model Structure Model Training						Core Layers
		Dense Layer				• •
						Convolut Layers
۶.						Pooling Layers

iv) The user needs to configure each of the dragged layers.

Model Structure - User I	taufaca		
	Layer Node Configuration	ns	⊙ ⊗
	units	3	
Dense Layer	activation	None 🔻	
	use_bias	True •	
	kernel_initializer	glorot_uniform 🔻	
	bias_initializer	Zeros 🔻	
	plas_initializer	Zeros	

#### v) Click the 'Next' option to proceed.



- vi) If users have chosen the '**User Interface**' option to create a model, then a script for the dragged components display on this page. However, the users need to edit the script using the Script Editor to proceed further in the creation of a model.
- vii) Validate the script. A success message should appear after script validation.
- viii) Click the 'Next' option to open the 'Model Training' tab.



Component	Console	Summary	Result	Visualization	Properties	÷ 1
General	Model Stru	ucture - Script B	ditor			
Model Structure Model Training	Validate successfully!	Python script	has been valida	ated	0	
>	2 import k 3 4 model = 6 model.a 7 bias_init 8 kernel_c 9 10 model.a	tializer = 'zeros', ker constraint = None, b dd(Layers.Conv1D	ers nits = 3, activatic nel_regularizer = plas_constraint = (filters = 3, kernel	None, bias_regularizer None, input_shape = (3, I_size = 3, strides = 1, pa	ue, kernel_initializer = 'gloro = None, activity_regularizer 3), output_shape = (3,3))) dding = 'valid', dilation_rate tializer = 'zeros', kernel_regu	= None, = 1, activation
					Previous	Next

# 12.2.4.2. Create Model using Script Editor

- i) Select the 'Script Editor' as a model creation option.
- ii) Click the '**Next**' option.

Component	Console	Summary	Re	sult	Visualization	Properties	<u>+</u> ↑	<u>+</u>
General	Model Selec	tion						
Model	Select	NN Model for	NM	N_TestMode	el 👻			
Model Training	Traini	ng						
	Re-tra exist)	in Model (if	۲	No	O Yes			
	Create	Model Using	0	User Interface	Script Editor	O Model Configuration		
•								
							Next	

- iii) The 'Model Structure' tab opens displaying the Script Editor.
- iv) Provide a relevant python script.
- v) Validate the script. The success message should appear after the script validation.
- vi) Click the 'Next' option to open the Model training tab.



Component	Console	Summary	Result	Visualization	Properties	<u>+</u> ⊥
General	Model Structure	- Script Edito	or			
Model Structure Model Training	Validate Py successfully!	rthon script has	been validateo	I	0	
-	<pre>2 import keras.la 3 4 model = Seque 6 6 model.add(Lay 7 glorot_uniform activity_regular</pre>	ntial() ers.Dense(units = ', bias_initializer = izer = None, kern	= 3, activation = = 'zeros', kernel	None, use_bias = True, regularizer = None, bia None, bias_constraint :		3),
- - »	activation = No kernel_regulariz bias_constraint	ers.Conv1D(filter ne, use_bias = Tr zer = None, bias_	ue, kernel_initia regularizer = No	lizer = 'glorot_uniform',	ing = 'valid', dilation_rate bias_initializer = 'zeros', = None, kernel_constrair	
	10				Previous	Next

# 12.2.4.3. Create Model using Model Configuration

This option can be used for the object deduction models only.

- i) Select the 'Model Configuration' as the Model Creation option.
- ii) Click the 'Next' option.

Component	Console	Summary	R	Result	Visu	alization		Properties	+++	<u>+</u>
General	Model Selectio	on								
Model Training	Select N	N Model for	N	N_TestModel		•				
	Training									
	Re-train exist)	Model (if	۲	No	0	Yes				
	Create N	Nodel Using	0	User Interface		Script Editor	0	Model Configuration		
•										
									Next	

- iii) Selecting the Model Configuration option redirects the user to the '**Model Training**' page with the Model Configuration fields displayed below
- iv) Configure the required fields.
  - a) Model Reference for Training: Select a reference model that can be used to refer to the inputs (At present, it displays the RCNN Inception model only).
  - b) Number of Classes: Provide the value (number) of distinct classes present in your training data.
  - c) Number of Steps: Define the number of steps required for training.
  - d) Initial Learning Rate: Provide the value of the learning rate to start the model training (it should be in 0.00 to 1.00 where 0 and 1 are included).
  - e) Momentum Optimizer Value: Provide value for the optimization function (it should be in 0.00 to 1.00 where 0 and 1 are included).



- f) Schedule Stepwise Learning Rate: enable this option if you wish to schedule the stepwise learning rate.
- g) After enabling the Schedule Stepwise Learning Rate, the user gets to configure the following options
  - 1. No of Steps
  - 2. Learning Rate
- v) Enable the 'Email Notification' option and provide the required information for the same.
- vi) Click the 'Start Model Training' option to begin with the model training.

Component	Console	Summary	Result	Visua	lization	Pro	perties				+	<u>+</u>
General	Model C	onfiguration										
Model Training		Model Reference fo	r Training		Faster RCNN	Incept	ion	-				
7	]	Number of Classes			3							
		Number of Steps			3							
		Initial Learning Rate			0.3							
		Momentum Optimi:	er Value		0.6							
		Schedule Stepwise I	earning Rate.	6	Z							
				Ν	No of Steps 1		Learning Ra 1	te	+			
<i></i>	L	tification Enable Email Notific	ation	C								
								Previo	us	Start Mode	el Traini	ng

Note:

- a. The 'Model Structure' tab does not appear if the selected option for creating the model is Model Configuration.
- b. If the selected model is already undergoing training, it throws an error message.

## 12.2.4.4. Model Training Tab

This section describes steps to select and interpret the variable files to proceed with the model training.

The user can interpret Logit File as independent variables data, which is preprocessed already, and Label File as target (or labeled) data. The selected model learns using the Label File data over the Logit File data and builds up weights internally, which can be used for prediction using the trained model.

- i) Navigate to the Model Training tab using the Model Training tree-node.
- ii) Configure the required fields to Train Model:
  - a. Select Logit Data File: Select the file with logit data using the drop-down option.
  - b. Select Label Data File: Select the file with labeled data using the drop-down option.
  - c. Enter Batch Size: Enter a value for batch size
  - d. Enter Epochs Value: Enter Epochs Value (the suggested value for this field is 4)



- e. Perform Validation Split: Select an option out of Yes/No
- f. Enter Validation Split Value: Enter a value indicating the validation split (the suggested value for this field is .3)
- g. Shuffle: Select an option out of True/False.
- h. Save Intermediate Checkpoint's Weights: Select an option out of **Yes/No**.

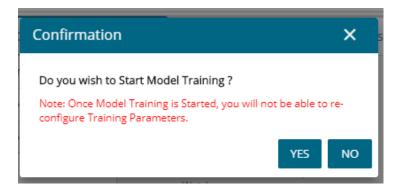
Component	Console	Summary	Result	Visualizatio	on Prope	rties
General	Train Mode	el				
Model Structure	Sele	ct Logit Data File		integrized_logi	it_data 🗸	
Model Training	Sele	ct Label Data File		labeled_data	•	
	Ente	er Batch Size		32		
	Ente	er Epochs Value		4		
	Perf	orm Validation Spli	t	Yes	O No	
	Ente	er Validation Split Va	alue	.3		
	Shu	ffle		True	O False	
•		e Intermediate Cheo ghts	ckpoint's	O Yes	No No	

- iii) Configure the following fields to send Email Notification for success or failure of the model training.
  - a. Enable Email Notification: Enable the box to get email notifications.
  - b. Email Address: Provide a valid email address where the notification can be sent.
  - c. Send Mail when Model Training gets Completed: Enable this option if you wish to get notified when the Model training gets completed.
  - d. Send Mail when Model Training gets failed: Enable this option if you wish to get notified when the Model training gets failed.
- iv) Click the 'Start Model Training' option to begin the training.



Component	Console	Summary	Result	Visualization	Prop	perties	+	<u>+</u>
General								
Model Structure								
Model Training	Email Notific	e Email Notification		9				
	Email	Address		bdbuser@bdb.ai				
		Mail When Model Tr Completed	aining					
· .	Send gets F	Mail When Model Tr ailed	aining 🛛	8				
					Previou	s Start I	Model Traini	ng

- v) A dialog box appears to confirm the action of Model Training.
- vi) Click the 'YES' option to confirm the model training.



- vii) A notification message appears asking the user to check the model status.
- viii) Once the model is trained successfully, the user can use the model for prediction purposes.



x) The 'Stop Training' option appears for the model that is undergoing training.

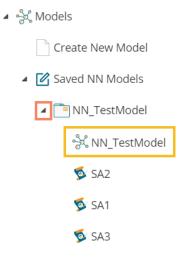


E C Search Tree Q	Component	Console Summary Result Visualization Properties	
Saved Workflows	General	Model Status	
👂 😝 Data Source	Supporting Files	C Refresh Status	
👂 🎲 Data Preparation	Summary		
Pre Packaged Models	Model Script	No Model Status Available, Please Train the Model for Model Status to be Avail	able
4 😽 Models	Model Status		
Create New Model	Tensor Board		
<ul> <li>Saved NN Models</li> </ul>			
▷ 🛅 NN_TestModel			
👂 🚳 Custom Scripts			
Model Training			

- xi) Click the 'Refresh Status' option if the user needs to refresh the model status.
- xii) The 'Model is Trained & Saved Successfully' message appears for the model once the training gets completed.

ble Summary Model Status CRefresh Status Epoch 1 val loss → 0.8208		Visualization Trained & Saved Succ	Properties					*	Ŧ
C Refresh Status		Trained & Saved Succ	essfully						
Epoch 1		Trained & Saved Succ	essfully						
Epoch 1		Trained & Saved Succ	essruny						
val_acc => 0.3542 loss => 0.8218									
acc => 0.6937									
loss => 0.6648									
<pre>val_loss =&gt; 0.6142 val_acc =&gt; 0.7292</pre>									
	<pre>val_loss =&gt; 0.5975 val_acc =&gt; 0.6458 loss =&gt; 0.6648 acc =&gt; 0.5946 Epoch 3 val_loss =&gt; 0.6142 val_acc =&gt; 0.7292</pre>	val_loss ⇒ 0.5975 val_acc ⇒ 0.6458 loss ⇒ 0.6648 acc ⇒ 0.5946 Epoch 3 val_loss ⇒ 0.6142 val_acc ⇒ 0.7292	val_loss => 0.5975 val_acc ⇒> 0.6458 loss => 0.6648 acc => 0.5946 Epoch 3 val_loss => 0.6142	val_loss ⇒ 0.5975 val_acc ⇒ 0.6458 loss ⇒ 0.6648 acc ⇒ 0.5946 Epoch 3 val_loss ⇒ 0.6142 val_acc ⇒ 0.7292	val_loss ⇒> 0.5975 val_acc ⇒> 0.6458 loss ⇒> 0.6648 acc ⇒> 0.5946 Epoch 3 val_loss ⇒> 0.6142 val_acc ⇒> 0.7292	<pre>val_cos =&gt; 0.5975 val_acc =&gt; 0.6458 loss =&gt; 0.6648 acc =&gt; 0.5946 Epoch 3 val_loss =&gt; 0.6142 val_acc =&gt; 0.7292</pre>	val_loss ⇒> 0.5975 val_acc ⇒> 0.6458 loss ⇒> 0.6648 acc ⇒> 0.5946 Epoch 3 val_loss ⇒> 0.6142 val_acc ⇒> 0.7292	val_loss ⇒> 0.5975 val_acc ⇒> 0.6458 loss ⇒> 0.6648 acc ⇒> 0.5946 Epoch 3 val_loss ⇒> 0.6142 val_acc ⇒> 0.7292	<pre>val_acc =&gt; 0.5975 val_acc =&gt; 0.6458 loss =&gt; 0.6648 acc =&gt; 0.5946 Epoch 3 val_loss =&gt; 0.6142 val_acc =&gt; 0.7292</pre>

xiii) After successful completion of the model training, the trained NN model gets added to the created model folder containing the same folder name.





#### Note:

- a. The selected Logit and Label data files should not be the same.
- b. Users can provide details of Batch Size, Epochs, Validation Split as per the model requirement.
- c. Users can track the status of the Model for each epoch, including visual tracking using Tensorboard when the model is undergoing the training process.
- d. Users can stop the model training in between during the period when the model training process is going on.
- e. Users cannot process a Neural Network Model for Model Training if it is already in between the training process.
- f. The user must provide specific parameter values for Model Training purposes.
- g. Since training a model is a time-consuming task, the user can set the Model for training and provide email details to get a notification when the training gets finished or if an error occurs.
- h. Click the 'Summary' tab to view the model summary using the 'View Model' option provided for the selected NN Models. The Summary appears for the trained model.

Component	Console	Summary	Result	Visualization	Prop	oerties	+	<u>+</u>
General	Summary							
Supporting Files								
Summary	Layer (typ	e)	Output Sha	ape	Param #			
	embedding_	1 (Embedding)	(None, 140	0, 128)	51199872			
Model Script	lstm_1 (LS	TM)	(None, 140	0, 64)	49408			
Model Status	lstm_2 (LS	TM)	(None, 140	0, 32)	12416			
Model Data		(Flatten)	(None, 448	30)	0			
Tensor Board	dense_1 (D	ense)	(None, 128	3)	573568			
	dense_2 (D	ense)	(None, 3)		387			
>	Trainable	ms: 51,835,651 params: 51,835,0 ble params: 0	551					

i. Click the '**Model Script**' tab to view the Model script using the '**View Model**' option provided for the Saved NN Models. The Model Script appears for the trained model.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> ↑	<u>+</u>				
General	Model Scri	pt									
Supporting Files	The reasonable is inport Sequential										
Summary	<pre>model = Sequential() model.add(Embedding(len(vocab), 128, input_length=140)) model.add(LSTM(64, return_sequences=True, dropout = 0.2)) model.add(LSTM(32, return_sequences=True, dropout = 0.2)) model.add(Flatten())</pre>										
Model Script											
Model Status											
Model Data											
Tensor Board											
	model.add( model.comp	'nadam', metrics=['a	ccuracy'])								



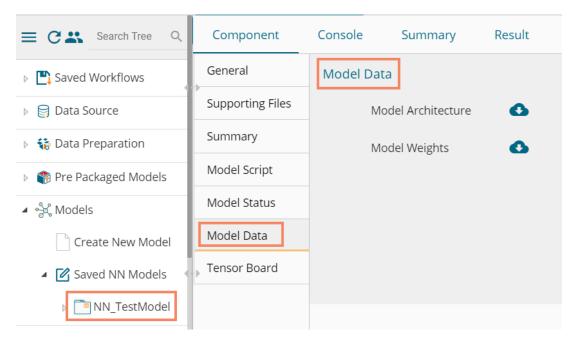
j. Please note that the above given 'Model Training' fields display only when the model creation option is either User Interface or Script Editor. The Model Training tab displays different fields when the model creation option is 'Model Configuration' (which has been already explained within section 10.2.4.3)

## 12.2.5. Model Data

The user can see the Model Data tab with the Model Architecture and Model Weights options, both provided with the download option.

**Model Architecture**: It is metadata for the selected model. It contains the details of layers and the configured parameters. The architecture file gets downloaded in the JSON format as it is a simplified way to recreate the model from JSON with Keras API.

**Model Weight**: The Model Weight option consists of Resultant assigned weights for each layer present in the model architecture during training and/or after the training is completed. The model weights file gets downloaded in the .h5/HDF5 as it is suitable to store multiple data types and extensive data. It can be loaded over a model using the Keras API.



## 12.2.6. Tensor Board

This tab displays live Tensor Board Visualization for the selected model (if enabled). The below image displays a sample visual for the reference of the user.



Component	Console	Summary	Result	Visualization	Properties
General	Те	nsorBoard			
Supporting Files			acc		
Summary	Ser	eothing	0.6 acc		
Model Script	Ho	rizontal Axis		159 158 158	•
Model Status		STUP RELATIVE	WALL 04	157	
Model Data			0		000
Tensor Board			loss		
			01	800 655 645 600 1.000 2.000 3	

# 12.3. Apply Model

This component is provided to generate predictions based on NN trained model. The user can view predicted column value for each label class.

Users can create an NN Apply Model in the following ways:

- Generate a model by pre-processing the selected data and training the model based on the created structure.
- Generate a new NN Apply Model using the saved NN model

The Apply Model within the Deep Learning Workspace consists of 2 input nodes and 1 output node.

- Input Nodes
  - Upper node Model/Training data
  - Lower node Testing data
- Output Node
  - Node Result data

Data Science Workbench 🛯
▶ 😝 Data Source
Apply M
Pre Packaged Models
> 🔆 Models
S Custom Scripts
100 Model Training
% Apply Model



- i) Drag the Apply Model component onto the workspace and connect it with a valid combination of data source and other components to create a valid workflow.
- ii) Click the dragged 'Apply Model' component.
- iii) Basic component details get displayed.
- iv) Click the '**Next**' option to move ahead.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Basic						
Advanced	Component Na	ame	Neural Network Ap	ply Model			
•	Alias		Apply Model0				
	Description		Optional				
						Next	

- v) The Advanced tab opens.
- vi) Select the required columns from the drop-down menu.
- vii) Click the 'Apply' option.

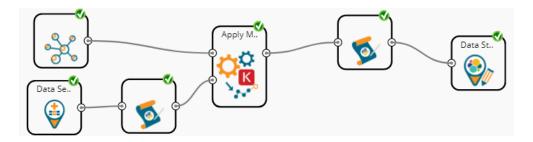
Component	Console	Summary	Result	Visualization	Properties		Ŧ	Ť
General	Advanced							
Advanced	1 checked			•		0		
•								
							Apply	y

- viii) Run the workflow after getting the success message.
- ix) The process status gets displayed under the 'Console' tab.

Component	Console	Summary	Result	Visualization	Properties
2/9/2019 - 16:39:	52 : Process a	dded to Queue			
02/09/2019 - 16:3	38:41 : Demo5	is started.			
02/09/2019 - 16:3	38:41 : Demo5	is completed.			
02/09/2019 - 16:3	38:41 :Data Se	ervice0 is started.			
02/09/2019 - 16:3	39:32 :Data Se	ervice0 is complet	ed.		
02/09/2019 - 16:3	39:32 : Custom	PythonScript - SA	4 - Integeriza	tion Process Script fo	or WF is started.
02/09/2019 - 16:3	39:33 : Custom	PythonScript - SA	4 - Integeriza	tion Process Script fo	or WF is completed.
02/09/2019 - 16:3	39:33 : Apply N	/lodel2 is started.			
02/09/2019 - 16:3	39:36 : Apply N	Iodel2 is complet	ed.		
02/09/2019 - 16:3	39:36 : Custom	PythonScript_19	- SA5 - Outpu	t Script for WF is star	ted.
02/09/2019 - 16:3	39:36 : Custom	PythonScript_19	- SA5 - Outpu	t Script for WF is com	pleted.
02/09/2019 - 16:3	39:36 :Data St	ore Writer5 is sta	rted.		
02/09/2019 - 16:3	39:38 : Data St	ore Writer5 is cor	npleted.		
02/09/2019 - 16:3	39:38 : Process	s Completed			



x) The completion of the Console process gets marked with green checkmarks on the top of the dragged components.



- xi) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged Apply Model component on the workspace.
  - **b.** Click the 'Result' tab.
- xii) The columns displaying numpied\_output probability get added to the Result view. The Apply Model displays the Result in the array format.

Component	Console	Summary	Result	Visualization	Properties		<del>*</del>
how 10 🔻	entries					Sear	ch:
text						Sentiments	numpied_output
Normally, a rest save two dollars would bring me	aurant worker wou AND get a mediur back to this restau	uld say "Állow me t n drink instead of ırant! When I brou	o save you some a small AND a co ght this to the er	! I ordered a bowl and e money by making th pokie". This is the type nployees attention, h Id have dramatically o	is a combo. You'll of service that e made no attempt	Positive	[0.999974250793457, 4.973092018190073e-06, 2.0761452105944045e-05]
bucket had 8 an included in the I not want to bac	d they forgot to ind bag It's very frustra k out to get the res hicken in Brunswid	clude the chocolat iting when you get t of your order. Le	e chip cookies. T home to discove arn to get your c	Today instead of gettin here was no napkins er your order is not co orders right. There are bigger pieces of chicke	or condiments prrect and you do other places to	Positive	[0.9999808073043823, 5.054907319390622e-07, 1.8671187717700377e-05]

Note:

- a. The user can connect the Apply model output to a related Python script to convert the predicted output from the array format to the predicted class Output.
- b. The Result data set of the model can be written to a database using a Data Writer.
- c. The Column header and data type of feature column should match for the saved model and testing data. If column headers and data types do not match, an alert message will be displayed.
- d. It is not mandatory for the testing data set to contain a label column.
- e. The user can view the model summary by clicking on the 'Summary' tab.

## **12.4.** Prediction using Trained Models

Users can use the Saved NN Model in a workflow as displayed below for the prediction purpose:

- i) Select and drag a Data Source onto the workspace.
- ii) Using Custom Python Script Component, create a script that can pre-process the data and transform the input Data Source data into a consumable form by the Apply Model component.
- iii) Drag a trained Neural Network Model and configure it.



- iv) Drag and Apply Model component. The Apply Model provided for the Deep Learning workspace is the same as the Apply Model component provided for the other workspaces; the only difference is in this, the user needs to select the Column Headers on which the Model predicts the values.
- v) After NN Apply Model, put a Custom Python Script to reverse the transform implemented by the previous script component turns the predicted values into the Predicted class Output.
- vi) The predicted output can be written to a Data Writer (in this case, it is the Data Store writer)
- vii) Run the workflow by clearing the previous cache.
- viii) The steps of the Console process get displayed under the 'Console' tab.
- ix) The completion of the Console process is marked by the green checkmarks on the top of the dragged components.

prediction_W	/F_check	Data Se	NN_Test.		pply M.	SA5 - 0.
Componen	t Cor	sole	Summary	Result	Visualization	Properties
04/09/2019 04/09/2019 04/09/2019 04/09/2019 04/09/2019	9 - 13:33:06 9 - 13:33:06 9 - 13:33:06 9 - 13:33:53 9 - 13:33:53	: NN_Testl : NN_Testl : Data Sen : Data Sen : CustomP		leted. red. - SA4 - Intege		ipt for WF is started. ipt for WF is completed.

#### x) The processed data appears under the 'Result' tab.

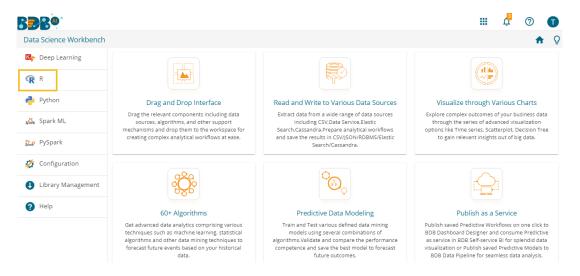
Component	Console	Summary	Result	Visualization	Propertie	s		+
Show 10 🔻	entries			Search:				
text						Sentiments	numpied_output	predicted_result
Normally, a rest You'll save two service that wou made no attemp	aurant worker wou dollars AND get a m uld bring me back to	ld say "Allow me to edium drink inste this restaurant! V le could have at le	o save you som ad of a small AN /hen I brought	n! I ordered a bowl and e money by making thi ND a cookie". This is the this to the employees a cookie, which would ha	s a combo. type of attention, he	Positive	[0.0017110684420913458, 0.9975250363349915, 0.0007638849201612175]	NEGATIVE
my bucket had a condiments incl correct and you There are other	8 and they forgot to luded in the bag It's do not want to bac	include the choco very frustrating w k out to get the re fried chicken in B	late chip cookie hen you get ho st of your order runswick that a	Today instead of gettin es. There was no napki me to discover your or r. Learn to get your ord re less expensive and a blishment	der is not ers right.	Positive	[0.6947669386863708, 0.30521681904792786, 1.620476905372925e-05]	POSITIVE
	ave ever had it has ; best semi-fast food			even their cookies are į	good! I must	Positive	[0.9267953038215637, 0.07320166379213333, 2.993991984112654e-06]	POSITIVE
away while we v	vere talking. Then w	e were supposed	to get a free ler	nat we asked for, and w monade, and the guy re comes with a 10 piece o	fused to give	Negative	[0.0019309261115267873, 0.9980691075325012,	NEGATIVE



# 13. R Workspace

This section of the document describes the R environment by focusing on the Statistical Analysis, Data Preparation connectors, Algorithms, Apply Model, Performance, and Custom R script components to build an R workflow under the Data Science environment.

The user can select the R Workspace from the Predictive landing page to access the R Environment under the BDB Data Science Workbench.



The user gets redirected to the following page by selecting the R Workspace:



# 13.1. Algorithms

Algorithms are a statistical set of rules that help users analyze vast quantities of numerical data and extract appropriate information out of it. The Data Science Workbench allows users to apply more than one algorithm to manage the enormous amount of data.

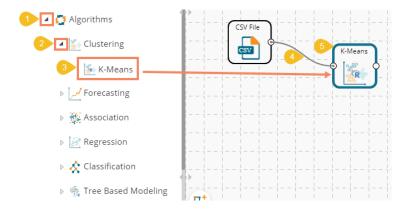
The following algorithms are available in the R Workspace:



4 🧕 Algorithms
▷ 🕍 Clustering
▷ 🛃 Forecasting
Association
▷ 🛃 Regression
🖻 🔆 Classification
> 🍇 Tree Based Modeling
Anomaly Detection
Dimensionality Reduction
⊳ •‡•• Ensemble Models
> 🕂 Instance Based
▷ 🛵 Regularization
💮 Structural Equation Modelling
👫 Automated ML
₩ SVD
ျာ <sup>ရွ</sup> WLS relative Std
🔔 WLSR Input Weights
af BLR
Ferformance Metrics

#### Step by Step Process to configure and connect an Algorithm in the R Workspace:

- i) Click the 'Algorithms' tree-node on the Predictive Analysis home page.
- ii) Click the Algorithm Category tree-node to display the available algorithm subcategories.
- iii) Select and drag an algorithm component onto the workspace.
- iv) Connect the algorithm component to a configured data source.
- v) Click on the algorithm component.



- vi) Configure the following '**Component**' fields for the dragged algorithm component.
- vii) Click the **'Apply'** option to save the information.



Component	Console	Summary	Result	Visualization	Properties	+ + <u>+</u>
General	Output Ir	formation				
Properties	Number Of	Clusters	5		0	
Advanced	Column S	Selection				
	Features		4 checked	•	0	
	New Colu	ımn Informatio	n			
	Cluster Nar	ne	ClusterNumb	per1	0	
>						
						Apply

- viii) Run the workflow.
- ix) The '**Console**' tab opens displaying the step-by-step completion of the process. The green marks on the top of the dragged components mark the completion of the Console process.

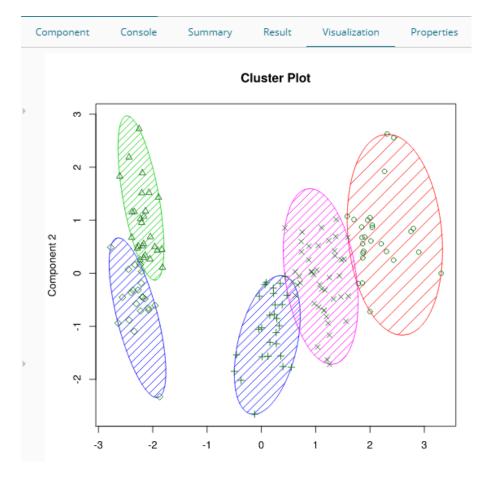
		▶ C +
Create New Workflow		
CSV	File	K-Means
Component	Console	Summary
< Component	: CSV0 is s	tarted
Component 18/09/2019 - 15:39:00	: CSV0 is s : CSV0 is c	tarted

- x) After the Console process gets completed, the user can view Result data using the 'Result' tab.
  - a. Click the algorithm component on the workspace.
  - b. Click the 'Result' tab.
- xi) The newly created Cluster Column gets added to the displayed Result dataset.



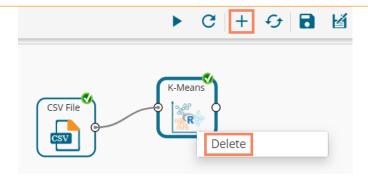
how 10 🔻 entrie	25				Search:
sepal_length	sepal_width	petal_length	petal_width	species	ClusterNumber1
5.1	3.5	1.4	0.2	setosa	2
4.9	3	1.4	0.2	setosa	5
4.7	3.2	1.3	0.2	setosa	5
4.6	3.1	1.5	0.2	setosa	5
5	3.6	1.4	0.2	setosa	2
5.4	3.9	1.7	0.4	setosa	2
4.6	3.4	1.4	0.3	setosa	5
5	3.4	1.5	0.2	setosa	2
4.4	2.9	1.4	0.2	setosa	5
4.9	3.1	1.5	0.1	setosa	5

#### xii) Click the 'Visualization' tab to see a graphical representation of the Result data.



xiii) Click the **'Delete'** option or the icon for the **'Create New Workflow'** option to remove the selected algorithm component from the workspace.





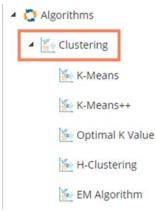
Note:

- a. The user can follow the steps mentioned above to configure all the available R- algorithms.
- b. The user can configure the alias name for the algorithm component via the 'General' tab.
- c. The basic configuration for all the algorithms is done through the '**Properties**' tab. The user is required to configure this tab while Applying an algorithm component manually.
- d. The user can avail of all the default values under the 'Advanced' tab. The user can manually set the 'Advanced' tab or modify the default values, only if the advanced level configuration is required.
- e. After execution, The user can click on the respective component to get data. The pipeline component does not have any Result set; it has only the summary to see. Users need to connect the pipeline components with an 'Apply Model' component and test dataset to view the Result.

## 13.1.1. Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

The Clustering tree node contains the below-given algorithms:



# 13.1.1.1. K-Means

K- means clustering is one of the most commonly used clustering methods. It clusters data points into a predefined number of clusters. It first assembles observations into 'K' groups, wherein 'K' is an input parameter. The algorithm then assigns each observation to a cluster based on the proximity of the observation.



#### Applying R-K Means to a Data Source

Users will be redirected to the '**Component**' tabs when Applying the '**R-K Means**' algorithm component to a configured data source.

- i) Drag the R-K Means to the Workspace and connect it to a configured Data Source.
- ii) The Component tabs get displayed on the View space.
- iii) Configure the following fields in the 'Properties' tab:
  - a. Output Information
    - i. **Number of Clusters**: Enter the number of groups for clustering. The default value for this field is 5. The range should be between 1 and the total number of clusters.
  - b. Column Selection
    - i. Features: Select the input columns with which you want to perform the Analysis
  - c. New Column Information
    - i. Cluster Name: Enter a name for the new column displaying cluster number.

Component	Console Summa	ry Result	Visualization	Properties	<del>+</del> †	⊥
General	Output Information					
Properties	Number Of Clusters	5	0			
Advanced	Column Selection					
	Features	4 checked	• <b>0</b>			
	New Column Informa	tion				
	Cluster Name	ClusterNumber	6			
•						
					Apply	

#### • Rules for Naming a New Column

- 1. Do not use space in the name of a new column. It should be a single word, or two words should be connected by an underscore (\_). E.g., SampleData or Sample\_Data.
- 2. Do not use any special symbol alone or with any character as the name of a new column. Eg. %, #, \$, @,\* or Sample# are not acceptable.
- 3. Do not use single or double quotes, dots, and brackets, to name a new column.
- **4.** Do not use numbers alone while naming a new column. Numbers can be used with at least one character of the alphabet, and the name should not begin with a numeral.
- 5. The name given to a new column should not exceed 50 characters.

Note: Users can access a list of rules for naming a new column by clicking the information

icon 💿 provided Next to the **'New Column Information'** tab.

- iv) Click the 'Advanced' tab (if required)
  - a. Configure the required 'Behavior' fields:
    - i. **Maximum Iterations**: Enter the number of iterations allowed for discovering clusters. (The default value for this field is 100).
    - ii. **Number of Initial Centroids**: Enter the number of random initial centroid sets for clustering (The default value for this field is 1).
    - iii. Algorithm type: Select an algorithm type from the drop-down menu



- iv. **Initial Cluster Center Seed**: Enter a number indicating the initial cluster center seed (The default value for this field is 10).
- v) Click the **'Apply'** option.

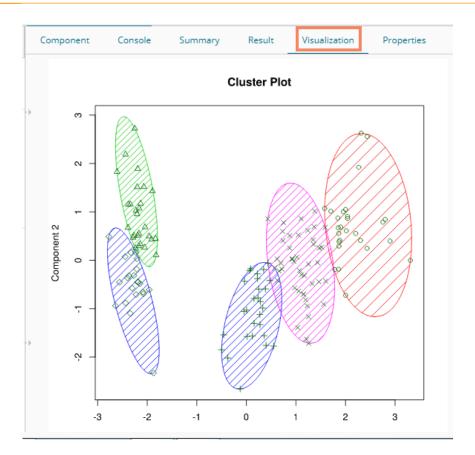
Component	Console	Summary	Result	Visualiza	ation	Properties	<del>+</del> †	<u>+</u>
General	Behavior							
Properties	Maximum Itera	tions	100					
Advanced	Number of intia	al	1					
	centroids							
	Algorithm Type		Hartigan-Wong	•				
	Initial Cluster C	enter	10					
•	Seed							
							Appl	y

- vi) Run the workflow after getting the success message.
- vii) The '**Console**' tab opens describing the progress of the process. The completion of the Console process gets marked by the green checkmarks on the top of the dragged component.
- viii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - **b.** Click the 'Result' tab.
- ix) A new column 'Cluster Number' gets displayed in the Result view.

Component	Console Su	mmary Resul	t Visualizatior	n Proper	ties 🕂
Show 10 • e	ntries			Search:	
sepal_length	sepal_width	petal_length	petal_width	species	ClusterNumber
5.1	3.5	1.4	0.2	setosa	2
4.9	3	1.4	0.2	setosa	5
4.7	3.2	1.3	0.2	setosa	5
4.6	3.1	1.5	0.2	setosa	5
5	3.6	1.4	0.2	setosa	2
5.4	3.9	1.7	0.4	setosa	2
4.6	3.4	1.4	0.3	setosa	5
5	3.4	1.5	0.2	setosa	2
4.4	2.9	1.4	0.2	setosa	5
4.9	3.1	1.5	0.1	setosa	5
Showing 1 to 10 of	150 entries		Previous 1	2 3 4	5 15 Next

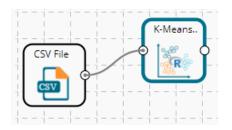
- x) Click the 'Visualization' tab.
- xi) The Result data gets displayed via the Cluster Plot Chart.





### 13.1.1.2. K-Means ++

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Feature Column: Select the feature column from the input data set.
  - b. Number of Clusters: Set the number of clusters to be created (The default value for this field is 5).
  - c. Max-Iteration: Enter the number of iterations allowed for discovering clusters. (The default value for this field is 20).
  - d. Seed Value: A random seed (or seed state, or just seed) is a number (or vector) used to initialize a pseudorandom number generator. Set the initial seed value (the default seed value for this field is 10).
- iv) Click the '**Apply**' option.



Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ŧ
General	Dynami	c Fields					
Custom Group	F	eature Column		5 checked	•		
	١	lumber Of Cluster		3			
	N	laxIteration		30			
*	s	eedValue		5			
						Арр	ly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

now 10	✓ entries				Search:	
Number	sepal_length	sepal_width	petal_length	petal_width	species	ClusterNumber
1	5.1	3.5	1.4	0.2	setosa	3
2	4.9	3	1.4	0.2	setosa	3
3	4.7	3.2	1.3	0.2	setosa	3
4	4.6	3.1	1.5	0.2	setosa	3
5	5	3.6	1.4	0.2	setosa	3
6	5.4	3.9	1.7	0.4	setosa	3
7	4.6	3.4	1.4	0.3	setosa	3
8	5	3.4	1.5	0.2	setosa	3
9	4.4	2.9	1.4	0.2	setosa	3
10	4.9	3.1	1.5	0.1	setosa	3

viii) Click the 'Summary' tab to open the model summary.

### www.bdb.ai



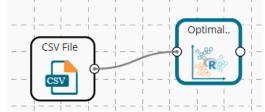
Co	ompone	ent		Cor	nsol	e		S	um	ma	ary		_	F	Res	ul	t			Vi	su	ıal	iza	tic	n			1	Properties	+	<u>+</u>
	K-means		Sum		-				_		si	ze	s S	0,	5(	),	50														
•	Cluster	r mean er sep		eng	th i	sen	al 1	d	th	net	al	le	nøt	h	net	al	ы	id	th												
	1 75.																	.3													
	2 125.																_	.0													
	3 25.																														
	Cluster	ing v	/ecto	or:																											
		333																													
	[38] 3																														
	[75] 1																														
	[112] 2		22	2 2	2 3	2 2	2 3	2 2	2	2 2	2	2	2 2	2 2	2	2	2	2	2 2	2 2	2	2	2	2	2	2	2 3	2 2	2		
	[149] 2	2																													
	Within	clust	ter s	um	of	squa	are	s b	y c	lus	ter																				
	[1] 104	43.12	2 104	56.	03 :	104	27.3	74																							
	(betwe	en_SS	5 / t	ota	1_5	5 =	81	3.9	%)																						
	Availab	le co	mpor	ent	s:																										
	[1] "d	luste	er"			"c(	ente	ers					tot	ss						"w	it	hi	ns:	s"							
	[5] "t	ot.wi	thir										si	e"						"i	te	r"									
	[9] "i																														
			End	l of	Sur	nmai	ry ·				-																				

ix) Click the 'Visualization' tab to see a scatter plot chart representing the processed data.



## 13.1.1.3. Optimal K Value

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



iii) Configure the 'Dynamic Fields' of the dragged algorithm component.a. Feature Column: Select the feature column from the input data set



### iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Dynamic	Fields					
Custom Group	Fe	ature Column		6 checked	•		
•							
						Appl	у

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.

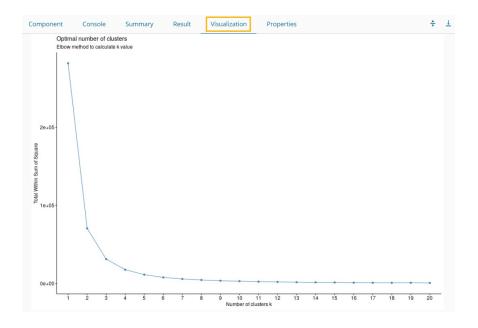


- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.



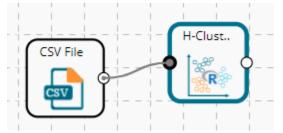
iow 10 🗸	entries			Search:	
Number	sepal_length	sepal_width	petal_length	petal_width	species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
3	5	3.4	1.5	0.2	setosa
Э	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

#### viii) Click the 'Summary' tab to open the model summary.



### 13.1.1.4. H-Clustering

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



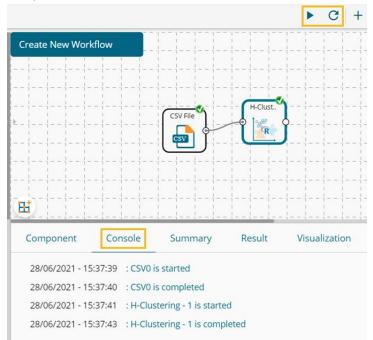
- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Features: Select the feature columns from the input data set
  - b. Number of Clusters: Set the number of clusters to be created (The default value for this field is 5).



- c. Method of Distance: Select a method of distance out of the given choices that include centroid, median, mcquitty, average, single, complete, ward.D2, ward.D
- d. Clustering Method: Select a clustering method out of the given choices that include euclidean, manhattan, maximum, binary, canberra, minkowski
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	Ŧ	Ť
General	Dynamic I	Fields					
Custom Group	Feat	ures	5 cł	necked	•		
	Nun	nber of Clusters	5				
	met	hod of Distance	cer	ntroid	~		
•	Clus	tering Method	eu	clidean	~		
						Apply	/

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result'** tab.

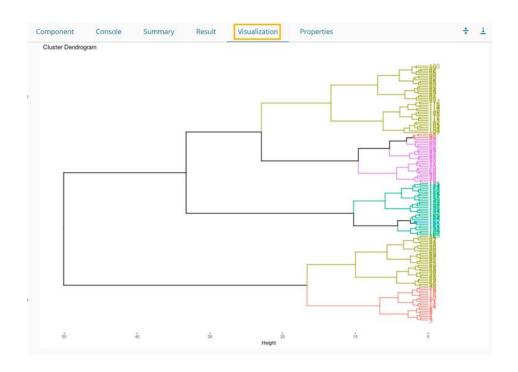


Component	Console	Summary Result	Visualization	Properties	÷ -
how 10	• entries			Search:	
Number	sepal_length	sepal_width	petal_length	petal_width	clusters
1	5.1	3.5	1.4	0.2	1
2	4.9	3	1.4	0.2	1
3	4.7	3.2	1.3	0.2	1
4	4.6	3.1	1.5	0.2	1
5	5	3.6	1.4	0.2	1
6	5.4	3.9	1.7	0.4	1
7	4.6	3.4	1.4	0.3	1
8	5	3.4	1.5	0.2	1
9	4.4	2.9	1.4	0.2	1
10	4.9	3.1	1.5	0.1	1

viii) Click the 'Summary' tab to open the model summary.



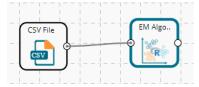
ix) Click the 'Visualization' tab to see the processed data represented through a Cluster Dendrogram chart.





## 13.1.1.5. EM Algorithm

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Feature Column: Select the feature columns from the input data set
  - b. Number of Cluster: Set the number of clusters to be created (The default value for this field is 5).
  - c. Seed Value: A random seed (or seed state, or just seed) is a number (or vector) used to initialize a pseudorandom number generator. Set the initial seed value (the default seed value for this field is 10).
- iv) Click the 'Apply' option.

Component	Console	Summary	Re	esult	Visualization		Properties	<u>+</u>	Ť
General	Dynamic Fi	elds							
Custom Group	Featur	e Column		5 checke	d	•			
	Numb	er Of Cluster		10					
	Seed \	/alue		1996					
•									
								Арр	ly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.

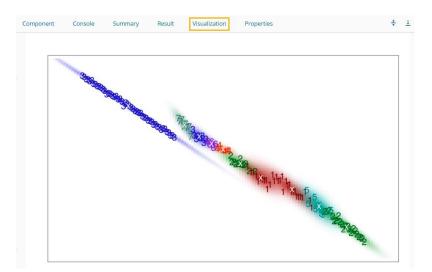




- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

how 10	✓ entries				Search:	
Number	sepal_length	sepal_width	petal_length	petal_width	species	ClusterNumber
1	5.1	3.5	1.4	0.2	setosa	10
2	4.9	3	1.4	0.2	setosa	10
3	4.7	3.2	1.3	0.2	setosa	10
4	4.6	3.1	1.5	0.2	setosa	10
5	5	3.6	1.4	0.2	setosa	10
6	5.4	3.9	1.7	0.4	setosa	10
7	4.6	3.4	1.4	0.3	setosa	10
8	5	3.4	1.5	0.2	setosa	10
9	4.4	2.9	1.4	0.2	setosa	10
10	4.9	3.1	1.5	0.1	setosa	10

- viii) Click the 'Summary' tab to open the model summary.
- ix) Click the 'Visualization' tab to see the processed data through a chart.



### 13.1.2. Forecasting

Forecasting is a method used extensively in time series analysis to predict a response variable, such as monthly profits, stock performance, or unemployment figures, for a specified period. Forecasts are based on patterns in existing data. For example, a warehouse manager can create a model of how much product to order for the next three months based on the previous 12 months of orders.

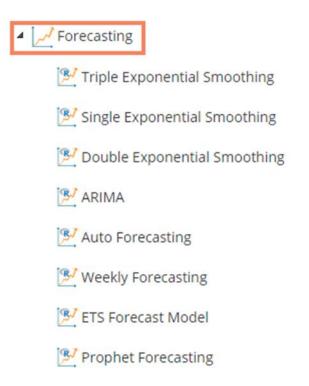
All the sub-categories of the Forecasting Algorithms provide two Output modes (to be set from the Properties tab):

1. Forecasting



2. Trend

The document describes all the available Forecasting algorithms considering both the output modes as possibilities.



# 13.1.2.1. Triple Exponential Smoothing

Triple exponential smoothing considers seasonal changes as well as trends (all of which are trends). Seasonality is defined to be the tendency of time-series data to exhibit behavior that repeats itself every L period, much like any harmonic function. The term season is used to represent the period before behavior begins to repeat itself. There are different types of seasonality: 'multiplicative' and 'additive' in nature, much like addition and multiplication are fundamental operations in mathematics.

i) Drag the Triple Exponential Smoothing component to the workspace and connect it to a configured data source.



💶 🧕 Algorithms	<b>.</b>
▷ 🎆 Clustering	
Forecasting	Triple
🔀 Triple Exponential Smoothing	CSV File
屋 Single Exponential Smoothing	
屋 Double Exponential Smoothing	
🖻 ARIMA	
🖻 Auto Forecasting	

- ii) Configure the following fields in the 'Properties' tab:
  - a. Output Information
    - i. **Output Mode**: Select a mode in which you want to display output data. The user gets two options for this field.
      - 1. **Trend**: Selecting this option displays source data along with predicted values for the given data set.
      - 2. **Forecast**: Selecting this option displays forecasted values for the given period. The forecasted values get appended to the target column when the '**Forecast**' output mode has been selected.
    - ii. **Period to Forecast**: Enter a period to forecast. This field appears only when the selected **'Output Mode'** option is **'Forecast.'**
  - b. Column Selection
    - i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (the first selected option gets selected by default. Only numerical columns are accepted.)
  - c. Input Data Handling
    - i. **Period:** Select a period of forecasting by choosing any one option from the drop-down menu.

Quarter	-
---------	---

Month

Custom

- ii. **Start Period:** Enter a value between 1 and the value specified for the selected option for the **'Period'** field.
- iii. **Start Year:** Enter a year from which you want the data entries to be considered. Enter **a** four-digit value for selecting a year (E.g., 2000)
- d. New Column Information



i. **Period Column Name:** Enter a name for the column containing a period value. (This field is predefined, but users can change the value if needed).

Component	Console	Summary	Result	Visualization	Properties		<del>*</del>	1
General	Output Inform	ation						
Properties	Output Mode		Forecast	-				
Advanced	Period To Forecas	t [	1					
	Column Selecti	ion						
	Target Variable	1	Beer_Sales	-	0			
	Input Data Har	ndling						
	Period		Quarter	-				
	Start Period		1					
	Start Year	3	2000					
	New Column Ir	nformation						
•	Period Column Na	ime (	Quarter		0			
						A	ply	

- iii) Click the 'Advanced' tab and configure, if required:
  - a. Configure the following 'Behavior' fields:
    - i. Alpha: Enter a valid double value in the given field for smoothing observations (Alpha Range: 0<alpha<=1)
    - ii. **Beta:** Enter a valid double value in the given field for finding trend parameters (Beta Range: 0-1)
    - iii. **Gamma:** Enter a valid double value in the given field for finding a seasonal trend parameter (Gamma Range: 0-1)
    - iv. **Seasonal:** Select a smoothing algorithm type from the drop-down list (Holtwinter's Exponential Smoothing algorithm)
    - v. **No. of Periodic Observation:** Enter the number of periodic observations required to start the calculation. The default value for this field is 2.

General	Behavior		
Properties	Alpha	.3	0
Advanced	Beta	.1	0
	Gamma	.1	0
	Seasonal	Additive -	
	No. of Periodic	2	0
	Observation		

- b. Configure the following 'Initial Values' information:
  - i. Level: Enter the initial value for the level. It is an optional field.
  - ii. Trend: Enter the initial value for finding trend parameters. It is an optional field.



- iii. **Season:** Enter initial values for finding seasonal parameters. It depends on the selected column. It is an optional field.
- iv. **Optimizer Inputs:** Enter the initial values given for alpha, beta, gamma required for the optimizer. It is an optional field.
- v. **Confidence**: Enter Confidence level for prediction intervals. It accepts only 0-99 and a comma-separated value. According to the number of comma-separated values, new low and high range columns get added to the Result dataset. (the default value for this field is 95)
- vi. Show Range: Select an option using the drop-down menu.
  - 1. True: By selecting this option, **Lower Range** and **Upper Range** get displayed in the Result and Visualization of the dataset.
  - 2. False: By selecting this option, Ranges do not get displayed in the dataset
- iv) Click the **'Apply'** option.

General Properties	Initial Values			
Advanced	Level	Optional		
Auvanceu	Trend	Optional		
	Season	Optional		
	Optimizer Inputs	Optional		
	Confidence	95		
•	Show Range	False 👻		
			Appl	y

- v) Run the workflow after getting the success message.
- vi) The user gets directed to the 'Console' tab displaying the ongoing process. The completion of the Console process gets marked by the green checkmarks on the top of the dragged component.

	Create New Work	flow		► C	+ &
	Ħ			ple	
	Component	Console	Summary	Result	Visualization
	20/09/2019 - 10:	20:32 : CSV1	is started		
	20/09/2019 - 10:	20:33 : CSV1	is completed		
•••	20/09/2019 - 10:	20:33 : Triple	e Exponential Smoo	othing0 is start	ed
	20/09/2019 - 10:	20:35 : Triple	e Exponential Smoo	othing0 is com	pleted

- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.

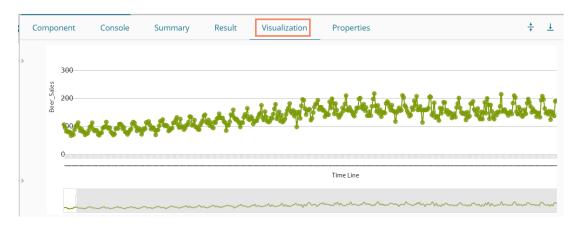


Component	Console Summa	ary Result Visualizati	on Properties	<u>+</u> ↑
Show 10 🔻	entries		Search:	
Year	Month	Beer_Sales	Quarter	
1965	January	93.2	Q1 2000	
1965	February	96	Q2 2000	
1965	March	95.2	Q3 2000	
1965	April	77.1	Q4 2000	
1965	May	70.9	Q1 2001	
1965	June	64.8	Q2 2001	
1965	July	70.1	Q3 2001	
1965	August	77.3	Q4 2001	
1965	September	79.5	Q1 2002	
1965	October	100.6	Q2 2002	

b. Click the 'Result' tab (In this case, the selected output mode is 'Forecasting')

### viii) Click the 'Visualization' tab.

### ix) The Result data will be displayed via the TimeLine Chart.



x) Click the '**Summary**' tab to view the model summary.



	Component	Console	Summary	Result	Visualization	Properties	*	<u>1</u>
•	Columns used							
	Call: HoltWinters(x	= tso, alpha riods = as.num	= as.numeric(0.3)	, beta = as	tive seasonal compone .numeric(0.1), gr optim.start = c())		, seasonal = c("additiv	
•	Coefficients: [,1] a 111.0213 b -3.1634 s1 -4.2978 s2 -1.4135 s3 12.6552 s4 -0.8968	End of Summary	,					

# 13.1.2.2. Single Exponential Smoothing

The Single Exponential Smoothing is the simplest of all the smoothing methods, also known as Simple Exponential Smoothing. This method is suitable for forecasting data with no trend or seasonal pattern.

i) Drag the Single Exponential Smoothing component to the workspace and connect it to a configured data source.



- ii) Configure the '**Properties**' tab.
  - a. Output Information
    - i. Output Mode: Select a mode in which you want to display output data
      - Trend: Selecting this option displays source data along with predicted values for the given data set. A new column 'Predicted Values' gets added in the Result view when the 'Trend' output mode has been selected.
      - 2. **Forecast**: Selecting this option displays forecasted values for the given period. The forecasted values get appended to the target column when the '**Forecast**' output mode has been selected.



ii. Period to Forecast: Enter a period to forecast. This field appears only when the selected 'Output Mode' option is 'Forecast.'

### b. Column Selection

- i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (the first option gets selected by default. Only numerical columns are accepted)
- c. Input Data Handling
  - i. **Period:** Select period of forecasting by choosing any one option from the drop-down menu
  - ii. Period Per Year: This field appears only when the selected 'Period' option is 'Custom.'
  - iii. Start Period: Enter a value between 1 and the value specified for the selected option for the 'Period' field
  - iv. **Start Year:** Enter a year from which you want the data entries to be considered. Enter **a** four-digit value for selecting a year (E.g., 2000)
- d. New Column Information
  - i. **Period Column Name:** Enter a name for the column containing a period value. (This field comes predefined, but the user can change the value if needed).

Component	Console Summary	y Result	Visualization	Properties	+	Ŧ
General	Output Information					
Properties	Output Mode	Forecast	-			
Advanced	Period To Forecast	1				
	Column Selection					
	Target Variable	Beer_Sales	• <b>0</b>			
	Input Data Handling					
	Period	Custom	•			
	Periods per year	4				
	Start Period	1				
	Start Year	2000				
•	New Column Informati	ion				
	Period Column Name	PeriodColumn	0			
					Apply	/

Note: The '**Period Per Year**' field gets displayed only when the selected value for the '**Period**' field is '**Custom**.'

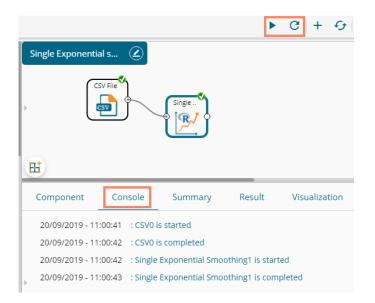
- iii) Click the 'Advanced' tab and configure if required.
  - a. Configure the following 'Behavior' fields:
    - i. Alpha: Enter a valid double value in the given field for smoothing observations. Alpha Range: 0<alpha<=1.
    - ii. **No. of Periodic Observation:** Enter the number of periodic observations required to start the calculation. The default value for this field is 2.
  - **b.** Configure the following 'Initial Values' information:
    - i. Level: Enter the initial value for the level. It is an optional field.



- ii. **Confidence**: Enter Confidence level for prediction intervals. It accepts only 0-99 and a comma-separated value. According to the number of comma-separated values, new low and high range columns get added to the Result dataset. (the default value for this field is 95)
- iii. Show Range: Select an option using the drop-down menu.
  - 1. True: By selecting this option, **Lower Range** and **Upper Range** get displayed in the Result and Visualization of the dataset.
  - 2. False: By selecting this option, Ranges do not get shown in the dataset.
- iv) Click the 'Apply' option.

Component	Console	Summary	e Result	Visuali	zation	Properties	<u>+</u> ↑	Ţ
General	Behavior							
Properties	Alpha		.3		0			
Advanced	No. of Periodic		2		0			
	Observation							
	Initial Values							
	Level		95					
	Confidence		95					
	Show Range		True	•				
P								
							Арр	ly

- v) Run the workflow after getting the success message.
- vi) The 'Console' tab opens, displaying the ongoing process. The completion of the Console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace
  - **b.** Click the 'Result' tab.



viii) Predicted values get appended to the target variable column when the selected output mode is Forecasting. The Lower Range and Upper Range columns display when the 'Show Range' field is marked 'True' from the Advanced tab.

how	10 • entries	5		Sea	rch:
Year	Month	Beer_Sales	PeriodColumn	Lower_Range_95_12	Upper_Range_95_12
1965	January	93.2	Q1 2000		
1965	February	96	Q2 2000		
1965	March	95.2	Q3 2000		
1965	April	77.1	Q4 2000		
1965	May	70.9	Q1 2001		
1965	June	64.8	Q2 2001		
1965	July	70.1	Q3 2001		
1965	August	77.3	Q4 2001		
1965	September	79.5	Q1 2002		
1965	October	100.6	Q2 2002		

- ix) Click the 'Visualization' tab.
- x) The Result data gets displayed via the **TimeLine** Chart.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
300							
Sales - 500				3 In Se Sentral	And the production of the first the test of the	t	
	~~~~	a a fairth	ŴŶ	Solution de .			
0					Time Line	<b>.</b>	
	~~~~~		~~~~~				

xi) Click the 'Summary' tab to view the model summary.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
Columns used i	in the algorit Beer_Sales	(double)		ithout seasonal compo	nent.		
Call:	= tso, alpha = 95)	-			start.periods = as.n	umeric	
Coefficients: [,1] a 116.3	ind of Summary	y					



# 13.1.2.3. Double Exponential Smoothing

The Single Exponential smoothing method cannot perform well when there is a trend in the data. In such circumstances, several methods were devised under the name Double Exponential Smoothing or Second-order Exponential Smoothing, which is the recursive application of an exponential filter twice. Therefore it was termed Double Exponential Smoothing. The basic idea behind double exponential smoothing is to introduce a term to consider the possibility of a series exhibiting some form of the trend. This slope component is itself updated via exponential smoothing.

i) Drag the Double Exponential Smoothing component to the workspace and connect it to a configured data source.



ii) Configure the 'Properties' tab

### a. Output Information

- i. Output Mode: Select a mode in which you want to display output data
  - Trend: Selecting this option displays source data along with predicted values for the given data set. A new column 'Predicted Values' gets added in the Result view when the 'Trend' output mode has been selected.
  - 2. **Forecast**: Selecting this option displays forecasted values for the given period. The forecasted values get appended to the target column when the '**Forecast**' output mode has been selected.
- ii. **Period to Forecast**: Enter a period to forecast. This field appears only when the selected **'Output Mode'** option is **'Forecast.'**

### b. Column Selection

- i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (The first selected option gets selected by default. Only numerical columns are accepted.)
- c. Input Data Handling
  - i. **Period:** Select a period of forecasting by choosing any one option from the drop-down menu.
  - ii. **Start Period:** Enter a value between 1 and the value specified for the selected option for the **'Period'** field
  - iii. **Start Year:** Enter a year from which you want the data entries to be considered. Enter **a** four-digit value for selecting a year (E.g., 2000)
- d. New Column Information



i. Period Column Name: Enter a name for the column containing the period value (This field is predefined, but users can change the value if needed)

Component	Console	Summary	Result	Visualization	Properties	÷ 1
General	Output Inf	ormation				
Properties	Output Mode	Output Mode		-		
Advanced	Period To Forecast		1			
	Column Se	Column Selection				
	Target Variab	ble	Beer_Sales	•	0	
	Input Data	Input Data Handling				
	Period		Month	•		
	Start Period		1			
	Start Year		2000			
	New Colur	nn Informatio	n			
•	Period Colun	nn Name	Months		0	
						Apply

Note: The user can click the 'Apply' option from the Properties tab if the configuration of the Advanced tab is not required.

#### iii) Click the 'Advanced' tab and configure if required

- a. Configure the following 'Behavior' fields:
  - i. Alpha: Enter a valid double value in the given field for smoothing observations (Alpha Range: 0<alpha<=1)
  - ii. **Beta:** Enter a valid double value in the given field for smoothing observations (Beta Range: 0-1)
  - iii. **No. of Periodic Observation:** Enter the number of periods observations required to start the calculation (The default value for this field is 2)
- **b.** Configure the following 'Initial Values' information:
  - i. Level: Enter the initial value for the level (It is an optional field)
  - ii. Trend: Enter the initial value for finding trend parameters (It is an optional field)
  - iii. **Optimizer Inputs:** Enter the initial values given for alpha and beta required for the optimizer (it is an optional field)
  - iv. **Confidence**: Enter Confidence level for prediction intervals. It accepts only 0-99 and comma-separated values. According to the number of commas separated values, new low and high range columns get added to the Result dataset (the default value for this field is 95).
  - v. Show Range: Select an option using the drop-down menu
    - True: By selecting this option 'Lower Range' and 'Upper Range' get displayed in the Result and Visualization of the dataset
    - 2. False: By selecting this option, Ranges do not get shown in the dataset
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Behavior						
Properties	Alpha	.3		0			
Advanced	Beta	.1		0			
	No. of Periodic	2		0			
	Observation						
	Initial Values						
	Level	Opti	onal				
	Trend	Opti	onal				
	Optimizer Inputs	Opti	onal				
•	Confidence	95					
	Show Range	True		-			
						Appl	y

- v) Run the workflow after getting the success message.
- vi) The 'Console' tab opens, displaying the ongoing process. The completion f the Console process gets marked by the green checkmarks on the top of the dragged components.

			► C +	- 🕁 🖪 🕍
D	ouble Exponential			
•	CSV File	}	Double	
H	1			
<	1 Component	Console	Summary	Result
<				Result
<	Component	: CSV0 is s	tarted	Result
<	Component 20/09/2019 - 12:25:24	: CSV0 is s : CSV0 is c	tarted ompleted	

- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

The Predicted values get appended to the target column in the Result data if the selected output mode is **Forecasting**.



now 10 🔻	entries		Search:
Year	Month	Beer_Sales	Months
1965	January	93.2	Jan 2000
1965	February	96	Feb 2000
1965	March	95.2	Mar 2000
1965	April	77.1	Apr 2000
1965	May	70.9	May 2000
1965	June	64.8	Jun 2000
1965	July	70.1	Jul 2000
1965	August	77.3	Aug 2000
1965	September	79.5	Sep 2000
1965	October	100.6	Oct 2000

- viii) Click the 'Visualization' tab.
- ix) The Result data will be displayed via the TimeLine chart.

Co	mponent	Console	Summary	Result	Visualization	Properties
	300 Sales 200 200	~~~~	~~~~~	Ŵ	ين الأيم عارية معارية م	not have be my had for the trade of the had a failed and a
						Time Line
						~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

x) Click the 'Summary' tab to view the model summary.

	Component Console	Summary	Result	Visualization	Properties	+	<u>+</u>
	Summary of the	model					
•	Columns used in the algorith						
	Holt-Winters exponential s	noothing with tr	end and witho	ut seasonal compon	ent.		
	Call: HoltWinters(x = tso, alpha = = as.numeric(2), optim.start		), beta = as.	numeric(0.1),	gamma = FALSE, star1	t.periods	
	Smoothing parameters:						
	alpha: 0.3						
	beta : 0.1						
	gamma: FALSE						
	Coefficients:						
	[,1]						
1	a 116.051						
	b -2.966						
	End of Summary						

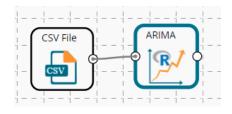
### 13.1.2.4. ARIMA

ARIMA returns the best ARIMA model according to either AIC, AICC, or BIC value. The function searches for a possible model within the order constraints provided.

i) Drag the ARIMA component to the workspace and connect it to a configured data source.



ii) Click on the algorithm component to get the configuration fields.



- iii) Configure the 'Properties' tab.
  - a. Output Information
    - i. Output Mode: Select a mode in which you want to display output data
      - 1. **Trend**: Selecting this option displays source data along with predicted values for the given data set. A new column '**Predicted Values**' gets added in the Result view when the '**Trend**' output mode has been selected.
      - 2. **Forecast**: Selecting this option displays forecasted values for the given period. The forecasted values get appended to the target column when the '**Forecast**' output mode has been selected.
    - ii. Period to Forecast: Enter a period to forecast. This field appears only when the selected 'Output Mode' option is 'Forecast.'
  - b. Column Selection
    - i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (the First option gets selected by default. Only numerical columns are accepted.

Component	Console	Summary	Result	Visualization	Properties
General	Output Inf	ormation			
Properties	Output Mode	•	Forecast	•	
Advanced	Period To For	recast	1		
	Column Se	lection			
	Target Variab	le	Beer_Sales	•	0

#### c. Input Data Handling

- i. **Period:** Select a period of forecasting by choosing any one option from the drop-down menu.
- ii. Period Per Year: This field appears only when the selected 'Period' option is 'Custom.'
- iii. Start Period: Enter a value between 1 and the value specified for the selected option for the 'Period' field
- iv. **Start Year:** Enter a year from which you want the data entries to be considered. Enter **a** four-digit value for selecting a year (E.g., 2000)

### d. New Column Information

i. **Period Column Name:** Enter a name for the column containing a period value (This field will be predefined, but users can change the value if needed).



iv) Enable the Manual Arima option by putting a checkmark in the given box.

v) The '**Next**' option appears on the page.

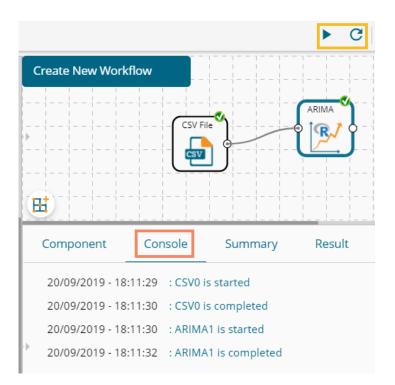
Properties	Input Data Handling				
	Period	Quarter 👻			
Advanced	Start Period	1			
	Start Year	2000			
	New Column Information				
	Period Column Name	QuarterlySales	0		
	Manual Arima 🔽				
				Next	Apply

- vi) Click the 'Advanced' tab and configure if required.
  - **a.** Configure the following '**Behavior**' fields:
    - i. Autoregressive order(p): It is a mandatory field; only integer values are accepted. The default value for this field is 0.
    - ii. Degree of differencing(d): It is a mandatory field; only integer values are accepted.The default value for this field is 0.
    - iii. Moving Average Order(q): It is a mandatory field; only integer values are accepted. The default value for this field is 0.
  - **b.** Configure the following 'Initial Values' information:
    - i. **Confidence**: Enter Confidence level for prediction intervals. It accepts only 0-99 and comma-separated values. According to the number of commas separated values, new low and high range columns get added to the Result dataset. (the default value for this field is 95)
    - ii. Show Range: Select an option using the drop-down menu.
      - 1. **True**: By selecting this option, **Lower Range** and **Upper Range** get displayed in the Result and Visualization of the dataset.
      - 2. False: By selecting this option, Ranges do not get shown in the dataset.
- vii) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> + .	Ţ
General	Behavior	1					
Properties	Auto regres	sive order (p)	0				
Advanced	Degree of d	lifferencing (d)	0				
	Moving Ave	rage order (q)	0				
	Initial Val	ues					
	Confidence		95				
	Show Range	e	False	•			
•							
						Apply	



- viii) Run the workflow after getting the success message.
- ix) The '**Console**' tab opens displaying the progress of the process. The completion of the Console process gets marked by the green marks on the top of the dragged components.



- x) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace
  - **b.** Click the '**Result**' tab.
- xi) Predicted values get appended to the target column in the Result data (The selected output mode is '**Forecasting**')

Component	Console	Summary Result	Visualization Properties $\frac{+}{+}$ $\pm$
Show 10 🔻	entries		Search:
Year	Month	Beer_Sales	QuarterlySales
2003	May	131	Q1 2115
2003	June	125	Q2 2115
2003	July	127	Q3 2115
2003	August	143	Q4 2115
2003	September	143	Q1 2116
2003	October	160	Q2 2116
2003	November	190	Q3 2116
2003	December	18	Q4 2116
		136	Q1 2117
showing 461 to 4	69 of 469 entries	Previous	1 43 44 45 46 47 Next

- xii) Click the 'Visualization' tab.
- xiii) The Result data will be displayed via the TimeLine chart.



Comp	onent	Console	Summary	Result	Visualization	Properties
Dama Calua	300 200 200		~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	v~v	in the test of the	an the dealer and a fingly in the stand and a final sector of
						Time Line
			~~~~~~	~~~~~		www.www.www.www.

xiv) Click the 'Summary' tab to view the model summary.

C	omponent	Console	Summary	Result	Visualization	Properties	<u>+</u> ↑	Ţ
•			model					
	Columns used	in the algorit						
		Beer_Sales	(double)					
	Call:							
	arima(x = tso	, order = c(0,	0, 0))					
	Coefficients:							
	interce							
	136.01							
	s.e. 1.58	67						
	sigma^2 estim	ated as 1178.	log likelihood	= -2318 85	aic = 4641 7			
	Signo 2 Cocim	acca as 11/0.	TOB TINCTIMOOD		uic - 4041.7			
		End of Summary						

**Note**: When the '**Manual Arima**' option is not enabled for the R-ARIMA algorithm, the '**Advanced**' tab does not display Behavior fields. The following images display, respectively, the '**Advanced**,' 'Result,' and '**Visualization**' tabs for the same dataset when the manual ARIMA option has been disabled.

### **Properties Tab**

Component	Console Summ	ary Result	Visualization	Properties	<del>*</del>	<u>+</u>
General	Column Selection					
Properties	Target Variable	Beer_Sales	•			
Advanced	Input Data Handling					1
	Period	Quarter	•			
	Start Period	1				
	Start Year	2000				
	New Column Inform	ation				
•	Period Column Name	QuarterlySales	0			
	Manual Arima 🗆					
					Apply	

#### **Advanced Tab**

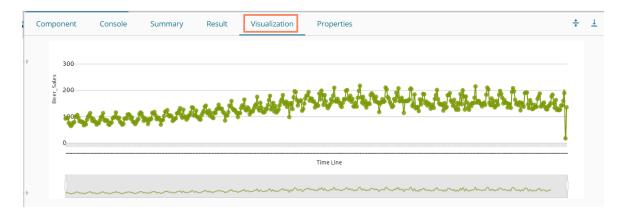


Component	Console	Summary	Result	Visualiz	ation	Properties	<u>+</u>	Ť
General	Initial Values	5						
Properties	perties Confidence		95					
Advanced	Show Range F		False	•				
							Ap	ply

### **Result Tab**

now 10 .	entries		Search:	
Year	Month	Beer_Sales	QuarterlySales	
2003	May	131	Q1 2115	
2003	June	125	Q2 2115	
2003	July	127	Q3 2115	
2003	August	143	Q4 2115	
2003	September	143	Q1 2116	
2003	October	160	Q2 2116	
2003	November	190	Q3 2116	
2003	December	18	Q4 2116	
		136	Q1 2117	

### Visualization Tab

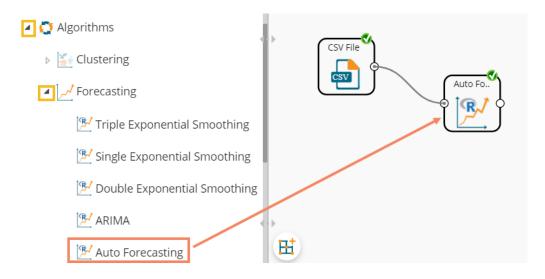


# 13.1.2.5. Auto Forecasting

The user can run the algorithm by adjusting smoothing parameters and other initial states variables to find the best AIC value.



i) Drag the R-Auto Forecasting component to the workspace and connect it to a configured data source.



ii) Configure the '**Properties**' tab.

### a. Output Information

- i. Output Mode: Select a mode in which you want to display output data
  - 1. **Trend**: Selecting this option displays source data along with predicted values for the given data set. A new column **'Predicted** Values' gets added in the Result view when the **'Trend'** output mode has been selected.
  - Forecast: Selecting this option displays forecasted values for the given period. Result values get appended to the target column when 'Forecast' output mode has been selected.
- ii. **Period to Forecast**: Enter a period to forecast. This field appears only when the selected **'Output Mode'** option is **'Forecast.'**
- b. Column Selection
  - i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (The first selected option gets selected by default. Only numerical columns are accepted.)
- c. Input Data Handling
  - i. **Period:** Select a period of forecasting by choosing any one option from the drop-down menu.
  - ii. Period Per Year: This field appears only when the selected 'Period' option is 'Custom.'
  - iii. **Start Period:** Enter a value between 1 and the value specified for the selected option for the **'Period'** field.
  - iv. **Start Year:** Enter **a** four-digit value for selecting a year from which you want the data entries to be considered (E.g., 2000).
- d. New Column Information
  - i. **Period Column Name:** Enter a name for the column containing the period value (This field will be predefined, but users can change the value if needed).



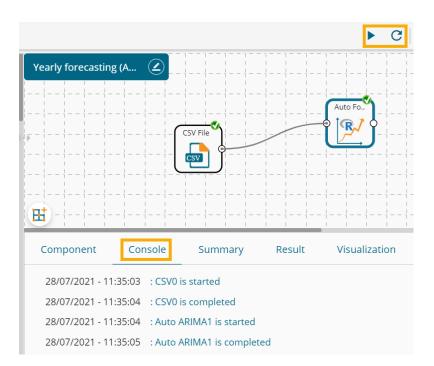
Component	Console Summary	Result Visualization	Properties	<u>+</u> <u>1</u>
General	Output Information			
Properties	Output Mode	Forecast		
Advanced	Period To Forecast	4		
	Column Selection			
	Target Variable	sales	• 0	
	Input Data Handling			
	Period	Quarter	*	
	Start Period	1		
	Start Year	2017		
	New Column Informatio	n		
	Period Column Name	Year1111	0	
				Apply

- iii) Click the 'Advanced' tab and configure if required:
  - a. Configure the following 'Behavior' fields:
    - i. **Seasonal:** Select a smoothing algorithm type from the drop-down menu (Holtwinter's Exponential Smoothing algorithm)
    - ii. **No. of Periodic Observation:** Enter the number of periodic observations required to start the calculation. The default value for this field is 2.
  - b. Configure the following 'Initial Values' fields:
    - i. Level: Enter the initial value for the level (It is an optional field)
    - ii. Trend: Enter the initial value for finding trend parameters (It is an optional field)
    - iii. **Season:** Enter initial values for finding seasonal parameters. It depends on the selected column. It is an optional field.
    - iv. **Optimizer Inputs:** Enter the initial values given for alpha and beta required for the optimizer (It is an optional field).
    - v. **Confidence**: Enter Confidence level for prediction intervals. It accepts only 0-99 and comma-separated value. According to the number of comma-separated values, new low and high range columns get added to the Result dataset (the default value for this field is 95).
    - vi. Show Range: Select an option using the drop-down menu.
      - 1. **True**: By selecting this option, 'Lower Range' and 'Upper Range' get displayed in the Result and Visualization of the dataset.
      - 2. False: By selecting this option, Ranges do not get displayed in the dataset.
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	*	<u>+</u>
General	Behavior						
Properties	Seasonal		Additive	-			
Advanced	No. of Periodic		2		3		
	Observation						
	Initial Values	;					
	Level		Optional				
	Trend		Optional				
	Season		Optional				
	Optimizer Inpu	ts	Optional				
	Confidence		95				
•	Show Range		False	-			
						Apply	

- v) Run the workflow after getting the success message.
- vi) The '**Console**' tab opens displaying the progress of the process. The completion of the Console process gets marked with green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
- viii) Predicted values get appended to the target column in the Result data (The selected output mode is '**Forecasting**').



	Component Console Su	Immary Result Visualization	Properties	<u>+</u> <u>↓</u>
	Show 10 v entries		Se	arch:
Þ	temperature	effective_price	sales	Year1111
	30	0.93	2.1	Q1 2017
	25	0.99	1.68	Q2 2017
	28	0.88	1.26	Q3 2017
	31	0.96	0.84	Q4 2017
	25	0.98	7.28	Q1 2018
	27	0.87	2.415	Q2 2018
	30	0.97	0	Q3 2018
	28	1.02	0.84	Q4 2018
	24	1.05	1.26	Q1 2019
Þ	30	0.94	2.1	Q2 2019
	Showing 1 to 10 of 103 entries	Previ	ous 1 2 3	4 5 11 Next

- ix) Click the 'Visualization' tab.
- x) The processed data will be displayed via the TimeLine chart.



xi) Click the 'Summary' tab to view the model summary.

C	Component	Console	Summary	Result	Visualization	Properties		++	Ŧ
×	Columns used i Holt-Winters	in the algorith sales (dou			ve seasonal comp	onent.			
	Call: HoltWinters(x Smoothing para alpha: 0.7935 beta : 0 gamma: 0.0555	ameters:	= NULL, beta = NUI	LL, gamma = M	IULL, seasona	l = c("additive"), start.periods = as.numer	ic(2), s.start = c())		
•	Coefficients: [,1] a 15.5148 b 0.1483 s1 -1.3328 s2 4.5166 s3 -1.5113 s4 -1.6843								
	6	and of Summary							

# **13.1.2.6.** Weekly Forecasting

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.





- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
- iv) Frequency: Provide a number to indicate the frequency for the data to get processed.
- v) Start Year: Enter a year from which you want the data entries to be considered. Enter a fourdigit value for selecting a year (E.g., 2017)
- vi) Start Period: Enter a value between 1 and the value specified for the selected option for the **'Period'** field.
- vii) Number of forecasting Period: Provide the number of the forecasting period.
- viii) Forecasting Column: Select a column from the input dataset.
- ix) Algorithm Name: Select an option from the given choices that include Dynamic Harmonic Regression, STL, and TBATS.
- x) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties		*	<u>+</u>
General	Dynan	nic Fields						
Custom Group		Frequency		52				
		Start Year		2017				
Start Period				1				
		Number of forecas	ting Period	1				
		Forecasting Colum	n	Close	~			
		Algorithm Name		DYNAMIC HARMOI	NIC REGRESSION ~			
•								
							Apply	y

- xi) Run the workflow after getting the success message.
- xii) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- xiii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - **b.** Click the 'Result' tab.

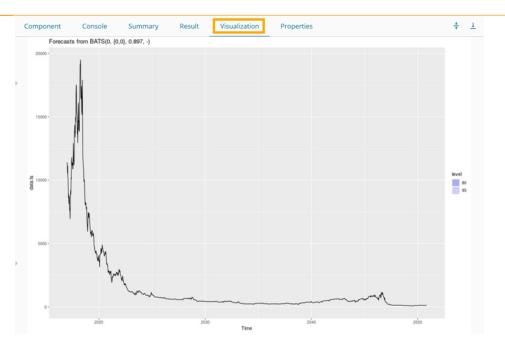
ihow 10 ·	<ul> <li>entries</li> </ul>					Search:	
Date_new	Open_new	High_new	Low_new	Close_new	Volume_new	Market_Cap_new	periodname_nev
20-Feb-18	11231.8	11958.5	11231.8	11403.7	9,926,540,000	189,536,000,000	1 - 2017
19-Feb-18	10552.6	11273.8	10513.2	11225.3	7,652,090,000	178,055,000,000	2 - 2017
18-Feb-18	11123.4	11349.8	10326	10551.8	8,744,010,000	187,663,000,000	3 - 2017
17-Feb-18	10207.5	11139.5	10149.4	11112.7	8,660,880,000	172,191,000,000	4 - 2017
16-Feb-18	10135.7	10324.1	9824.82	10233.9	7,296,160,000	170,960,000,000	5 - 2017
15-Feb-18	9488.32	10234.8	9395.58	10166.4	9,062,540,000	160,025,000,000	6 - 2017
14-Feb-18	8599.92	9518.54	8599.92	9494.63	7,909,820,000	145,023,000,000	7 - 2017
13-Feb-18	8926.72	8958.47	8455.41	8598.31	5,696,720,000	150,516,000,000	8 - 2017
12-Feb-18	8141.43	8985.92	8141.43	8926.57	6,256,440,000	137,258,000,000	9 - 2017
11-Feb-18	8616.13	8616.13	7931.1	8129.97	6,122,190,000	145,245,000,000	10 - 2017

xiv) Click the 'Summary' tab to open the model summary.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
Training set	ME -3.189917 230. ACF1	model RMSE МАЕ 1793 62.93477 -0	MPE M	APE MASE 823 0.1038116			

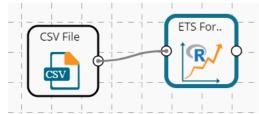
xv) Click the 'Visualization' tab to see the processed data through a chart.





# **13.1.2.7.** ETS Forecasting Model

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

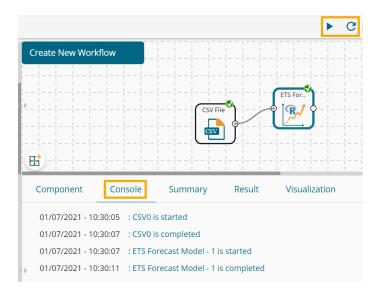


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - **a.** Forecasting Column: Select a column from the input dataset.
  - **b.** Frequency: Provide a number to indicate the frequency for the data to get process.
  - **c.** Start Year: Enter a year from which you want the data entries to be considered. Enter a four-digit value for selecting a year (E.g., 2017)
  - **d.** Start Period: Enter a value between 1 and the value specified for the selected option for the **'Period'** field.
  - e. Number of Forecasting Period: Provide the number of forecasting periods.
- iv) Click the '**Apply**' option.



Component	Console	Summary	Result	Visualization	Properties		+	<u>+</u>
General	Dynamic	Fields						
Custom Group	Fo	recasting Column		Close	~			
	Fre	equency		12				
	Sta	art Year		2017				
	Sta	art Period		1				
	Nu	Imber of Forecast	ting Period	365				
•								
							Apply	

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

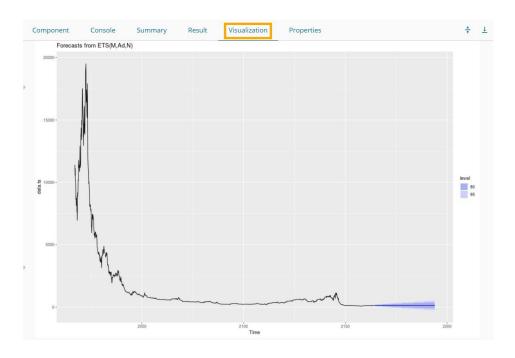
Component	Console	Summary	Result	Visualization	Properties		+
how 10 🗸	<ul> <li>entries</li> </ul>					Search:	
Date_new	Open_new	High_new	Low_new	Close_new	Volume_new	Market_Cap_new	PeriodName_new
20-Feb-18	11231.8	11958.5	11231.8	11403.7	9,926,540,000	189,536,000,000	Jan 2017
19-Feb-18	10552.6	11273.8	10513.2	11225.3	7,652,090,000	178,055,000,000	Feb 2017
18-Feb-18	11123.4	11349.8	10326	10551.8	8,744,010,000	187,663,000,000	Mar 2017
17-Feb-18	10207.5	11139.5	10149.4	11112.7	8,660,880,000	172,191,000,000	Apr 2017
16-Feb-18	10135.7	10324.1	9824.82	10233.9	7,296,160,000	170,960,000,000	May 2017
15-Feb-18	9488.32	10234.8	9395.58	10166.4	9,062,540,000	160,025,000,000	Jun 2017
14-Feb-18	8599.92	9518.54	8599.92	9494.63	7,909,820,000	145,023,000,000	Jul 2017
13-Feb-18	8926.72	8958.47	8455.41	8598.31	5,696,720,000	150,516,000,000	Aug 2017
12-Feb-18	8141.43	8985.92	8141.43	8926.57	6,256,440,000	137,258,000,000	Sep 2017
11-Feb-18	8616.13	8616.13	7931.1	8129.97	6,122,190,000	145,245,000,000	Oct 2017



viii) Click the 'Summary' tab to open the model summary.

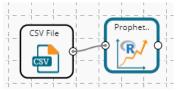


ix) Click the 'Visualization' tab to see the processed data through a chart.



# 13.1.2.8. Prophet Forecasting

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



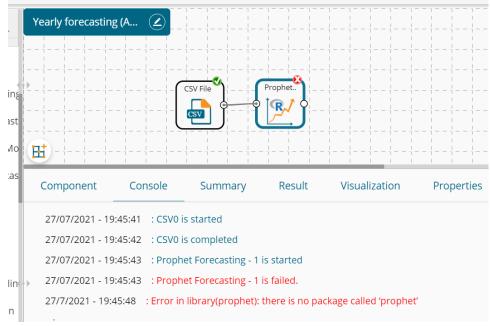
- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - **a.** Seasonality: Select an option from the given choices that include additive and multiplicative options.
  - **b.** Period: Enter a value between 1 and the value specified for the selected option for the **'Period'** field.
  - c. Frequency: Choose an option to set the frequency for running the data out of the given choices that include day, month, or yearly.



#### iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>↓</u> 1	1
General	Dynamic Fie	lds					
Custom Group	seasona	lity	additive	~			
•	period		1				
	frequen	cy	day	~			
						Apply	

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
- viii) Click the 'Summary' tab to open the model summary.

# **13.1.2.9.** Forecasting Algorithms with 'Trend' Output Mode:

A new column '**Predicted Values**' gets added to the Result view when '**Trend**' is selected as an output mode.

#### 1. Triple Exponential Smoothing

- i) Drag the Forecasting algorithm to the workspace and connect it with the configured data source.
- ii) Configure the '**Properties**' tab for the Forecasting Algorithm component, keeping '**Trend**' as the '**Output Mode**.'

#### a. Output Information

i. **Output Mode**: Select a mode in which you want to display output data



1. **Trend**: Selecting this option displays source data along with predicted values for the given data set. A new column displaying the predicted values gets added in the Result view when the '**Trend**' output mode has been selected.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Output Inf	ormation					
Properties	Output Mode	2	Trend				

#### b. Column Selection

- i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (The first selected option gets selected by default. Only numerical columns are accepted.)
- c. Input Data Handling
  - i. **Period:** Select a period of forecasting by choosing any one option from the drop-down menu.
  - ii. **Period Per Year**: This field appears only when the selected '**Period**' option is '**Custom.'**
  - iii. **Start Period:** Enter a value between 1 and the value specified for the selected option for the **'Period'** field
  - iv. **Start Year:** Enter a year from which you want the data entries to be considered. Enter **a** four-digit value for selecting a year (E.g., 2000)

#### d. New Column Information

- i. **Predicted Column Name:** Enter a name for the column containing predicted values (This field is predefined. It gets displayed if the selected '**Output Mode**' is '**Trend**').
- ii. **Period Column Name:** Enter a name for the column containing a period value. (This field is predefined, but users can change the value if needed).

Component	Console Sum	mary	Result	Visu	alization	Properties	+	Ŧ
General	Column Selection							
Properties	Target Variable	Beer_S	Sales	•	0			
Advanced	Input Data Handling	g						
	Period	Quarte	er	•				
	Start Period	1						
	Start Year	2000						
	New Column Inform	nation						
	Predicted Column	Quarte	erlySales		0			
• •	Name							
	Period Column	Quarte	erO		0			
	Name					_		_
							Apply	<b>′</b>

- iii) Click the 'Advanced' tab and configure it.a. Configure the following 'Behavior' fields:
- a. Compute the following behavior



- i. Alpha: Enter a valid double value in the given field for smoothing observations. (Alpha Range: 0<alpha<=1.)
- ii. **Beta:** Enter a valid double value in the given field for finding trend parameters. (Beta Range: 0-1.)
- iii. **Gamma:** Enter a valid double value in the given field for finding seasonal trend parameters. (Gamma Range: 0-1.)
- iv. **Seasonal:** Select a smoothing algorithm type from the drop-down list (Holtwinter's Exponential Smoothing algorithm)
- v. **No. of Periodic Observation:** Enter the number of periodic observations required to start the calculation. The default value for this field is 2.
- b. Configure the following 'Initial Values' information:
  - i. Level: Enter the initial value for the level. It is an optional field.
  - ii. Trend: Enter the initial value for finding trend parameters. It is an optional field.
  - iii. **Season:** Enter initial values for finding seasonal parameters. It depends on the selected column. It is an optional field.
  - iv. **Optimizer Inputs:** Enter the initial values given for alpha, beta, gamma required for the optimizer. It is an optional field.
- iv) Click the **'Apply'** option.

Component	Console	Summary	Result	Visualization	Properties		*	<u>+</u>
General	Behavior							
Properties	Alpha		.3		0			
Advanced	Beta		.1		0			
	Gamma		.1		0			
	Seasonal		Additive	•				
	No. of Period	ic	2		0			
	Observation							
	Initial Value	es						
	Level		Optional					
	Trend		Optional					
•	Season		Optional					
	Optimizer Inp	outs	Optional					
						_		
							Apply	y

- v) Run the workflow and open the 'Result' tab after the Console process gets completed
  - **a.** Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

In this case, the QuarterlySales column displays the predicted values in the Result tab.



	Component	Console	Summary	Result	: Vi	sualization	Prop	erties	+	<u>+</u>
	Show 10	• entries				S	earch:			
•	Year	Month	Beer_Sales		Quarte	r0	Quarter	ySales		
	1965	January	93.2							
	1965	February	96							
	1965	March	95.2							
	1965	April	77.1							
	1965	May	70.9		Q1 2001		85.22			
	1965	June	64.8		Q2 2001		71.75			
	1965	July	70.1		Q3 2001		76.84			
	1965	August	77.3		Q4 2001		56.81			
•	1965	September	79.5		Q1 2002		56.81			
	1965	October	100.6		Q2 2002		55.85			
	Showing 1 to 1	0 of 468 entries		Previo	us 1	2 3	4 5	47	Ne	xt

## vi) Click the 'Visualization' tab.

vii) The Result data gets displayed via the TimeLine Chart.



viii) Click the 'Summary' tab to view the model summary.



	Component	Console	Summary	Result	Visualization	Properties	<del>*</del>	1
•		in the algorith	model m (double)					
	Holt-Winters	exponential sm	wothing with tre	nd and addit	cive seasonal compo	nent.		
	HoltWinters(x				numeric(0.1), ( = c(), optim.star	gamma = as.numeric(0.1), t = c())	seasonal = c	
	Smoothing par- alpha: 0.3 beta : 0.1 gamma: 0.1	ameters:						
	Coefficients: [,1] a 111.0213 b -3.1634							
•	s1 -4.2978 s2 -1.4135 s3 12.6552 s4 -0.8968							
		End of Summary						

Note:

- a. **'Properties'** and **'General'** sections remain the same for all the Forecasting sub-algorithms.
- b. The 'Advanced' tab displays different fields as per the Forecasting sub-types. Hence,
   'Advanced' fields for all the sub-types are explained over here. Predicted values get appended to the target column in the Result view for all the 'Forecasting' algorithms.
  - 1. Single Exponential Smoothing
  - i) Configure the following Properties fields with 'Trend' as the selected 'Output Mode' option.
    - a. Output Information
      - i. Output Mode: Select a mode in which you want to display output data
        - 1. **Trend**: Selecting this option displays source data along with predicted values for the given data set. A new column displaying the predicted values gets added in the Result view when the '**Trend**' output mode has been selected.
    - b. Column Selection
      - i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (The first selected option gets selected by default. Only numerical columns are accepted.)

Component	Console	Summary	Result	Visualization	Properties	*	<u>+</u>
General	Output In	formation					
Properties	Output Mod	le	Trend	-			
Advanced	Column S	election					
	Target Varia	ble	Beer_Sales	-	0		

## c. Input Data Handling

- i. **Period:** Select period of forecasting by choosing any one option from the drop-down menu.
- ii. Period Per Year: This field appears only when the selected 'Period' option is

#### www.bdb.ai



'Custom.'

- iii. **Start Period:** Enter a value between 1 and the value specified for the selected option for **'Period'** field
- iv. **Start Year:** Enter a four-digit value for selecting a year from which you want the data entries to be considered (E.g., 2000)
- d. New Column Information
  - i. **Predicted Column Name:** Enter a name for the column containing predicted values (This field is predefined and displayed if the selected Output Mode is '**Trend**').
  - iii. **Period Column Name:** Enter a name for the column containing a period value. (This field is predefined, but users can change the value if needed).

General	Input Data Handling	· · · · · · · · · · · · · · · · · · ·		
Properties	Period	Custom 👻		
Advanced	Periods per year	4		
	Start Period	1		
	Start Year	2000		
	New Column Informat	ion		
	Predicted Column	PredictedValues1	0	
	Name			
•	Period Column Name	PeriodColumn1	θ	
				Apply

- ii) Configure the required 'Advanced' fields:
  - a. Configure the following 'Behavior' fields:
    - i. **Alpha:** Enter a valid double value in the given field for smoothing observations. (Alpha Range: 0<alpha<=1.)
    - ii. **No. of Periodic Observation:** Enter the number of periodic observations required to start the calculation. The default value for this field is 2.
  - b. Configure the following 'Initial Values' information:
    - i. Level: Enter the initial value for the level. It is an optional field.
- iii) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Behavior						
Properties	Alpha		3	0			
Advanced	No. of Periodic	2	2	0			
	Observation						
	Initial Values						
	Level	g	95				
• •							
						Appl	У

iv) Run the workflow and open the 'Result' tab after the Console process gets completed
 a. Click the dragged algorithm component from the workspace and then click

b. Click the 'Result' tab.

	Componen	t Console	Summary	Result	Properties 🕂
	Show 10	▼ entries			Search:
Þ	Year	Month	Beer_Sales	PeriodColumn1	PredictedValues1
	1965	January	93.2		
	1965	February	96	Q2 2000	95
	1965	March	95.2	Q3 2000	95.3
	1965	April	77.1	Q4 2000	95.27
	1965	May	70.9	Q1 2001	89.82
	1965	June	64.8	Q2 2001	84.14
	1965	July	70.1	Q3 2001	78.34
	1965	August	77.3	Q4 2001	75.87
	1965	September	79.5	Q1 2002	76.3
	1965	October	100.6	Q2 2002	77.26
		October 10 of 468 entries	100.6		77.26 3 4 5 47 N

- v) Click the 'Visualization' tab.
- vi) The Result data gets displayed via the Time Series Chart.





vii) Click the 'Summary' tab to view the model summary.

1	Component	Console	Summary	Result	Visualization	Properties		<del>*</del>	<u>+</u>
	Columns used i								
	Call:		-		ithout seasonal compo	onent. start.periods = as	numeric(2) 1 st	art	
	smoothing para alpha: 0.3 beta : FALSE	-		,, , , , , , , , , , , , , , , , , , , ,					
	gamma: FALSE Coefficients: [,1]								
	a 116.3	nd of Summary	/					_	

## 2. Double Exponential Smoothing

- i) Select the 'Trend' option from the 'Output Mode' drop-down menu.
- ii) Configure the following fields in the 'Properties' tab:
  - a. Output Information
    - i. Output Mode: Select a mode in which you want to display output data
      - 1. **Trend**: Selecting this option displays source data along with predicted values for the given data set. A new column displaying the predicted values gets added in the Result view when the 'Trend' output mode has been selected.
  - b. Column Selection
    - i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (The first selected option gets selected by default. Only numerical columns are accepted.)
  - c. Input Data Handling
    - i. **Period:** Select a period of forecasting by choosing any one option from the drop-down menu.
    - ii. **Start Period:** Enter a value between 1 and the value specified for the selected option for the **'Period'** field
    - iii. **Start Year:** Enter a year from which you want the data entries to be considered. Enter **a** four-digit value for selecting a year (E.g., 2000)
  - d. New Column Information
    - i. **Predicted Column Name:** Enter a name for the column containing predicted values (This field is predefined and displayed if the selected Output Mode is '**Trend**').
    - iv. **Period Column Name:** Enter a name for the column containing a period value. (This field is predefined, but users can change the value if needed).



Component	Console	Summary	Result	Visualization	Properties		+	<u>+</u>
General	Output Info	rmation						
Properties	Output Mode		Trend	•				
Advanced	Column Sele	ection						
	Target Variable	•	Beer_Sales	-	0			
	Input Data H	Handling						
	Period		Month	-				
	Start Period		1					
	Start Year		2000					
	New Colum	n Information	I					
	Predicted Colu	mn Name	PredictedValue	s	0			
• >	Period Column	Name	Months		0			
						_		
							Apply	

- iii) Click the 'Advanced' tab and configure
  - a. Configure the following 'Behavior' fields:
    - i. **Alpha:** Enter a valid double value in the given field for smoothing observations. (Alpha Range: 0<alpha<=1.)
    - ii. **Beta:** Enter a valid double value in the given field for finding trend parameters. (Beta Range: 0-1.)
    - iii. **No. of Periodic Observation:** Enter the number of periodic observations required to start the calculation. The default value for this field is 2.
  - b. Configure the following 'Initial Values' information:
    - i. Level: Enter the initial value for the level. It is an optional field.
    - ii. Trend: Enter the initial value for finding trend parameters. It is an optional field.
    - iii. **Optimizer Inputs:** Enter the initial values given for alpha, beta, gamma required for the optimizer. It is an optional field.
- iv) Click the '**Apply**' option.

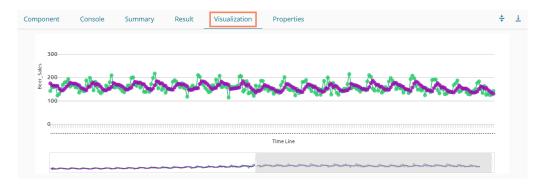
Component	Console	Summary	Result	Visualization	Properties	+	Ŧ
General	Behavior						
Properties	Alpha	.3		0			
Advanced	Beta	.1		0			
	No. of Periodic	2		0			
	Observation						
	Initial Values						
	Level	Optic	onal				
	Trend	Optic	onal				
	Optimizer Inputs	Optic	onal				
•							
						Apply	/



- v) Run the workflow and open the 'Result' tab after the Console process gets completeda. Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.

10 now	<ul> <li>entries</li> </ul>			Search:
Year	Month	Beer_Sales	Months	PredictedValues
1965	January	93.2		
1965	February	96		
1965	March	95.2	Mar 2000	98.8
1965	April	77.1	Apr 2000	100.41
1965	May	70.9	May 2000	95.41
1965	June	64.8	Jun 2000	89.32
1965	July	70.1	Jul 2000	82.48
1965	August	77.3	Aug 2000	78.92
1965	September	79.5	Sep 2000	78.53
1965	October	100.6	Oct 2000	78.95

- vi) Click the 'Visualization' tab.
- vii) The Result data gets displayed via the TimeLine Chart.



## 3. R-ARIMA

- i) Select the 'Trend' option from the 'Output Mode' drop-down menu.
- ii) Configure the following fields in the 'Properties' tab:
  - a. Output Information
    - i. Output Mode: Select a mode in which you want to display output data
      - Trend: Selecting this option displays source data along with predicted values for the given data set. A new column 'Predicted Values' gets added in the Result view when the 'Trend' output mode has been selected.
      - Forecast: Selecting this option displays forecasted values for the given period. The Result values are appended to the target column when 'Forecast' output mode has been selected.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Output Infor	mation					
Properties	Output Mode		Trend				

### b. Column Selection

- i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (The first selected option gets selected by default. Only numerical columns are accepted.)
- c. Input Data Handling
  - i. **Period:** Select a period of forecasting by choosing any one option from the drop-down menu.
  - ii. Period Per Year: This field appears only when the selected 'Period' option is 'Custom.'
  - iii. **Start Period:** Enter a value between 1 and the value specified for the selected option for the **'Period'** field
  - iv. **Start Year:** Enter a year from which you want the data entries to be considered. Enter **a** four-digit value for selecting a year (E.g., 2000)

#### d. New Column Information

- i. **Predicted Column Name:** Enter a name for the column containing predicted values (This field is predefined and displayed if the selected Output Mode is '**Trend**')
- ii. **Period Column Name:** Enter a name for the column containing the period value (This field will be predefined, but users can change the value if needed).
- iii. Manual Arima: Enable this option to get Behaviour fields in the Advanced tab. If the Manual Arima option is enabled, then the 'Next' option appears on the Properties configuration page, and the user can click it to configure the Advanced fields.

Component	Console Summar	y Result	Visualization	Properties	÷	+ -	Ť
General	Column Selection						
Properties	Target Variable	Beer_Sales	•	0			
Advanced	Input Data Handling						
	Period	Quarter	•	·			
	Start Period	1					
	Start Year	2000					
1	New Column Inform	ation					
	Predicted Column Name	PredictedVal	ues	Θ			
•	Period Column Name	QuarterlySal	es	Θ			
	Manual Arima 🗹						
				Ne	ext Aj	pply	

#### Properties tab with Manual Arima option Disabled



Component	Console	Summary	Result	Visualization	Properties	<del>+</del> +	<u>+</u>
General	Output Inf	ormation					
Properties	Output Mode	:	Trend	-			
	Column Se	lection					
	Target Variab	le	Beer_Sales	-	0		
	Input Data	Handling					
	Period		Quarter	-			
	Start Period		1				
	Start Year		2000				
	New Colun	nn Informatio	n				
	Predicted Col	umn Name	PredictedValue	s1	0		
• •	Period Colum	in Name	QuarterlySales		0		
	Manual Arim	a 🗆					
						pply	
						фріу	

- iii) Click the 'Advanced' tab and configure it.
  - a. Configure the following 'Behavior' fields:
    - i. **Alpha:** Enter a valid double value in the given field for smoothing observations (Alpha Range: 0<alpha<=1)
    - ii. **Beta:** Enter a valid double value in the given field for finding trend parameters (Beta Range: 0-1)
    - iii. **Gamma:** Enter a valid double value in the given field for finding a seasonal trend parameter (Gamma Range: 0-1)
    - iv. **Seasonal:** Select a smoothing algorithm type from the drop-down list (Holtwinter's Exponential Smoothing algorithm)
    - v. **No. of Periodic Observation:** Enter the number of periodic observations required to start the calculation (The default value for this field is 2)
  - b. Configure the following 'Initial Values' information:
    - i. Level: Enter the initial value for the level. It is an optional field.
    - ii. Trend: Enter the initial value for finding trend parameters. It is an optional field.
    - iii. **Season:** Enter initial values for finding seasonal parameters. It depends on the selected column. It is an optional field.
    - iv. **Optimizer Inputs:** Enter the initial values given for alpha, beta, gamma required for the optimizer. It is an optional field.
- iv) Click the 'Apply' option. Advanced Tab when Manual Arima is enabled



Component	Console	Summary	Result	Visualization	Properties	4	+	<u>+</u>
General	Behavior	]						
Properties	Auto regres	sive order (p)	5					
Advanced	Degree of d	ifferencing (d)	4					
	Moving Ave	rage order (q)	2					
•								
						A	pply	

Note: The Advanced tab does not appear if the Manual Arima option is disabled.

- v) Run the workflow and open the ' ${\bf Result'}$  tab after the Console process gets completed
  - **a.** Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.
  - c. A new column displaying the predicted values gets added to the Result view.

#### The following is the 'Result' tab display when 'Manual Arima' is Enabled

	Componen	t Console	Summary	Result	Visualization	Ρ	roperties	<u>+</u> <u>-</u>
:	Show 10	▼ entries					Search:	
÷	Year	Month	Beer_Sales	Q	uarterlySales		PredictedValues	]
	1965	January	93.2	Q1	2000		93.19	
	1965	February	96	Q2	2000		96.06	
	1965	March	95.2	Q3	2000		95.11	
	1965	April	77.1	Q4	2000		77.50	
	1965	May	70.9	Q1	2001		61.80	
	1965	June	64.8	Q2	2001		67.81	
	1965	July	70.1	Q3	2001		69.05	
	1965	August	77.3	Q4	2001		85.85	
	1965	September	79.5	Q1	2002		90.91	
	1965	October	100.6	Q2	2002		101.79	
	Showing 1 to	10 of 468 entries			Previous 1	2 3	3454	7 Next

#### vi) Click the 'Visualization' tab.

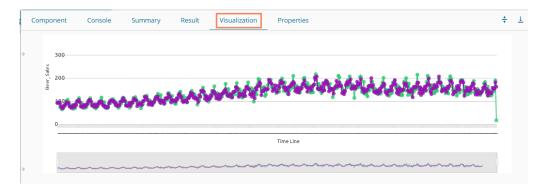
vii) The Result data gets displayed via the TimeLine Chart.



Component	Console	Summary	Result .	Visualization	Properties	+	1
300 Sales 200 Sales 200	~~~~	<b>~~~</b> ^	***	MANN	water		
					Time Line		
			~~~~	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~			

The following are the 'Result' and 'Visualization' tabs for the selected dataset when 'Manual Arima' is Disabled.

now 10	<ul> <li>entries</li> </ul>			Search:
Year	Month	Beer_Sales	QuarterlySales	PredictedValues
1965	January	93.2	Q1 2000	93.11
1965	February	96	Q2 2000	94.15
1965	March	95.2	Q3 2000	95.59
1965	April	77.1	Q4 2000	89.02
1965	May	70.9	Q1 2001	76.01
1965	June	64.8	Q2 2001	71.38
1965	July	70.1	Q3 2001	70.38
1965	August	77.3	Q4 2001	81.12
1965	September	79.5	Q1 2002	84.25
1965	October	100.6	Q2 2002	88.42



#### 4. R-Auto Forecasting

- i) Select the 'Trend' option from the 'Output Mode' drop-down menu.
- ii) Configure the following fields in the 'Properties' tab:

## a. Output Information

- i. Output Mode: Select a mode in which you want to display output data
  - 1. **Trend**: Selecting this option displays source data along with predicted values for the given data set. A new column '**Predicted Values**' gets added in the Result view when the '**Trend**' output mode has been selected.



- Forecast: Selecting this option displays forecasted values for the given period. Results get appended to the target column when the 'Forecast' output mode has been selected.
- b. Column Selection
  - i. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (The first selected option gets selected by default. Only numerical columns are accepted.)
- c. Input Data Handling
  - i. **Period:** Select the period of forecasting by choosing any one option from the dropdown menu.
  - ii. **Period Per Year**: This field appears only when the selected '**Period**' option is '**Custom.'**
  - iii. **Start Period:** Enter a value between 1 and the value specified for the selected option for the **'Period'** field
  - iv. **Start Year:** Enter a year from which you want the data entries to be considered. Enter **a** four-digit value for selecting a year (E.g., 2000)

### d. New Column Information

- i. **Predicted Column Name:** Enter a name for the column containing predicted values (This field is predefined and displayed only if the selected Output Mode is '**Trend**').
- ii. **Period Column Name:** Enter a name for the column containing the period value (This field will be predefined, but users can change the value if needed).

Component	Console Summary Res	ult Visualization Properti	es	÷ <u>1</u>
General	Output Information			
Properties	Output Mode	Trend	*	
Advanced	Column Selection			
	Target Variable	sales	. 0	
	Input Data Handling			
	Period	Quarter	•	
	Start Period	1		
	Start Year	2017		
	New Column Information			
	Predicted Column Name	PredictedValues1	0	
	Period Column Name	Year1111	0	
				Apply

- iii) Click the 'Advanced' tab and configure
  - a. Configure the following 'Behavior' fields:
    - i. **Alpha:** Enter a valid double value in the given field for smoothing observations. (Alpha Range: 0<alpha<=1.)
    - ii. **Beta:** Enter a valid double value in the given field for finding trend parameters. (Beta Range: 0-1.)
    - iii. **Gamma:** Enter a valid double value in the given field for finding seasonal trend parameters. (Gamma Range: 0-1.)
    - iv. **Seasonal:** Select a smoothing algorithm type from the drop-down list (Holtwinter's Exponential Smoothing algorithm)
    - v. **No. of Periodic Observation:** Enter the number of periodic observations required to start the calculation. The default value for this field is 2.
  - b. Configure the following 'Initial Values' information:
    - i. Level: Enter the initial value for the level. It is an optional field.



- ii. **Trend:** Enter the initial value for finding trend parameters. It is an optional field.
- iii. **Season:** Enter initial values for finding seasonal parameters. It depends on the selected column. It is an optional field.
- iv. **Optimizer Inputs:** Enter the initial values given for alpha, beta, gamma required for the optimizer. It is an optional field.
- iv) Click the 'Apply' option.

Component	Console Summar	y Result \	/isualization	Properties	+	<u>+</u>
General	Behavior					
Properties	Seasonal	Additive	•			
Advanced	No. of Periodic	2	3			
	Observation					
	Initial Values					
	Level	Optional				
	Trend	Optional				
	Season	Optional				
	Optimizer Inputs	Optional				
				_		
					Apply	

- viii) Run the workflow and open the 'Result' tab after the Console process gets completed
  - **a.** Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.
  - c. A new column with the predicted values gets added to the Result data.

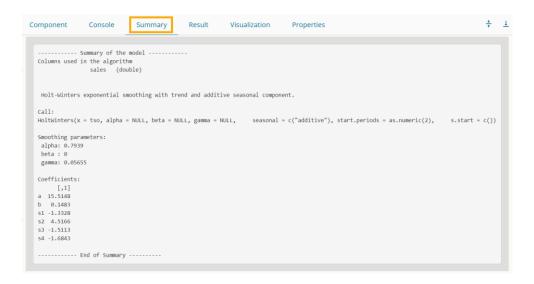
ow 10 🗸 entries				Search:
temperature	effective_price	sales	Year1111	PredictedValues1
0	0.93	2.1		
5	0.99	1.68		
8	0.88	1.26		
1	0.96	0.84		
5	0.98	7.28	Q1 2018	6.5172
7	0.87	2.415	Q2 2018	2.5636
0	0.97	0	Q3 2018	1.9552
В	1.02	0.84	Q4 2018	-0.6082
4	1.05	1.26	Q1 2019	7.2044
0	0.94	2.1	Q2 2019	-2.0849

- v) Click the 'Visualization' tab.
- vi) The Result data gets displayed via the TimeLine chart.



Component	Console	Summary	Result	Visualization	Properties			*	<u>+</u>
100 80 50 60 40 20 0	\ree <sup>2</sup> +er <sup>2</sup> *i	diret todi	-44 <sup>0</sup> 544 <sup>0</sup> 4	An	Time Line		toman Mare		
	~~~~					/	the		

Note: Click the '**Summary**' tab to view the model summary for the Forecasting models with '**Trend**' as the output mode.



# 13.1.3. Association

This algorithm generates association rules discovering the recurrent patterns in large transactional data sets. It tries to understand the future trends of customers based on their previous purchases and assists the vendors to associate items or services together.

# 13.1.3.1. Market Basket Analysis

i) Drag the Market Basket Analysis component to the workspace and connect it with a configured data source.



- ii) Configure the following fields in the 'Properties' tab:
  - a. Output Information
    - i. Output Mode: Select a mode of display for output data



- 1. Selecting the 'Rules' option displays rules for the selected dataset.
- 2. Selecting the **'Transaction'** option displays the transaction IDs for the selected dataset.
- b. Input Data Information
  - i. **Input Data Format**: Select an input data format out of the following choices via the drop-down menu:
    - 1. Tabular
    - 2. Transactions
      - As per the selected '**Input Data Format**,' two types of the result view appear.
  - ii. **Item Columns:** Select the item columns on which you want to Apply association rules/analysis. Choose at least one option from the drop-down menu. This field displays numerical and strings columns. It cannot display Date columns.
  - iii. Transaction Id Column: Select the column containing Transaction Ids to which you can apply the algorithm. (This field gets added when the selected 'Input Data Information' is 'Transactions')

**Note: 'Transaction Id Column'** field appears when the **'Transactions'** option has been selected from the **'Input Data Format'** drop-down menu.

- c. Behavior
  - i. **Support:** Enter a value for the minimum support of an item. The default value for this field is 0.1
  - ii. **Confidence**: Select a value for the minimum confidence of the association (The default value for this field is 0.8)

Component	Console	Summary	Result	Visualization	Properties	<del>*</del>	Ŧ
General	Output Info	rmation					
Properties	Output Mode		Rules	•			
Advanced	Input Data l	nformation					
	Input Data For	mat	Tabular	•			
	Item Column(s	Item Column(s) 6 d		•	0		
	Behavior						
	Support		0.1		0		
•	Confidence		0.8		0		
						Ар	ply

Properties fields with 'Transactions' as 'Input Data Information'



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Output Infor	mation					
Properties	Output Mode		Transactions	-			
Advanced	Input Data Ir	nformation					
	Input Data Forr	nat	Transactions	•			
	Item Column		Products	•	0		
	Transaction Id	column	TID	-	0		
	Behavior						
	Support		0.1		0		
•	Confidence		0.8		0		
						A	pply

- iii) Click the 'Advanced' tab and configure if required:
  - a. Output Appearance
    - i. Lhs Item(s): Enter item tags separated by a comma which should display on the lefthand side of rules or item sets
    - ii. **Rhs Item(s):** Enter item tags separated by a comma which should display on the righthand side of rules or item sets
    - iii. Both Item(s): Enter item tags separated by a comma which should display on both sides of rules or item sets
    - iv. None Item(s): Enter item tags separated by a comma which need not display in the rules or item sets
    - v. **Default Appearance**: Select the default appearance of the items out of the abovegiven choices using a drop-down menu
    - vi. Min Length: Set a minimum length value. The default value for this field is 1.
    - vii. Max Length: Set a maximum length value. The default value for this field is 10.

General	Output Appearance		
Properties	Lhs Item(s)	Optional	0
Advanced	Rhs Item(s)	Optional	0
	Both Item(s)	Optional	0
	None Item(s)	Optional	0
	Default Appearance	Both 👻	
	Min. Length	1	
•	Max. Length	10	

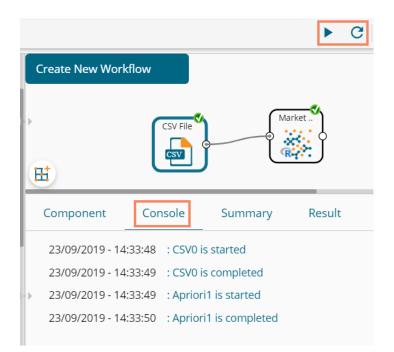
- b. Performance
  - i. **Sort Type**: Select a sort type using the drop-down menu for sorting items based on their frequency.
  - ii. **Filter Criteria**: Enter an indicating numerical value for filtering unused items from Transactions. The default value for this field is 0.1.
  - iii. **Use Tree Structure**: Selecting **the' True'** option from the drop-down menu organizes the transaction as a prefix tree.



- iv. **Use Heapsort**: Selecting the **'True'** option from the drop-down menu uses heapsort against quicksort for sorting transactions.
- v. **Optimize Memory**: Selecting **the' True'** option from the drop-down menu minimizes memory usage instead of maximizing speed.
- vi. Load Transaction into Memory: Selecting 'True' from the drop-down menu loads transactions into memory.

Component	Console Sun	nmary	Result	Visualization	Properties	<u>+</u> ↑	<u>+</u>
General	Performance						
Properties	Sort Type	Ascen	ding Transactior	n Size <del>.</del>			
Advanced	Filter Criteria	0.1					
	Use Tree Structure	True		•			
	Use Heapsort	True		•			1
	Optimize Memory	False		•			
	Load Transaction into	True		•			
•	memory						
						Apply	y

- iv) Click the 'Apply' option.
- v) Run the workflow after getting a success message.
- vi) The user gets directed to the 'Console' tab displaying the progress of the process.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
- viii) Two types of Result view gets displayed:



a. 'Rules' gets displayed as a first column in the Result data (When the selected 'Output Mode' option is 'Rules').

Component	Console	Summary	Result	Visualization	Properti	es	<del>+</del>
Show 10 •	entries					Search:	
Rules					Support	Confidence	Lift
{Affluence=Low}	=> {MetroPolitan:	=Yes}			0.12	1	1.66666666666666
{Affluence=Low}	=> {SKYBox=Sky+	HD 2TB}			0.12	1	1.51515151515152
{Affluence=Very I	Low} => {MetroPo	litan=No}		0.1	0.83333333333333333	2.0833333333333333	
{Affluence=Mid L	ow} => {MetroPol	itan=Yes}			0.12	0.857142857142857	1.42857142857143
{Affluence=Mid L	ow} => {SKYBox=	Sky+HD 2TB}			0.12	0.857142857142857	1.2987012987013
{Demographiclife	estyle=Liberal Opi	nion} => {Househo	ldComposition=	Men only HH}	0.12	0.857142857142857	2.52100840336134
{Demographiclife	estyle=Liberal Opi	nion} => {MetroPol	itan=Yes}		0.12	0.857142857142857	1.42857142857143
{Demographiclife	estyle=Liberal Opi	nion} => {SKYBox=	5ky+HD 2TB}		0.12	0.857142857142857	1.2987012987013
{Affluence=Mid}	=> {MetroPolitan=	=No}		0.12	0.857142857142857	2.14285714285714	
{Demographiclife	estyle=Terraced M	lelting Pot} => {Hou	iseholdCompos	ition=Men only HH}	0.14	0.875	2.57352941176471

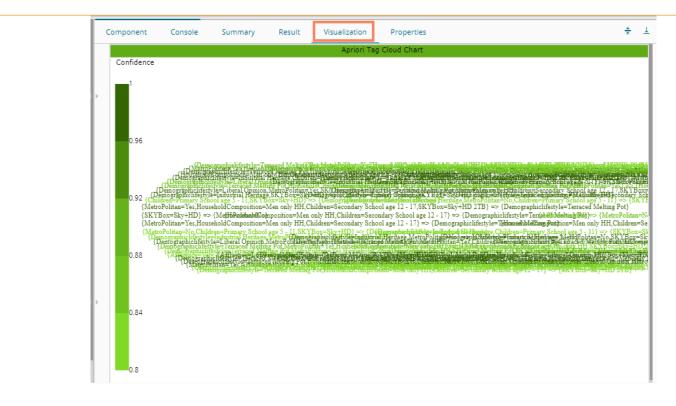
b. 'Transaction\_Id' will be displayed as the second column in the Result data (When the selected 'Output Mode' option is 'Transaction').

The matching rules for the selected items get displayed through the **'Matching\_Rules'** column.

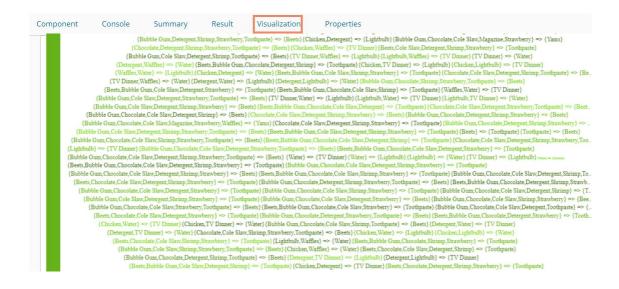
Console	Summary	Result	Visualization	Properties	<del>*</del>	_
entries				Search:		
Trans	action_ld		Matching_F	tules		
396			103			
434						
486			1455			
576			1392			
664			1176			
700			382			
	entries <b>Trans</b> 396 434 486 486 576 664	entries Transaction_Id 396 434 486 576 664	entries       396       434       486       576       664	Image: strain of the strain	entries     Search:       Transaction_ld     Matching_Rules       396     103       434	entries     Search:       Transaction_Id     Matching_Rules       396     103       434

- ix) Click the 'Visualization' tab.
- x) The Result data will be displayed via the Apriori Tag Cloud chart.
  - a. The Visualization tab for the 'Rules' output mode





#### b. Visualization tab for the 'Transactions' output mode



# 13.1.4. Regression Analysis

This algorithm is used to determine how an individual variable influences another variable using an exponential function. It finds a trend in the dataset Applying univariate regression analysis. The Regression Analysis tree node has the following algorithms:

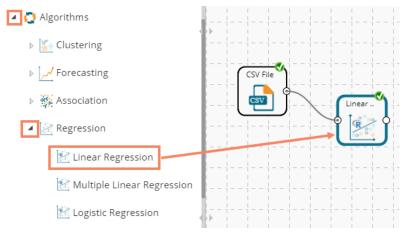


# Regression

- 🛃 Linear Regression
- Multiple Linear Regression
- 🛃 Logistic Regression
- Ordinal Regression
- FFQ Regression
- 🛃 Stepwise Regression

## 13.1.4.1. Linear Regression

i) Drag the Linear Regression component to the workspace and connect it with a configured data source.



- ii) Configure the following fields in the 'Properties' tab:
  - a. Column Selection
    - i. **Dependent Column**: Select the target column on which the regression analysis gets applied
    - ii. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column
  - b. New Column Information
    - i. **Predicted Column Name**: Enter a name for the new column containing the predicted values
  - c. Model Tuning
    - i. Enable Validation: Use a checkmark to enable the validation tab.
    - ii. XG Boosting: Use a checkmark in the box to enable XG Boosting Scenario-1- Validation and XG Boosting are enabled



Component	Console 9	Summary	Result	Visualiz	ation	Properties	<del>*</del>	<u>+</u>
General	Column selectio	n						
Properties	Dependent Column	sepa	l_length	•	0			
Validation	Independent Colun	nn petal	_length	•	0			
Advanced	New Column In	formation						
	Predicted Column	Pred	ictedValues		0			
	Name							
	Model Tuning							
	Enable Validation	1						
	XGBoosting							
							Apply	/

Scenario-2- Validation and XG Boosting are disabled

Component	Console	Summary	Result	Visualization	Properties	+	Ŧ
General	Column sele	ction					
Properties	Dependent Col	umn se	epal_length	•	9		
Advanced	Independent C	olumn p	etal_length	•	9		
	New Columr	n Information					
	Predicted Colu	mn P	redictedValues		9		
	Name						
	Model Tunin	g					
	Enable Validati	on 🗌					
	XGBoosting						
<u>&gt;</u>							
						Apply	/

Scenario-3- Validation is enabled, but XG Boosting is disabled



Component	Console Summ	ary Result	Visualization	Properties	<del>*</del>	<u>+</u>			
General	Column selection								
Properties	Dependent Column	sepal_length	• 0						
Validation	Independent Column	petal_length	- <b>O</b>						
Advanced	New Column Informa	ew Column Information							
	Predicted Column	PredictedValues	0						
	Name	Jame							
	Model Tuning								
	Enable Validation								
•	XGBoosting								
					Apply	/			

Scenario-4- Validation is disabled, but XG Boosting is enabled

Component	Console	Summary	Result	Visu	alization	Properties	<del>*</del>	<u>+</u>
General	Column selec	tion						
Properties	Dependent Colu	mn s	epal_length	-	0			
Advanced	Independent Col	umn 🛛	etal_length	•	0			
	New Column	Informatic	n					
	Predicted Colum	n f	PredictedValues		0			
	Name	_						
	Model Tuning	;						
•	Enable Validation	n 🗌						
r	XGBoosting	1						
							Apply	y

- iii) Click the 'Validation' tab and configure it:
  - a. Model Selection (when XG Boosting is enabled)
    - i. Number of folds: Enter a number deciding the creation of folds in a model.

Component	Console	Summary	Result	Visualization	Properties	⊥ ⊺
General	Model Sele	ection				
Properties	Number of fo	olds 3	3			
Validation						
Advanced						_
						Apply

#### Validation tab when XG Boosting is disabled



- a. Model Selection
  - i. Model Selection Method: Select a Model Method using the drop-down menu
  - ii. **Number of folds:** Enter a number deciding the creation of folds in a model

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> ↑	<u>+</u>
General	Model Selectio	on					
Properties	Model Selection	Cross	validation	•			
Validation	Method						
Advanced	Number of folds	3					
L							
•							
						Apply	/

iv) Click the 'Advanced' tab and configure if required:

#### Advanced tab when XG Boosting and Validation are disabled

- a. Input Data Handling
  - i. **Missing Values**: Select a method to deal with missing values from the drop-down menu
    - 1. **Ignore**: Select this option to skip the records containing missing values from the dependent and independent columns.
    - 2. **Keep**: Select this option to retain the records containing missing values while performing the calculation.
    - 3. **Stop**: Select this option to stop the algorithm application if a value is missing in any column.
- b. Behavior
  - i. Allow Singular Fit: Select an option for providing value to the Boolean Column
    - 1. **True:** Select this option to ignore aliased coefficients from the coefficient covariance matrix.
    - 2. False: Select this option to show an error in a model containing aliased coefficients
  - ii. **Contrasts**: Select this option to display a list of contrast items that can be used for some variables in the model. The available options are:
    - 1. Contr. Treatment
    - 2. Contr.poly
    - 3. Contr.sum
    - 4. Contr.helmert
  - iii. **Confidence Level**: Enter a value specifying accuracy (Confidence Level) of predictions for the algorithm. This field takes 0.95 as the default value.
  - iv. Click the 'Apply' option.



Component	Console Sun	nmary	Result	Visualization	Properties	<del>*</del>	<u>+</u>
General	Input Data Handli	ng					
Properties	Missing values	Igno	re	-			
Advanced	Behavior						
	Allow Singular Fit	True		-			
	Contrasts	None	e Selected	•			
	Confidence Level	0.95			0		
•							
							-
						Apply	

## Advanced Tab when XG Boosting is disabled, but Validation is enabled

- a. Intercept Parameter
  - i. Intercept Value: Enter an intercept value
  - ii. Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	± ∓
General	Intercept Pa	rameter				
Properties	Intercept Value	2	0.1			
Validation						
Advanced						
						Apply

Advanced Tab when XG Boosting and Validation is enabled or XG Boosting is enabled, but Validation is disabled

- a. Boosting Parameter
  - i. Number of Iterations: Enter the number of iterations.
  - ii. Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> .	Ť
General	Boosting Pa	rameter					
Properties	Number Of Ite	rations 3					
Validation							
Advanced							
						Apply	

**Note:** The model containing aliased coefficients signifies that the square matrix x\*x is singular.

v) Run the workflow after getting the success message.



vi) The '**Console**' tab opens, displaying the process. The completion of the console process gets marked by the green checkmarks at the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
    - i. A new column '**Predicted Values1**' gets added to the Result data displaying the predicted values.

Result when Validation and XG Boosting are disabled.

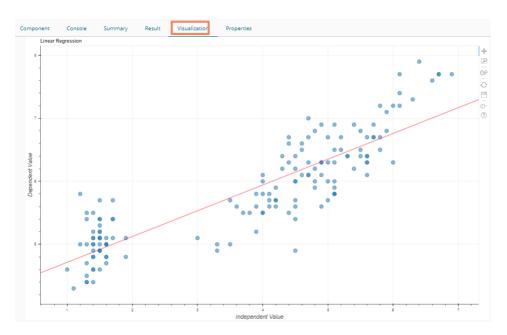
	Component	Console Su	mmary Resul	t Visualizatio	on Prop	erties <u>+</u> ⊥
	Show 10 • ent	tries			Search:	
Þ	sepal_length	sepal_width	petal_length	petal_width	species	PredictedValues
	5.1	3.5	1.4	0.2	setosa	4.87834171414709
	4.9	3	1.4	0.2	setosa	4.87834171414709
	4.7	3.2	1.3	0.2	setosa	4.8374291243003
	4.6	3.1	1.5	0.2	setosa	4.91925430399387
	5	3.6	1.4	0.2	setosa	4.87834171414709
	5.4	3.9	1.7	0.4	setosa	5.00107948368745
	4.6	3.4	1.4	0.3	setosa	4.87834171414709
•	5	3.4	1.5	0.2	setosa	4.91925430399387
	4.4	2.9	1.4	0.2	setosa	4.87834171414709
	4.9	3.1	1.5	0.1	setosa	4.91925430399387
	Showing 1 to 10 of 1	50 entries	F	Previous 1 2	2 3 4	5 15 Next

Result when XG Boosting enabled, and Validation enabled or disabled (No visualization is available for this situation).



	Component	Console Su	immary Result	Visualization	Proper	ties <u>+</u> ⊥
	Show 10 🔻 ent	tries			Search	:
Þ	sepal_length	sepal_width	petal_length	petal_width	species	PredictedValues
	5.1	3.5	1.4	0.2	setosa	3.50660634040833
	4.9	3	1.4	0.2	setosa	3.50660634040833
	4.7	3.2	1.3	0.2	setosa	3.50660634040833
	4.6	3.1	1.5	0.2	setosa	3.50660634040833
	5	3.6	1.4	0.2	setosa	3.50660634040833
	5.4	3.9	1.7	0.4	setosa	3.50660634040833
	4.6	3.4	1.4	0.3	setosa	3.50660634040833
Þ	5	3.4	1.5	0.2	setosa	3.50660634040833
	4.4	2.9	1.4	0.2	setosa	3.50660634040833
	4.9	3.1	1.5	0.1	setosa	3.50660634040833
	Showing 1 to 10 of 1	50 entries		Previous 1	2 3 4	5 15 Next

- viii) Click the 'Visualization' tab.
- ix) The Result data gets displayed via the Scatter Plot with Regression line chart.



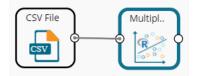
**Note:** 'Behavior' fields provided under the 'Advanced' section differs as per the algorithm sub-type. 'Input

**Data Handling**' remains the same for all the provided Regression types. Hence, only the 'Advanced' tab is explained below for the remaining R sub-algorithms provided under 'Regression.'

# 13.1.4.2. Multiple Linear Regression

i) Drag the Multiple Linear Regression component to the workspace and connect it with a configured data source.





- ii) Configure the 'Properties' tab.
  - a. Column Selection
    - i. **Dependent Column**: Select the target column on which the regression analysis gets applied
    - ii. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column
  - b. New Column Information
    - i. **Predicted Column Name**: Enter a name for the new column containing the predicted values
  - c. Model Tuning
    - i. Enable Validation: Use a checkmark to enable the validation tab
    - ii. XG Boosting: Use a checkmark in the box to enable XG Boosting

cenario 1: Validation is enabled, and XG Boosting is disabled
---

Component	Console	Summary	Result	Visua	lization	Properties	+	<u>+</u>
General	Column selec	tion						
Properties	Dependent Colu	umn uso	d_billing	-	0			
Validation	Independent Co	lumn 6 c	hecked	•	0			
Advanced	New Column	Information						
	Predicted Colun	nn Pre	edictedValues		0			
	Name	_						
	Model Tuning	g						
	Enable Validatio	n 🕑						
	XGBoosting							
•								
							Apply	/

Scenario 2: Validation and XG Boosting are enabled



Component	Console	Summary	Result	Visualization	Properties	<u>+</u> ↑	<u>+</u>		
General	Column sele	tion							
Properties	Dependent Colu	umn us	d_billing	•	9				
Validation	Independent Co	lumn 6 c	checked	•	9				
Advanced	New Column	v Column Information							
	Predicted Colun	nn Pre	edictedValues		9				
	Name								
	Model Tunin	g							
	Enable Validatio	n 🕑							
	XGBoosting	•							
>									
						Appl	y		

Scenario 3: When Validation is disabled, but XG Boosting is enabled.

Component	Console	Summary	Result	Visualizatio	n Properties	+	Ŧ
General	Column selec	ction					
Properties	Dependent Colu	umn uso	d_billing	-	0		
Advanced	Independent Co	olumn 6 c	hecked	•	0		
	New Column	Information					
	Predicted Colun	nn Pre	dictedValues		0		
	Name						
	Model Tuning	g					
	Enable Validatio	n 🗌					
	XGBoosting						
•							
						_	
						Apply	/

Scenario 4: When Validation and XG Boosting are disabled.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Column selec	tion					
Properties	Dependent Colu	ımn usd	_billing	• <b>0</b>			
Advanced	Independent Co	lumn 6 ch	ecked	- <sup>0</sup>			
	New Column	Information					
	Predicted Colun	nn Pred	dictedValues	0			
	Name						
	Model Tuning	g					
	Enable Validatio	n 🗌					
	XGBoosting						
>							
						Apply	y

- iii) Validation
  - a. Validation Model Selection when XG Boosting is disabled
    - i. **Model Selection Method:** Select a model selection method using the drop-down menu.
    - ii. Number of folds: Enter a value for the number of folds.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Model Selectio	n					
Properties	Model Selection	Cross	s validation	•			
Validation	Method						
Advanced	Number of folds	3					
						Apply	/

- b. Validation Model Selection when XG Boosting is enabled
  - i. Number of folds: Enter a value for the number of folds.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Model Sele	ction					
Properties	Number of fol	ds 3					
Validation							
Advanced						Apply	y



iv) Click the 'Advanced' tab and configure if required: When Validation and XG Boosting are disabled

## a. Input Data Handling

- i. Missing Values: Select a method to deal with missing values (via the drop-down menu).
  - 1. **Ignore**: Select this option to skip the records containing missing values from the dependent and independent columns.
  - 2. **Keep**: Select this option to retain the records containing missing values while performing the calculation.
  - 3. **Stop**: Select this option to stop the algorithm application if a value is missing in any column.
- b. Behavior
  - i. **Confidence Level**: Enter a value specifying accuracy (confidence level) of Predictions for the algorithm. This field takes 0.95 as the default value.

Component	Console	Summar	У	Result	Visu	alization	Properties	<u>+</u> ↑	<u>+</u>
General	Input Data Ha	ndling							
Properties	Missing values		Ignore		-				
Advanced	Behavior								
	Confidence Level		0.95			0			
	Contrasts		None Se	elected	•				
•									
							_		_
								Appl	у

# When Validation is enabled and XG Boosting disabled

#### a. Intercept Parameter

- i. Intercept Value: Enter an intercept value.
- ii. Click the **'Apply'** option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> T
General	Intercept	Parameter				
Properties	Intercept Va	lue	0.1			
Validation						
Advanced						
						Apply

# When XG Boosting is enabled with either Validation is enabled or disabled

#### a. Boosting Parameter

- i. No. of Iterations: Enter number suggesting no. of iterations.
- ii. Click the **'Apply'** option.



Component	Console	Summary	Result	Visualization	Properties	± T
General	Boosting F	Parameter				
Properties	Number Of I	terations	3			
Validation						
Advanced	_					Apply

- v) Run the workflow after getting the success message.
- vi) The '**Console**' tab opens displaying the steps of the process. The completion of the console process is marked by the green checkmarks on the top of the dragged components.

ľ				▶ C +
1	Multiple Linear R	egr 🖉		
E	±		Multipl	
	Component	Console	Summary	Result
•	24/09/2019 - 13 24/09/2019 - 13 24/09/2019 - 13	:14:27 : CSV( :14:27 : Mult		

- vii) The processed data gets displayed under the '**Result**' tab (a new column gets added to the result data) with the following possibilities:
- viii) A new column is added to the Result data.
  - a. Result when XG Boosting is disabled.



	Component	Console	Summ	nary Result	Visualization	Properties				÷ ±
	Show 10 🔻	entries							Search:	
	usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joinin;
Þ	4000	Male	Indeed	15	1	Management, Selenium	Athenahealth	1	1800000	02-07-2018
	4000	Male	Orgspire	10	2	Selenium	Support.com	2	1500000	12-01-2018
	2600	Male	Orgspire	4	3	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000	18-07-1980
	2300	Female	Referral	5	4	Selenium	Inventateq	4	650000	18-03-2018
	1750	Male	Referral	3	5	Selenium	Tekinspy	5	520000	15-04-1972
	0	Male	BMS Innolabs	4	6	Java	CGI Information Systems	6	980000	20-05-2018
	0	Male	Orgspire	3	7	AWS	Cognizant Technology solutions	7	650000	10-06-2018
	0	Male	BMS Innolabs	3	8	Java+UI	HCL Technologies	8	845000	20-05-2018
Þ	2000	Male	Referral	2	9	Selenium	Support.com	9	520000	20-02-2017
	0	Male	SkillRecruit	2	10	XLS, Report	Altisource	10	650000	06-02-2017
	Showing 1 to 10	of 224 entries	5	1	1	1	Previous 1	2	3 4 5	23 Next

**b.** Result when XG Boosting is enabled, and Validation is enabled or disabled (No Visualization is available for this Result data)

now 10 •	entries							Search:	
usd_billing	gender	source	experience_Year	candidate_id	skills	previous_organisation	id	offered_ctc	expected_joir
4000	Male	Indeed	15	1	Management, Selenium	Athenahealth	1	1800000	02-07-2018
4000	Male	Orgspire	10	2	Selenium	Support.com	2	1500000	12-01-2018
2600	Male	Orgspire	4	з	Java+UI	Accenture Solutions Pvt. Ltd	3	1024000	18-07-1980
2300	Female	Referral	5	4	Selenium	Inventateq	4	650000	18-03-2018
1750	Male	Referral	з	5	Selenium	Tekinspy	5	520000	15-04-1972
0	Male	BMS Innolabs	4	6	java	CGI Information Systems	6	980000	20-05-2018
0	Male	Orgspire	3	7	AWS	Cognizant Technology solutions	7	650000	10-06-2018
0	Male	BMS Innolabs	3	8	Java+UI	HCL Technologies	8	845000	20-05-2018
2000	Male	Referral	2	9	Selenium	Support.com	9	520000	20-02-2017
D	Male	SkillRecruit	2	10	XLS, Report	Altisource	10	650000	06-02-2017

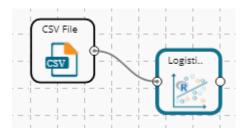
- ix) Click the 'Visualization' tab.
- x) The Scatterplot with Regression Line Chart appears to display the Result data when the XG Boosting is disabled.





# 13.1.4.3. Logistic Regression

i) Drag the Logistic Regression component to the workspace and connect it with a configure data source.



- ii) Configure the 'Properties' tab.
  - a. Column Selection
    - i. **Dependent Column**: Select the target column on which the regression analysis gets applied
    - ii. **Independent Column**: Select the required input columns against which the regression analysis to the target column gets applied
  - b. New Column Information
    - i. **Predicted Column Name**: Enter a name for the new column containing the predicted values
  - c. Model Tuning
    - i. Enable Validation: Use a checkmark to enable the validation tab
    - ii. XG Boosting: Use a checkmark in the box to enable XG Boosting

Scenario 1: XG Boosting and Validation are disabled.



Component	Console	Summary	Result	Visualization	Properties	+	Ŧ
General	Column sele	ction					
Properties	Dependent Col	umn ad	mit	• 0			
Advanced	Independent Co	olumn 3 d	hecked	•			
	New Columr	Information					
	Predicted Colur	nn Pre	edictedValues	0			
	Name						
	Model Tunin	g					
	Enable Validatio	on 🗌					
	XGBoosting						
•							
						Apply	/

Scenario 2: When Validation is enabled, and XG Boosting is disabled.

Component	Console	Summary	Result	Visualiz	ation	Properties	+	<u>+</u>
General	Column sele	ction						
Properties	Dependent Col	umn	admit	•	0			
Validation	Independent Co	olumn	3 checked	•	0			
Advanced	New Columr	n Informatio	'n					
	Predicted Colur	mn	PredictedValues		0			
	Name							
	Model Tunin	g						
	Enable Validatio	on 🗸						
	XGBoosting		]					
•								
							Ар	oly

Scenario 3: When Validation is disabled, and XG Boosting is enabled.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Column selec	tion					
Properties	Dependent Colu	ımn adr	nit	• 0			
Advanced	Independent Co	lumn 3 cl	hecked	<b>•</b> 0			
	New Column	Information					
	Predicted Colun	n Pre	dictedValues	0			
	Name						
	Model Tuning	g					
	Enable Validatio	n 🗌					
	XGBoosting						
>							
						Apply	/

### Scenario 4: Validation and XG Boosting are enabled

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> ↑	<u>+</u>
General	Column sel	ection					
		admit	•	0			
Validation	Independent (	Column	3 checked	•	0		
Advanced	New Colum	n Informatio	'n				
	Predicted Colu	ımn	PredictedValues1		0		
	Name						
	Model Tuni	ng					
•	Enable Validat	ion					
r	XGBoosting						
						Арр	ly

### iii) Validation Tab

#### a. Validation tab when XG Boosting is disabled. Model Selection

- i. **Model Selection Method:** Select a model selection method from the drop-down menu.
- ii. Number of folds: Enter a value for the number of folds.



Component	Console	Summary	Result	Visualization	Properties	<u>↓</u>	<u>+</u>
General	Model Select	ion					
Properties	Model Selection	Cross	validation	•			
Validation	Method						
Advanced	Number of folds	3					
•							
						Apply	y

### b. Validation tab when XG Boosting is enabled Model Selection

i. Number of folds: Enter a value for the number of folds.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Model Select	tion					
Properties	Number of fold	s 3					
Validation							
Advanced						Арр	ly

#### iv) Click the 'Advanced' tab and configure if required:

#### Advanced Tab when Validation and XG Boosting are disabled

#### a. Input Data Handling

- i. Missing Values
  - 1. **Ignore**: Selecting this option will skip the records containing missing values in the columns
  - 2. **Keep:** Select this option to retain the records containing missing values while performing the calculation
  - 3. **Stop:** Select this option to stop (not allow) the records containing missing values while performing the calculation

#### b. Behavior

- i. Family: Select an option from the drop-down list
  - 1. Binomial
  - 2. Poisson
  - 3. Gaussian
  - 4. Gamma
  - 5. Quasi
  - 6. Quasi-Poisson
  - 7. Quasibinomial



ii. **Maximum No. of Iterations:** Enter a valid integer value allowed to calculate the algorithm coefficient. The default value for this field is 25.

Component	Console	Summa	iry	Result	Vis	ualization	Properties	<del>+</del> †	Ŧ
General	Input Data Har	ndling							
Properties	Missing values		Кеер		•				
Advanced	Behavior								
	Family		Binomia	I	•				
	Maximum No of		25						
	Iterations								
• >									
								Appl	y

#### Advanced Tab with Validation enabled and XG Boosting disabled

### a. Input Data Handling

- i. Missing Values:
  - 1. **Ignore**: Select this option to skip the records containing missing values in the columns
  - 2. **Keep:** Select this option to retain the records containing missing values while performing the calculation
  - 3. **Stop:** Select this option to **stop (not allow)** the records containing missing values while performing the calculation

### b. Behavior

- i. Contrast: Select an option from the following list
  - 1. None Selected
  - 2. Contr.treatment
  - 3. Contr.poly
  - 4. Contr.sum
  - 5. Contr.helmert

Component	Console	Summary	Result	Visualizatio	on Properties	<u>+</u>	Ť
General	Input Data I	Handling					
Properties	Missing values		Кеер	•			
Validation	Behavior						
Advanced	Contrasts		Select	•			
						Арр	ly



Advanced tab when XG Boosting is enabled and Validation is enabled or disabled

- a. Boosting Parameter
  - i. No. of Iterations: Enter a number suggesting no. of Iterations

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Boosting Par	ameter					
Properties	No Of Iteration	s 3					
Validation							
Advanced						Арр	ly

- v) Click the 'Apply' option.
- vi) Run the workflow.
- vii) The '**Console**' tab opens, displaying the stepwise process. The completion of the console process gets marked by the green checkmarks at the top of the dragged components.

			► C
Linear Regression			
		Logisti.	
Component	Console	Summary	Result
25/09/2019 - 11: 25/09/2019 - 11:			
25/09/2019 - 11: 25/09/2019 - 11:	0	tic Regression1 is s tic Regression1 is c	

- viii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.
- ix) A new column is inserted into the Result Data.

**Result when XG Boosting is disabled** 



	Component	Console S	Summary	Result Visua	lization F	Properties	+	<u>+</u>
	Show 10 • entr	ries			Searc	h:		
•	admit	gre	gpa	rank	PredictedVa	lues		
	0	380	3.61	3	0.18955274392	27614		
	1	660	3.67	3	0.31778073651	5971		
	1	800	4	1	0.71781360690	)4384		
	1	640	3.19	4	0.14894919378	38017		
	0	520	2.93	4	0.09795420358	353394		
	1	760	3	2	0.37867847044	12818		
	1	560	2.98	1	0.39904112751	1822		
	0	400	3.08	2	0.22117613133	39986		
ĺ	1	540	3.39	3	0.22152034675	5047		
	0	700	3.92	2	0.52050192101	3081		
	Showing 1 to 10 of 40	0 entries		Previous 1	2 3 4	4 5 40	Nex	‹t

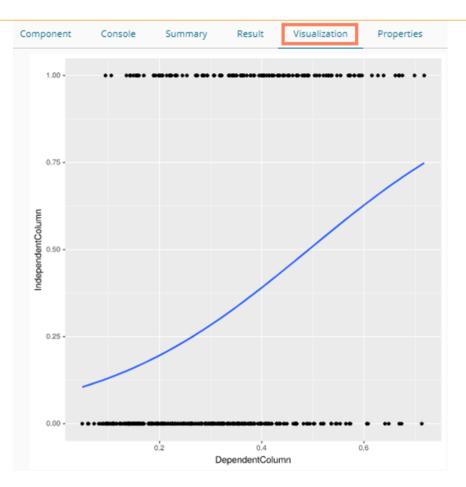
### Result when XG Boosting is enabled

	Component	Console	Summary	Result Visu	alization	Properties $\frac{+}{+}$ $\perp$
	Show 10 🔹 ent	ries			Search	h:
Þ	admit	gre	gpa	rank	PredictedV	alues
	0	380	3.61	3	0.3301004469	39468
	1	660	3.67	3	0.5486503839	94928
	1	800	4	1	0.5851519703	386505
	1	640	3.19	4	0.3215677738	31897
	0	520	2.93	4	0.3278822004	179507
	1	760	3	2	0.6093657612	28006
	1	560	2.98	1	0.3448157310	)48584
Þ	0	400	3.08	2	0.2813929319	38171
	1	540	3.39	3	0.2473349422	21642
	0	700	3.92	2	0.4254856109	061914
	Showing 1 to 10 of 40	00 entries		Previous 1	2 3 4	5 40 Next

### x) Click the 'Visualization' tab.

xi) The Result data gets displayed via the chart displaying the Scatter Plot with a Regression Line.

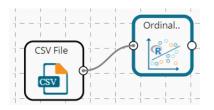




Note: No Visualization is available for the models in which XG Boosting is enabled.

# 13.1.4.1. Ordinal Regression

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

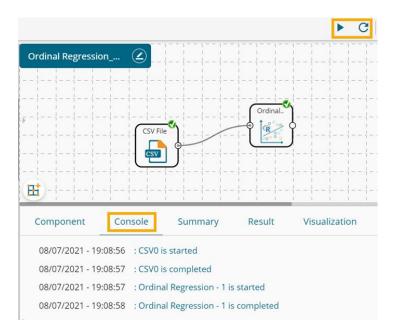


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - **a. Dependent Column**: Select the target column on which the regression analysis gets applied.
  - **b. Independent Column**: Select the required input columns against which the analysis gets applied to the target column.
  - c. Method: Select a method from the given choices that include Probit, loglog, and cloglog
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties		<u>+</u>	Ť
General	Dynamic Fi	elds						
Custom Group	Depe	ndent Column	sex		~			
	[Facto	or Type]						
>	Indep	endent Column	8 che	cked	•			
	Meth	od	prob	it	~			
						_		
							Apply	

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

how	10 🗸 entri	es							Search:
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues_Ordinal_Regression
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	М
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	М
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	F
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	М
1	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	I
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	F
М	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	М
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F

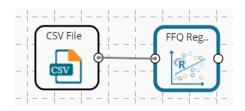
viii) Click the 'Summary' tab to open the model summary.



<pre></pre>	Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
End of Summary	<pre>call: polr(formula = Hess = TRU Coefficients: sexI 10.8758327089 weight_whole -0.2851927400 Intercepts: F I 5.18081 16.133 Residual Devia AIC: 24.00 Warning: did m</pre>	data_1[, deper E, method = met 22.3433726626 weight_shuckec -0.0932266631 I M 0995 nce: 3.651249e- ot converge as	ndentColumn] ~ ., thod) 4 length 5 0.6947355690 4 weight_viscera 1 -0.1055925451 -06 iteration limit	data = data_1 diameter 0.5729872794 weight_shell -0.0004051714	height 0.1371469132 rings			

### 13.1.4.2. FFQ Regression

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

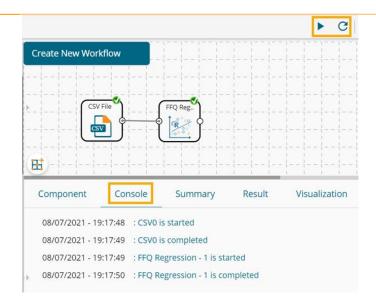


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the analysis gets applied to the target column.
  - c. **Quantile Value (0-1)**: Select a number between 0 to 1 to configure the Quantile Value.
- iv) Click the '**Apply**' option.

Component	Console	Summary	Res	ult	Visualization	Pro	operties	<u>1</u>	. T	
General	Dynamic F	Fields								
Custom Group	Dep	endent Column		sex		~				
•	Inde	pendent Column		8 chec	ked	-				
	Qua	ntile Value(0 - 1)		0.5						
								Ap	oply	

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

ow	10 🗸 entri	es							Search:
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues_FFQ_Regression
4	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	3
4	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	3
	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1
4	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	3
	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	2
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	2
	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	1
	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	1
4	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	3
	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	1

viii) Click the 'Summary' tab to open the model summary.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
No. of varia	est(x = ndata,	r of trees: 500 each split: 2		, importance = TRUE)			

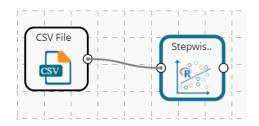
ix) Click the 'Visualization' tab to see the processed data through a chart.





# 13.1.4.3. Stepwise Regression

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



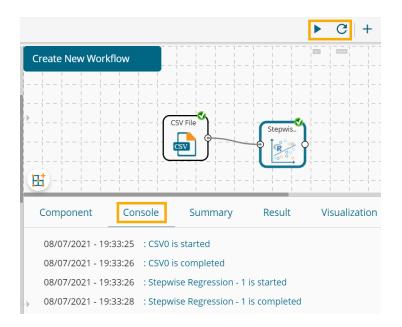
- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the analysis gets applied to the target column.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Dynamic Fi	elds					
Custom Group	Depen	ident Column	sex	~			
•	Indepe	endent Column	8 checked	-			
						Appl	ly

v) Run the workflow after getting the success message.



vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - Component Console Summary Result Visualization Properties Show 10 v entries Search: PredictedValues Number sepal\_length petal\_length 5.1 3.5 1.4 0.2 25.8232916972813 setosa 4.9 1.4 0.2 2 3 setosa 25.8232916972813 4.7 3.2 1.3 0.2 26.3284349742835 3 setosa 4 4.6 3.1 1.5 0.2 setosa 25.3181484202791 5 5 3.6 1.4 0.2 setosa 25.8232916972813 6 5.4 3.9 0.4 setosa 24.3078618662747 4.6 3.4 1.4 0.3 setosa 25.8232916972813

1.5

1.4

1.5

b. Click the 'Result' tab.

8

9

10

5

4.4

4.9

Showing 1 to 10 of 150 entries

3.4

2.9

3.1



0.2

0.2

0.1

setosa

setosa

setosa

25.3181484202791

25.8232916972813

25.3181484202791

Previous 1 2 3 4 5 ... 15

÷ 1

Next



# 13.1.5. Classification

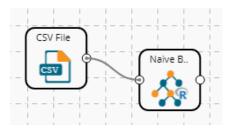
This algorithm categorizes a new observation by a trained set of data that contains observations from the known category. It compares each new observation to previous observations using means of similarity or distance.

# 13.1.5.1. Naive Bayes

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to the presence of any other feature. For example, a fruit may be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, these properties independently contribute to the probability that this fruit is an apple, and that is why it is known as **Naive**.

Naive Bayes is a leaf node under Classification algorithms under the Algorithm tree node. The component consists of one node for reading data from a data source and another one for giving the Result.

i) Drag the Naive Bayes component to the workspace and connect it with a configured data source.



ii) Configure the following fields in the 'Properties' tab:

### a. Column Selection

- i. **Feature**: Select input columns from the drop-down menu to which the target variable can be compared to performing the analysis.
- ii. Target Variable: Select the target column for which the analysis is Performed.
- b. Output Information
  - i. Show Probability: Select an option out of True or False (Selecting the 'True' option displays the Probability Column Name field under the 'New Column Information' section).
- c. New Column Information
  - i. **Predicted Column Name**: Enter a name for the new column containing the predicted values.
  - ii. **Probability Column Name**: Enter a name for the new column containing the probability values.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Column S	election					
Properties	Feature		8 checked	•	0		
Advanced	Target Varia	ble	sex	-	0		
	Output In	formation					
	Show Proba	bility	True	•			
	New Colu	mn Informatio	n				
	Predicted Co	olumn Name	PredictedVal	ues	0		
•	Probability (	Column Name	Probability		0		
	Enable Valid	lation 🗆					
						Apply	,

**d.** Enable Validation: Enable validation by a checkmark in the given box.

- iii) Click the 'Validation' tab and configure it, if it has been enabled from the Properties tab
  - a. Model Selection
    - i. Model Selection Method: Select a modeling method using the drop-down menu.
      - 1. Cross-Validation
      - 2. BootStrap
      - 3. Repeated Cross-Validation
      - 4. Leave One Out Cross-Validation
    - ii. The number of folds: Enter a numerical value for the number of folds.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Model Sel	lection					
Properties	Model Selec	tion Method	Cross validati	ion 👻			
Validation	Number of f	folds	3				
Advanced							
						Apply	/

- iv) Click the 'Advanced' tab and configure if required.
- Advanced Tab when 'Validation' is Disabled:

### a. Input Data Handling

- i. Missing Values: Select a method to deal with missing values from the drop-down menu.
  - 1. **Ignore**: Selecting this option will skip the records containing missing values in the columns.
  - 2. **Keep:** Selecting this option will retain the records containing missing values while performing the calculation.
- Laplace Smoothing: Enter the smoothing constant for smoothing observations.
   The smoothing constant must be a double value greater than 0. Entering 0 will disable Laplace smoothing.

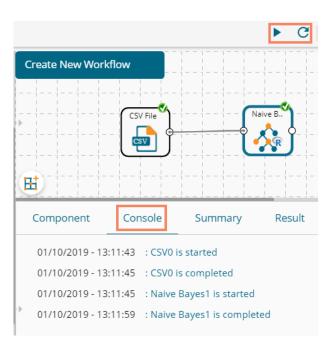


Component	Console	Summary	Result	Visualization	Properties	<u>⊥</u> <u>⊤</u>
General	Input Dat	a Handling				
Properties	Missing val	ues	Ignore	•		
Advanced	Laplace Sm	oothing	0			
						Apply

- Advanced Tab when 'Validation' is Enabled:
- a. Input Data Handling
  - i. Laplace Smoothing: Enter the smoothing constant for smoothing observations. The smoothing constant must be a double value greater than 0. Entering 0 disables Laplace smoothing.
  - ii. Kernel: Select an option using the drop-down menu.
    - 1. True
    - 2. False
  - iii. Band Width: Enter a bandwidth value (the Default value for this field is 0.1).
  - iv. Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Input Data	a Handling					
Properties	Laplace Smo	oothing	0				
Validation	Kernel		True	•			
Advanced	Band Width		0.1				
						Appl	ly

- v) Run the workflow and after getting the success message.
- vi) The '**Console**' tab opens displaying the steps of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





### vii) Click the '**Result**' tab to display the dataset in the result view.

i. Result View when Validation is disabled.

now	10 • entr	ries							Search	á 📃
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	Probability
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	1	[7e-04,0.9963,0.003]
N	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1	[0,1,0]
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1	[0.2073,0.4623,0.330
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	1	[0.0017,0.9895,0.008
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1	[0,1,0]
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1	[0,0.9998,2e+04]
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	М	[0.4222,0.0276,0.550
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	м	[0.39,0.1305,0.4795]
N	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	1	[0.0041,0.9804,0.015
-	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F	[0.5101,0.0039,0.486

### ii. Result View when Validation is Enabled.

Show	10 • ent	ries							Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	Probability
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	1	[7e-04,0.9953,0.004]
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1	[0,0.9999,1e-04]
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1	[0.189,0.6747,0.1363]
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	1	[0.0079,0.9857,0.006
I.	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1	[0,0.9999,1e-04]
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1	[0,1,0]
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F	[0.5632,0.1197,0.317
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	F	[0.4052,0.3552,0.239
M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	1	[0.0039,0.9906,0.005
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F	[0.7526,0.0021,0.245

viii) Click the 'Summary' tab to see the detailed Model Summary.



(	Component	Console	Summary	Result	Visualization	Properties	<del>*</del>	Ţ
	9	Summary of the	model					
	1.Independent	Columns:						
	heigh weigh weigh weigh	e) ter (double) t (double) t_whole (double t_shucked (double t_viscera (double t_shell (double	ble) ble)					
	2.Dependent Co	olumn:						
	sex (stri	ng)						
	3. Model Call	:						
	naiveBayes.de	fault(x = df,	y = sex, laplac	e = 0, na.act:	ion = na.omit)			
		End of Summary						

#### Note:

- a. The '**Visualization**' tab does not display any graphical representation for the Naive Bayes Results in data.
- b. The 'Validation' tab provides multiple options under the 'Model Selection Method' dropdown menu.

All the available Model Selection Methods are described below:

### i. Cross-Validation

The user needs to configure the 'Number of folds' if Cross-Validation is selected as the Model Selection Method.

Component	Console	Summary	Result	Visualization	Properties	<u>1</u>	Ť
General	Model Sel	ection					
Properties	Model Selec	tion Method	Cross validation	-			
Validation	Number of t	folds	3				
Advanced							
						Ар	ply

#### ii. Bootstrap

The user needs to configure the 'Number of resamples' if 'Bootstrap' is selected as the Model Selection Method.



Component	Console	Summary	Result	Visualization	Properties	± T
General	Model Sel	ection				
Properties	Model Select	tion Method	Boot Strap	•		
Validation	Number of F	Resamples	3			
Advanced						
						Apply

#### iii. Repeated Cross-Validation

The user needs to configure the **Number of repeats**, and **the Number of folds** fields if the selected modeling method is **Repeated Cross-Validation**.

Component	Console	Summary	Result	Visualization	Properties	<u>⊥</u> ⊺
General	Model Sel	ection				
Properties	Model Selec	tion Method	Repeated Cros	ss Validation 👻		
Validation	Number of f	olds	3			
Advanced	Number of F	Repeats	3			
						Apply

# iv. Leave One Out Cross-Validation

Users do not get any other field to configure if the selected model method is **Leave one out cross-validation.** 

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Model Sel	lection					
Properties Validation Advanced	Model Selec	tion Method	Leave one out	cross validation 👻			
						Apply	

### 13.1.6. Tree-Based Modeling

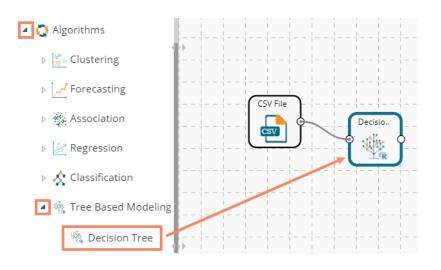
The Tree-Based Modeling Decision Tree can be configured using two algorithm types from the '**Properties**' tab.

Check out the below-given description of the configuration details:

# 13.1.6.1. Classification as Algorithm Type for Decision Tree

i) Drag the Decision Tree component to the workspace and connect it with a configured data source.





ii) Configure the '**Properties**' tab:

### a. Output Information

- i. **Algorithm Type:** Select an algorithm type from the drop-down menu.
  - 1. **Classification**: Select this option if users want to pass the dependent column as the categorical values.
  - 2. **Regression:** Select this option if users want to pass the dependent column as numerical values.
- ii. **Show Probability**: Select an option from the drop-down menu to create a new column for indicating the chance factor involved in the probability.
  - 1. **True:** Select this option to display a new column in the output data with probability values.
  - 2. False: Select this option to display any probability value in the output data.

### b. Column Selection

- i. **Features**: Select input columns from the drop-down list to which the target column needs to compare performing the analysis.
- ii. Target Variable: Select the target column for which the analysis is performed.
- c. New Column Information
  - i. **Predicted Column Name**: Enter a name for the new column containing the predicted values.
  - ii. **Probability Column Name:** Enter a name for the new column containing the probability values.
- d. Model Tuning
  - i. **Enable Validation:** Enable validation as a model tuning option by a checkmark in the given box.
  - ii. **XG Boosting:** Enable validation as a model tuning option by a checkmark in the given box.

#### Properties Tab when Model Tunning is not Enabled



Component	Console	Summary	Result	Visualization	Properties		*	<u>+</u>
General	Output Info	ormation						
Properties	Algorithm Typ	e	Classification	-				
Advanced	Show Probabi	ility	True	-				
	Column Sel	lection						
	Features		7 checked	-	0			
	Target Variab	le	sex	-	0			
	New Colum	nn Information						
	Predicted Col	umn Name	PredictedValue	s0	0			
	Probability Co	olumn Name	Probability0		0			
	Model Tuni	ing				-		
•	Enable Valida	tion						
	XGBoosting	ĺ						
						_		
							Apply	

Properties Tab when Validation is Enabled as Model Tuning

Component	Console	Summary	Result	Visualization	Properties	+
General	Output Info	ormation				
Properties	Algorithm Typ	pe	Classification	-		
» Advanced	Show Probab	ility	True	-		
Validation	Column Se	lection				
	Features		7 checked	•	0	
	Target Variab	le	sex	-	0	
	New Colum	nn Informatior	ı			
	Predicted Col	umn Name	PredictedValu	es0	0	
	Probability Co	olumn Name	Probability0		0	
	Model Tuni	ing				
>	Enable Valida	ition	1			
						Apply

Properties Tab when XG Boosting is Enabled as Model Tuning



Component	Console	Summary	Result	Visualizatior	Properties	<u>+</u> ⊥
General	Output Inform	ation				
Properties	Algorithm Type	Classi	ification	•		
Advanced	Column Select	ion				
	Features	7 che	cked	• 3	l.	
	Target Variable	sex		• 0		
	New Column I	nformation				
	Predicted Column	Predic	tedValues0	0		
	Name					
	Model Tuning					
•	XGBoosting					
						Apply

Note: The 'Show Probability' field appears only if, 'Classification' option is selected via the 'Algorithm Type' drop-down menu.

iii) Click the 'Advanced' tab and configure if required:

• Advanced Tab when both the Model Tuning options are Disabled

### a. Input Data Handling

- i. Missing Values: Select a method to deal with missing values from the drop-down list.
  - 1. **Rpart:** Select this option to get the estimated missing values for the dependent column based on the independent columns.
  - 2. **Ignore:** Select this option to skip the records containing missing values in the columns.
  - 3. **Keep:** Select this option to retain the records containing missing values while performing the calculation.
  - 4. **Stop:** Select this option to stop the algorithm application if a value is missing in any column.

### b. Tree Pruning

- i. **Minimum Split:** It indicates a minimum number of observations within a single node for a split to be attempted. The default value for this field is 10.
- ii. **Complexity Parameter:** This parameter is primarily used to save computing time by pruning off splits that are not worthwhile. Any split which does not improve the fit by a factor of the complex parameter is pruned off performing cross-validation, hence the program does not pursue it. The default value for this field is 0.05.
- iii. Maximum Depth: It sets the maximum depth of any node of the final tree keeping the depth count for root node 0. It is an optional field (It is recommended to set Maximum Depth value less than 30 rpart for 32 bit-machines.)
- c. Behavior



i. **Split Criteria:** It is an optional field that depends on the selected algorithm type from the **'Properties'** tab. (This field appears only when the selected algorithm type is **'Classification'**).

The splitting index can be:

- 1. **Gini:** Select this option to measure inequality among values of randomly chosen elements from a set.
- 2. **Information:** Select this option to get information about the variables used in the algorithm.
- ii. **Cross-Validation:** It indicates the number of cross-validations that were performed to check the accuracy of the analysis method.
- iii. Prior Probability: It is an optional field. This field is dependent on the other data values mentioned in the selected dataset. (This field appears when the selected algorithm type is 'Classification').

### d. Surrogate Information

- i. Use Surrogate: Select one option from the drop-down menu.
  - 1. **Display Only:** Select this option to display only the observation, but not split it further.
  - 2. **Use Surrogate:** Select this option to search surrogate values for the missing values to split the observation. Two fields are displayed:
    - a. Surrogate Style: Select a style using the drop-down menu.
    - b. Maximum Surrogate: Set the maximum surrogate value.
  - 3. **Stop if missing:** Select this option to choose an action based on the nature of the majority observations. If values are missed for all the observations, then they will stop splitting further.

Component	Console	Summary	Result	Visualization	Properties	*	Ţ
General	Input Data H	Handling					
Properties	Missing values		Rpart	•			
Advanced	Tree Prunin	g					
	Minimum Split		10				
	Maximum Dep	th	Optional				
	Behavior						
	Split Criteria		Gini	•			
	Cross Validatio	n	Optional		]		
	Prior Probabili	ty	Optional				
	Surrogate Ir	nformation					
•	Use Surrogate		Use surrogate	•			
	Surrogate Style	e	Use total corre	ct classification 🕞			
	Maximum Sur	rogate	Optional				
					4	pply	

- Advanced Tab when 'Validation' is enabled:
  - a. Tree Pruning:
    - i. **Complexity Parameter**: This parameter is primarily used to save computing time by pruning off splits that are not worthwhile. Any split which does not improve the fit by a factor of the complex parameter is pruned off performing cross-validation, hence the program does not pursue it. The default value for this field is 0.05.



Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Tree Pruning						
Properties	Complexity	.005					
Advanced	Parameter						
Validation							
						Apply	

- iv) Click the 'Validation' tab and configure the required fields
  - **a.** Model Selection Method: Select a method using the drop-down menu. Users need to configure the other fields based on the selected model method.
    - i. Cross-Validation

The user needs to configure the 'Number of folds' if the selected model method is Cross Validation.

Component	Console	Summary	Result	Visualization	Properties	<u>⊥</u> ⊺
General	Model Selection	on				
Properties	Model Selection	Cros	s validation	•		
Advanced	Method					
Validation	Number of folds	3				
						Apply

### ii. Bootstrap

The user needs to configure the 'Number of resamples' (the Default value for this field

is

5) if the selected model method is 'Bootstrap.'

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Model Select	ion					
Properties	Model Selection	Вос	tstrap	•			
Advanced	Method						
Validation	Number of resa	imples 5					
						Apply	

### iii. Repeated Cross-Validation

The user needs to configure the 'Number of repeats' and 'Number of folds' if the selected method is 'Repeated Cross-Validation.'



Component	Console	Summary	Result	Visualization	Properties	+ +	Ŧ
General	Model Selectio	n					
Properties	Model Selection	Repe	ated cross valid	ation 👻			
Advanced	Method						
Validation	Number of repeat	ts 5					
	Number of folds	3					
•							
						Apply	

iv. Leave One Out Cross-Validation

The user does not get any other field to configure if the selected model method is **Leave one out cross-validation**.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Model Select	ion					
Properties	Model Selection	Leave	e one out cross	validat. <del></del>			
Advanced	Method						
Validation						Apply	

### • Advanced Tab when 'XG Boosting' is enabled

#### a. Boosting Parameter

- i. Number of Iterations: Enter a number suggesting the Number of Iterations
- ii. Number of Classes: Enter a number suggesting the Number of Classes
- v) Click the '**Apply**' option after configuring the required Properties, Advanced, and/or Validation fields as per your selection of the model.

Component	Console S	Summary	Result	Visualization	Properties	± ↑	
General	Boosting Parame	ter					
Properties	Number Of Iterations	5 3					
Advanced	Number Of Classes	10		0			
r							
						Apply	

- vi) Run the workflow after getting the success message.
- vii) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



			► C
Create New Workflow			<b>.</b>
CSV I			
E			
	onsole	Summary	Result
			Result
Component Co	: CSV1	is started	Result
Component Co 03/10/2019 - 10:31:00	: CSV1	is started	Result

- viii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
    - i. Result view when both the Model Tuning options are disabled

Show	10 <b>v</b> ent	ries							Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	Probability
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	1	[0.1532,0.6312,0.2156]
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1	[0.1532,0.6312,0.2156]
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1	[0.1532,0.6312,0.2156]
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	1	[0.1532,0.6312,0.2156]
I.	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1	[0.1532,0.6312,0.2156]
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1	[0.1532,0.6312,0.2156]
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	1	[0.1532,0.6312,0.2156]
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	м	[0.3411,0.227,0.4319]
М	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	1	[0.1532,0.6312,0.2156]
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	М	[0.3411,0.227,0.4319]

#### ii. Result view when 'Validation' is enabled

how	10 🔻 ent	ries							Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	Probability
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	1	[0.1532,0.6312,0.215
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1	[0.1532,0.6312,0.215
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1	[0.1532,0.6312,0.215
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	1	[0.1532,0.6312,0.2156
1	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1	[0.1532,0.6312,0.2156
1	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1	[0.1532,0.6312,0.2156
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	1	[0.1532,0.6312,0.2156
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	м	[0.3411,0.227,0.4319]
М	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	1	[0.1532,0.6312,0.215
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	м	[0.3411,0.227,0.4319

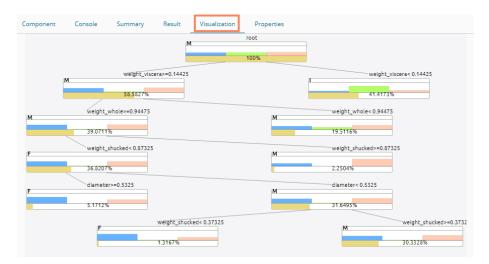
iii. Result view when 'XG Boosting' is enabled



how	10 <b>v</b> en	tries					Se	earch:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValu
M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	1
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	М
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	М
M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	1
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F

Note: The Probability column displays data in the Array format when Validation is enabled.

- ix) Click the 'Visualization' tab.
- x) The Result data gets displayed via the tree chart.a. Visualization tab when no Model Tuning option is enabled



b. Visualization tab when Validation is enabled

Component	Console	Summary	Result	Visualization	Properties	
				м	root	
					100%	
,	и	weight_visc	era>=0.1442	5		weight_viscera< 0.14425
í í	•	58.5827%				41.4173%
	weight_w	hole>=0.94475			M	weight_whole< 0.94475
M	39.0711%	,			M	19.5116%
	weight_sl	hucked< 0.87325				weight_shucked>=0.87325
F	36.8207%				M	2.2504%
	diameter	>=0.5325				diameter< 0.5325
F	5.1712%				м	31.6495%
-	5.171270		weight chu	cked< 0.37325		weight_shucked>=0.373
	F			0.57525		M
			1.3167%			30.3328%
	-		weight_vis	scera>=0.29925		Weight_viscera< 0.299
			8.2356%			22.0972%
	weight_	whole< 1.441				weight_whole>=1.441
F	4.5487%				м	3.6869%



# 13.1.6.2. Regression as Algorithm Type for Decision Tree

- i) Drag the Decision Tree component to the workspace and connect it to a configured data source.
- ii) Configure the following fields in the 'Properties' tab:
  - a. Output Information
    - i. **Algorithm Type:** Select an algorithm type from the drop-down menu.
      - 1. **Classification**: Select this option if users want to pass the dependent column as the categorical values.
      - 2. **Regression:** Select this option if users want to pass the dependent column as numerical values.
  - **b.** Column Selection
    - i. **Features**: Select input columns from the drop-down list to which the target column can be compared to performing the analysis.
    - ii. Target Variable: Select the target column for which the analysis is performed.
  - c. New Column Information
    - i. **Predicted Column Name**: Enter a name for the new column containing the predicted values.
    - ii. **Probability Column Name:** Enter a name for the new column containing the probability values.
  - d. Model Tuning
    - i. Enable Validation: Enable validation by a checkmark in the given box.
    - ii. **XG Boosting:** Enable XG Boosting with a checkmark in the given box.

Component	Console	Summary	Result	Visualization	Properties	+	Ŧ
General	Output Infor	mation					
Properties	Algorithm Type	Reg	ression	•			
Advanced	Column Sele	ction					
	Independent Co	olumns 7 ch	necked	- 0			
	Dependent Col	umns ring	s	- 8			
	New Columr	Information					
	Predicted Colur	nn Pred	dictedValues0	0			
	Name						
	Model Tunin	g					
•	Enable Validatio	on 🗌					
	XGBoosting						
						Apply	

Note: Other possible scenarios to configure the Properties tab can be when either of the Model Tuning option is enabled.

- iii) Click the 'Advanced' tab and configure if required:
  - Advanced Tab when both the Model Tuning options are disabled:



- a. Input Data Handling
  - i. **Missing Values:** Select a method to deal with missing values from the drop-down list.
    - 1. **Rpart:** Select this option to estimate the missing values for the dependent column based on the independent columns.
    - 2. **Ignore:** Select this option to skip the records containing missing values in the columns.
    - 3. **Keep:** Select this option to retain the records containing missing values while performing the calculation.
    - 4. **Stop:** Select this option to stop the algorithm application if a value is missing in any column.

### b. Tree Pruning

- i. **Minimum Split:** It indicates a minimum number of observations within a single node for a split to be attempted. The default value for this field is 10.
- ii. **Complexity Parameter:** This parameter is primarily used to save computing time by pruning off splits that are not worthwhile. Any split which does not improve the fit by a factor of the complex parameter is pruned off performing cross validation, hence the program does not pursue it. The default value for this field is 0.05.
- iii. **Maximum Depth:** It sets the maximum depth of any node of the final tree keeping the depth count for root node 0. It is an optional field (It is recommended to set Maximum Depth value less than 30 rpart for 32 bit-machines.)
- c. Behavior
  - i. **Split Criteria:** It is an optional field that depends on the selected algorithm type from the **'Properties'** tab. (This field appears when the selected algorithm type is **'Classification'**).

The splitting index can be:

- 1. **Gini:** Select this option to measure inequality among values of randomly chosen elements from a set.
- 2. **Information:** Select this option to get information about the variables used in the algorithm.
- ii. **Cross-Validation:** It indicates the number of cross-validations that were performed to

check the accuracy of the analysis method.

iii. **Prior Probability:** It is an optional field. This field is dependent on the other data values mentioned in the selected dataset. (This field appears when the selected algorithm type is '**Classification**').

### d. Surrogate Information

- i. **Use Surrogate:** Select one option from the drop-down menu.
  - 1. **Display Only:** Select this option to display only the observation, but not split it further.
  - 2. **Use Surrogate:** Select this option to search surrogate values for the missing values to split the observation. Two fields are displayed:
    - a. Surrogate Style: Select a style using the drop-down menu.
    - b. Maximum Surrogate: Set the maximum surrogate value.
  - 3. **Stop if missing:** Select this option to choose an action based on the nature of the majority observations. If values are missed for all the observations, then they will stop splitting further.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Input Data H	andling					
Properties	Missing values	R	part	-			
Advanced	Tree Pruning						
	Minimum Split	10	D				
	Maximum Dept	h O	ptional				
	Behavior						
	Cross Validation	0	ptional				
	Surrogate Inf	ormation					
	Use Surrogate	U	se surrogate	-			
>	Surrogate Style	U	se total correct clas:	sifica <del></del>			
	Maximum Surro	gate O	ptional				
						Apply	,

### • Advanced Tab when 'Validation' is enabled:

### a. Tree Pruning:

i. Complexity Parameter: This parameter is primarily used to save computing time by pruning off splits that are not worthwhile. Any split which does not improve the fit by a factor of the complex parameter is pruned off performing cross-validation, hence the program does not pursue it. The default value for this field is 0.05.

Component	Console	Summary	Result	Visualization	Properties	<u>⊥</u> <u>⊤</u>
General	Tree Prunir	ıg				
Properties	Complexity Pa	arameter .00	5			
Advanced						
Validation						
						Apply

iv) Click the '**Validation**' tab and configure the required fields. The user can refer to the description provided under section 12.2.6.1 to configure the Validation tab.

### • Advanced Tab when XG Boosting is Enabled

#### a. Boosting Parameter

- i. Number of Iterations: Enter a number suggesting the Number of Iterations
- ii. Number of Classes: Enter a number indicating the Number of Classes

Component	Console	Summary	Result	Visualization	Properties	<u>⊥</u> <u>⊤</u>
General	Boosting Pa	rameter				
Properties	Number Of Ite	rations 3				
Advanced	Number Of Cla	asses 10		0		
,						
						Apply



- v) Click the 'Apply' option.
- vi) Run the workflow after getting the success message.
- vii) The 'Console' tab opens.

		►	G +
Decision Tree			
•			
	CSV File	Decisio	ļ
<b>H</b>		₩ R	Í
Component	Console	Summary	Result
03/10/2019 - 12	:15:40 : CSV1 i	s started	
03/10/2019 - 12	:15:41 : CSV1 is	s completed	
03/10/2019 - 12	:15:41 : CNR T	ree0 is started	
03/10/2019 - 12	:15:43 : CNR Ti	ree0 is completed	

- viii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.
    - i. The Result tab when both the Model Tuning options are disabled

now	10 v ent	ries						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	8.770609
N	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	7.551181
-	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	9.553571
Л	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	8.770609
	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	6.283951
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	8.770609
	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	13.160338
-	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	12.745902
Л	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	8.770609
	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	13.160338

ii. The Result tab when the 'Validation' option is enabled



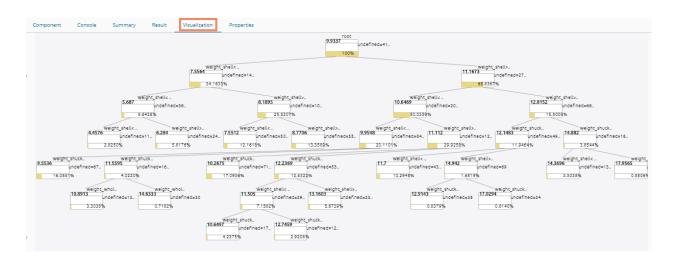
how	10 • ent	ries							Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	Probability
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	1	[0.1532,0.6312,0.215
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1	[0.1532,0.6312,0.215
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1	[0.1532,0.6312,0.215
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	1	[0.1532,0.6312,0.215
I.	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1	[0.1532,0.6312,0.215
I.	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1	[0.1532,0.6312,0.215
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	1	[0.1532,0.6312,0.215
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	м	[0.3411,0.227,0.4319
М	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	1	[0.1532,0.6312,0.215
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	м	[0.3411,0.227,0.4319

#### iii. Result view when 'XG Boosting' is enabled

how	10 🔻 ent	tries					Se	earch:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValue
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	1
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	М
1	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	М
М	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	1
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F

Note: The Probability column is displayed in the Array format while enabling the 'Validation' option.

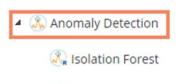
- ix) Click the 'Visualization' tab.
- x) The Result data will be displayed via the tree chart.
   (The following visualization displays processed data when no Model Tuning option is enabled.)





# 13.1.7. Anomaly Detection

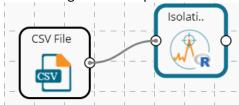
The Anomaly Detection algorithm has the following algorithms:



Anomaly Detection

# 13.1.7.1. Isolation Forest

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

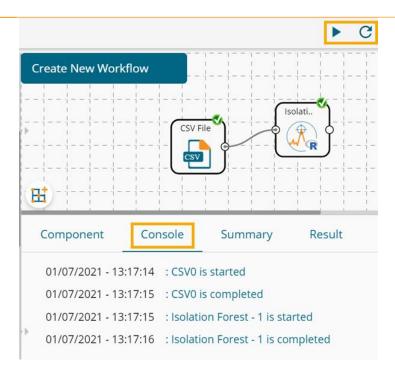


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Features: Select the input columns with which you want to perform the Analysis
  - b. Quantile Value (0-1): Select a number between 0 to 1 to configure the Quantile Value.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>⊥</u> <u>⊺</u>
General	Dynamic Fie	elds				
Custom Group	Featur	es	8 checked	•		
•	Quanti	le value (0 - 1)	0.95			
						Apply

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

Show 10 v entries Search:									
essure_height	Wind_speed	Humidity	Temperature_Sandburg	Temperature_ElMonte	Inversion_base_height	Pressure_gradient	Inversion_temperature	Visibility	Anomalydetec
0	8	20	54		5000	-15	30.56	200	FALSE
0	6		38			-14		300	FALSE
0	4	28	40		2693	-25	47.66	250	FALSE
0	3	37	45		590	-24	55.04	100	FALSE
0	3	51	54	45.32	1450	25	57.02	60	FALSE
0	4	69	35	49.64	1568	15	53.78	60	FALSE
D	6	19	45	46.4	2631	-33	54.14	100	FALSE
0	3	25	55	52.7	554	-28	64.76	250	FALSE
D	3	73	41	48.02	2083	23	52.52	120	FALSE
0	3	59	44		2654	-2	48.38	120	FALSE

viii) Click the 'Summary' tab to open the model summary.



C	Componer	nt	Console	Summary	Result	Visualization	Properties
	177	0.28418	31				
	178	0.28418	31				
	179	0.28418	31				
	180	0.59706	57				
	181	0.28418	31				
	182	0.28418	31				
	183	0.28418	31				
	184	0.28418	31				
	185	0.28418	31				
	186	0.28418	31				
	187	0.28418	31				
	188	0.88118	27				
	189	0.82717	80				
	190	0.28418	31				
	191	0.28418	31				
	192	0.28418	31				
	193	0.28418	31				
	194	0.28418	31				
	195	0.28418	31				
	196	0.28418	31				
	197	0.28418	31				
	198	0.28418	31				

# 13.1.7.2. Anomaly Detection

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
- iv) Provide the following details to configure the Dynamic Fields:
  - a. No. of Periods
  - b. Column Name
  - c. Maximum Number of Anomalies (as % "0.01-0.49)
  - d. Direction
- v) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Dynamic	Fields					
Custom Group	No	o of Periods		10			
	Co	lumn Name		ozone_reading	~		
	Ma	aximum Number Of		0.40			
	An	iomalies (as % "0.01-					
	0.4	49)					
	Di	rection		both	~		
•							
						 Apply	v



- vi) Run the workflow after getting the success message.
- vii) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- viii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

show 10 🗸	entries			Search:	
ature_ElMonte	Inversion_base_height	Pressure_gradient	Inversion_temperature	Visibility	OutliersDetected
	5000	-15	30.56	200	FALSE
		-14		300	FALSE
	2693	-25	47.66	250	FALSE
	590	-24	55.04	100	FALSE
	1450	25	57.02	60	FALSE
	1568	15	53.78	60	FALSE
	2631	-33	54.14	100	FALSE
	554	-28	64.76	250	FALSE
	2083	23	52.52	120	FALSE
	2654	-2	48.38	120	FALSE





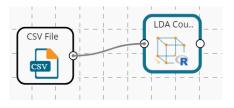
# 13.1.8. Dimensionality Reduction

The Dimensionality Reduction algorithm has the following algorithms:

Dimensionality Reduction
EDA Count
LDA Feature Selection
Factor Analysis
PCA
Boruta FeatureSelect

## 13.1.8.1. LDA Count

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

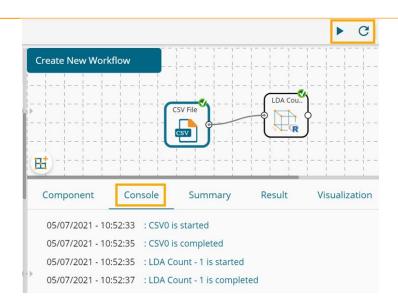


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the analysis gets applied to the target column.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>↓</u> T
General	Dynamic F	ïelds				
Custom Group	Depe	endent Column	sex		~	
>	Inde	pendent Column	8 chec	ked	•	
						Apply

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

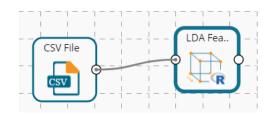
Show 10 v entries Search:								
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	ring
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16
M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19

Compon	ent Console	Summary	Result	Visualization	Properties	÷ <u>+</u>
	Summary of th					
		02				
	0.119193510 0.398128					
	2.121285219 -0.289739					
	0.931010201 0.310578					
	0.307868432 0.077195					
	2.380895339 -0.419890 1.892206698 -0.034617					
	L.892206698 -0.034617 L.002066708 -0.070073					
	.868878353 0.419804					
	0.152125754 0.525397					
	L.363113450 -0.740598					
	0.190332518 -1.296687					
	0.058561289 0.723544					
	0.553041944 0.086805					
	.580343333 -0.486692					
15 0	0.685478380 1.111974	52				
16 -6	0.512769054 0.989493	40				
17 2	2.126243826 0.347006	19				
 18 🤅	0.684345100 -0.090547	25				
19 1	1.557475179 1.316197	79				
	L.228519789 -0.218763					
21 1	1.334979910 -0.622894	85				
	1.641311377 -0.762758					
	0.869266061 -0.299864					
24 -6	0.290727841 -0.426519	81				



## 13.1.8.2. LDA Feature Selection

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the analysis gets applied to the target column.
- iv) Click the 'Apply' option.

General	Dynamic Fields		
Custom Group	Dependent Column	sex	~
•	Independent Column	8 checked	•
			Apply
			трру

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





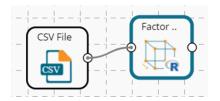
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

Component Console	Summary Result	Visualization Properties 🕂 🛓			
Show 10 v entries		Search:			
dependent	LD1	LD2			
М	-0.119193509632358	0.398128167632963			
Μ	2.12128521863818	-0.2897397258665			
F	-0.931010200821666	0.310578922579137			
Μ	-0.307868431626052	0.0771955526076383			
1	2.38089533864659	-0.419890125434961			
I	1.89220669787132	-0.0346173682605865			
F	-1.00206670791673	-0.0700733857472464			
F	-0.868878353441633	0.419804353815104			
Μ	-0.15212575350618	0.525397968383977			
F	-1.36311344951122	-0.740598132605857			
Showing 1 to 10 of 100 entries	Pr	revious 1 2 3 4 5 10 Next			

omponent	Console	ummary Result	Visualization	Properties	* <del>*</del>
	Summary of the mode				
Call:	summary of the mode	1			
	~ ., data = data 1	.)			
	· –				
Prior probabi	lities of groups:				
FI					
0.44 0.12 0.4	4				
-					
Group means:	diamoton boight	weight whole weight	churched und able ude e		
	.4294318 0.14568182		4856818 0.19511		
		0.2018750 0.0			
		0.7250227 0.2			
weight_shel	l rings				
F 0.2727272	7 11.840909				
	3 6.416667				
M 0.2337500	0 10.636364				
Coefficients	of linear discrimin	ants:			
coerriceienco	LD1	LD2			
length	6.98856475 -9.	7879992			
	-35.91089554 36.				
	-18.00989320 -39.				
	3.79995685 -22.				
	d -1.32033731 17.				
	a 1.65836534 12. 4.01369614 35.				
rings	-0.07906803 -0.				
1 1163	-0.07500005 -0.	1145542			
Proportion of	trace:				
LD1 LD2					
0.9314 0.0686	End of Summary				

#### 13.1.8.3. Factor Analysis

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.





- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
- iv) Provide the value of Factors and Variables.
- v) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Dynamic F	Fields					
Custom Group	Fact	ors [ (V-F)^2 > (V +	F): 3				
•	V(va	riables) and F(fact	ors)]				
						Apply	,

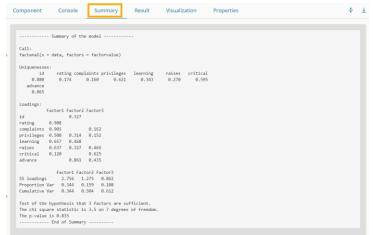
- vi) Run the workflow after getting the success message.
- vii) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.

			► C
Factor Analysis_Sam			·
		Factor	<b>y</b>
	CSV File		
			<b>/</b>
🛱 -	·		$\frac{1}{2} - \frac{1}{2} - \frac{1}$
<b>H</b>			
	onsole Summary	Result	Visualization
	onsole Summary	Result	Visualization
Component Co	onsole Summary : : CSV0 is started	Result	Visualization
Component Co 09/07/2021 - 11:02:35	5 : CSV0 is started 5 : CSV0 is completed		Visualization

- viii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

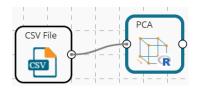
10 •     entries     Search:							h:
id	rating	complaints	privileges	learning	raises	critical	advance
	43	51	30	39	61	92	45
	63	64	51	54	63	73	47
	71	70	68	69	76	86	48
	61	63	45	47	54	84	35
	81	78	56	66	71	83	47
	43	55	49	44	54	49	34
	58	67	42	56	66	68	35
	71	75	50	55	70	66	41
	72	82	72	67	71	83	31
0	67	61	45	47	62	80	41





#### 13.1.8.4. PCA

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

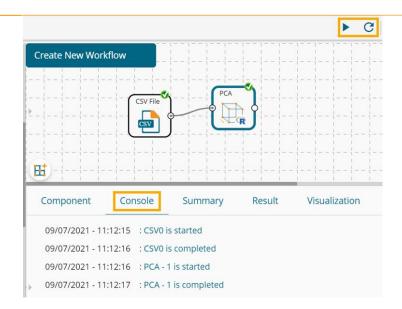


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component as given below:
  - a. Features [Numeric Columns]
  - b. Number of PCA Components
- iv) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Dynami	c Fields					
Custom Group	F	eatures [Numeric C	[olumns]	8 checked	•		
►	1	Number of PCA Con	nponents	3			
						Арр	ly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

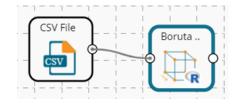
ow 10 v entries		Search:
PC1	PC2	PC3
2.56590832791396	2.91011071665735	1.67557407671481
0.44161156945008	0.484477363154375	1.05997026283436
1.88405069512985	1.1311608595477	1.441810925659
1.51461157668581	0.556010458525973	1.58306882206549
1.88868524344925	0.324132545428099	1.40896610055505
2.82650104808833	-1.1873025371556	0.0934302900685209
0.944519421119523	-0.437065378148994	0.70285593338795
).35249067333804	-0.758054986262042	0.65512483683614
1.66926594189581	-0.685799958961177	1.57212979870521
0.802885462810068	0.531777026382975	0.628419856751585

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
Created from Pre-processin - Box-Cox t - centered - ignored ( - principal - scaled (& Lambda estima 0.7, 1.8, 1.4 PCA used 3 co	30 samples and g: ransformation (8) o) component sig )	<ul> <li>(8)</li> <li>nal extraction (3)</li> <li>x transformation</li> <li>4, 2, -0.6</li> <li>ecified</li> </ul>	3)				

## 13.1.8.5. Boruta Feature Select



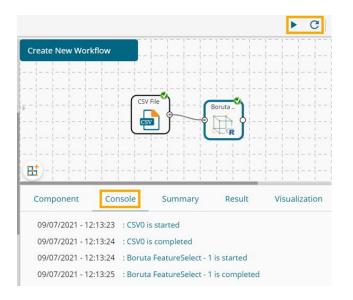
- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the analysis gets applied to the target column.
- iv) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	± Ŧ
General	Dynamic I	Fields				
Custom Group	Dep	endent Column	critic	al	~	
	Inde	pendent Column	7 che	cked	•	
•						
						Apply

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.



ow	10 v entries					Search:	
id	rating	complaints	privileges	learning	raises	critical	advance
	43	51	30	39	61	92	45
	63	64	51	54	63	73	47
	71	70	68	69	76	86	48
	61	63	45	47	54	84	35
	81	78	56	66	71	83	47
	43	55	49	44	54	49	34
	58	67	42	56	66	68	35
	71	75	50	55	70	66	41
	72	82	72	67	71	83	31
0	67	61	45	47	62	80	41

	Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
•	Boruta perform 4 attributes rating; 4 attributes	ned 17 iterati confirmed imp	e model ions in 0.1698139 portant: complain important: advance /	secs. ts, learning,				

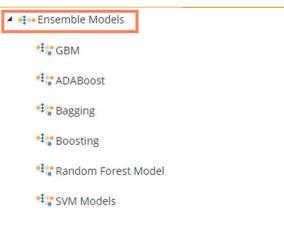
ix) Click the '**Visualization**' tab to see the processed data through a chart having a candlestick pattern.



# 13.1.9. Ensemble Models

The Ensemble Models tree node contains the following algorithms:





## 13.1.9.1. GBM

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component as given below:
  - a. Dependent Column
  - b. Independent Column
  - c. Distribution
  - d. Quantile Value (0-1)
  - e. Iterations
  - f. Depth (1-50)
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visu	ualization	Properties			*	<u>+</u>
General	Dynam	nic Fields								
Custom Group		Dependent Co	umn		sex		~			
		Independent C	olumn		8 checked		•			
		Distribution			quantile		~			
		Quantile Value	(0 - 1)		0.50					
		Iterations			10					
		Depth (1 - 50)			2					
r										
									Appl	ly

v) Run the workflow after getting the success message.



vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

ow	10 🗸 en	tries						Sear	ch:
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues_GB
Л	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	2.00995511979025
A	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	2.00995511979025
	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1.99699904087819
N	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	2.00995511979025
	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	2.0009870578642
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	2.0009870578642
	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	1.99699904087819
	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	1.99699904087819
Л	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	2.00995511979025
	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	1.99699904087819

viii) Click the 'Summary' tab to open the model summary.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
<pre>gbm(formula =</pre>	data_1[, c(d lpha), data = = 0.001, tra: osted model w: were perform s-validation : -set iteration predictors of	<pre>data_1, n.trees = in.fraction = 0.5, ith quantile loss ed. iteration was 10. n was 10. which 1 had non-:</pre>	<ul> <li>., distribution</li> <li>NoOfTree, distribution</li> <li>cv.folds = function.</li> </ul>	ution = list(name = interaction.depth = 2, verbose = FALSE) :e.	depth,		



# 13.1.9.2. ADABoost

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

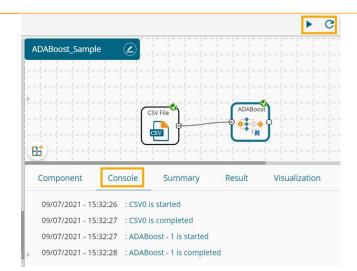


- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component as given below:
  - a. Dependent Column [Factors not more than two]
  - b. Independent Column
  - c. Loss
  - d. Type (The provided choice for this field is discrete, real, and gentle)
  - e. Iterations
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	4 †	4
General	Dynami	c Fields					
Custom Group	C	Dependent Column [		Sex	~		
	F	actors not more than	n two ]				
	1	ndependent Column		11 checked	•		
	L	oss		exponential	~		
	1	уре		discrete	~		
	1	terations		10			
						_	
						Арр	oly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

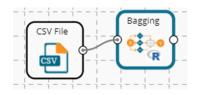
ow 10 🗸	entries											Search:	
Passengerid	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	PredictedValues_ADABoost	
	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	C92	s	male	
	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	с	female	
	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/02. 3101282	7.925		s	female	
	1	1	Futrelle, Mrs. Jacques Heath (Lify May Peel)	female	35	1	0	113803	53.1	C123	s	female	
	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		s	male	
	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q	male	
	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	s	male	
	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		s	male	
	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		s	female	
0	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		с	female	

(	Component	Console	Summary	Result	Visualization	Properties	*	<u>+</u>
	Call:	Summary of the	e model					
×		dependentColu eration, type	umn)] ~ ., data = = type)	data1, loss	= loss,			
			liscrete Iterati	ion: 10				
	Fir	on Matrix for	Data:					
		male male 314 0 0 577						
	Train Error:	0						
	Out-Of-Bag Er	ror: 0 itera	ation= 6					
			ber of iterations	:				
	train.err1 tr 1	ain.kap1 1						
ŀ		End of Summary	,					



## 13.1.9.3. Bagging

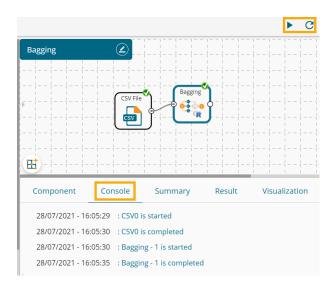
- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component as given below:
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the analysis gets applied to the target column.
  - c. Iterations: The number of passes over the training data (aka epochs)
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ŧ
General	Dynam	ic Fields					
Custom Group		Dependent Column		sex	~		
		Independent Colum	ı	8 checked	-		
	1	Iterations		10			
r							
						Арр	oly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

how	10 v ent	tries						Se	earch:
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues_Bagging
м	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	F
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	М
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	F
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	F
1	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	L
1	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	М
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	F
М	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	F
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F

	Component Console Summary Result Visualization Properties	÷	<u>1</u>
	Summary of the model \$formula dependent ~ .		
	\$trees \$trees[1]] n= 100		
	node), split, n, loss, yval, (yprob) * denotes terminal node		
	<ol> <li>root 100 49 F (0.51000000 0.11000000 0.38000000)</li> <li>weight_whole&gt;=0.5165 73 24 F (0.67123288 0.02739726 0.30136986) *</li> <li>weight_whole</li> <li>0.5165 27 11 M (0.07407407 0.3333333 0.59259259) *</li> </ol>		
	\$trees[[2]] n= 100		
	node), split, n, loss, yval, (yprob) * denotes terminal node		
	1) root 100 55 F (0.45000000 0.13000000 0.4200000) 2) length>-0.34 91 46 F (0.49450549 0.04395604 0.46153846) * 3) length< 0.34 9 0 I (0.00000000 0.000000000 0.00000000) *		
•	\$tres[[3]] n= 100		
	node), split, n, loss, yval, (yprob) * denotes terminal node		
	1) root 100 49 M (0.39000000 0.10000000 0.51000000) ) wither functorial 0.0015 0 0 T (0.00000000 1.00000000 0.000000000) *		

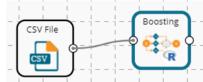
ix) Click the 'Visualization' tab to open the processed data through a chart.





#### 13.1.9.4. Boosting

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

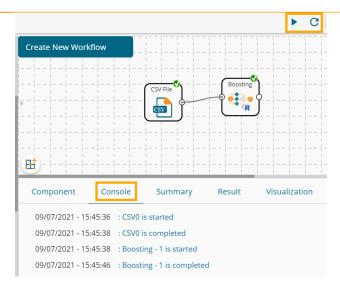


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the analysis gets applied to the target column.
  - c. Iterations: The number of passes over the training data (aka epochs)
  - d. Coefficient Learn
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Dynami	: Fields					
Custom Group	D	ependent Column		sex	~		
	Ir	ndependent Columr	1	8 checked	-		
>	lt	erations		10			
	C	pefficient Learn		Breiman	~		
						Appl	у

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

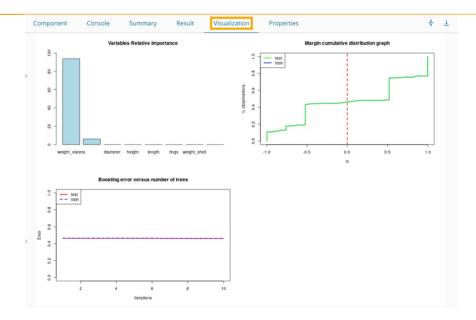
now	10 🗸 ent	tries						9	Search:
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues_Boosting
4	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	1
1	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1
	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	м
4	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	1
	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1
	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	м
	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	м
1	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	1
	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	м

viii) Click the 'Summary' tab to open the model summary.

	Component Console	Summary	Result	Visualization	Properties	+	<u>+</u>
Þ	Summary of th \$formula dependent ∼ .	ne model					
	<pre>\$trees \$trees[[1]] n= 4177</pre>						
	node), split, n, loss, yva * denotes terminal n						
	1) root 4177 2662 M (0.300 2) weight_viscera>=0.139 3) weight_viscera< 0.139	525 2538 1351 M (0	.4164697 0.1				
	\$trees[[2]] n= 4177						
Þ	node), split, n, loss, yva * denotes terminal n						
	1) root 4177 2681 M (0.324 2) weight_viscera>=0.145 3) weight_viscera< 0.145	925 2416 1320 M (0	.4457781 0.10				

ix) Click the 'Visualization' tab to see the result data through a chart.





## 13.1.9.5. Random Forest Model

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

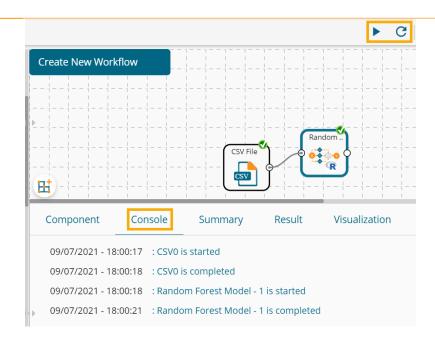


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the analysis gets applied to the target column.
  - c. Number of Tree: Provide a digit to set the number of trees.
- iv) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	<u>1</u>	Ť
General	Dynamic	Fields					
Custom Group	Dep	pendent Column		sex	~		
•	Ine	pendent Column		8 checked	•		
	Nu	mber of Tree		500			
						Арр	ıly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





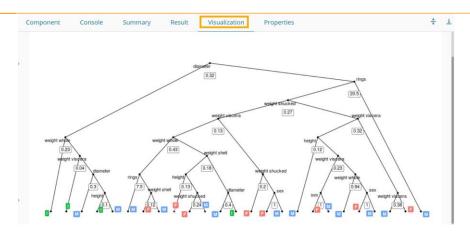
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

ow	10 🗸 entri	25							Search:
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues_Random_Forest
Л	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	M
N	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	M
	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	F
Л	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	M
	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1
	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F
	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	F
4	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	M
	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F

omponent	Console	Summary	Result	Visualization	Properties	Ť
	Summary of th	e model				
Call:						
	t(formula - da	to 1[ c(dependent	tColump)] .	data - data 1	ntree = numbers tree, replace = TRUE, na.a	ction - no foil)
randomrores		idom forest: class		., uata - uata_i,	ficiee - fumbers_cree, reprace - fkbe, fa.a	ccion - na.raii)
		r of trees: 500	111Cación			
No. of varia	bles tried at					
		coon oparer o				
OOB	estimate of e	error rate: 2%				
Confusion ma	trix:					
FIMC	lass.error					
F44 0 0	0.0000000					
I 0 10 2	0.1666667					
M 0 0 44	0.0000000					
	End of Summar					

ix) Click the 'Visualization' tab to see the result data through a chart.





#### 13.1.9.6. SVM Models

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

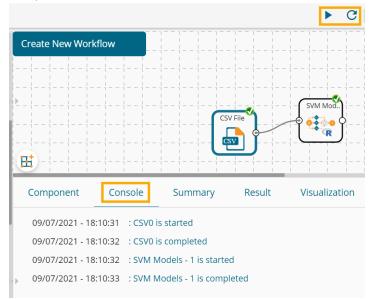


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component as given below:
  - a. Dependent Column
  - b. Independent Column
  - c. Type: This field contains the following options:
    - i. C-classification
    - ii. nu-classification
    - iii. eps-regression
    - iv. nu-regression
  - d. Kernel: This field contains the following options:
    - i. Linear
    - ii. polynomial
    - iii. radial
    - iv. sigmoid
  - e. Degree Polynomial
  - f. Gamma Factor
  - g. Cost Factor
  - h. Nu Factor (0.1-0.9)
  - i. Cross Validation
- iv) Click the 'Apply' option.



Component	Console	Summary	Resu	lt Visualization	Propert	ties	+	<u>+</u>
General	Dynamic	Fields						
Custom Group	Dej	pendent Column		sex	~			
	Ind	ependent Column		8 checked	•			
	Тур	e		C-classification	~			
	Ker	mel		radial	~			
	Deg	gree Polynomial		3				
	Gai	mma Factor		0.5				
	Cos	st Factor		1				
<i>.</i>	Nu	Factor (0.1 - 0.9)		0.5				
	Cro	oss Validation		1				
							Appi	y

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

ow	10 🗸 entries	5					Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings
N	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
A	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
4	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10
	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8
	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20
	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16
4	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9
	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19

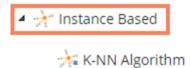


Component	Console	Summary	Result	Visualization	Properties	+ +	
	Summary of th	e data					
sex	length	diameter	height	weight_whole			
F:44 Min.	:0.2050 M	lin. :0.1500	Min. :0.0450	Min. :0.0420			
I:12 1st (	Qu.:0.4500 1	st Qu.:0.3538	1st Qu.:0.1050	1st Qu.:0.4581			
M:44 Media	an :0.5225 M	Nedian :0.4075	Median :0.1350	Median :0.6707			
			Mean :0.1313				
	•			3rd Qu.:0.9524			
Max.			Max. :0.2000				
		viscera weigh		rings			
Min. :0.0		0.01250 Min.		: 4.00			
1st Qu.:0.1 Median :0.2		0.09538 1st Qu 0.14850 Median		Qu.: 8.00 an :10.00			
Mean :0.2				:10.66			
3rd Qu.:0.3				)u.:13.00			
Max. :0.8	-	-		:21.00			
indice for the		0112250 110/01					
	End of Summar	v					

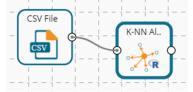
#### 13.1.10. Instance Based

The Instance-based algorithms also known as memory-based algorithms compare new problem instances with the instances already seen in training and have been stored in memory. These algorithms are sometimes referred to as 'Lazy' computation gets postponed until a new instance is observed.

The Data Science Workbench provides the K-Nearest Neighbor under the R Workspace as displayed in the below-given image:



- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

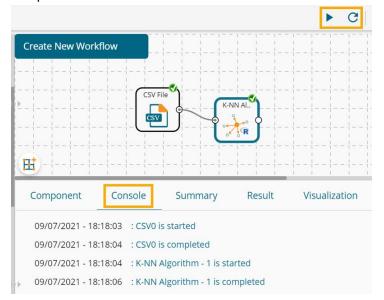


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Dynam	ic Fields					
Custom Group		Dependent Column		sex	~		
		independent Columr	ı	8 checked	•		
		Validation Method		boot	~		
		Number of Folds/ Ite	rations/	3			
		Resamples					
		Seed		10			
		Tuning Parameter		10			
•							
						Apply	y

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



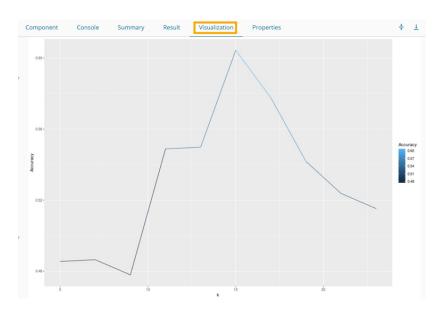
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

ow	10 🗸 entrie	s							Search:
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues_K_NN
	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	М
	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	М
	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	F
	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	M
	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	M
	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F
	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	F
	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	М
	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F



	Compo	nent	Console	Summary	Result	Visualization	Properties	*	1
*	k-Nea 100 s 8 p	Sum rest Neighb amples redictor lasses: 'F'	pors	model					
	Resam	pling: Boot ry of sampl	strapped (3 le sizes: 10						L
*	5 7 9 11 13 15	Accuracy 0.4854987 0.4864733 0.4779385 0.5488383 0.5498261 0.6043807 0.5775776	0.1218291 0.1263040 0.1051273 0.2276930 0.2384848 0.3242221						
	19 21	0.5775776 0.5417654 0.5237474 0.5152258	0.1998356 0.1733571						

ix) Click the 'Visualization' tab to open the result data through a chart.



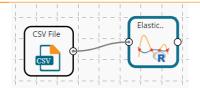
#### 13.1.11. Regularization

The Regularization tree node contains the Elastic Net algorithm as shown in the following image:



- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.

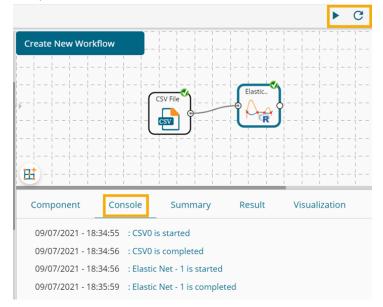




- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Dependent Column
  - b. Independent Column
  - c. Lambda
  - d. Alpha
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ŧ
General	Dynamic	: Fields					
Custom Group	D	ependent Column		sex	~		
	Ir	dependent Column		8 checked	•		
>	Li	ambda		10			
	A	lpha		10			
						Аррі	v
						Appl	y

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



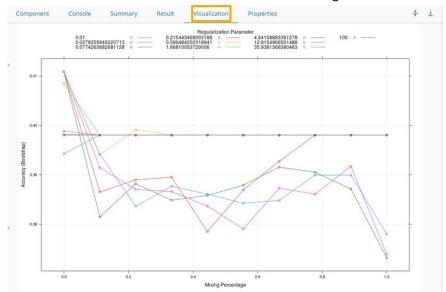
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.



how	10 🗸 en	tries							Search:
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues_Elastic_Ne
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	м
м	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	м
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	F
м	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	м
1	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1
	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	F
м	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	м
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F

Component	Console	Summa	ary Result	t Visualizat	ion Properti	es	4	÷	<u>+</u>
glmnet	Summary of th	e model							
100 samples 8 predicto 3 classes:	r 'F', 'I', 'M'								
Summary of s	Bootstrapped (	00, 100, 10	0, 100, 100, 10 meters:	0,					
alpha	lambda	Accuracy	Карра						
0.0000000	0.01000000	0.4108882	-0.0069577786						
0.0000000	0.02782559 0.07742637		-0.0069577786 -0.0069577786						
0.0000000	0.21544347		-0.0069577786						
0.0000000	0.59948425		-0.0069577786						
0.0000000	1.66810054		-0.0101689784						
0.0000000	4.64158883	0.3988589	-0.0161730586						
0.0000000	12.91549665	0.3942779	-0.0102017073						
0.0000000	35,93813664	0.3981550	0.0006112151						
0.0000000	100.00000000		0.0000000000						
0.1111111	0.01000000		-0.0307307000						
0.1111111	0.02782559		-0.0355612090						
0.1111111	0.07742637	0.3814413	-0.0491611678						

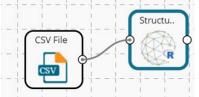
ix) Click the 'Visualization' tab to see the result data through a chart.





# 13.1.12. Structural Equation Modelling

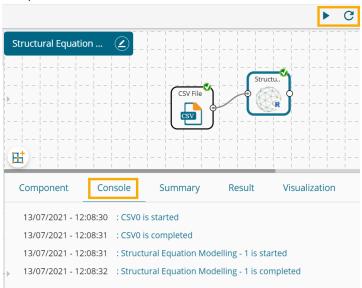
- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the model field provided under the '**Dynamic Fields**' of the dragged algorithm component.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Dynamic	Fields					
Custom Group	Mc	del		mValues=~ rating + cor rms =~ privileges + lear			
>							
						Apply	,

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.



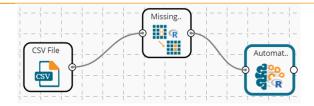
ow [	10 🗸 entri	es				Search:	
id	rating	complaints	privileges	learning	raises	critical	advance
	43	51	30	39	61	92	45
	63	64	51	54	63	73	47
	71	70	68	69	76	86	48
	61	63	45	47	54	84	35
	81	78	56	66	71	83	47
	43	55	49	44	54	49	34
	58	67	42	56	66	68	35
	71	75	50	55	70	66	41
	72	82	72	67	71	83	31
0	67	61	45	47	62	80	41

omponent	C	onsole	Sum	mary	Re	sult	Visi	ualizatio	n	Propertie
		nary of the	model -							
\$ Parameter						÷.				
	is op		est					std.all		
		rating					10.561		0.882	
		complaints							0.935	
		privileges						0.646		
		learning								
		rating								
6 complaint										
7 privilege										
		learning semValues								
10 norm		norms								
II semvalue	5	norms	70.983	28.023	1 2.4/9	0.015	0.800	0.805	0.800	
\$`Fit Measur										
		ar		fmin			hisq		d	
	9.6			.841			.462		1.00	
	pval		seline.c		1	aselin		baseli		
	0.1		63	.859			.000		0.00	
		:fi		tli			nnfi		rf	
	0.5			.848			.848		0.76	
		ıfi		pnfi			ifi		rn.	
	0.5			.160			.977		8.97	
		ogl unrest					aic		bi	
	439.1			.963		896			908.99	
	ntot	10 T -		bic2			nsea	rmsea.	.ci.lowe	
	30.6			.978			.221		0.00	
rmsea.c	1.upp 0.5		msea.pv				nar	13	tr_nomea	
	1.200			.130			.075		4.07	
			srmr_ber		rmr_bent				crm	
	0.6			.028	50mm m		.028		0.03	
Crmi	_nome 0.6			0.028	srar_n		mean .028		cn_0 47.81	
	cn_ 81.8			gfi .962			agfi .621		pgf. 0.09	
		nfi		ecvi		0	.021		0.09	0
	0.5			.682						
\$ Modificati		ndices' rhs	mi	epr	sect.la	sept	all sen	c.nox		
20 complaint										
		learning								
19 complaint										
		privileges								
	End	of Summary								
	EUG	or sumdry								

# 13.1.13. Automated ML

- i) Drag the algorithm component to the workspace and connect it with a configured data source (In this case the workflow contains a Missing Value Replacement component to prepare the selected input data).
- ii) Click the algorithm component.

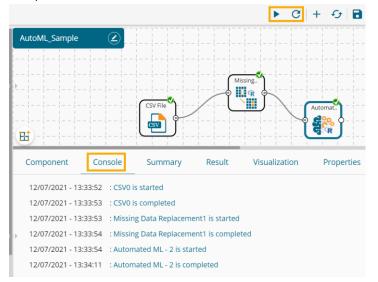




- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Propert	ies	<del>*</del>	<u>+</u>
General	Dynamic	Fields						
Custom Group	De	pendent Column		Sex	~			
	See	ed		10				
	Ma	ximum Models to Ru	n	5				
	N-F	olds		5				
	Ma	ximum Run Time		30				
	Sec	onds (1-3600)						
•								
							Apply	/

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

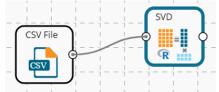


Show 10 🗸	entries										Search:	
Passengerid	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	predicte
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	C92	s	male
2	1	1	Curnings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	с	female
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/02. 3101282	7.925	B96 B98	s	male
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S	female
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05	B96 B98	s	male
6	0	3	Moran, Mr. James	male	29.6991176470588	0	0	330877	8.4583	B96 B98	Q	male
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S	male
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	B96 B98	s	female
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	B96 B98	S	female
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	B96 B98	c	female

	Component Console Summary Result Visualization Properties	+	<u>+</u>
	Summary of the model		
Þ	[[1]] Model Details:		
	H2OBinomialModel: xgbocst Model ID: XGBocst_3_AutcML_20210730_111718 Model Summary: number of trees		ľ
	numcer_or_trees 1 77		
	H2OBinomialMetrics: xgboost ** Reported on training data. **		
	MSE: 0.09787179 RMSE: 0.3128447		
	LogLoss: 0.3324787 Mean Per-Class Error: 0.127441		
	AUC: 0.9366203 pr_auc: 0.9528376 Gini: 0.8732407		
Þ	Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold: female male Error Rate		
	female 259 55 0.175159 =55/314 male 46 531 0.079723 =46/577		
	Totals 305 586 0.113356 =101/891		
	Maximum Metrics: Maximum metrics at their respective thresholds		

#### 13.1.14. SVD

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



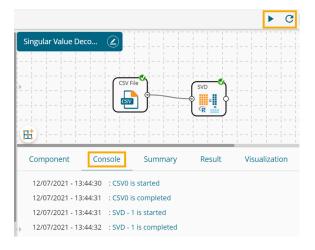
- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Dynamic	Fields					
Custom Group							
						Арр	oly

v) Run the workflow after getting the success message.



vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

iow [	10 🗸 entrie	es				Search:	
id	rating	complaints	privileges	learning	raises	critical	advance
1	43	51	30	39	61	92	45
2	63	64	51	54	63	73	47
3	71	70	68	69	76	86	48
1	61	63	45	47	54	84	35
5	81	78	56	66	71	83	47
5	43	55	49	44	54	49	34
7	58	67	42	56	66	68	35
3	71	75	50	55	70	66	41
)	72	82	72	67	71	83	31
10	67	61	45	47	62	80	41

#### viii) Click the 'Summary' tab to open the model summary.

Component	Console Su	mmary Result	Visualization	Properties	<u>+</u> <u>↓</u>
S	Summary of the model V2	V3	V4		
Min. :0.000 1st Ou.:0.092			Min. :0.000100 1st Qu.:0.009075		
Median :0.319	2 Median :0.09835	Median :0.14445	Median :0.072450		
Mean :0.352 3rd Qu.:0.605			Mean :0.120792 3rd Qu.:0.132300		
Max. :0.955 V5	8 Max. :0.67340 V6	Max. :0.54260 V7	Max. :0.704300 V8		
Min. :0.000 1st Qu.:0.004					
Median :0.021	.550 Median :0.023	450 Median :0.0167	0 Median :0		
Mean :0.081 3rd Qu.:0.121					
Max. :0.541	.600 Max. :0.384 nd of Summary		0 Max. :0		
	ind of Sumilary				



## 13.1.15. WLS Relative Std

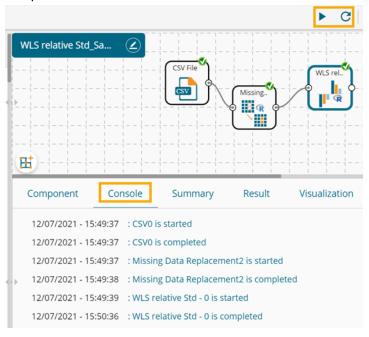
- i) Drag the algorithm component to the workspace and connect it with a configured data source (In this case the Missing Value Replacement component is connected to prepare the input data).
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column.
- iv) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> T
General	Dynamic F	Fields				
Custom Group	Dep	endent Column	Age		~	
>	Inde	pendent Column	11 c	hecked	•	
						Apply

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



vii) Follow the below-given steps to display the Result view:



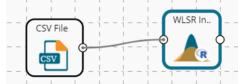
- a. Click the dragged algorithm component onto the workspace.
- b. Click the 'Result' tab.

ow 10 🗸	entries										Sea	rch:
Passengerid	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	PredictedValues_WLS
	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	C92	s	22
	1	1	Cumings. Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	с	37.99999999999999
	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/02. 3101282	7.925	B96 B98	s	26
	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	s	35
	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05	B96 B98	s	35
	0	3	Moran. Mr. James	male	29.6991176470588	0	0	330877	8.4583	B96 B98	Q	29.6991176470588
	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	s	54
	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075	B96 B98	s	2
	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333	B96 B98	s	27
0	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708	B96 B98	с	14

	Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
*	call:	· Summary of the	model					
			dentColumn] ~ .,	data = data_	1, weights = wts)			
	coerricience					(Intercept)		
						-3.358e-16		
						Age 1.000e+00		
						PassengerId		
						-1.610e-17		
						Survived		
						-4.620e-15		
						Pclass		
						-1.693e-15		
				N	lameAbbott, Mr. Ros			
				Namath	bott, Mrs. Stantor	1.381e-14		
				NaileAL	bott, Mis. Stanto	-8.269e-16		
Þ					NameAbelsor	, Mr. Samuel		
						-4.053e-15		
				NameAbelson	, Mrs. Samuel (Har			
						1.384e-14		
				Name	Adahl, Mr. Mauritz	Nils Martin		
						-1.442e-15		

## 13.1.16. WLSR Input Weights

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column.



- c. Weighted Column: Select a column from the input dataset.
- d. Weighted Type: Select a Weighted Type out of the given options that are Square Response and Normal Response.
- iv) Click the '**Apply**' option.

Component	Console	Summary	Re	esult	Visualization		Properties	<u>+</u>	Ť
General	Dynamic Fi	elds							
Custom Group	Deper	ndent Column		learning		~			
	Indep	endent Column		7 checked	ł	•			
*	Weigh	nted Column		rating		~			
	Weigh	nted Type		Square R	esponse	~			
								Арр	ly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.

				► C +
WLSR input Weights		 		
		File	WLSR In	
	nsole	Summary	Result	Visualization
12/07/2021 - 16:08:24	: CSV0 is	started		
12/07/2021 - 16:08:25	: CSV0 is	s completed		
12/07/2021 - 16:08:25	: WLSR I	nput Weights - 1	is started	
12/07/2021 - 16:08:26	: WLSR I	nput Weights - 1	is completed	

- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

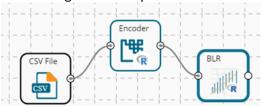
10 v entries Search:												
id	rating	complaints	privileges	learning	raises	critical	advance	PredictedValues_WLSR				
1	43	51	30	39	61	92	45	39				
2	63	64	51	54	63	73	47	54				
3	71	70	68	69	76	86	48	69				
4	61	63	45	47	54	84	35	47				
5	81	78	56	66	71	83	47	66				
6	43	55	49	44	54	49	34	44				
7	58	67	42	56	66	68	35	56				
8	71	75	50	55	70	66	41	55				
9	72	82	72	67	71	83	31	67				
10	67	61	45	47	62	80	41	47				



Component	Console	Summary	Resu	lt Visua	lization	Properties	<del>*</del>	<u>+</u>
Call: lm(formula = Coefficients: (Intercept) -1.970e-14 raises 1.590e-17	data_1[, depe learning 1.000e+00 critical	-6.160e-17 advance 2.397e-17	., data = o rating	complaints	privileges			

#### 13.1.17. BLR

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. **Dependent Column**: Select the target column on which the regression analysis gets applied.
  - b. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column.
- iv) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> T
General	Dynamic					
Custom Group	Dep	endent Column	MAI	NTAINER_Engineering	~	
	Inde	ependent Column	2 ch	ecked	•	
•						
						Apply

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





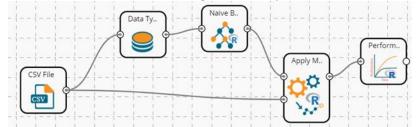
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

how 10 🔹	<ul> <li>entries</li> </ul>				Sear	rch:
LATITUDE	LONGITUDE	NAME	MAINTAINER	MAINTAINER_Engineering	MAINTAINER_Parks	PredictedValues_BL
49.234947	-123.027254	Fountain locationAberdeen	Parks	0	1	0.191841996194677
49.275858	-123.024141	Fountain locationAdanac	Parks	0	1	0.192221046580136
49.285554	-123.142392	Fountain locationAlexandria	Parks	0	1	0.193754226052746
49.26098	-123.184759	Fountain locationAlmond	Parks	0	1	0.194017775250904
49.278811	-123.106194	Fountain locationAndy Livingstone	Parks	0	1	0.193246442028759
49.277951	-123.103575	Fountain locationAndy Livingstone	Parks	0	1	0.193205911454429
49.262128	-123.155318	Fountain locationArbutus Greenway	Parks	0	1	0.193672349634539
49.248679	-123.157991	Fountain locationArbutus Village	Parks	0	1	0.193567751622521
49.285919	-123.120939	Fountain locationArt Phillips Park	Parks	0	1	0.193497718129018
49.245614	-123.175342	Fountain locationBalaclava	Parks	0	1	0.193746993688921

Note: No Summary is available for this component node.

# 13.1.18. Performance Metrics

- i) Drag the Performance metrics component and connect it to the workflow containing an Apply Model component as displayed in the image given below.
- ii) Click the Performance Metrics algorithm component.

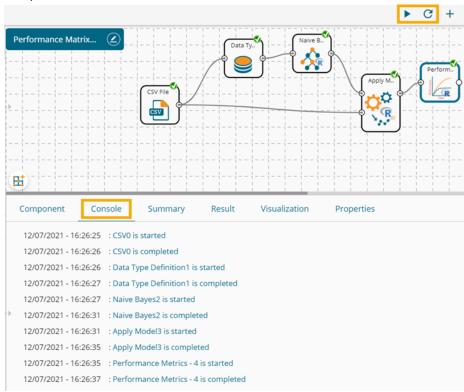


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Actual Value: Select a column from the input data to find the actual value.
  - b. Predicted Value: Select a column from the input data to get the predicted value.
  - c. Type: Select a type out of Classification or Regression options.
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Dynamic	Fields					
Custom Group	Act	ual Value		sex	~		
	Pre	dicted Value		PredictedValues2	~		
Þ	Тур	e		classification	~		
						Apply	,

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

now	10 🗸 ent	ries							Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues2	Probability2
м	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	1	[0.0139,0.9395,0.046
м	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	1	[0,1,0]
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	1	[0.1687,0.5336,0.297
м	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	1	[0.0019,0.9881,0.01]
	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	1	[0,1,0]
1	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	1	[0,0.9999,1e-04]
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	F	[0.5849,0,0.4151]
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	F	[0.5148,0.004,0.4812]
м	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	1	[0.0029,0.9849,0.012]
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	F	[0.6375,0.0.3625]



viii) Click the 'Summary' tab to open the model summary.

Component Conso	le S	Summary	Result	Visualization	Properties	÷
Summary c Confusion Matrix and S						
Reference						
Prediction F I	м					
F 851 261						
I 120 1088						
M 903 388	237					
Overall Statistics						
	y : 0.520					
		57, 0.5362)				
No Information Rat						
P-Value [Acc > NIF	[] : < 2.2	e-16				
Kapp	a : 0.291	8				
Mcneman's Test P-Valu	ie : < 2.2	e-16				
Statistics by Class:						
(	lass: F C	lass: I Cla	is: M			
Sensitivity	0.4541	0.6264 0.4	1873			
	0.8020	0.8959 0.0	54248			
		0.8107 0.1				
	0.6436	0.7711 0.8	37580			
Prevalence		0.4158 0.1				
		0.2605 0.0				
Detection Prevalence						
Balanced Accuracy		0.7611 0.9	53060			
End of Su	mmary					

# 13.2. Apply Model

This component is provided to generate predictions based on the trained model. The user can view the predicted column value and probability of each label class by using the Apply Model component.

The user can create a model in the following ways:

- Generate a model using an algorithm
- Generate a model using the saved models

The Apply Model consists of 2 input nodes and 1 output node.

- Input Nodes
  - Upper node Model/Training data
  - Lower node Testing data
- Output Node
  - Node Result data

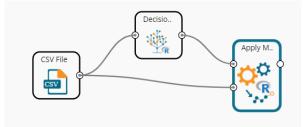
The Apply Model component provided under R, Python, and Spark can be configured using the same set of steps within the respected Workbenches, so this component is only described for R.

- i) Drag the 'Apply Model' component to the workspace.
- ii) The Apply Model has two input components and one output component.



BBB@	
Data Science Workben	nch 🥷
E C Search Tree	Q Create New Workflow
▶ ∑ Statistical Analysis	• •
🕨 🙀 Data Preparation	
🕨 🧖 Algorithms	Apply M.
😤 Apply Model	
Performance	
🖻 🚢 Data Writer	

- iii) Connect the Apply Model component with a valid combination of Data source and algorithm (Configure the data source and algorithm components. In this case, the used algorithm is Decision Tree.)
- iv) Click the 'Apply Model' component.



- v) Basic component details get displayed.
  - a. Component Name: It displays the predefined name of the component
  - b. Alias Name: It displays a predefined name that suggests the component's position in the workflow
- vi) Click the 'Apply' option.

Component C	onsole	Summary	Result	Visualization	Properti	ies	Ŧ	Ť
General	Basic							
	Compone	nt Name	R App	ly Model				
>	Alias		Apply	Model2				
	Descriptio	n	Option	al				
							Apply	

Note: The number given to the Apply Model signifies its place in the workflow. E.g., R Apply Model2 in the below-given image suggests that it is in the third position in the workflow.

- vii) Run the workflow.
- viii) The '**Console**' tab opens displaying the progress of the process. Completion of the console process gets marked by the green checkmarks on the top of the dragged components.



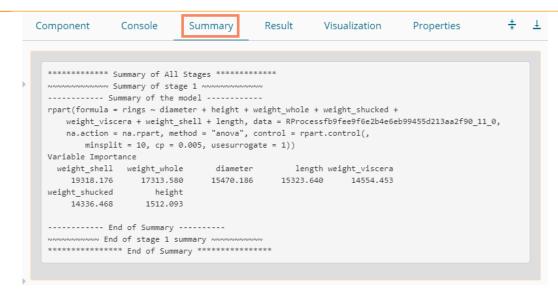
		- 1 1 1 1	• • • • • • • • •	C + 4
Apply Model WF				
			Decisio	
				Apply M.
	CSV File			
		<b>J</b>		
H				
Component	Console	Summary	Result	Visualization
Component	Console 5:14:52 : CSV1		Result	Visualization
Component	5:14:52 : CSV1		Result	Visualization
Component 03/10/2019 - 1	5:14:52 : CSV1 5:14:53 : CSV1	is started	Result	Visualization
Component 03/10/2019 - 1 03/10/2019 - 1 03/10/2019 - 1	5:14:52 : CSV1 5:14:53 : CSV1 5:14:54 : CNR	is started is completed		Visualization
Component 03/10/2019 - 1 03/10/2019 - 1 03/10/2019 - 1	5:14:52 : CSV1 5:14:53 : CSV1 5:14:54 : CNR 5:14:55 : CNR	l is started l is completed Tree0 is started	4	Visualization

- ix) Follow the below-given steps to display the Result view:
  - a. Click the dragged R Apply Model component on the workspace.
  - b. Click the 'Result' tab.

Comp	onent	Console	Summary	Result	Visualization Pr	operties			+ *
Show	10 • en	tries						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
М	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15	8.77060931899642
М	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7	7.5511811023622
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9	9.55357142857143
М	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10	8.77060931899642
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7	6.28395061728395
1	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8	8.77060931899642
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20	13.1603375527426
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16	12.7459016393443
М	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9	8.77060931899642
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19	13.1603375527426

x) Click the 'Summary' tab to view the model summary.





Note:

- a. The Result dataset of the model can be written to a database using a Data Writer.
- b. Column header and data type of feature column for both the saved model and testing data should match. If column headers and data types do not match, an alert message gets displayed.
- c. It is not mandatory for the testing data set to contain a label column.

### 13.3. Performance

The user can evaluate model performance through a list of parameters using the performance component. The user can use the R Performance components only for the classification algorithms.

The Performance component is provided as a leaf node under the Performance tree node. It contains 3 input nodes that can be used to compare up to 3 models. Each node has a static name like model\_0, model\_1, and model\_2. Based on the connection to the node model, the summary can be viewed with respective names.

The performance component can be of the following formats:

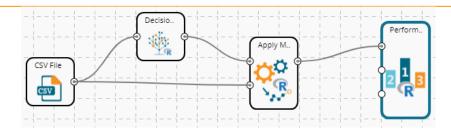
- 1. Binary Classification: Used when the label has two classes
- 2. Multi Classification: Used when the label has 3 or more beta values
- 3. Regression Metrics: Used when the regression algorithm is used in the workflow

In the case of multiple models, all the model statistics get displayed in the summary of performance (up to 3 models can be compared).

#### Steps to Connect a Performance component (to a model)

i) Drag the Performance component to the workspace and connect to a valid workflow (In this example, a workflow created with the Decision Tree algorithm has been used).





- ii) Configure the 'Properties' tab.
  - a. **Performance Type**: Select an option using the drop-down menu.
    - i. Binary Classification: Use this option when the label has two classes.
    - ii. MultiClassification Metrics(Default option): Use this option when the label has 3 or more beta values.
    - iii. Regression Metrics: Use this option when the Apply model in the workflow is trained using the Regression Algorithm.

iii) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	R-Perform	nance					
Properties	Performanc	се Туре	MultiClass	sification Metrics			
			Search				
			🗸 Mu	ltiClassification Metri	ics		
			Binary (	Classification Metrics			
			Regress	sion Metrics			
>							
						Appl	y

The user gets different outcomes based on the selected Performance types as described below:

### • Multi Classification Metrics

- 1. Navigate to the '**Properties**' tab of the R-Performance component.
- 2. Select the 'Multi-Classification Metrics' Performance type via the drop-down list.
- 3. Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	R-Performa	ince					
Properties	Performance	Туре М	ultiClassification	Metrics 👻			
•							
						Apply	

4. Run the workflow.



5. The '**Console**' tab opens, displaying steps of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



6. The user can view the summary by clicking the '**Summary**' tab (First click the performance component and then click on the '**Summary**' tab).

The following details get displayed by clicking on the 'Summary' tab:

### a. Confusion Metrix and Statistics

- i. The Confusion Matrix of each model gets displayed.
- ii. The column consists of Actual labels and the row consists of Predicted labels.

#### b. Overall Statistics

- i. Overall statistics of each model can be viewed in a tabular format
- ii. Each model displays the following statistics columns
  - 1. Accuracy
    - 2. 95% CI
    - 3. No Information Rate
    - 4. P-value
    - 5. Kappa
  - 6. Mcnemar's Test P-Value

#### c. Statistics by Class

- i. Label-wise the following statistics can be shown:
  - 1. Sensitivity
  - 2. Specificity
  - 3. Pos Pred Value



- 4. Neg Pred Value
- 5. Prevalence
- 6. Detection Rate
- 7. Detection Prevalence
- 8. Balanced Accuracy

Component	Console	Summary	Result	Visualization	Properties	*	Ţ
Overall Stat	istics						
	Accuracy : 0 95% CI : (0 mation Rate : 0. [Acc > NIR] : 1						
	Kappa : 0						
Mcnemar's T	est P-Value : NA	4					
Statistics b	y Class:						
	Class:	8.77060931899642	Class: 7.551	1811023622			
Sensitivity		NA		NA			
Specificity		0.8664		0.8784			
Pos Pred Val	ue	NA		NA			
Neg Pred Val	ue	NA		NA			
Prevalence		0.0000		0.0000			
Detection Ra	te	0.0000		0.0000			
Detection Pr	evalence	0.1336		0.1216			
Balanced Acc	uracy	NA		NA			
	Class:	9.55357142857143	Class: 6.283	95061728395			
Sensitivity		NA		NA			
Specificity		0.8391		0.94182			
Pos Pred Val	ue	NA		NA			
Neg Pred Val	ue	NA		NA			
Prevalence		0.0000		0.00000			
Detection Ra	te	0.0000		0.00000			
Detection Pr	evalence	0.1609		0.05818			
Balanced Acc	uracy	NA		NA			

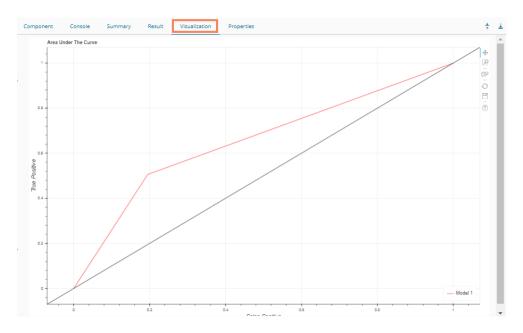
- Binary Classification Metrics
  - 1. Navigate to the '**Properties**' tab of the R-Performance component.
  - 2. Select the '**Binary Classification Metrics**' Performance type via the drop-down menu. (Select columns with binary attributes from the dataset).
  - 3. Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	Ŧ	Ť
General	R-Performa	nce					
Properties	Performance 1	Гуре	Binary Classificati	on Metrics 🔽			
Þ							
						Apply	/

- 4. Run the workflow.
- 5. The '**Console**' tab opens, displaying the steps of the process, and the completion gets marked by the green checkmarks on the top of the dragged components.



6. Click the '**Visualization**' tab to see the graphical representation of the process data (No data displays under the '**Result**' tab for the Binary Classification Metrics).



7. Click the 'Summary' tab to see the model comparison summary.

	Component	Console	Summary	Result	Visualization	Properties	*	Ŧ
		Summary of Mo	del Comparision -					
		Performance o	of first model					
Þ	Confusion Mat	trix and Stati	stics					
	0 1							
	0 442 169							
	1 107 173							
		Accuracy :						
			(0.6587, 0.7205)					
		nation Rate : [Acc > NIR] :						
	P-Value [	[ACC > NIK] :	2.5898-00					
		Kappa :	0.322					
	Mcnemar's Te	est P-Value :	0.0002409					
		Sensitivity :						
		Specificity :						
		Pred Value :						
		Pred Value : Prevalence :						
		ection Rate :						
		Prevalence :						
		ed Accuracy :						
Þ		-						
	'Posit	tive' Class :	0					
		End	-					
		End of Summar	y					
								-

- Regression Metrics
  - 1. Navigate to the '**Properties**' tab of the R-performance component.
  - 2. Select the '**Regression Metrics**' Performance Type via the drop-down menu. (Make sure that the workflow chosen for Performance check has Regression Algorithm).
  - 3. Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	R-Performa	nce					
Properties	Performance T	ype	ression Metrics	-			
Þ							
						Apply	

- 4. Run the workflow.
- 5. The console tab gets displayed with steps of the process completion. The process completion is also suggested through the green marks on the top of the dragged components.

		► C + 5 🖬
Performance_Regres		$\begin{array}{cccccccccccccccccccccccccccccccccccc$
	CSV File	Apply M. Perform.
Component Con	isole Summary Result	Visualization Properties
07/10/2019 - 11:41:28 07/10/2019 - 11:41:28 07/10/2019 - 11:41:28 07/10/2019 - 11:41:30 07/10/2019 - 11:41:30 07/10/2019 - 11:41:31 07/10/2019 - 11:41:31 07/10/2019 - 11:41:32	: CSV0 is started : CSV0 is completed : Linear Regression1 is started : Linear Regression1 is completed : Apply Model2 is started : Apply Model2 is completed : Performance3 is started	

6. Click the 'Summary' tab to view the model comparison summary.

	Component	Console	Summary	Result	Visualization	Properties	+	Ŧ
Þ		-	del Comparision f first model					
		ME	RMSE MAE 755701 0.4546393 -	MPE				
		End	-					
		End of Summar	у					

Note:

- a. In the case of multiple models, all the model statistics get displayed in the summary tab of the performance component (up to 3 models can be compared).
- b. The 'Result' tab for Binary Classification Performance (Binary Classification)



# 13.4. Custom Scripts (R Scripts)

The user can create and add customized R algorithm components by using the '**Custom Scripts**' component. The created scripts get stored in the '**Saved Scripts'** option.

# **13.4.1.** Creating a New Script

- i) Click the 'Custom Scripts' tree node from the tree menu.
- ii) Click the 'Create New Script' component.
- iii) The 'General' tab opens, displaying the Basic information for the script component.

### a. Basic

- i. **Component Name:** Enter a name or title that you wish to give a created R script.
- ii. **Component Type:** Default Component type gets displayed in this field.
- iii. **Description:** Describe the Component (It is an optional field).
- iv) Click the **'Next'** option.

> 🎽 Data Writer	Component	Console	Summary	Result	Visualization	Properties	Ŧ	Ť
💶 🔯 Custom Scripts	General	Basic						
Create New Script	Script	Component N	Name	Sample Script				
Saved Scripts	Settings	Component T	Гуре	Algorithms				
▶ 📆 Scheduler		Description		Optional				
Saved Models							Next	

- v) The 'Script' tab opens.
- vi) Provide the following information as required:

#### a. Script Editor

- i. Provide a relevant script in the given space on the 'Script Editor' page.
- ii. Click the 'Validate' option.
- iii. Configure the **'Primary Function Details'** to embed the customized script into the function.
  - 1. **Primary Function Name:** Select the name of the created function from the drop down menu.
  - 2. Input Data Frame: Select a dataset (that has been used above) from a drop-down menu.
  - 3. Output Data Frame: Enter a choice to which the data gets passed.
  - 4. **Model Variable Name:** Enter the output model variable (This field appears only when the model summary has been enabled).
- iv. If you need a Visualization chart for ensuring data, tick the 'Show Visualization' checkbox.v. If you need to show the summary, tick the 'Show Summary' checkbox.
- vii) Click the '**Next'** option.



Component	Console	Summary	Result	Visualization	Properties	*	<u>+</u>
General Script	Script Editor	·	olumn NoOfClum	ter Mayltoration coord	(j		
Settings	2 { 3 library(L; 4 library(f; 5 set.sed 6 data_1           6 data_1           7 if(ncol(d)           8 {           9 colnar           10 }           11 else           12 {           13 colnar           14 }           14 dataclus	CORS) pc) (seedValue) as.data_frame(data ata_1)!=1) nes(data_1)<-Featu nes(data_1)[1]<-Featu	[,FeatureColumn reColumn atureColumn		r.max =MaxIteration,nstart = 5)		
Þ	main € Shov	ary Inp	taFrame	Output DataFrame out	Summary Variable Name summary		
					Previous	Next	

- viii) The 'Settings' tab opens.
- ix) Configure the following fields:
  - a. Output Table Definition

The Output TableDefinition option helps to configure some output columns, column headers, and data types.

- i. **Consider all columns from the previous component:** To display all columns of the prior component.
- ii. Consider None: To display no column from the previous component.
- iii. Data Type: Select a data type for the newly created column using the drop-down list.
- iv. New Predicted Column Name: Enter an appropriate name for the new predicted column.
- v. To remove the added row containing **'Data Type'** and **'New Predicted Column** Name.'
- vi. ••• To add a new row containing 'Data Type' and 'New Predicted Column Name.'

#### b. Property View Definition

- i. Function Parameters: Actual names of parameters configured in the script.
- ii. **Property Display Name**: Parameter name to be displayed while configuring saved R script as a component.
- iii. Control Type: User can select out of the following options:
  - 1. Text box,
  - 2. Drop-down menu,
  - 3. Column Selector (single),
  - 4. Column Selector (multiple)
  - Settings option
- iv. Settings option : To set the display for mandatory fields and validate data type for the input column. This field is associated with function parameters.
- x) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties		*	<u>+</u>
General								_
Script	Funct	ion Parameter Defi	nition					
Settings		Function Parameters	Prope Name	erty Display	Control Type			
		FeatureColumn	Feat	ureColumn	Column selector	<b>v</b> 3	¢	
		NoOfCluster	NoO	fCluster	TextBox	v 3	¢	
					3			
		MaxIteration	Max	Iteration	TextBox	v 3	¢	
					30			
		seedValue	seed	Value	TextBox	•	¢	
>					5			
					Prev	vious	Apply	

- xi) A success message appears to confirm the creation of the new script.
- xii) The newly created script gets saved under the 'Saved Scripts' options.

E C Search Tree Q	Component	Console	Summary	Result		Visualization	Properties			*	<u>+</u>
	General										
Data Preparation	Script	Functi	on Parameter De	efinition							
Algorithms	Settings		Function Paramet	ers	Prope	erty Display Name	Control Type				
😤 Apply Model	Jettings		FeatureColumn		Feat	tureColumn	Column selector(№	Y	•		
📲 Performance			NoOfCluster		NoO	OfCluster	TextBox	•	•		
👂 🐣 Data Writer							3				
🔺 🕵 Custom Scripts								11			
Create New Script			MaxIteration		Max	Iteration	TextBox	•	¢		
▲ Saved Scripts	•						30	11			
😼 Sample Script			seedValue		seed	dValue	TextBox	Ŧ	٥		
<b>§</b> K-Means++_copy_1554375646						The R-script is Sav	red.				

#### **Guidelines for Writing an R- Script**

- 1. R- script needs to be written inside a valid R function. i.e., The entire code body should be inside the curly braces of the function.
- 2. The R-script should have at least one main function. Multiple functions are acceptable, and one function can call another function, but it should be written above the calling function body. (If called function is an outer function) alternatively, above the calling statement (if called function is an inner function).
- 3. Any extra packages that are required to run your R script must be installed on the R-server, and it should be loaded using the library ('library\_name') statement before calling the associated function in your script.
- 4. The R-script should return data in the form of a list only, containing the data frame and model (if used).
- 5. In the return statement, only a data frame can be assigned to the variable 'out.' This data frame supports all structures like list, string, vector, matrix, table.
- 6. If the **'Show Visualization'** field is marked as **'yes'** during the creation of the component, then there should be a plot created in the R-script, and if the **'Show Summary'** field is marked, then the structures list should have the **'model'** variable.



7. Empty cells, (NULL), (null), NULL, null, /N, NA, N/A are considered as unwanted values and replaced by "NaN" in case of double, long, short, float, byte, integer, and "NA" in case of boolean, string, so instead of using these values in R code use "NaN" or "NA" according to the data type of input data.

#### Note:

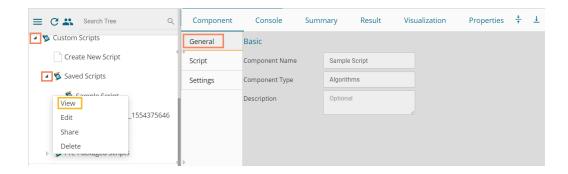
- **a.** Click the **'Information**' button <sup>1</sup> to get the list mentioned above of rules for R-script.
- b. 'Model Variable Name' can be enabled only after selecting the 'Show Summary' option.
- c. Select the 'Show Summary' and 'Show Visualization' options only if the R-script carries both items.
- **d.** All the supported date data types are listed in date formats in the data type definition; all other date formats are considered as a string data type.
- e. Mssql data types are considered string data types.
- f. If the input and output components have a different structure, it will not subset or row bind with the "Consider All" option, Users must change to "Consider None" and give different column names for the output to make it run successfully.

# 13.4.2. Saved Scripts

This section describes options that can be applied to a saved R Script.

# 13.4.2.1. Viewing a Saved R Script

- i) Select a saved Script from the list of 'Saved Scripts'
- ii) Open the context menu by using the right-click.
- iii) Select the 'View' option.
- iv) The 'General' tab opens for the selected saved script.



# 13.4.2.2. Editing a Saved R Script

- i) Select a saved Script from the list of 'Saved Scripts'
- ii) Open the context menu of the selected script by using the right-click.
- iii) Select the '**Edit**' option.
- iv) The General tab opens, displaying the Basic component information.
- v) The user can edit the required fields provided under the displayed script component tabs (General, Script, and Settings tabs).



E C Search Tree Q	Component	Console Su	mmary Result	Visualization	Properties 🕂 🕹
Custom Scripts	General	Basic			
Create New Script	Script	Component Name	Sample Script		
💶 💆 Saved Scripts	Settings	Component Type	Algorithms		
Connolo Contro View Edit _1554375646 Share Delete		Description	Optional		
▶ 🔠 Scheduler					
Saved Models					Next

Note: The '**Next**' and '**Apply**' options get displayed for the various tabs of the selected script component.

# 13.4.2.3. Sharing a Saved R Script

This feature gives users the ability to share a custom R script with other users and groups. The following options are available to share a custom R script:

- 1. **Share With**: This option allows the user to share a custom script with the selected users or user groups. Any changes made to the script get transferred to all the users with whom it has been shared.
  - i) Right-click on a saved script from the list of 'Saved Scripts'
  - ii) Select the 'Share' option from the context menu.
  - iii) The 'Share With' option gets displayed (by default)
  - iv) Select either 'Group' or 'Users'
    - a. By selecting a group, all group members inside the group get listed. The users can be excluded by not selecting them from the group.
    - b. Users can be excluded by not selecting a username from the list when the '**User**' option has been selected.
  - v) Select a specific user or group from the list by using a checkmark in the box.
  - vi) Click the 'Apply' option.

E C 🗳 Search Tree	Q Component	Console	Summary	Result	Visualization	Properties	+	Ŧ
Custom Scripts	3 Share With 4	Group	Users					
Create New Script	Copy to	Search						
▲ 📡 Saved Scripts		Select All Gro	oup				C	C
View		Admin Role				5	C	8
Edit _15543756	46	Viewer Role					<b>~</b> 0	2
2 Share Delete		UserGroup					<b>*</b> 6	8
		Kpmg Group					<b>*</b> 0	2
Scheduler		testUserGrou	up				<b>v</b> (	2
Baved Models						6	Apply	y

vii) The selected saved R script gets shared with the chosen user(s)/group(s).

2. Copy To: This option creates a copy and shares a copy of the custom R script with the selected



users and user groups. Any changes to the original custom R script after sharing will not show up for the users that received the shared file via the '**Copy To**' option.

- i) Use right-click on a saved R script from the 'Saved Scripts' list.
- ii) Select the 'Share' option from the context menu.
- iii) Select the 'Copy to' option for sharing the script.
- iv) The copied custom R script name gets displayed in a box.
- v) Select either the 'Group' or 'Users' tab.
  - a. By selecting a group, all group members inside the group get listed. Users can be excluded by not selecting them from the group.
  - b. Users can be excluded by not selecting a username from the list when the '**Users**' option has been selected.
- vi) Select a specific group or user from the list by using a checkmark in the box.
- vii) Click the 'Apply' option.

= C 🗳 Search Tree	Q Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
Saved Scripts	Share With Copy to 5	custom r script name	4 copyOfSar	nple Scrij				
View Edit y_155437	5646	, Select All Group					0	
2 Share y Delete		Admin Role Viewer Role				6	ی ۲ 0	11
▶ 📸 Scheduler		UserGroup					<b>v</b> 0	
Baved Models							Apply	

viii) The selected saved R script gets copied to the selected user(s)/group(s).

# 13.4.2.4. Deleting a Saved R Script

- i) Select a Script from the list of 'Saved R-Script'
- ii) Right-click on the selected R Script.
- iii) A context menu will open.
- iv) Select the 'Delete' option.

ſ	View	
	Edit 4-	1
	Share y	
	Delete	
P	> TTC T ucituged scriptS	

- v) A pop-up window appears to assure the deletion.
- vi) Click the '**Ok**' option.

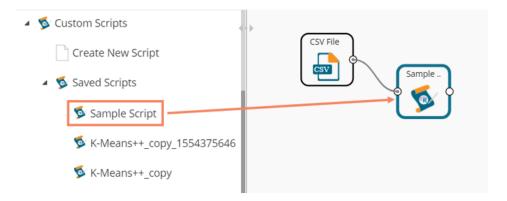


Delete Custom R Script		×
Do you want to delete Custom R Script ?		
	Ok	Cancel

vii) The selected R-Script gets deleted.

# 13.4.2.5. Connecting Saved R Script with a Data Source

- i) Click the **'Custom Script'** tree node.
- ii) Select and drag a saved R-script to the workspace.
- iii) Connect the Script component to a configured data source.
- iv) Click the dragged script component to get the configuration fields.



- v) Configure the Dynamic Fields.
- vi) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<del>+</del>	⊥
General	Dynamic Fie	elds					
Custom Group	FeatureColumn		1 checke	ed 🗸			
	NoOfCluster		3				
	MaxIteration		30				
	seedValue		5				
Þ							
						Appl	y

vii) Run the workflow after getting the success message.



viii) The console tab appears displaying steps of the process. The completion of the console process gets marked by green marks at the top of the dragged components.

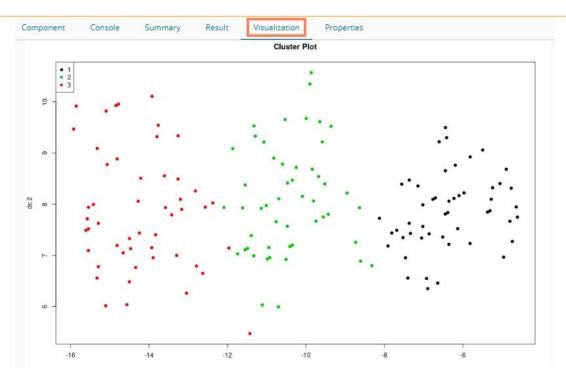
			Þ	• C + 5
Create New Wo	rkflow			
Þ	CSV File	Sample		
H				
Component	Console	Summary	Result	Visualization
09/10/2019 - 1	4:54:21 : CSV0 i	is started		
09/10/2019 - 1	4:54:23 : CSV0 i	is completed		
09/10/2019 - 1	4:54:23 : Custo	m R Script1 is star	ted	

- ix) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.

Compon	ient Console	Summary Re	sult Visualizat	ion Propertie	S	<u>+</u> <u>⊥</u>
Show 1	0 • entries				Search:	
Numbe	er SepalLength	SepalWidth	PetalLength	PetalWidth	Species	ClusterNumber
1	5.1	3.5	1.4	0.2	setosa	1
2	4.9	3	1.4	0.2	setosa	1
3	4.7	3.2	1.3	0.2	setosa	1
4	4.6	3.1	1.5	0.2	setosa	1
5	5	3.6	1.4	0.2	setosa	1
6	5.4	3.9	1.7	0.4	setosa	1
7	4.6	3.4	1.4	0.3	setosa	1
8	5	3.4	1.5	0.2	setosa	1
9	4.4	2.9	1.4	0.2	setosa	1
▶ 10	4.9	3.1	1.5	0.1	setosa	1
Showing 1	to 10 of 150 entries			Previous 1	2 3 4	5 15 Next

x) Click the **'Visualization'** tab to see the result data presented through the Cluster Plot chart.



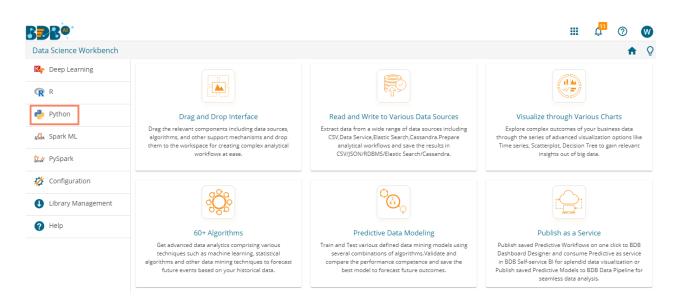


#### Note:

a. The above-given process is displayed for a CSV data source. A similar set of steps can be followed for other data source types.

# 14. Python Workspace

The user can select the Python Workspace from the Predictive landing page to access the Python Environment under the Data Science Workbench.



The following screen opens loading the Python Workbench



BBB®					۹.	I 🗜 🖓	A
Data Science Workbench	÷	The Python Workbe	nch is loaded.	► C	2 + 9 8	a ≄ (← C)	Q
≡ C 👪 Search Tree Q	Create New Workflow		1 1				
🐐 🛄 Saved Workflows							
🕞 Data Source							
Σ Statistical Analysis							
🎲 Data Preparation							
Q Algorithms							
😤 Apply Model							
😼 Performance							
🐣 Data Writer							
😼 Custom Scripts							
Jupyter Notebooks	⊞						
Scheduler	Component Console S	Summary Result	Visualization	Properties [	DataInsight	+ +	Ť

# 14.1. Algorithms

### 14.1.1. Forecasting

The forecasting modeling method is used extensively in time series analysis to predict a response variable, such as monthly profits, stock performance, or unemployment figures, for a specified period. Forecasts are based on patterns in existing data. For example, a warehouse manager can create a model of how much product to order for the next three months based on the previous 12 months of orders.

All the sub-categories of the Forecasting Algorithms provide two Output modes (to be set from the Properties tab):

- 1. Forecasting
- 2. Trend

The document describes all the available Forecasting algorithms considering both the output modes as possibilities.

### 14.1.1.1. SARIMAX

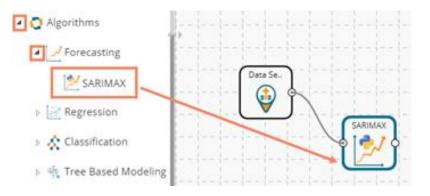
Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors Variables X (SARIMAX) is an extension of SARIMA (Seasonal ARIMA) and ARIMA model that explicitly supports univariate time series data with a seasonal component along with the inclusion of exogenous variables X.

It adds three new hyperparameters to specify the autoregression (AR), differencing (I), and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

www.bdb.ai



- i) Drag the SARIMAX component to the workspace and connect it to a configured data source.
- ii) Click on the dragged SARIMAX component to get the component properties fields.



- iii) The user gets Properties fields based on the selected Output Mode (Forecast/ Trend)
  - Properties with Forecast Output Mode
    - a. Output Information
      - i. Output Mode: Select a mode in which you want to display output data. The user gets two options for this field.
        - 1. **Trend**: Selecting this option displays source data along with predicted values for the given data set.
        - 2. Forecast: Selecting this option displays forecasted values for the given period. Results data gets appended to the target column when 'Forecast' output mode has been selected.
      - ii. **Period to Forecast**: Enter a period to forecast. This field appears only when the selected **'Output Mode'** option is **'Forecast.'**
    - b. Column Selection
      - i. Feature: Select the feature columns using the drop-down menu.
      - ii. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (The first selected option gets selected by default. Only numerical columns are accepted.)
    - c. Input Data Handling
      - i. **Steps to Increment:** Provide a number to decide increment and a period by choosing any one option from the drop-down menu.
      - ii. Start Date: Select a start date using the calendar or set it manually.
      - iii. Start Time: Enter the definite time to start the process in the hh/mm/ss format
    - d. New Column Information
      - i. **Period Column Name:** Enter a name for the column containing a period value. (This field is predefined, but the user can change the value if needed)



Component	Console	Summary	Result	Visualization	Properties	÷ 1
General	Output Inform	nation				
Properties	Output Mode		Forecast	-		
Advanced	Period To Foreca	est	150			
	Column Selec	tion				
	Feature		4 checked	•	0	
	Target Variable		Number	-	0	
	Input Data Ha	andling				
	Steps to Increme	ent	2	Min -		
	Start Date		06-08-2019			
	Start Time		07:57:03			
•	New Column	Information	ı			
	Period Column N	Name	PeriodValues		0	
						-
						Apply

- Properties with Trend Output Mode
  - a. Output Information
    - i. Output Mode: Select a mode in which you want to display output data. The user gets two options for this field.
      - 1. **Trend**: Selecting this option displays source data along with predicted values for the given dataset.
      - Forecast: Selecting this option displays forecasted values for the given period. Results get appended to the target column when the 'Forecast' output mode has been selected.
    - ii. **Period to Forecast**: Enter a period to forecast. This field appears only when '**Forecast**' is selected as an **'Output Mode'** option.
  - b. Column Selection
    - i. Feature: Select the feature columns using the drop-down menu.
    - ii. **Target Variable:** Select the target variable for which you want to Apply forecasting analysis (The first selected option gets selected by default. Only numerical columns are accepted.)
  - c. Input Data Handling
    - i. **Steps to Increment:** Provide a number to decide increment and a period by choosing any one option from the drop-down menu.
    - ii. Start Date: Select a start date using the calendar or set it manually.
    - iii. Start Time: Enter the definite time to start the process in the hh/mm/ss format
  - d. New Column Information
    - i. **Predicted Column Name**: Enter a name for the column containing the Predicted Values (The title for this field comes pre-defined, but the users can change the value if needed).
    - ii. **Period Column Name:** Enter a name for the column containing a period value (This field comes predefined, but users can change the value if needed).



Component	Console Summary	Result	Visualization	Properties	<u>+</u> ⊥			
General	Output Information							
Properties	Output Mode	Trend						
Advanced	Column Selection	Column Selection						
	Feature	3 checked	-	0				
	Target Variable	Number	-	0				
	Input Data Handling	iput Data Handling						
	Steps to Increment	4	Month 👻					
	Start Date	06-08-2019						
	Start Time	16:45:00						
	New Column Informatio	on						
÷	Predicted Column Name	PredictedValue	es	0				
	Period Column Name	PeriodValues		0				
					Apply			

- iv) Click the 'Advanced' tab and configure, if required:
  - a. Configure the following 'Seasonal Order' fields
    - i. AR Parameter
    - ii. Difference
    - iii. MA Parameter
    - iv. Season
  - b. Configure the following 'Trend Order' information
    - i. AR Parameter
    - ii. Difference
    - iii. MA Parameter
- v) Click the **'Apply'** option.

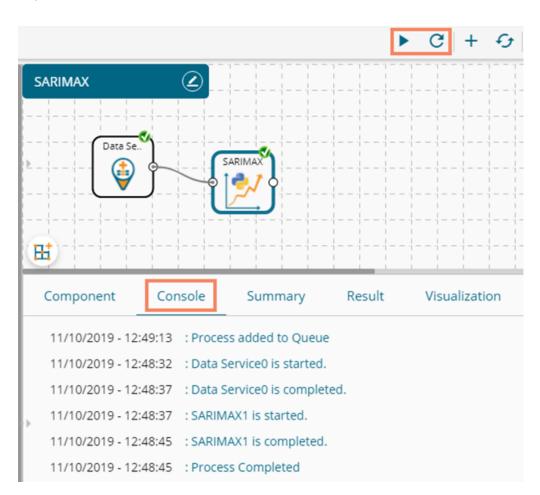
Component	Console	Summary	Result	Visualiza	ation	Properties	<del>+</del>	<u>+</u>
General	Seasonal Ord	der						
Properties	AR Parameter		Optional					
Advanced	Difference		Optional					
	MA Parameter		Optional					
	Season		0					
-	Trend Order							
	AR Parameter		Optional					
• •	Difference		Optional					
r	MA Parameter		Optional					
							Appl	y

Note: The 'Advanced' tab remains the same for any output mode.

### www.bdb.ai



- vi) Run the workflow after getting the success message.
- vii) The user gets directed to the '**Console**' tab displaying the ongoing process. The completion of the Console process gets marked by the green checkmarks on the top of the dragged component.



- viii) View the processed data by clicking the dragged SARIMAX component and then clicking the '**Result**' tab.
  - a) Result tab with 'Forecast' as the output mode



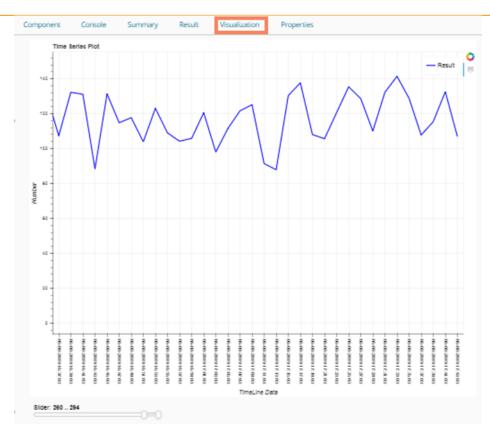
C	Component	Console	Summary	Result V	'isualization	Properties	<u>*</u> <u>+</u>
Sh	10 10	<ul> <li>entries</li> </ul>				Search:	
	Number	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	PeriodValues
	1	5.1	3.5	1.4	0.2	setosa	06-08-2019 07:57:03
1	2	4.9	3	1.4	0.2	setosa	06-08-2019 07:59:03
-	3	4.7	3.2	1.3	0.2	setosa	06-08-2019 08:01:03
4	4	4.6	3.1	1.5	0.2	setosa	06-08-2019 08:03:03
1	5	5	3.6	1.4	0.2	setosa	06-08-2019 08:05:03
6	6	5.4	3.9	1.7	0.4	setosa	06-08-2019 08:07:03
	11	5.4	3.7	1.5	0.2	setosa	06-08-2019 08:09:03
	12	4.8	3.4	1.6	0.2	setosa	06-08-2019 08:11:03
	13	4.8	3	1.4	0.1	setosa	06-08-2019 08:13:03
	14	4.3	3	1.1	0.1	setosa	06-08-2019 08:15:03
Sh	nowing 1 to 1	0 of 294 entries		Prev	ious 1 2	3 4	5 30 Next

### b) Result tab with 'Trend' as the output mode

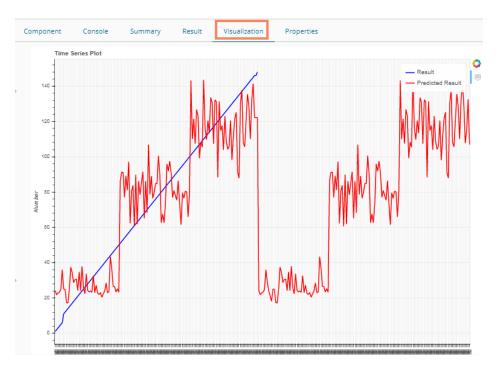
Component	Console	Summary	Result	Visualization	Properties		÷ <u>1</u>
Show 10	▼ entries					Search:	
Number	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	PredictedValues	PeriodValues
1.0	5.1	3.5	1.4	0.2	setosa	25.053	06-08-2019 16:45:00
2.0	4.9	3	1.4	0.2	setosa	16.681	06-12-2019 16:45:00
3.0	4.7	3.2	1.3	0.2	setosa	22.763	06-04-2020 16:45:00
4.0	4.6	3.1	1.5	0.2	setosa	28.926	06-08-2020 16:45:00
5.0	5	3.6	1.4	0.2	setosa	29.731	06-12-2020 16:45:00
6.0	5.4	3.9	1.7	0.4	setosa	38.57	06-04-2021 16:45:00
11.0	5.4	3.7	1.5	0.2	setosa	26.955	06-08-2021 16:45:00
12.0	4.8	3.4	1.6	0.2	setosa	35.507	06-12-2021 16:45:00
13.0	4.8	3	1.4	0.1	setosa	18.827	06-04-2022 16:45:00
14.0	4.3	3	1.1	0.1	setosa	19.732	06-08-2022 16:45:00
Showing 1 to '	10 of 294 entries				Previous	1 2 3 4	5 30 Next

- ix) Click the '**Visualization**' tab to open the graphical representation of the processed data through a time series chart.
  - a) Visualization of the processed data with 'Forecast' as output mode.





### b) Visualization of the processed data with Trend as output mode



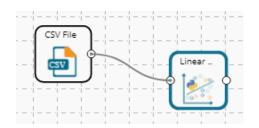
# 14.1.2. Regression

This algorithm is used to determine how an individual variable influences another variable using an exponential function. It finds a trend in the dataset Applying univariate regression analysis.



### 14.1.2.1. Linear Regression

i) Drag the Linear Regression component to the workspace and connect it to a configured data source.



ii) Configure the following fields in the 'Properties' tab:

#### a. Column Selection

- i. **Dependent Column**: Select the target column on which the regression analysis gets applied
- ii. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column
- b. New Column Information
  - i. **Predicted Column Name**: Enter a name for the new column containing the predicted values.

Component	Console	Summary	Result	Visualization	Properties		<u>↓</u> ↑	<u>+</u>
General	Column se	election						
Properties	Independent	t Column	SepalLength	•	0			
Advanced	Dependent (	Column	SepalWidth	-	0			
	New Colu	mn Informatic	in					
	Predicted Co	olumn	PredictedValues		0			
	Name							
>								
						_		
							Apply	y

iii) Click the 'Advanced' tab and configure if required:

#### a. Behavior

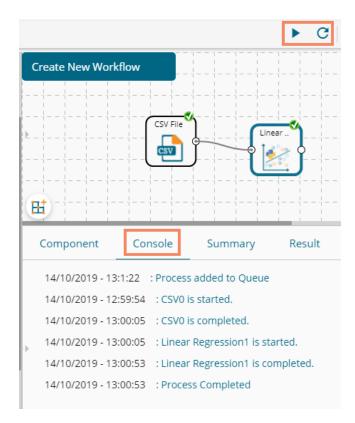
- i. **Fit Intercept**: This option is used to select whether to calculate the intercept for the selected model or not
  - 1. True: By selecting this option intercept gets calculated (It is the default selection)
  - 2. False: By selecting this option intercept does not get calculated
- ii. Normalize: This option is used to select whether to normalize the feature column or not
  - 1. **True**: If the selected Normalize option is **True**, the feature column gets the selected normalization option.
  - 2. False: If the Normalize option is False, the feature column cannot be normalized (It is the default option).
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualizatior	n Properties	<u>+</u>	Ť
General	Behavior						
Properties	Fit Intercept	Tru	le	-			
Advanced	Normalize	Fa	lse	•			
7							
						Apply	/

**Note:** The model containing aliased coefficients signifies that the square matrix x\*x is singular.

- v) Run the workflow after getting the success message.
- vi) The user gets the process status under the 'Console' tab.

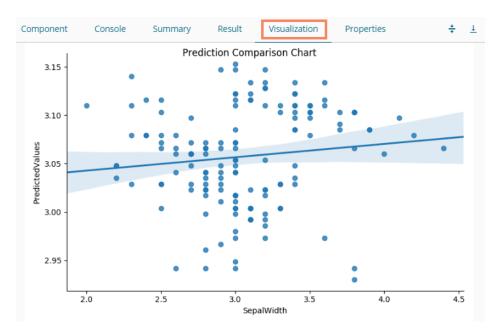


- vii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.
    - i. A new column displaying the predicted values gets added to the result view.



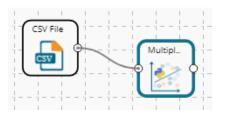
	Component	Console	Summary	Result	Visualization	Properties	; <u>+</u> <u>⊥</u>
	Show 10	▼ entries				Search:	
Þ	Number	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	PredictedValues
	1	5.1	3.5	1.4	0.2	setosa	3.1
	2	4.9	3	1.4	0.2	setosa	3.12
	3	4.7	3.2	1.3	0.2	setosa	3.13
	4	4.6	3.1	1.5	0.2	setosa	3.13
	5	5	3.6	1.4	0.2	setosa	3.11
	6	5.4	3.9	1.7	0.4	setosa	3.08
	7	4.6	3.4	1.4	0.3	setosa	3.13
	8	5	3.4	1.5	0.2	setosa	3.11
,	9	4.4	2.9	1.4	0.2	setosa	3.15
	10	4.9	3.1	1.5	0.1	setosa	3.12
	Showing 1 to 1	10 of 150 entries		Previou	1 2	3 4 5	15 Next

- viii) Click the 'Visualization' tab.
- ix) The processed data gets displayed via the Prediction Comparison chart with a Regression line.



# 14.1.2.2. Multiple Linear Regression

i) Drag the R-Multiple Linear Regression component to the workspace and connect it with a configured data source.





#### ii) Configure the 'Properties' tab as displayed below:

Component	Console S	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Column selectio	n					
Properties	Independent Colum	nn 4 che	cked	• 0			
Advanced	Dependent Column	Numb	ber	<b>.</b> •			
	New Column Int	formation					
	Predicted Column	Predi	ctedValues	0			
	Name						
•							
						Apply	/

- iii) Click the 'Advanced' tab and configure if required:
  - a. Behavior
    - i. **Fit Intercept**: This option is used to select whether to calculate the intercept for the selected model or not
      - 1. True: By selecting this option intercept gets calculated (It is the default selection)
      - 2. False: By selecting this option intercept gets calculated
    - ii. **Normalize**: This option is used to select whether to normalize the feature column or not
      - 1. True: If Normalize option is 'True,' it normalizes the feature column
      - 2. False: If Normalize option is '**False**,' the feature column does not take the normalization value (It is the default option)
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>⊥</u> <u>⊺</u>
General	Behavior					
Properties	Fit Intercept		True	•		
Advanced	Normalize		False	•		
						Apply

- v) Run the workflow after getting the success message.
- vi) The '**Console**' tab opens displaying the progress of the process. The completed console process gets marked by the green checkmarks on the top of the dragged components.



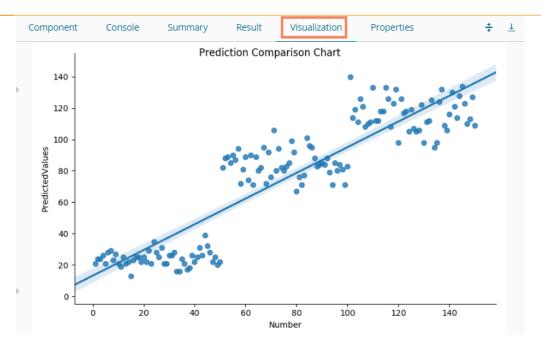
Create New Work	flow		Multipl.	+ C +
Component	Console	Summary	Result	Visualization
16/10/2019 - 10: 16/10/2019 - 10: 16/10/2019 - 10: 16/10/2019 - 10:	57:37 : CSV0 57:42 : CSV0	ess added to Queu ) is started. ) is completed. iple Linear Regress		

- vii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.
    - i. A new column containing the Predicted Values gets added to the Result data.

10 10	<ul> <li>entries</li> </ul>				Search:	
Number	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	PredictedValu
1	5.1	3.5	1.4	0.2	setosa	21
2	4.9	3	1.4	0.2	setosa	24
3	4.7	3.2	1.3	0.2	setosa	24
4	4.6	3.1	1.5	0.2	setosa	26
5	5	3.6	1.4	0.2	setosa	21
5	5.4	3.9	1.7	0.4	setosa	28
7	4.6	3.4	1.4	0.3	setosa	29
В	5	3.4	1.5	0.2	setosa	23
9	4.4	2.9	1.4	0.2	setosa	27
10	4.9	3.1	1.5	0.1	setosa	21

- viii) Click the 'Visualization' tab.
- ix) The Result data gets displayed via the Prediction Comparison Chart with a Regression line.





# 14.1.2.3. Logistic Regression

i) Drag the R-Multiple Linear Regression component to the workspace and connect it with a configured data source.

CSV File		Logisti	
	J	00000	J

ii) Configure the 'Properties' tab as displayed below:

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Column se	election					
Properties	Independent	t Column	4 checked	-	0		
Advanced	Dependent (	Iolumn	species	•	0		
	New Colur	mn Informatio	n				
	Predicted Co	lumn Name	PredictedValues		0		
▶							
						Apply	,



iii) Click the 'Advanced' tab and configure if required:

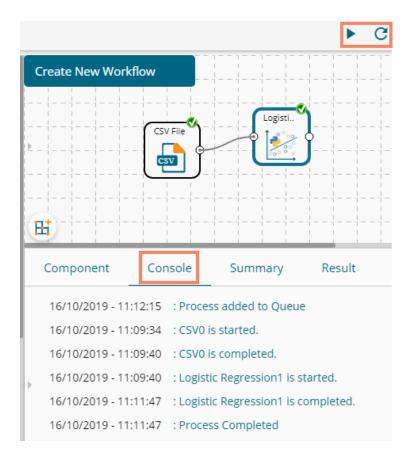
#### a. Input Data Handling

- i. **Missing Values**: Select a method to deal with missing values (via the drop-down menu)
  - 1. **Fit Transform**: Selecting this option will consider the records containing missing values from the independent columns
  - 2. **Stop**: Selecting this option stops the application of the algorithm if a value is missing in any column
- **b.** Behavior: The fields provided under this section are used to improve model accuracy
  - i. Weight: This field can have either 'None' or 'Balanced' as value. The default value for this field is 'None.'
  - ii. Class Penalty: This field can have a value of either 'L1' or 'L2'. The default value for this field is 'L2'.
  - iii. Maximum No. of Iterations: Enter a valid integer value allowed to calculate the algorithm coefficient. The default value for this field is 100.
  - iv. Solver: The following options get listed for this field
    - 1. Newton-CG,
    - 2. Lib-Linear (It is the default value for this field)
    - 3. LBFGS
    - 4. SAG
  - v. Dual: It can have a Boolean value (The default value for this field is 'False')
  - vi. Tolerance: It can have a double type value (The default value for this field is 0.0001)
  - vii. Fit Intercept: It has two options 'True' and 'False.' By selecting 'True,' it calculates the intercept for the selected model (The default value for this field is 'True')
  - viii. Intercept Scaling: It can have a double type value (The default value for this field is 1.0)
  - ix. Inverse Regularization: This field can only take value in double type (The default value for this field is 1.0)
  - x. Random State: This field can only take integer values (The default value for this field is 12)
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	*	<u>1</u>
General	Behavior						
Properties	Class Weigh	t	None		-		
Advanced	Penalty		L2		-		
	Maximum N	lo of Iterations	100				
	Solver		Lib Linear		-		
	Dual		True		-		
	Tolerance		0.0001				
	Fit Intercept		True		-		
	Intercept Sc	aling	1				
	Inverse Reg	ularization	1.0				
×	Random Sta	ite	12				
							_
						Appl	у



- v) Run the workflow after getting the success message.
- vi) The 'Console' tab opens, displaying steps of the ongoing process.



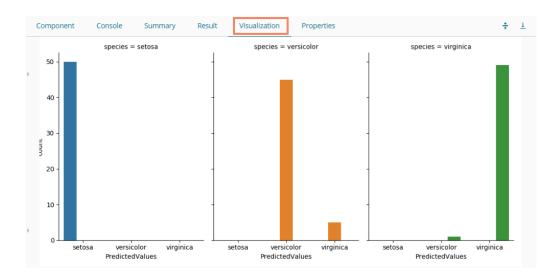
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
- viii) A new column containing Predicted values gets added to the Result data.



	Component	Console Sur	nmary Result	Visualizatior	n Proper	ties 🔆 🛨		
Show 10 v entries Search:								
Þ	sepal_length	sepal_width	petal_length	petal_width	species	PredictedValues		
	5.1	3.5	1.4	0.2	setosa	setosa		
	4.9	3	1.4	0.2	setosa	setosa		
	4.7	3.2	1.3	0.2	setosa	setosa		
	4.6	3.1	1.5	0.2	setosa	setosa		
	5	3.6	1.4	0.2	setosa	setosa		
	5.4	3.9	1.7	0.4	setosa	setosa		
	4.6	3.4	1.4	0.3	setosa	setosa		
	5	3.4	1.5	0.2	setosa	setosa		
	4.4	2.9	1.4	0.2	setosa	setosa		
Þ	4.9	3.1	1.5	0.1	setosa	setosa		
	Showing 1 to 10 of 1	50 entries		Previous 1	2 3 4	5 15 Next		

### ix) Click the 'Visualization' tab.

x) The processed data gets displayed via the Comparative Column chart.



xi) Click the 'Summary' tab to view the model summary.



	Component	Console	Summary	Result	Visualization	Properties	*	1
		Summary of th	ne model					
•	sepal petal petal 2.Dependent C	_length (f _width (f _length (f _width (f						
						, intercept_scaling=1.0 r, tol=0.0001, verbose=		
•	Accuracy Repo 0.96	rt: End of Summar	у					

### 14.1.2.4. Elastic Net

i) Drag the Elastic Net component to the workspace and connect it to a configured data source.

	Elastic
CSV File	> ↑
	-

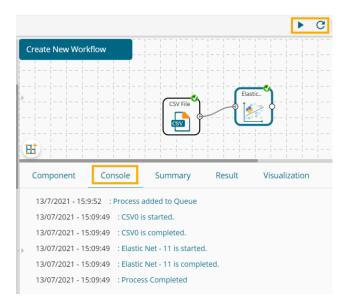
- ii) Configure the following fields in the 'Custom Group' tab:
  - a. Dynamic Fields
    - i. **Dependent Column**: Select the target column on which the regression analysis gets applied
    - ii. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column.
    - iii. Alpha: Provide value for the alpha field
    - iv. **I1 Ration [0..1]:** Provide a number to decide the ratio (It should be between 0 to 1)
    - v. Maximum Iterations: Provide the maximum number of iterations.
    - vi. Tol Values: Provide the Tol values.
    - vii. Selection: Select one option out of the 'Random' or 'Cyclic' options.
- iii) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	DataInsight	*	<u>+</u>
General	Dynam	nic Fields						
Custom Group	_	Independent Colum	ın	5 checked	•			
		Dependent Column		Number	~			
		Alpha		1.0				
		11 Ratio[0.,1]		0.5				
		Maximum Iteration	S	1000				
		Tol Value		0.0001				
		selection		cyclic	~			
							Apply	

**Note:** The model containing aliased coefficients signifies that the square matrix x\*x is singular.

- iv) Run the workflow after getting the success message.
- v) The user gets the process status under the 'Console' tab.

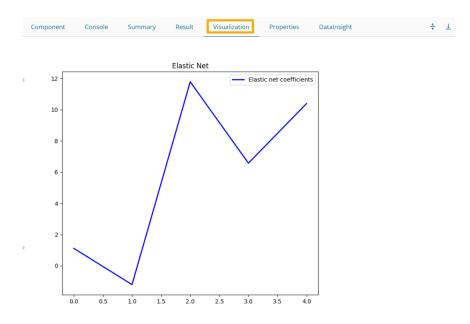


- vi) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

Component	Console	Summary Result	Visualization	Properties	DataInsight	<del>+</del>
Show 10 🗸	• entries				Sea	rch:
Number	sepal_length	sepal_width	petal_length	petal_width	species	PredictedValues
1	5.1	3.5	1.4	0.2	setosa	29.409
2	4.9	3	1.4	0.2	setosa	29.789
3	4.7	3.2	1.3	0.2	setosa	28.148
4	4.6	3.1	1.5	0.2	setosa	30.513
5	5	3.6	1.4	0.2	setosa	29.177
6	5.4	3.9	1.7	0.4	setosa	34.107
7	4.6	3.4	1.4	0.3	setosa	29.63
8	5	3.4	1.5	0.2	setosa	30.596
9	4.4	2.9	1.4	0.2	setosa	29.354
10	4.9	3.1	1.5	0.1	setosa	30.19



vii) A new column displaying the predicted values gets added to the result view. viii) Click the '**Visualization**' tab.

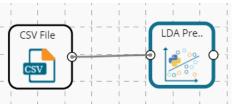


ix) The processed data gets displayed via the Prediction Comparison chart with a Regression line.x) Click the 'Summary' tab to display the summary.

mponent	Console	Summary	Result	Visualization	Properties	DataInsight	<del>4</del> †
	Summary Of	The ElasticNet -		-			
Call:-							
max_ite	r=1000, normal	_X=True, fit_inte lize=False, posit election='cyclic'	ive=False, p				
	)f Determinatio						
0.817							
	Error (MAE):	-					
15.614							
	Error (MSE):-						
343.173							
	ared Error (R)	MSE):-					
18.5249							
	[						

### 14.1.2.5. LDA Prediction

i) Drag the LDA Prediction component to the workspace and connect it to a configured data source.

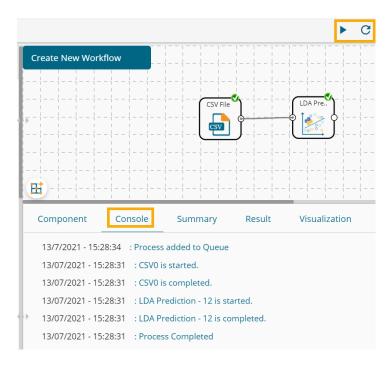




- ii) The 'Custom Group' tab opens.
  - a. Dynamic Fields
    - i. **Dependent Column**: Select the target column on which the regression analysis gets applied.
    - ii. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column.
- iii) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General Custom Group		t <b>Fields</b> Independent Column Iependent Column		5 checked species	•			
							Apply	у

iv) Run the workflow after getting the success message.



v) The user gets the process status under the 'Console' tab.

- vi) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.
  - c. A new column displaying the predicted values gets added to the result view.



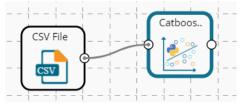
	Component	Console	Summary F	Result	zation Prope	erties Da	atalnsight 🕂 🕹
	Show 10	✓ entries				Search:	
Þ	Number	sepal_length	sepal_width	petal_length	petal_width	species	PredictedValues
	1	5.1	3.5	1.4	0.2	setosa	setosa
	2	4.9	3	1.4	0.2	setosa	setosa
	3	4.7	3.2	1.3	0.2	setosa	setosa
	4	4.6	3.1	1.5	0.2	setosa	setosa
	5	5	3.6	1.4	0.2	setosa	setosa
	6	5.4	3.9	1.7	0.4	setosa	setosa
	7	4.6	3.4	1.4	0.3	setosa	setosa
	8	5	3.4	1.5	0.2	setosa	setosa
	9	4.4	2.9	1.4	0.2	setosa	setosa
Þ	10	4.9	3.1	1.5	0.1	setosa	setosa
	Showing 1 to 10	) of 150 entries		F	Previous 1	2 3 4	5 15 Next

#### vii) Click the 'Summary' tab to display the summary.

	Component	Console	Summary	Result	Visualization	Properties	DataInsight	÷ <u>+</u>
•		Summary Of	The LDA					
	 LinearDiscrimi	solver='svd'	(n_components=None , store_covariance		ne, shrinkage=None, =0.0001)			
	3.12882807e-0	9 4.85069285 9 3.60504250	e-08 1.72392591e-00 e-09 1.94998225e-00 e-08 1.93932582e-00	3 1.2688129	2e-08			
	6.39801049e-1 2.18691670e-0	l0 6.97974439 8 6.78352086	'e-08 7.24319214e-1: e-09 3.17350624e-09 ie-09 1.13564654e-09 e-07 5.12418212e-09	9 3.0987036 9 1.4340589	6e-09 3e-07			
•	4.66048493e-1 2.32249368e-0 7.21759475e-0	10 1.64615229 99 6.42206142 99 3.04834850	e-07 1.47463214e-03 e-10 6.13014944e-03 e-08 1.14347026e-03 e-06 5.05343200e-03	3 1.1345680 7 1.7906225 3 8.0418927	8e-08 7e-08 6e-08			
	4.06871126e-0 9.99999946e-0 9.99999788e-0	99 2.34853567 91 9.999999996 91 1.00000000	<pre>le=07 5.53641620e-09 'e=08 1.00000000e+00 ie=01 9.99999895e-0 be+00 9.99999996e-0 ie=01 1.00000000e+00</pre>	0 9.99999999 1 9.99999995 1 9.99999999	9e-01 0e-01 4e-01			

# 14.1.2.6. Catboost Regression

i) Drag the Catboost Regression component to the workspace and connect it to a configured data source.



ii) The 'Custom Group' tab opens.

#### a. Dynamic Fields

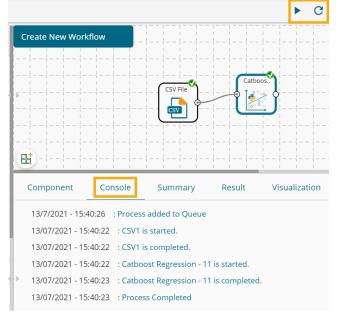
- i. **Dependent Column**: Select the target column on which the regression analysis gets applied.
- ii. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column.
- iii. Iterations: Provide iteration value.



- iv. Depth [1-16]: Provide a depth value.
- v. Learning Rate: Provide learning rate value.
- vi. Seed Value: Provide a number to generate the seed value.
- vii. Evaluation Metric: Select an option out of the given choices: RMSE, LogLinQuantile, MAPE, Poisson, Quantile, MSLE, R2, MedianAbsoluteError.
- viii. Boosting Type: Select the boosting type out of the 'Ordered' and 'Plain' options.
- ix. One-Hot Encoding: Provide value to encode the integer representation in the model.
- iii) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	+	<u>+</u>
General Custom Group	Dynam	nic Fields						
		Independent Colum	ns	5 checked	-			
		Dependent Column		Number	~			
		Iterations		50				
		Depth [1-16]		3				
		Learning Rate		0.1				
		Seed Value		10				
		Evaluation Metric		RMSE	~			
		Boosting Type		Ordered	~			
•		One-Hot Encoding		2				
							Арг	oly

- iv) Run the workflow after getting the success message.
- v) The user gets the process status under the 'Console' tab.



- vi) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.
  - c. A new column displaying the predicted values gets added to the result view.



ow 10	✓ entries				Search:	
Number	sepal_length	sepal_width	petal_length	petal_width	species	PredictedValues
	5.1	3.5	1.4	0.2	setosa	24.609
	4.9	3	1.4	0.2	setosa	26.503
	4.7	3.2	1.3	0.2	setosa	26.25
	4.6	3.1	1.5	0.2	setosa	25.225
	5	3.6	1.4	0.2	setosa	24.264
	5.4	3.9	1.7	0.4	setosa	25.969
	4.6	3.4	1.4	0.3	setosa	21.761
	5	3.4	1.5	0.2	setosa	24.884
	4.4	2.9	1.4	0.2	setosa	22.185
0	4.9	3.1	1.5	0.1	setosa	26.16

vii) Click the 'Summary' tab to display the summary.

	Component	Console	Summary	Result	Visualization	Properties	DataInsight	+++	<u>+</u>
		Su	mmary of the CatBo	ost Regress	ion				
ŀ									
	Parameters as	signed to the	model:-						
					False, 'eval_metric' E', 'depth': 3, 'lea		ot_max_size': 2, 'all 'iterations': 50}	low_const	
	Feature Import	tance:-							
	Footur	re Importanc							
	0 sepal_lengt								
		th 6.80618							
	2 petal_lengt	th 50.92153	4						
	3 petal_widt	th 19.97849	4						
	4 specie	15.81667	9						
	RMSE Score:-								
F.									
	14.94258562626	36736							
			End Of Summary						

### 14.1.3. Classification

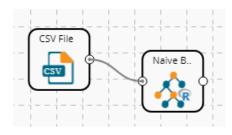
### 14.1.3.1. Naive Bayes

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to the presence of any other feature. For example, a fruit may be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, these properties independently contribute to the probability that this fruit is an apple, and that is why it is known as **Naive**.

Naive Bayes is a leaf node under Classification algorithms under the Algorithm tree node. The component consists of one node for reading data from a data source and another one for giving the Result.

i) Drag the R-Naive Bayes component to the workspace and connect it with a configured data source.





- ii) Configure the following fields in the 'Properties' tab:
  - a. Column Selection
    - i. **Feature**: Select input columns from the drop-down menu to which the target variable can be compared to performing the analysis.
    - ii. **Target Variable**: Select the target column for which the analysis is Performed.
  - b. Output Information
    - i. **Show Probability:** Select an option out of True or False (Selecting the 'True' option displays the Probability Column Name field under the 'New Column Information' section).
  - c. New Column Information
    - i. **Probability Column Name**: Enter a name for the new column containing the probability values.
    - ii. **Predicted Column Name**: Enter a name for the new column containing the predicted values.
  - **d. Hyperparameter Tuning:** Apply Hyperparameter Tuning for the model by using a checkmark in the given box.
  - e. Enable Validation: Enable validation by using a checkmark in the given box.

There are three scenarios for the Properties tab to get configured:

**1.** Hyperparameter Tuning and Validation are disabled.

Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Output Inf	ormation					
Properties	Show Probab	oility	True	•			
Advanced	Column Se	lection					
	Feature		7 checked	•	0		
	Target Variab	ble	sex	•	0		
	New Colun	nn Informatior	1				
	Probability C	olumn	ProbabilityValue	s	0		
	Name						
	Predicted Co	lumn Name	PredictedValues		0		
Þ	Hyperparam	eter Tuning 🗆					
	Enable Valida	ation 🗆					
						Apply	

2. Hyperparameter Tuning is Applied



Component	Console	Summary	Result	Visualization	Properties	+	<u>+</u>
General	Column Se	election					
Properties	Feature		7 checked	•	0		
Advanced	Target Varial	ble	sex	-	0		
	New Colur	nn Information	n				
	Predicted Co	lumn Name	PredictedValues		0		
	Hyperparam	eter Tuning 🗹					
)-							
						Apply	,

### 3. Validation is enabled

Component	Console Summary	Result	Visualization	Properties	÷	+	<u>+</u>
General	Output Information						
Properties	Show Probability	True	•				
Validation	Column Selection						
Advanced	Feature	7 checked	•	0			
	Target Variable	sex	•	0			
	New Column Informatio	on					
	Probability Column	ProbabilityValues		0			
	Name						
	Predicted Column	PredictedValues		0			
F	Name						
	Enable Validation 🗹						
					A	pply	

- iii) Click the 'Validation' tab to configure, if it has been enabled from the Component Properties tab. The 'Validation' tab provides multiple options under the 'Model Type Values' drop-down menu. The user can select any one of the available options to configure the Validation tab.
  - a. Model Selection
    - i. Stratified K-fold Validation

The user needs to configure the **'Number of Folds'** fields if the selected Model Type Value is **Stratified K-fold Validation.** 



Component	Console	Summary	Result	Visualization	Properties	+	$\pm$
General	Model Selec	tion					
Properties	ModalTypeVal	ues	Stratified K-Fold v	validation 💌			
Validation	Number of Fo	lds	Search				
Advanced		-					
			<ul> <li>Stratified</li> </ul>	K-Fold valida			
			K-Fold validati	on			
			Leave one out	cross validat			
>							
						Apply	

#### ii. K-fold Validation

The user needs to configure the **'Number of k-folds'** field if the selected Model Type Values is **K-Fold Validation.** 

Component	Console	Summary	Result	Visualizatio	on Properties	<u>+</u>	Ť
General	Model Sele	ection					
Properties	ModalTypeVa	lues	K-Fold validation	-			
Validation	Number of k-	folds	3				
Advanced							
						Apply	/

### iii. Leave One Out Cross-Validation

The user gets to configure no other fields when the selected Model Type Values option is **Leave One Out Cross-Validation**.

Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Model Selec	tion					
Properties	ModalTypeVal	ues 🚺	eave one out cro	ss validatio. <del></del>			
Validation							
Advanced							
						Арр	oly

- iv) Click the 'Advanced' tab and configure if required.
- Advanced Tab when 'Hyperparameter Tuning' is Enabled a. Hyperparameter Tuning



- i. **Maximum Evaluation:** Enter optimal evaluation value for defining hyperparameters to search for the ideal model architecture. The default value for this field is 10.
- ii. **Fit Increment:** Provide increment value for Hyperparameter model tuning. The default value for this field is 1.
- iii. **Trial Timeout:** Set value for the trial timeout field by providing a number. The default value for this field is 30.
- iv. Seed: Provide value to configure the seed field. The default value for this field is 123.
- v. Click the **'Apply'** option.

Component	Console	Summary	/ R	esult	Visualiza	ation	Properties	<u>+</u> ↑	Ţ
General	Hyperparam	eter Tunin	σ						
Properties	Maximum evalu		10						
Advanced	Fit Increment	adion	1						
	Trial Timeout		30						
	Seed		123						
Þ									
								Apply	

- Advanced Tab when 'Validation' is Enabled
  - a. Behavior
    - i. **Laplace Constant:** Enter the smoothing constant for smoothing observations. The smoothing constant must be a double value greater than 0. Entering 0 disables Laplace smoothing.
    - ii. Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	<u>⊥</u> <u>⊺</u>
General	Behaviour					
Properties	LaplaceConst	ant	0.000000001			
Validation						
Advanced						
						Apply

Note: The same field appears when Validation and Hyperparameter Tuning are disabled.

- v) Run the workflow and after getting the success message.
- vi) The '**Console**' tab opens displaying the steps of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





vii) Click the 'Result' tab to display the dataset in the result view.

#### i. Result View with Validation disabled.

how	10 <b>v</b> en	tries					Se	arch:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValue
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	1
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	1
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	1
I	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
I	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	М
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	М
M	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	F

#### ii. Result View with Validation enabled.

how	10 v ent	ries							Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	ProbabilityValu
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	1	[0.0, 1.0, 0.0]
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1	[0.0, 1.0, 0.0]
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	1	[0.21, 0.46, 0.33]
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	1	[0.0, 0.99, 0.01]
I	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1	[0.0, 1.0, 0.0]
1	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1	[0.0, 1.0, 0.0]
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	М	[0.42, 0.03, 0.55]
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	М	[0.39, 0.13, 0.48]
М	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1	[0.0, 0.98, 0.02]
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	м	[0.51, 0.0, 0.49]

iii. Result View with Validation and Hyperparameter disabled

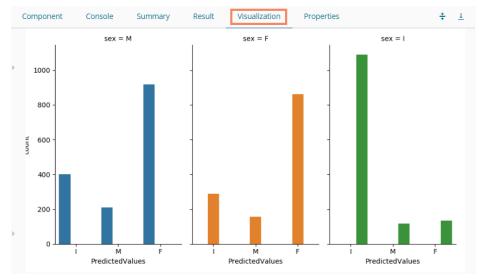
#### Copyright © 2015-21 BDB

#### www.bdb.ai



Comp	onent	Console	Summary	Result	Visualization P	roperties				÷ ±
Show	10 <b>v</b> en	itries							Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	ProbabilityValues
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	1	[0.0, 1.0, 0.0]
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1	[0.0, 1.0, 0.0]
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	1	[0.21, 0.46, 0.33]
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	1	[0.0, 0.99, 0.01]
1	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1	[0.0, 1.0, 0.0]
1	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1	[0.0, 1.0, 0.0]
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	М	[0.42, 0.03, 0.55]
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	М	[0.39, 0.13, 0.48]
М	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1	[0.0, 0.98, 0.02]
> F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	F	[0.51, 0.0, 0.49]
Showin	g 1 to 10 of 4	l,177 entries					Previ	ious 1	2 3 4	5 418 Next

viii) Click the '**Visualization**' tab to see the processed data in the comparative Column charts (the current visualization displays the processed data when '**Validation**' is enabled).



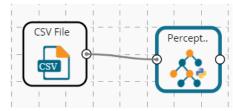
ix) Click the 'Summary' tab to see the detailed Model Summary.

Component	Console	Summary	Result	Visualization	Properties	*
	Summary of the	model				
1.Independent	Columns					
	th (float64)					
	. ,	.oat64)				
heigh	nt (float64)					
	nt_whole (fl					
	nt_shucked (fl					
	nt_viscera (fl					
2.Dependent (	nt_shell (fl	.oat64)				
	(object)					
	()/					
6-33						
Call:	ions-None van	_smoothing=1e-09	\ \			
Ganzzraiinn(bi	1015-100110, Val	_smoorning-ie-05	)			
Accuracy Repo	ort:					
0.517						



### 14.1.3.2. Perceptron

i) Drag the Perceptron component to the workspace and connect it to a configured data source.

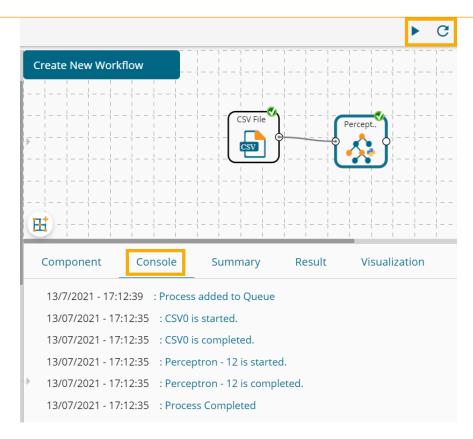


- ii) The 'Custom Group' tab opens.
  - a. Provide the required details for the Dynamic Fields
    - i. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column.
    - ii. **Dependent Column**: Select the target column on which the regression analysis gets applied.
    - iii. Penalty Term
    - iv. Alpha Value
    - v. Maximum Iterations (aka epochs)
    - vi. Eta Value (>0)
    - vii. Stopping Criterion
    - viii. Random State Value
    - ix. Fit Intercept
- iii) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	*	<u>+</u>
General	Dyna	mic Fields						
Custom Group	- Dyna							
		Independent Co	olumns	8 checked		-		
		Dependent Col	umn	sex		~		
		Penalty Term		None		~		
		Alpha Value		0.0001				
		Maximum Itera	tions(aka epochs	5) 1000				
		Eta Value (>0)		0.1				
		Stopping Criter	ion	0.001				
		Random State \	/alue	0				
•		Fit Intercept		True		~		
							Apply	y

- iv) Run the workflow after getting the success message.
- v) The user gets the process status under the 'Console' tab.





- vi) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
  - c. A new column displaying the predicted values gets added to the result view.

ow	10 🗸 ent	ries						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValue
1	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	М
1	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1
	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	F
1	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	М
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	М
	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	М
	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	М
	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	м

vii) Click the 'Summary' tab to display the summary.



omponent	Console	Summ	ary	Result	Visualization	Properties	Datalnsig
	Su	mmary of th	e Percep	tron			
Call:-							
fit_intero n_jobs=Nor	ept=True, ne, penalty	max_iter=10 ='None', ra	00, n_it ndom_sta	er=None, n_	ng=False, eta0=0.1 iter_no_change=5, le=True, tol=0.001 se)		
Weights assigned	to the fe	atures:-					
-0.42512021 -0			0.2831	.7657 -0.034	56567 -0.21467564		
			0.0845	8547 -0.239	30588 -0.20973282		
0.32512244 -0							
[ 0.0714035 0	.55412112	-0.02575803	-0.8466	6606 -0.144	0015 0.14822842		
0.34014346 0	.15980541]	]					
Accuracy:-							
pr	recision	recall f1	-score	support			
F	0.64	0.66	0.65	44			
I	0.77	0.83	0.80	12			
м	0.60	0.57	0.58	44			
micro avg	0.64	0.64	0.64	100			
macro avg	0.67	0.69	0.68	100			
weighted avg	0.64	0.64	0.64	100			
Accuracy Score:-	0.64						
		End Of Su	mmary				

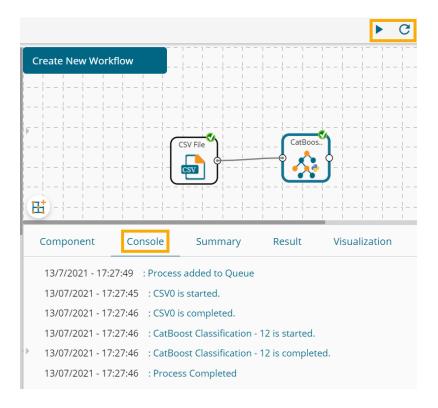
### 14.1.3.3. Catboost Classification

- i) Drag the Catboost Classification component to the workspace and connect it to a configured data source.
- ii) The 'Custom Group' tab opens.
  - a. Dynamic Fields
    - i. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column.
    - ii. **Dependent Column**: Select the target column on which the regression analysis gets applied.
    - iii. Iteration: Set the iteration value
    - iv. Depth [1-16]: Set depth. it should be between 1 to 16 (wherein 1 and 16 numbers are also included)
    - v. Learning Rate: Provide Learning rate value
    - vi. Loss: Select an option from out of Logloss or Multiclass
    - vii. Evaluation Metric: Set evaluation metric out of the given choices: Logloss, Crossentropy, Precision, F1, Recall, BalancedAccuracy, BalancedErrorRate
    - viii. Seed Value: Set a seed value.
    - ix. Boosting Type: Select a boosting type out of the given choices: Ordered and Plain.
    - x. Prediction Type: Select a Prediction Type option out of Class, RowFormulaVal, and Probability.
    - xi. One-Hot Encoding: Provide a value to encode the data.
- iii) Click the '**Apply**' option.



Component	Console	Summary	Result	Visualization	Properties	DataInsight		*	<u>+</u>
General	Dynam	nic Fields							
Custom Group		Independent Colum	าท	8 checked	-				
		Dependent Column	I	sex	~				
		Iteration		50					
		Depth [1-16]		3					
		Learning Rate		0.1					
		Loss		Logloss	~				
		Evaluation Metric		Logloss	~				
		Seed Value		50					
		Boosting Type		Ordered	~				
÷		Prediction Type		Class	~				
		One-Hot Encoding		2					
							A	\pply	,

- iv) Run the workflow after getting the success message.
- v) The user gets the process status under the 'Console' tab.

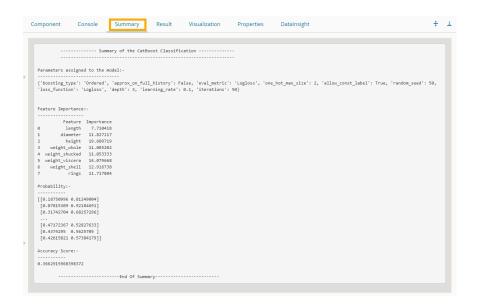


- vi) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
  - c. A new column displaying the predicted values gets added to the result view.



now	10 🗸 en	tries						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValue
1	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	1
1	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1
	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	F
1	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	1
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	F
	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	F
1	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1
	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	F

vii) Click the 'Summary' tab to display the summary.



#### 14.1.4. Tree-Based modeling

The Tree Based Modeling Random Forest can be configured using two algorithm types from the '**Properties**' tab.

Check out the below-given description of the configuration details:

### 14.1.4.1. Random Forest

### 14.1.4.1.1. Random Forest with Classification as Algorithm Type

i) Drag the Random Forest component to the workspace and connect it with a configured data source.





ii) Configure the 'Properties' tab:

### a. Output Information

- i. **Algorithm Type:** Select an algorithm type from the drop-down menu.
  - 1. **Classification**: Select this option if users want to pass the dependent column as the categorical values.
  - 2. **Regression:** Select this option if users want to pass the dependent column as numerical values.
- ii. **Show Probability**: Select an option from the drop-down menu to create a new column for indicating the chance factor involved in the probability.
  - 1. **True:** Select this option to display a new column in the output data with probability values.
  - 2. **False:** Select this option to display any probability value in the output data.
- b. Column Selection
  - i. **Features**: Select input columns from the drop-down list to which the target column needs to compare performing the analysis.
  - ii. **Target Variable**: Select the target column for which the analysis is performed.
- c. New Column Information
  - i. **Predicted Column Name**: Enter a name for the new column containing the predicted values.
  - ii. **Probability Column Name:** Enter a name for the new column containing the probability values.
- d. Model Tuning
  - i. **Enable Validation:** Enable validation as a model tuning option by a checkmark in the given box.
  - ii. **XG Boosting:** Enable validation as a model tuning option by a checkmark in the given box.

### Properties Tab when Model Tunning is not Enabled



Component	Console	Summary	Result	Visualization	Properties	÷ 1
General	Output Inf	ormation				
Properties	Algorithm Ty	pe	Classification	-		
Advanced	Show Probal	bility	True	•		
	Column Se	election				
	Features		4 checked	•	0	
	Target Varial	ble	gender	-	0	
	New Colur	nn Informatior	n			
	Predicted Co	lumn Name	PredictedValu	ies	0	
	Probability C	olumn Name	Probability		0	
	Model Tur	ning				
>	Enable Valid	ation				
	Hyperparam	eter Tuning				
						Apply

### Properties Tab when Validation is Enabled as Model Tuning

Component	Console	Summary	Result	Visualization	Properties		*	<u>+</u>
General	Output Inf	ormation						
Properties	Algorithm Ty	rpe	Classification	-				
Advanced	Show Probal	bility	True	-				
Validation	Column Se	election						
	Features		4 checked	-	0			
	Target Varial	ble	gender	-	0			
	New Colur	mn Informatior	ı					
	Predicted Co	olumn Name	PredictedValue	s	0			
	Probability C	olumn	Probability		0			
	Name							
Þ	Model Tur	ning						
	Enable Valid	ation	•					
						_		
							Apply	′

### Properties Tab when Hyperparameter Tuning is Enabled as Model Tuning

Component	Console	Summary	Result	Visualization	Properties	<u>+</u> <u>+</u>
General	Output Inf	ormation				
Properties	Algorithm Ty	pe	Classification	•		
Advanced	Column Se	election				
	Features		4 checked	-	0	
	Target Variat	ble	gender	•	0	
	New Colur	nn Informatior	n			
	Predicted Co	lumn	PredictedValues		0	
	Name					
<i></i>	Model Tur	iing				
	Hyperparam	eter Tuning				
						Apply



#### Note:

- a. The 'Show Probability' field appears only if, 'Classification' option is selected via the 'Algorithm Type' drop-down menu.
- b. The 'Show Probability' field disappears in the following scenarios:
  - i. If the selected Algorithm Type is Regression
  - ii. If the selected Model Tuning option is Hyperparameter Tuning.
- iii) Click the 'Advanced' tab and configure if required:
  - Advanced Tab when both the Model Tuning options are Disabled a. Tree Pruning
    - i. **No. of Trees:** It is a numerical value that defines the structural size of your tree. The higher number of trees gives you better performance but makes your code slower.
    - ii. **Maximum Depth:** It sets the maximum depth of any node of the final tree keeping the depth count for root node 0. It is an optional field (It is recommended to set Maximum Depth value less than 30 rpart for 32 bit-machines.)
    - iii. Min Sample Split: It indicates a minimum number of observations within a single node for a split to be attempted. The default value for this field is 10.
    - iv. **Min Sample Leaf:** Leaf is the end node of a decision tree. A smaller leaf makes the model more prone to capturing noise in train data.
    - Max Leaf Node: Select an option from the given choices: 'int' or 'None' (The field is optional, and the default option for the field is 'None').
    - vi. **Random State:** This parameter makes a solution easy to replicate. A definite value of random\_state produces the same results if given with the same parameters and training data. The default value for this field is **None**.
    - b. Behavior
      - Criteria: It is an optional field that depends on the selected algorithm type from the 'Properties' tab. The splitting index can be:
        - 1. **Gini:** Select this option to measure inequality among values of randomly chosen elements from a set.
        - 2. **Entropy:** Select this option to measure impurities for exploratory analysis.
      - ii. **Bootstrap:** Select an option from the drop-down menu out of True/False (the default value for this field is '**True**').
- iv) Click the 'Apply' option.

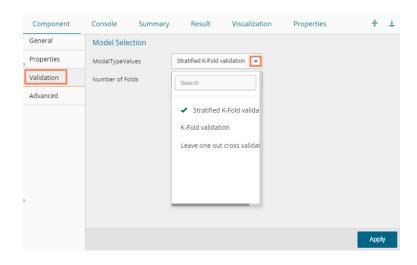


Component	Console	Summary	Result	Visualizatio	on Propertie	s 🕂	<u>+</u>
General	Tree Pruning	]					
Properties	No of Trees	10					
Advanced	Maximum Depth	n Nor	ie				
	Min Sample Spli	t 2					
	Min Sample Lea	f 1					
	Max Leaf Node	Nor	ne				
	Random State	Nor	ie				
	Behavior						
•	Criteria	Gin		-			
	Bootstrap	Tru	е	-			
						Appl	у

Note: The 'Advanced' tab remains the same as displayed when both the model tuning options are disabled or when Validation is enabled.

- v) Click the 'Validation' tab and configure the required fields.
- vi) Click the 'Validation' tab to configure, if it has been enabled from the Component Properties tab. The 'Validation' tab provides multiple options under the 'Model Type Values' drop-down menu. The user can select any one of the available options to configure the Validation tab.
  - a. Model Selection
    - i. Stratified K-fold Validation

The user needs to configure the **'Number of Folds'** fields if the selected Model Type Value is **Stratified K-fold Validation**.



ii. K-fold Validation

The user needs to configure the **'Number of k-folds'** field if the selected Model Type Values are **K-Fold Validation.** 



Component	Console	Summary	Result	Visualizatio	n Properties	<u>+</u>	Ť
General	Model Sele	ection					
Properties	ModalTypeVa	alues	K-Fold validation	•			
Validation	Number of k-	folds	3				
Advanced							
						Apply	

#### iii. Leave One Out Cross-Validation

The user gets to configure no other fields when the selected Model Type Values option is **Leave One Out Cross-Validation**.

Component	Console	Summary	Result	Visualization	Properties	⊥⊤
General	Model Sele	ction				
Properties Validation Advanced	ModalTypeVa	lues	eave one out c	oss validatio <del>.</del>		
						Apply

- Advanced Tab when Hyperparameter Tuning is enabled a. Hyperparameter Tuning
  - i. Maximum evaluation: Provide a numerical value set to indicate the maximum value of model evaluation (The default value for this field is 10).
  - ii. Fit Increment: Provide a numerical value set as the increment to model fitting (The default value for this field is 1).
  - Trial Timeout: Provide a numerical value set for the process timeout (usually in seconds (The default value for this field is 30).
  - Seed: A numerical value set as the initialization state of a pseudo-random number generator (the default value for this field is 123).
- vii) Click the 'Apply' option to configure the 'Advanced' tab (if required).

Component	Console	Summary	Result	Visua	lization	Properties	*	Ŧ
General	Hyperparamet	er Tuning						
Properties	Maximum evaluat	ion 10						
Advanced	Fit Increment	1						
	Trial Timeout	30						
	Seed	123	3					
•								
							Apply	/

viii) Run the workflow after getting the success message.



ix) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- x) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

now	10 🔻 ent	ries							Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	Probability
M	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	М	[0.0, 0.4, 0.6]
M	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	м	[0.16, 0.0, 0.84]
-	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	м	[0.31, 0.1, 0.59]
A.	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	м	[0.0, 0.3, 0.7]
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1	[0.1, 0.9, 0.0]
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1	[0.0, 1.0, 0.0]
	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	F	[0.9, 0.1, 0.0]
	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	F	[0.8, 0.1, 0.1]
1	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	М	[0.0, 0.0, 1.0]
	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	F	[0.85, 0.0, 0.15]

i. Result view when both the Model Tuning options are disabled.

ii. Result view with the 'Validation' option enabled.

how	10 v ent	tries							Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	Probability
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	1	[0.1, 0.1, 0.8]
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	м	[0.27, 0.0, 0.73]
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	м	[0.47, 0.0, 0.53]
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	1	[0.25, 0.25, 0.5]
1	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1	[0.0, 1.0, 0.0]
I.	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1	[0.0, 1.0, 0.0]
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	F	[0.6, 0.3, 0.1]
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	F	[0.8, 0.13, 0.07]
М	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	м	[0.0, 0.1, 0.9]
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	F	[0.9, 0.0, 0.1]

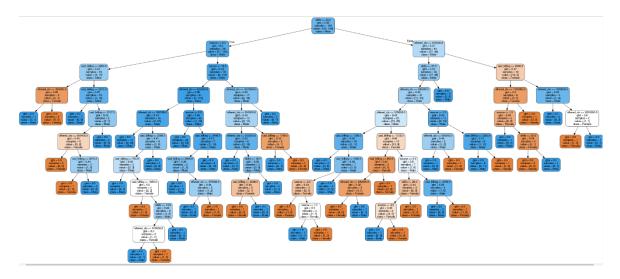
iii. Result view with the 'Hyperparameter Tuning' option enabled



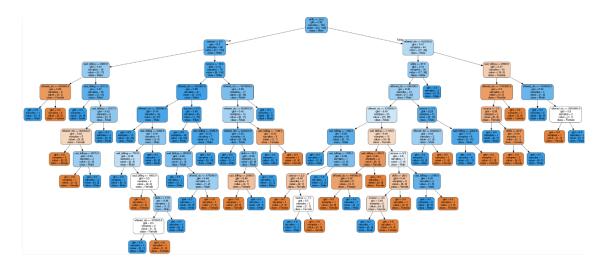
ow	10 🔻 entri	es						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	I
	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	T
	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	М
	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	T
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	М
	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	М
	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1
	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	М

Note: The Probability column displays data in the Array format when Validation is enabled.

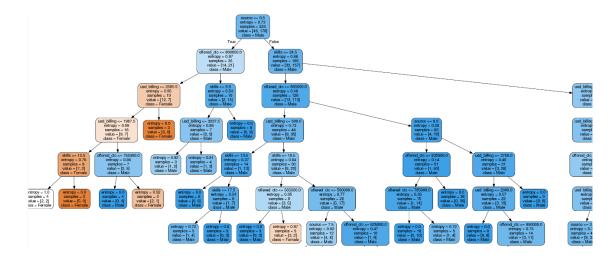
- xi) Click the 'Visualization' tab.
- xii) The Result data gets displayed via the tree chart.
  - a. Visualization when no Model Tuning option is enabled



**b.** Visualization when the 'Validation' option is enabled







c. Visualization when Hyperparameter Tuning is enabled

xiii) Click the 'Summary' tab to open the model summary.

omponent	Console	Summary	Result	Visualization	Properties	4
	Summary of the	e model				
sourc	t Columns billing (ir ce (object) ls (object) ned_ctc (ir					
2.Dependent ( gende	Columns er (object)					
5 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	Classifier(boot max_depth=None, min_impurity_de min_samples_lea min_weight_frac	, max_features='a ecrease=0.0, min_ af=1, min_samples ction_leaf=0.0, n e, random_state=N	uto', max_lea impurity_spl: _split=2, _estimators=:	it=None, 10, n_jobs=None,	,	
4.Feature Imp	portance : [	[('usd_billing',	0.26477), (':	source', 0.19355),	('skills', 0.17194), ('d	offered_ctc', 0.36974
5.Accuracy Sr	core : 1	1.0				

# 14.1.4.1.2. Regression as Algorithm Type for Random Forest

- i) Drag the Decision Tree component to the workspace and connect it to a configured data source.
- ii) Configure the following fields in the 'Properties' tab:
  - a. Output Information
    - i. **Algorithm Type:** Select an algorithm type from the drop-down menu.
      - 1. **Classification**: Select this option if users want to pass the dependent column as the categorical values.
      - 2. **Regression:** Select this option if users want to pass the dependent column as numerical values.
  - b. Column Selection
    - i. **Independent Columns**: Select input columns from the drop-down list to which the target column can be compared to perform the analysis.



- ii. **Dependent Columns**: Select the target column for which the analysis is performed.
- c. New Column Information
  - i. **Predicted Column Name**: Enter a name for the new column containing the predicted values.
- d. Model Tuning
  - i. **Enable Validation:** Enable validation by a checkmark in the given box.
  - ii. **Hyperparameter Tuning:** Enable the Hyperparameter Tuning option by a checkmark in the given box.

Component	Console	Summary	Result	Visualization	Properties	4	<u>+</u>
General	Output In	formation					
Properties	Algorithm Tj	/pe	Regression	-			
Advanced	Column S	election					
	Independen	t Columns	1 checked	•	0		
	Dependent	Columns	Open	•	0		
	New Colu	mn Informatior	i				
	Predicted Co	olumn Name	PredictedValues	3	0		
	Model Tur	ning					
•	Enable Valid	ation (					
	Hyperparam	neter Tuning					
						_	
						A	ply

Note: Other possible scenarios to configure the Properties tab can be when either of the Model Tuning options is enabled.

- iii) Click the 'Advanced' tab and configure if required:
  - Advanced Tab when both the Model Tuning options are Disabled a. Tree Pruning
    - i. **No. of Trees:** It is a numerical value that defines the structural size of your tree. The higher number of trees gives you better performance but makes your code slower.
    - ii. **Maximum Depth:** It sets the maximum depth of any node of the final tree keeping the depth count for root node 0. It is an optional field (It is recommended to set Maximum Depth value less than 30 rpart for 32 bit-machines.)
    - Min Sample Split: It indicates a minimum number of observations within a single node for a split to be attempted. The default value for this field is 10.
    - iv. **Min Sample Leaf:** Leaf is the end node of a decision tree. A smaller leaf makes the model more prone to capturing noise in train data.
    - v. Max Leaf Node: Select an option from the given choices: 'int' or 'None' (The field is optional, and the default option for the field is 'None').
    - vi. **Random State:** This parameter makes a solution easy to replicate. A definite value of random\_state produces the same results if given with the same parameters and training data. The default value for this field is **None**.
    - b. Behavior



i. **Criteria:** It is an optional field that depends on the selected algorithm type from the **'Properties'**.

The available splitting index options are:

- 1. **MSE**
- 2. **MAE**
- ii. **Bootstrap:** Select an option from the drop-down menu out of True/False (the default value for this field is '**True**').

Component	Console	Summary	Result	Visualizatio	on Properties	*	<u>+</u>
General	Tree Pruning	g					
Properties	No of Trees		10				
Advanced	Maximum Dep	th	None				
	Min Sample Sp	lit	2				
	Min Sample Le	af	1				
	Max Leaf Node		None				
	Random State		None				
	Behavior						
	Criteria		MSE	•			
ŕ	Bootstrap		True	•			
						Apply	y

Note: The Advanced tab remains the same when 'Validation' is enabled.

- iv) Click the 'Validation' tab to configure, if it has been enabled from the Component Properties tab. The 'Validation' tab provides multiple options under the 'Model Type Values' drop-down menu. The user can select any one of the available options to configure the Validation tab.
  - a. Model Selection
    - i. K-fold Validation

The user needs to configure the 'Number of k-folds' field if the selected option for the 'Model Type Values' is K-Fold Validation.

Component	Console	Summary	Result	Visualizatio	on Properties	<u>1</u>	Ť
General	Model Sele	ection					
Properties	ModalTypeV	alues	K-Fold validation	•			
Validation	Number of k	-folds	3				
Advanced							
						Appl	y

#### ii. Leave One Out Cross-Validation

The user gets to configure no other fields when the selected Model Type Values option is **Leave One Out Cross-Validation**.

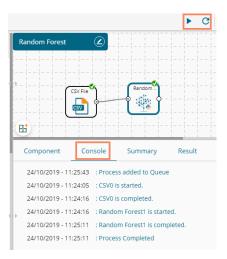
Component	Console	Summary	Result	Visualization	Properties	<u>+</u>	Ť
General	Model Sele	ection					
Properties	ModalTypeVa	alues	eave one out cr	oss validatio. <del></del>			
Validation							
Advanced							
						Appl	y



- Advanced Tab when Hyperparameter Tuning is Enabled a. Hyperparameter Tuning
  - i. Maximum evaluation: Provide a numerical value set to indicate the maximum value of model evaluation (The default value for this field is 10).
  - ii. Fit Increment: Provide a numerical value set as the increment to model fitting (The default value for this field is 1).
  - Trial Timeout: Provide a numerical value set for the process timeout (usually in seconds (The default value for this field is 30).
  - iv. Seed: A numerical value set as the initialization state of a pseudo-random number generator (the default value for this field is 123).

	Component	Console	Summary	Result	Visualiz	ation	Properties	+++++++++++++++++++++++++++++++++++++++	Ŧ
	General	Hyperparam	eter Tuning						
Þ	Properties	Maximum evalu	ation 10						
	Advanced	Fit Increment	1						
		Trial Timeout	30						
		Seed	123						
								Appl	ly 🛛

- v) Click the **'Apply'** option after configuring the Properties, Advanced (if required), and validation (if enabled) tabs.
- vi) Run the workflow after getting the success message.
- vii) The 'Console' tab opens.



- viii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
    - i. Result View when both the Model Tuning options are disabled



ow 10 🔻	entries						Search:	
Timestamp	Open	High	Low	Close	BTC	currency	WeightedPrice	PredictedValues
499155260	296127	296558	296016	296540	1.159	343244.138	296257.672	296277.042
499155320	296539	296769	296060	296679	11.116	3295332.006	296462.514	296590
499155380	296060	296090	296060	296060	5.527	1636491.185	296063.836	296253.929
499155440	296060	296260	296015	296015	8.414	2491620.368	296125.668	295878.567
499155500	296361	296540	296155	296155	3.993	1183291.629	296340.786	296290.3
499155560	296360	296360	296060	296060	4.113	1218324.398	296216.135	296253.929
499155620	296360	296460	296014	296450	24.563	7273386.537	296110.238	296249.658
499155680	296360	296671	296001	296001	10.75	3186951.403	296460.003	296163.288
499155740	296279	296500	296093	296150	7.031	2083921.622	296396.323	296390.5
499155800	296150	296231	296122	296122	1.372	406497.426	296172.988	296063.5

### ii. Result view when the 'Validation' option is enabled

	Component	Console	e Su	immary	Result	Vis	sualization	Properties	<u>+</u> <u>⊥</u>
	Show 10 🔻	entries						Search:	
Þ	Timestamp	Open	High	Low	Close	BTC	currency	WeightedPrice	PredictedValues
	1499155260	296127	296558	296016	296540	1.159	343244.138	296257.672	610999
	1499155320	296539	296769	296060	296679	11.116	3295332.006	296462.514	610999
	1499155380	296060	296090	296060	296060	5.527	1636491.185	296063.836	610999
	1499155440	296060	296260	296015	296015	8.414	2491620.368	296125.668	610999
	1499155500	296361	296540	296155	296155	3.993	1183291.629	296340.786	610999
	1499155560	296360	296360	296060	296060	4.113	1218324.398	296216.135	610999
	1499155620	296360	296460	296014	296450	24.563	7273386.537	296110.238	610999
	1499155680	296360	296671	296001	296001	10.75	3186951.403	296460.003	610999
	1499155740	296279	296500	296093	296150	7.031	2083921.622	296396.323	610999
Þ	1499155800	296150	296231	296122	296122	1.372	406497.426	296172.988	610999
	Showing 1 to 10 c	f 5,556 entr	ies			F	Previous 1	2 3 4	5 556 Next

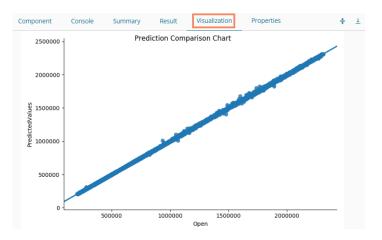
### iii. Result view when 'Hyperparameter Tuning' is enabled.

now 10 🔻	entries			Result			Search:	
Timestamp	Open	High	Low	Close	BTC	currency	WeightedPrice	PredictedValues
1499155260	296127	296558	296016	296540	1.159	343244.138	296257.672	303322.419
1499155320	296539	296769	296060	296679	11.116	3295332.006	296462.514	303322.419
1499155380	296060	296090	296060	296060	5.527	1636491.185	296063.836	303322.419
1499155440	296060	296260	296015	296015	8.414	2491620.368	296125.668	303322.419
1499155500	296361	296540	296155	296155	3.993	1183291.629	296340.786	303322.419
1499155560	296360	296360	296060	296060	4.113	1218324.398	296216.135	303322.419
1499155620	296360	296460	296014	296450	24.563	7273386.537	296110.238	303322.419
1499155680	296360	296671	296001	296001	10.75	3186951.403	296460.003	303322.419
1499155740	296279	296500	296093	296150	7.031	2083921.622	296396.323	303322.419
1499155800	296150	296231	296122	296122	1.372	406497.426	296172.988	303322.419

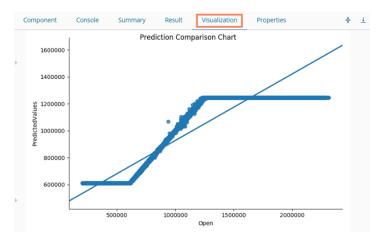


Note: The Probability column is displayed in the Array format while enabling the **'Validation'** option.

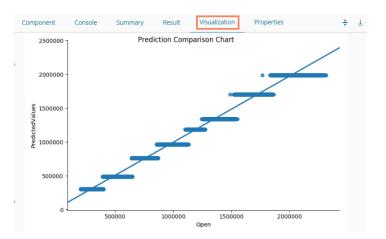
- ix) Click the 'Visualization' tab.
- x) The Result data gets displayed via the tree chart (The following visualization displays result in data when no Model Tuning option is enabled).
  - **a.** Visualization tab when no Model Tuning option is enabled.



**b.** Visualization tab when Validation is enabled.



c. Visualization tab when Hyperparameter Tuning is enabled.





# 14.1.4.2. Decision Tree 14.1.4.2.1. Decision Tree with Classification Model Type

- i) Drag the Decision Tree component to the workspace and connect it to a configured data source.
- ii) Click the Decision Tree algorithm component.



iii) The 'Custom Group' tab opens.

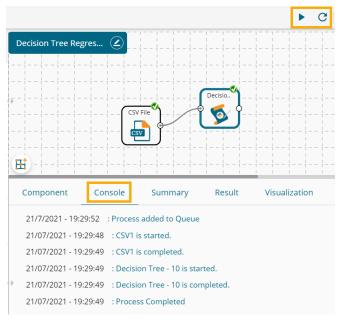
### a. Dynamic Fields

- i. Type of Model: Select a model out of the given choices.
  - **1. Classification**: Select this option if users want to pass the dependent column as the categorical values.
  - **2. Regression:** Select this option if users want to pass the dependent column as numerical values.
- ii. **Dependent Column**: Select the target column on which the regression analysis gets applied.
- iii. **Independent Column**: Select the required input columns against which the regression analysis gets applied to the target column.
- iv. Criterion ['mse' for Regresssion]: Select an option to measure inequality among values of randomly chosen elements from a set out of gini, entropy, or mse.
- Maximum Depth: It sets the maximum depth of any node of the final tree keeping the depth count for root node 0. It is an optional field (It is recommended to set Maximum Depth valueless than 30 rpart for 32 bit-machines).
- vi. Minimum Sample for Split: It indicates a minimum number of observations within a single node for a split to be attempted. The default value for this field is 10.
- vii. Minimum Sample for Leaf node: It indicates the minimum number of the sample for a leaf node.
- viii. Maximum Leaf Nodes: It indicates the maximum number of the sample for a leaf node.
- ix. Random Seed: A random seed (or seed state, or just seed) is a number (or vector) used to initialize a pseudorandom number generator.
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	DataInsight	÷	<u>+</u>
General Custom Group	Dynar	mic Fields						
		Type Of Model		Classification		~		
		Dependent Colu	mn	sex		~		
		Independent Col	umn	8 checked		-		
		Criterion ["mse"	for Regression]	gini		~		
		Maximum Depth		10				
		Minimum Sampl	e For Split	10				
		Minimum Sampl	e For Leaf	10				
		Node						
		Maximum Leaf N	lodes	10				
		Random Seed		10				
							Арр	ły

- v) Run the workflow after getting the success message.
- vi) The user gets the process status under the 'Console' tab.



- vii) Follow the below-given steps to display the Result view:
  - Click the dragged algorithm component onto the workspace. a.
  - A new column displaying the predicted values gets added to the result b. view.

how	10 v en	tries					Se	arch:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValue
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	F
м	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	1
м	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	F
1	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
1	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	F
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	м
м	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	м



viii) Click the 'Summary' tab to display the summary.

Component	Console	Sumn	hary	Result	Visualization	Properties	DataInsig
	Summary Of	F The Decis	ion Tree	Classifier			
Call:-							
DecisionTreeCla	ssifier(cla	ass_weight=	None, cr:	iterion='gin	i', max_depth=10,		
max	_features=1	None, max_1	eaf_node:	5=10,			
				mpurity_spli	t=None,		
		eaf=10, mir					
	_weight_fra itter='best		=0.0, pro	esort=False,	random_state=10,		
Prediction Prob	ability:-						
	-						
[0.27576602 0.8	402439 0.6	50144928	. 0.0484	115 0.04841	15 0. ]		
Confusion Matri							
[[ 276 108 92							
[ 100 961 28 [ 199 200 112							
Classification	Report:-						
р	recision	recall f	1-score	support			
F	0.48	0.21	0.29	1307			
I	0.76	0.72	0.74	1342			
м	0.48	0.74	0.58	1528			
micro avg	0.57	0.57	0.57	4177			
macro avg	0.57	0.56	0.54	4177			
weighted avg	0.57	0.57	0.54	4177			
Accuracy:-							
0.5664352406033	930						
0.0004352406055	0.35						

14.1.4.2.2. Decision Tree with Regression Model Type

- i) Drag the Decision Tree component to the workspace and connect it to a configured data source.
- ii) Click the Decision Tree algorithm component.



- iii) The 'Custom Group' tab opens. Configure the Dynamic Fields with Regression selected as the Model type.
  - **a.** Make sure that the selected Criterion option is 'mse'

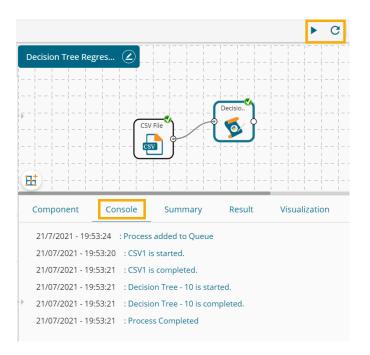
Note: All the other fields configuration remains the same.

iv) Click the '**Apply**' option.



Component	Console	Summary	Result	Visualization	Properties	DataInsight	+++++++++++++++++++++++++++++++++++++++	<u>+</u>
General Custom Group	Dynar	nic Fields						
custom broup	_	Type Of Model		Regression		~		
		Dependent Colu	imn	length		~		
		Independent Co	lumn	8 checked		•		
		Criterion ["mse"	for Regression]	mse		~		
		Maximum Dept	h	10				
		Minimum Samp	le For Split	10				
		Minimum Samp	le For Leaf Node	10				
		Maximum Leaf	Nodes	10				
÷		Random Seed		10				
							Appł	y 🛛

- v) Run the workflow after getting the success message.
- vi) The user gets the process status under the '**Console**' tab.



- vii) Follow the below-given steps to display the Result view:
  - **a.** Click the dragged algorithm component onto the workspace.
  - **b.** Click the '**Result**' tab.
  - **c.** A new column displaying the predicted values gets added to the result view.



10 v   entries   Search:										
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues	
И	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	0.462	
4	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	0.347	
	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	0.521	
4	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	0.462	
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	0.347	
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	0.397	
	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	0.521	
	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	0.521	
1	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	0.462	
	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	0.572	

# viii) Click the '**Summary**' tab to display the summary.

Component	Console	Summary	Result	Visualization	Properties	DataInsight
			-			
	Summary	Of The Decision Tre				
Call:-						
DecisionTree	Regressor(cr	iterion='mse', max	depth=10, max	features=None,		
		s=10, min_impurity				
n	in_impurity_	split=None, min_sam	nples_leaf=10,			
		plit=10, min_weight				
F	presort=False	, random_state=10,	splitter='bes	t')		
Feature Impo	ortance:-					
Fe	ature Impor	tance				
0		99648				
1 dia	meter 0.	00000				
2 F	eight 0.	00352				
3 weight_	whole 0.	00000				
4 weight_sh		00000				
5 weight_vi	scera 0.	00000				
6 weight	shell 0.	00000				
7	rings 0.	00000				
Mean Absolut	e Error (MAE	):-				
0.02						
Mean Squared	Error (MSE)	:-				
	· · · · · · · · · · · · · · · · · · ·					
0.001						
	uared Error	(RMSE):-				
0.0253						
		End Of Summary				

## 14.1.5. Anomaly Detection

Anomaly Detection is a data mining algorithm type that identifies data points, events, and/or observations that deviate from the normal behavior of a dataset.

The Anomaly Detection tree node contains two sub-algorithms as shown in the image given below:

- 1. Interquartile Range
- 2. Isolation Forest
  - Anomaly Detection

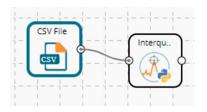
🚯 Interquartile Range

🚯 Isolation Forest



# 14.1.5.1. Interquartile Range

i) Drag the Interquartile algorithm component to the workspace and connect it with a configured data source.

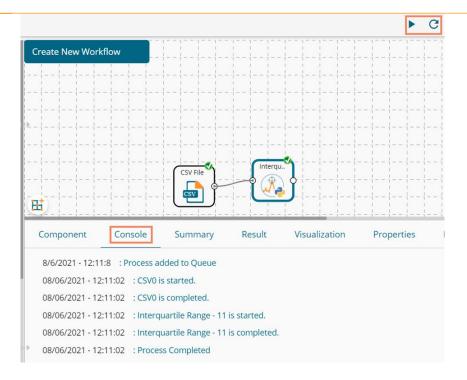


- ii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Select a feature column.
  - b. Set the fence coefficient.
  - c. Select an option to either show or remove outliers.
- iii) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	Ŧ	Ť
General Custom Group	Dynar	nic Fields						
		Feature		ozone_reading		~		
		Fence Coefficien	t	1.5				
>		Show/Remove O	utlier	True				
								_
							Apply	y

- iv) Run the workflow after getting the success message.
- v) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vi) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

how 10												Search:			
Month	Day_of_month	Day_of_week	ozone_reading	pressure_height	Wind_speed	Humidity	Temperature_Sandburg	Temperature_ElMonte	Inversion_base_height	Pressure_gradient	Inversion_temperature	Visibility	OutlierDete		
1	1	4	3.01	5480.0	1	20.0	54		5000.0	-15.0	30.56	200	FALSE		
1	2	5	3.2	5660.0	6		38			-14.0		300	FALSE		
1	3	6	2.7	5710.0	4	28.0	40		2693.0	-25.0	47.66	250	FALSE		
1	4	7	5.18	5700.0	3	37.0	45		590.0	-24.0	55.04	100	FALSE		
1	5	1	5.34	5760.0	3	51.0	54	45.32	1450.0	25.0	57.02	60	FALSE		
1	6	2	5.77	5720.0	4	69.0	35	49.64	1568.0	15.0	53.78	60	FALSE		
1	7	3	3.69	5790.0	6	19.0	45	45.4	2631.0	-33.0	54.14	100	FALSE		
1	8	4	3.09	5790.0	3	25.0	55	52.7	554.0	-28.0	64.76	250	FALSE		
1	9	5	5.76	5700.0	3	73.0	41	48.02	2083.0	23.0	52.52	120	FALSE		
1	10	6	6.94	5700.0	3	59.0	44		2654.0	-2.0	48.38	120	FALSE		

vii) Click the '**Visualization**' tab. The Result data gets displayed via the Boxplot chart displaying the outlier values. It is also known as the Prediction Comparison chart. The following chart shows predicted values by experience.



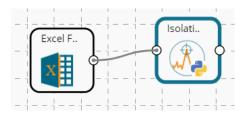


viii)	Click the	'Summary'	tab to	open t	the model	summary.
-------	-----------	-----------	--------	--------	-----------	----------

Component	Console	Summary	Result	Visualization	Properties	DataInsight
		InterQuartile Ran				
First Quartil	e Occured at:					
nan						
Third Quartil	e Occured at:					
nan						
Fence Coeffie	cient Value:					
1.5						
Lower Fence \						
nan Upper Fence V	(a)					
nan						
Total Number	of Outliers:					
е						
Median of the	Column Selec	ted:				
nan						
Standard Devi	ation of the	Column Selected:				
7.90289659727	7894					
Number of Val	ues Considere	d:				
366						
		End Of Summary				

# 14.1.5.2. Isolation Forest

- i) Drag the Isolation Forest algorithm component to the workspace and connect it with a configured data source.
- ii) Click the Isolation Forest component.





- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Select a feature column.
  - b. Set the contamination value (the default value for this field is 0.1).
  - c. Set the value for the Estimators.
  - d. Set value for the Maximum Samples (the default value for this field is 50).
  - e. Set a Seed Value (the default value for this field is 5).
  - f. Select an option to either show or remove outliers.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Propertie	s DataInsight	+	<u>+</u>
General	Dynamic	: Fields						
	Fe	eature Columns		1 checked	-			
	C	ontamination(0., 0.5)		0.1				
	E	stimators		100				
	M	laximum Samples		50				
	S	eed Value		5				
÷								
							Apply	/

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.



	Component Console Summary	Result Visualizat	tion Properties DataInsight 🕂 🕹
	Show 10 v entries		Search:
÷	Spanned_days	Depth	OutlierDetected
	1	1000	YES
	1.003	1000.86	YES
	1.006	1001.682	YES
	1.007	1002.495	YES
	1.01	1003.322	YES
	1.013	1004.197	YES
	1.015	1005.062	YES
	1.017	1005.914	YES
	1.018	1006.758	YES
Þ	1.021	1007.6	YES
	Showing 1 to 10 of 10,213 entries	Previou	is 1 2 3 4 5 1022 Next

Component	Console	Summary	Result	Visualization	Properties	DataInsight	+
	Summary Of	The Isolation Fo	rest				
Call:-							
IsolationF	orest(behaviour=	'old', bootstrap=	False, contar	ination=0.1,			
ma	x_features=1.0,	max_samples=50, n	_estimators=1	.00, n_jobs=None,			
	ux_features=1.0, undom_state=5, ve		_estimators=1	.00, n_jobs=None,			
ra	indom_state=5, ve		_estimators=1	00, n_jobs=None,			
	indom_state=5, ve		_estimators=1	00, n_jobs=None,			
ra	undom_state=5, ve Outliers:-		_estimators=1	00, n_jobs=None,			
ra	indom_state=5, ve		_estimators=1	00, n_jobs=None,			
ra	undom_state=5, ve outliers:-  Depth		_estimators=2	.00, n_jobs=None,			
Group By C	undom_state=5, ve outliers:-  Depth		_estimators=:	00, n_jobs=None,			

# 14.1.6. Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

The Clustering algorithm in the Python Workspace contains the following algorithms:

- 1. K-Means++
- 2. H-Clustering
- 3. K-Means
- 4. Optimal K Value

4	🕌 Clustering
	K-Means++
	🔣 H-Clustering
	K-Means
	🔣 Optimal K Value



## 14.1.6.1. K-Means ++

The K-Means ++ algorithm ensures a smarter initialization of the centroids and improves the quality of the clustering. Apart from initialization, the rest of the algorithm is the same as the standard K-means algorithm. That is K-means++ is the standard K-means algorithm coupled with a smarter initialization of the centroids.

- i) Drag the K-Means++ algorithm component to the workspace and connect it with a configured data source.
- ii) Click the K-Means++ algorithm component.

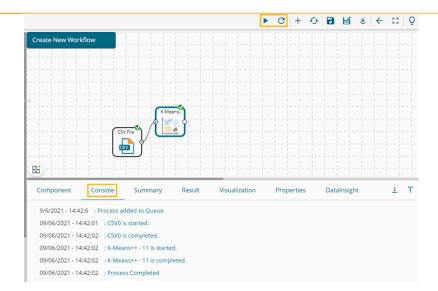


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Feature Column: Select the feature columns from the input data set
  - b. Seed Value: A random seed (or seed state, or just seed) is a number (or vector) used to initialize a pseudorandom number generator. Set the initial seed value (the default seed value for this field is 500).
  - c. Number of Iteration: Set the number of Iteration to run the data set repetitively (the default value for this field is 10).
  - d. Maximum Iteration: Enter the number of iterations allowed for discovering clusters. (The default value for this field is 1000).
  - e. Number of Clusters: Set the number of clusters to be created.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>1</u>	Ť
General	Dynamic	: Fields						
Custom Group		eature Column		5 checked	•			
	S	eed Value		500				
	N	umber Of Iteration		10				
>	M	laximum Iteration		1000				
	N	umber Of Cluster		3				
							Apply	y

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

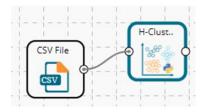
now 10	✓ entries				Search:	
Number	sepal_length	sepal_width	petal_length	petal_width	species	ClusterNumber
91	5.5	2.6	4.4	1.2	versicolor	2
92	6.1	3	4.6	1.4	versicolor	2
93	5.8	2.6	4	1.2	versicolor	2
94	5	2.3	3.3	1	versicolor	2
95	5.6	2.7	4.2	1.3	versicolor	2
96	5.7	3	4.2	1.2	versicolor	2
97	5.7	2.9	4.2	1.3	versicolor	2
98	6.2	2.9	4.3	1.3	versicolor	2
99	5.1	2.5	3	1.1	versicolor	2
100	5.7	2.8	4.1	1.3	versicolor	2



-						-			Su	mm	ary	/	of	t	he	К	-M	ear	ns-	++	c1	u:	sti	er	in	3											
Ca	11.																																				
Ca.																																					
КМ					~ ~		+1				+ -		-									4					2				·						
KPR																					, p																
																								om	pu	ce.	_a.	15	ca	nc	es	-	au	to	,		
	r	ar	ac	om	_s	τ	at	e	=N	on	e,	τ	01	=0	.0	00:	1,	Ve	eri	00:	se=	0	)														
cl	1161			10	an																																
	131	e		ic																																	
								NI.		ho			<b>a</b> n	-1	1	0.00	-+1			200	al		i di	+h		101		1	1.0		+h			+ -	1	uri.	d+1
cl			-Ma		ba			14	um	De		2	ep	a 1		en	Su		21	=p			IU	Car	2	Je.	La.		1C	ing	cii		pe	La.	-	W II	uu
0	15 (	.cı	141	A111	De				2	-	5					5.6	20	6				5	1	10					1	4	64					0.	24
1											5																									2.1	
2									1	5.	5					5.9	931	b				2	. /	10					4	- 2	60					1.	32
Cl	ust	- 01	. i i	10	N	10	ct	0	<b>.</b>																												
			-	6		Ĩ																															
[0]	0	0	a	0	0		a	a	0	a	0	a	a	a	a	a	a	a	0	0	0	a	a	0	0	0	a	a	a	a	a	0	0	0	a	a	a
																					2																
																					2																
																					1																
			1	1	1		1	T	1	1	1	1	1	1	T	1	1	1	T	1	T	T	1	1	1	T	1	1	1	1	1	1	T	1	1	1	T
1	1]																																				
Co					~	0	4	0	1	c+	<u></u>	~		÷																							
CO	51.6		a	.e	2	0		-	Tu	St	en	-	en	Le																							
	E ar		-01			7	-	-	_		6			4.4							ð.2		41														
	[2:																						-				. 1										
				1.1																	5.5				2.1	920	2]										
				. /	5.	2	96	2	5.	93	0	1	۷.	//	0		+	208	9	8	1.3	21	0]														
		of.	SI	111	ar	0	d	D	is	ta	nci		0	f	Sa	mn	le	. 1	to	+1	hei	r	c	10	50	:+	C	1	<t.< td=""><td>er</td><td>C</td><td>en</td><td>te</td><td></td><td>_</td><td></td><td></td></t.<>	er	C	en	te		_		
Su	n c													•										~ ~							~						
Su	n c																					-												-			

#### 14.1.6.2. H-Clustering

- i) Drag the H-Clustering algorithm component to the workspace and connect it with a configured data source.
- ii) Click the H-Clustering algorithm component.

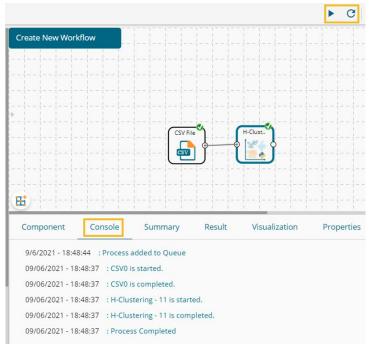


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Feature Columns: Select the feature columns from the input dataset
  - b. Number of Clusters: Set the number of clusters to be created
  - c. Affinity: Select an affinity option out of the given choices (Euclidean, l1, l2, manhattan, cosine, precomputed)
  - d. Linkage: Select a linkage option out of the given choices (average, ward, complete)
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General Custom Group	Dynam	nic Fields						
		Feature Columns		5 checked	•			
		Number of Clusters		5				
		Affinity		euclidean	~			
×		Linkage		average	~			
							Appl	y

- v) Run the workflow after getting the success message.
- The Console tab opens displaying the step-by-step completion of the process. The completion vi) of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

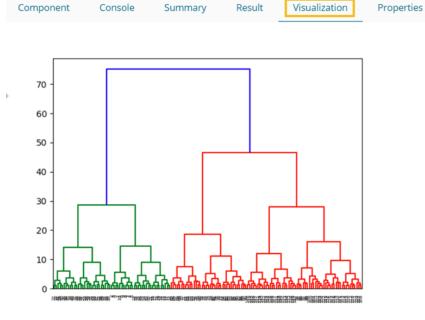
10 10	✓ entries				Search:	
Number	sepal_length	sepal_width	petal_length	petal_width	species	ClusterNumber
1	5.1	3.5	1.4	0.2	setosa	1
2	4.9	3	1.4	0.2	setosa	1
3	4.7	3.2	1.3	0.2	setosa	1
1	4.6	3.1	1.5	0.2	setosa	1
5	5	3.6	1.4	0.2	setosa	1
5	5.4	3.9	1.7	0.4	setosa	1
7	4.6	3.4	1.4	0.3	setosa	1
3	5	3.4	1.5	0.2	setosa	1
9	4.4	2.9	1.4	0.2	setosa	1
10	4.9	3.1	1.5	0.1	setosa	1



DataInsight

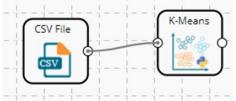
omponent	Conso	le Sumn	nary Re	sult Visu	alization	Properties	DataInsight	÷	
	Summar	y of the Hiera	rchical Clust	ering					
Call:-									
Call:-									
	vaClustanin	g(affinity='eu	clidean' com	nute full tree	-'auto'				
ABBIOMELACI		ty=None, linka			auco,				
		=5, pooling_fu							
	n_crusters		ne= ueprecace	.u )					
Cluster Mea	ans:-								
	Number	sepal_length	sepal_width	petal_length	petal_width				
ClusterNumb	ber								
0	76.0	5.905405	2.729730	4.245946	1.316216				
1	15.0	5.037931	3.458621	1.468966	0.248276				
2	110.5	6.425000	2.912500	5.343750	1.887500				
3	43.5	5.303571	3.260714	2.232143	0.539286				
4	138.5	6.575000	3.012500	5.441667	2.016667				
		End of Su	mmary						

#### ix) Click the 'Visualization' tab to open the result view via chart.



#### 14.1.6.3. K-Means

- x) Drag the K-Means algorithm component to the workspace and connect it with a configured data source.
- xi) Click the K-Means algorithm component.



- xii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Feature Columns: Select the feature columns from the input data set



- b. Seed: A random seed (or seed state, or just seed) is a number (or vector) used to initialize a pseudorandom number generator. Set the initial seed value (the default seed value for this field is 10).
- c. Number of Iteration: Set the number of Iteration to run the data set repetitively (the default value for this field is 10).
- d. Maximum Iteration: Enter the number of iterations allowed for discovering clusters. (The default value for this field is 20).
- e. Number of Clusters: Set the number of clusters to be created (The default value for this field is 5).
- xiii) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>1</u>	Ť
General Custom Group	Dynai	mic Fields						
		Feature Columns		5 checked		•		
		Seed		10				
		Number of Iterat	ion	10				
•		Maximum Numb	er of Iteration	20				
		Number of Cluste	ers	5				
							Appl	y

- xiv) Run the workflow after getting the success message.
- xv) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



xvi) Follow the below-given steps to display the Result view:



- a. Click the dragged algorithm component onto the workspace.
- b. Click the 'Result' tab.

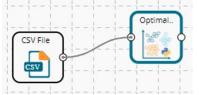
ow 10 🔹	<ul> <li>entries</li> </ul>				Sea	irch:
Number	sepal_length	sepal_width	petal_length	petal_width	species	ClusterNumber
	5.1	3.5	1.4	0.2	setosa	3
2	4.9	3	1.4	0.2	setosa	3
3	4.7	3.2	1.3	0.2	setosa	3
4	4.6	3.1	1.5	0.2	setosa	3
5	5	3.6	1.4	0.2	setosa	3
6	5.4	3.9	1.7	0.4	setosa	3
7	4.6	3.4	1.4	0.3	setosa	3
3	5	3.4	1.5	0.2	setosa	3
9	4.4	2.9	1.4	0.2	setosa	3
10	4.9	3.1	1.5	0.1	setosa	3

	Sum	m 2	-		f	+	ha	v.				c	1	c+4	-	ind													
											12						5												
Call:-																													
KMeans(algorith	nm='a	ut	0	,	C	op	y_:	×='	Tri	ue,	, :	in	it	='}	<-1	nea	an	s+-	۲.	, 1	na	<	it	er	=2	0,			
n_clusters:	=5, n	_j	ni	it	=10	а,	n	j	ob	s=1	Vor	ne	, ;	pre	eco	omp	but	te_	_d:	is	tar	nce	es	='a	au	to	٠,		
random_stat	e=No	ne	,	to	01	=0	.0	00	1,	V	erl	bo	se	=0)	)														
Cluster Means:																													
ciuster means.																													
	Numb	or		=	en:	-1	1.	en	-+1	h	-	en	-1	147	d	th		het	ta i		1.01		⊦h		ne'	+ a		wi.	dt
ClusterNumber	TYCHILD	CI		20	-P			e ing	Su		34	chi			Lu		. '	)e	La.	-	Tei	18		1	Je.	La.		N II	
0	104					6	2	361	56	7		,		866	56	67				-	286	200	aa			1	7	66	66
1	45									2				220							37							93	
2	135							870				2.983871 5.												96					
3	15									7				450							47							46	
4	74									9																		43	
Clustering Vect	or:-																												
[3 3 3 3 3 3 3	3 3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	1	1
1111111																													
4 4 4 4 4 4 4																													
0000000	0 2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2 2]																													
Coordinates of	Clus	te	r	C	ent	te	rs	: -																					
[104.500	6.2	37			2.1	86	7	:	5.0	886	Э		1.3	767	7]														
[45.000	5.3	55	5		3.	22	1		2.3	372	2		a.:	593	3]														
[135.00	90			4	5.	58	7	:	2.9	984	4	1	5.4	468	3	1	L.9	997	7]										
[15.50	5.0	27		-	з.4	45	0		1.4	473	3		ð.:	247	7]														
[74.500	5.9	70	•		2.7	75	7		4	297	7		1.3	343	3]														
Sum of Squared	Dist	ar	ice	25	0	f	Sai	np	le	s t	to	t	he:	ir	C	10:	se	st	C	Lu	ste	er	C	en	te	r:	_		

#### 14.1.6.4. Optimal K Value



- i) Drag the Optimal K Value algorithm component to the workspace and connect it with a configured data source.
- ii) Click the Optimal K Value algorithm component.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.a. Feature Columns: Select the feature columns from the input data set
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u> 7	Ť
General Custom Group	Dynamic	Fields ature Columns		5 checked	×			
							Apply	1

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



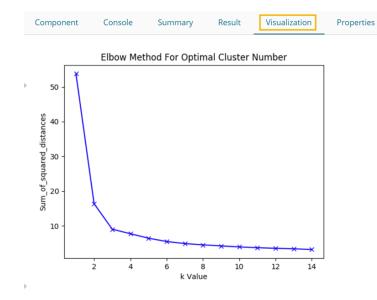
vii) Follow the below-given steps to display the Result view:a. Click the dragged algorithm component onto the workspace.



#### b. Click the 'Result' tab.

10 v	entries			Search:	
Number	sepal_length	sepal_width	petal_length	petal_width	species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
В	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

#### viii) Click the 'Visualization' tab to open the result data in the chart form.



# 14.1.7. Dimensionality Reduction

The Dimensionality Reduction algorithm has the following sub-types in the Python Workspace as given below:

Dimensionality Reduction

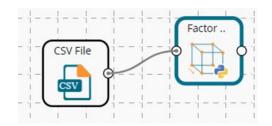
Ex Factor Analysis

- LDA Feature Selection
- Principle Component Analysis

# 14.1.7.1. Factor Analysis



- i) Drag the Factor Analysis component to the canvas and connect it to a configured data source to get the data.
- ii) Click the Factor Analysis algorithm component to get the configuration fields.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Feature Columns: Select required columns from the input data source against which the analysis gets applied.
  - b. Number of Components: Provide the number of components to be taken (the default number for this field is 1).
  - c. Tol Value: Set the stopping tolerance value for log-likelihood increase. Datatype by default is float (the default value for this field is 1e-2 or 0.01)
  - d. Maximum Iterations: Set value for the maximum iterations allowed to the selected data source (the default value for this field is 100).
  - e. SVD Method: Select an option out of the given choices auto/full/randomized/arpack
  - f. Random State: Provide random state value (the default value for this field is 5)
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	4 †	<u>1</u>
General	Duparr	nic Fields						
Custom Group	Dynan							
		Feature Column		8 checked	•			
		Number Of Compo	onents	1				
		Tol Value		0.01				
		Maximum Iteration	15	100				
		SVD Method		randomized	~			
		Random State		5				
Þ								
							40	olu
							Ap	ріу

- v) Run the Workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
- viii) Click the dragged algorithm component onto the workspace.
- ix) Click the 'Result' tab.

now	10 🗸 ent	ries				Sea	arch:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	ring
M	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15
M	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9
M	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16
M	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19



Component	Console	Sum	mary	Result	Visualization	Properties	DataInsight			
Call:-										
FactorAnalysis(c noise_va tol=0.01	riance_in			ax_iter=100, r e=5, svd_metho						
Uniquenesses:-										
length dia	neter	height we	ight_whole	weight_shuci	ked weight_visce	ra \n0 0.13645	8 0.136187 0.323128	0.002424	0.061 <mark>4</mark> 83	0.06493
weight_shell 0 0.087698	rings 0.706254									
Loadings:-										
length diam	eter h	eight wei	ght_whole	weight_shucke	ed weight_viscer	a \n0 0.92927	0.929415 0.822722	0.998787	8,968771	0.966988
weight_shell 0 0.955145	rings 0.541983									
Covariance:-										
		diameter			weight_shucked	\nlength	1.000000 0.863677	8.764531	0.928143	0.900249
		1.000000		0.928288	0.900390					
		0.928288		1.000000	0.967596					
weight_shucked				8.967596	1.000000					
weight_viscera				0.965815	0.936798					
		0.8987726		0.953987	0.925316					
		8.583728		0.541326	0.525058					
	unicht uf	scera wei	oht shall	rings						
length		98593	0.887587							
diameter		98734		0.503728						
height		95562	0.785819							
		95562	0.785819							
weight_whole										
weight_shucked		36790	8.925316							
weight_viscera		00000	0.923614							
weight_shell rings		23614 24091	1.000000							
Average Log-Like										
-4.7819										

# 14.1.7.2. LDA Feature Selection

- i) Drag the LDA Feature Selection component to the canvas and connect it to a configured data source to get the data.
- ii) Click the LDA Feature Selection algorithm component to get the configuration fields.

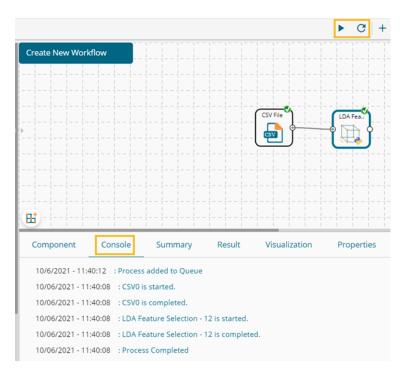
·	 
CSV File	LDA Fea.
· -	

- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Independent Columns: Select the required input columns against which the analysis gets applied to the target column.
  - b. Dependent Column: Select the target column on which the regression analysis gets applied.
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>1</u>	Ť
General Custom Group	Dynam	ic Fields						
	_	Independent Colum	in	8 checked	•			
		Dependent Column		sex	~			
>								
							Apply	/

- v) Run the Workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result'** tab.

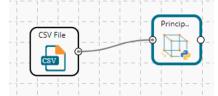
Component Console Summary	Result Visualization	Properties	DataInsight 🕂 🛓
Show 10 👻 entries			Search:
LDA1	LDA2		
0.07	0.985		
1.671	0.14		
0.201	-1.112		
0.329	-0.244		
1.733	0.324		
1.474	-0.135		
-0.856	-0.276		
-0.465	0.186		
0.614	-0.46		
-0.985	0.268		
Showing 1 to 10 of 4,177 entries	Pre	vious 1 2	3 4 5 418 Next



Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>1</u>	Ť
		The LDA						
Call:-								
LinearDiscri		<pre>(n_components=Non , store_covarianc</pre>		ne, shrinkage=None, =0.0001)				
	End OF	Summary						

# 14.1.7.3. Principle Component Analysis

- i) Drag the Principle Component Analysis component to the canvas and connect it to a configured data source to get the data.
- ii) Click the Principle Component Analysis algorithm component to get the configuration fields.

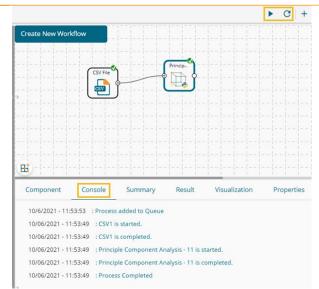


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Select Feature Columns: Select required columns from the input data source against which the analysis gets applied.
  - b. Seed Value: Provide a number (or vector) to initialize the process of the pseudo-random number generator (the default value for this field is 10).
  - c. SVD Solver: Select an option out of the given choices- auto/full/randomized/arpack.
  - d. Number of Iteration: Set a number to indicate the iterations for the model (the default number for this field is 10).
- iv) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General Custom Group	Dynami	ic Fields						
		Feature Columns		8 checked	•			
		Number of Compon	ients	2				
		Seed Value		10				
>		SVD Solver		auto	~			
		Number of Iteration	IS	10				
							Арр	ly

- v) Run the Workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.







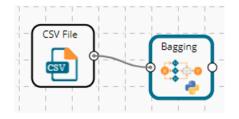
Note: There is no processed input data to display under the 'Result' tab.

## 14.1.8. Ensemble Models

The Ensemble Models Algorithm contains the following sub-types under the Python Workspace:

## 14.1.8.1. Bagging

- i) Drag the Bagging algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to get the configuration fields.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Independent Columns: Select the required input columns against which the analysis gets applied to the target column.



- b. Dependent Column: Select the target column on which the analysis gets applied.
- c. Algorithm Type: Select an algorithm type out of Classification or Regression.
- d. No. of Estimators: Set a number indicating the number of estimators. The default value for this field is
- e. Maximum Samples: Provide a number indicating the maximum samples. The default value for this field is
- f. Maximum Features: Provide a number indicating the Maximum Features. The
- g. Bootstrap: Configure with the 'True' option to select Bootstrap as a modeling method. Configure with the 'False' option if do not wish to select it. The default value for this field is 'True'.
- h. Bootstrap Features: Mention 'True' or 'False' to configure this field. the default option for this field is 'False'.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualiz	ation	Properties	DataInsight	Ŧ	Ť
General	Dyna	amic Fields							
Custom Group		Independent (	Columns		8 checked	d	-		
		Dependent Co	olumn		sex		~		
		Algorithm Typ	e		Classific	ation	~		
		No Of Estimat	ors		10				
Þ		Maximum San	nples		1.0				
		Maximum Fea	itures		1.0				
		Bootstrap			True				
		Bootstrap Fea	tures		False				
								Apply	

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

how	10 👻 entr	ries						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	м
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	м
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	F
M	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	м
1	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	F
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	F
M	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	м
	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	F

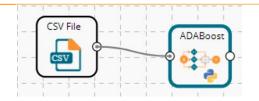
	nent	Console	Sun	nmary	Result	Visualization
		Summary C	( T) D			
		Summary C	Ine Bag	ging Class	1+1er	
Call:	-					
Baggin	ngClassi	lfier(base_es	timator=N	lone, boots	trap='True',	
	boot	strap_featur	es='False	', max_fea	tures=1.0, m	ax_samples=1.0,
	n_es	stimators=10,	n_jobs=N	lone, oob_s	core=False,	random_state=None
	verb	oose=0, warm_	start=Fal	se)		
Confu	sion Mat	rix:-				
[[129]	2 8	7]				
[ (	6 1332	4]				
[ 1]	7 19 1	492]]				
Class	ificatio	on Report:-				
		precision	recall	f1-score	support	
		precision	recall	f1-score	support	
	F		recall 0.99		1307	
	F	0.98		0.99	0.000	
		0.98	0.99	0.99 0.99	1307	
	I	0.98 0.98	0.99 0.99	0.99 0.99	1307 1342	
	I	0.98 0.98 0.99	0.99 0.99	0.99 0.99 0.98	1307 1342	
	I M	0.98 0.98 0.99 0.99	0.99 0.99 0.98	0.99 0.99 0.98 0.99	1307 1342 1528	
	I M cro avg	0.98 0.98 0.99 0.99 0.99	0.99 0.99 0.98 0.99	0.99 0.99 0.98 0.99 0.99	1307 1342 1528 4177	
	I M cro avg cro avg	0.98 0.98 0.99 0.99 0.99	0.99 0.99 0.98 0.99 0.99	0.99 0.99 0.98 0.99 0.99	1307 1342 1528 4177 4177	
	I M cro avg cro avg	0.98 0.98 0.99 0.99 0.99	0.99 0.99 0.98 0.99 0.99	0.99 0.99 0.98 0.99 0.99	1307 1342 1528 4177 4177	
	I M cro avg cro avg ted avg	0.98 0.98 0.99 0.99 0.99	0.99 0.99 0.98 0.99 0.99	0.99 0.99 0.98 0.99 0.99	1307 1342 1528 4177 4177	
mi mae weight	I M cro avg cro avg ted avg	0.98 0.98 0.99 0.99 0.99	0.99 0.99 0.98 0.99 0.99	0.99 0.99 0.98 0.99 0.99	1307 1342 1528 4177 4177	
mid maa weight Accura	I M cro avg cro avg ted avg	0.98 0.98 0.99 0.99 0.99 0.99	0.99 0.99 0.98 0.99 0.99	0.99 0.99 0.98 0.99 0.99	1307 1342 1528 4177 4177	
mid maa weight Accura	I M cro avg cro avg ted avg acy:-	0.98 0.98 0.99 0.99 0.99 0.99	0.99 0.99 0.98 0.99 0.99	0.99 0.99 0.98 0.99 0.99	1307 1342 1528 4177 4177	
mid maa weight Accura	I M cro avg cro avg ted avg acy:-	0.98 0.98 0.99 0.99 0.99 0.99	0.99 0.99 0.98 0.99 0.99 0.99	0.99 0.99 0.98 0.99 0.99	1307 1342 1528 4177 4177	

# 14.1.8.2. ADABoost

# 14.1.8.2.1. Algorithm Type Classification

- i) Drag the ADABoost algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to get the configuration fields.
- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.



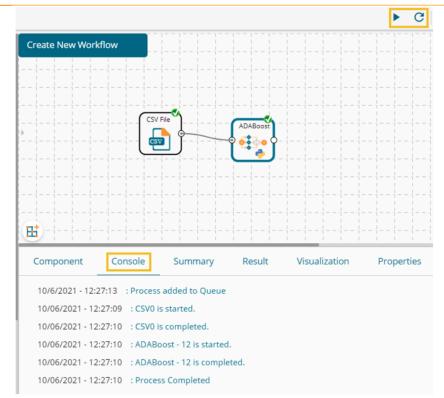


- a. Independent Columns: Select the required input columns against which the analysis gets applied to the target column.
- b. Dependent Column: Select the target column on which the analysis gets applied.
- c. **Types of Algorithms:** The selected algorithm type is **Classification.**
- d. Loss Function: Select an option out linear, square, or exponential
- e. Algorithm: Select an option out of SAMME.R and SAMME (the default option for this field is SAMME.R)
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General Custom Group	Dynan	nic Fields						
		Independent Colu	mns	8 checked	•			
		Dependent Colum	n	sex	Ŷ			
		Type of Algorithm		Classification	Ŷ			
Þ		Loss Function		linear	Ŷ			
		Algorithm		SAMME.R	Ŷ			
							Apply	

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

ow	10 👻 ent	ries						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
4	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	1
4	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1
	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	1
1	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	1
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	1
	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	1
1	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1
	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	м



	C	omponent	Console	Sum	nmary	Result	Visualization
			Summary O	f The ADA	Boost Class	sifier	
		Call:-					
Þ							
		AdaBoostClass lea	ifier(algori mning_rate=1				
		Prediction Pr					
		[0.33723868 0		33733617	0.3226	3232 0.32585	214 0.18951892]
		Confusion Mat	rix:-				
		[[ 366 246	695]				
		[ 32 1097	213]				
		[ 303 367	858]]				
		Classificatio	n Report:-				
			precision	recall	f1-score	support	
		F	0.52	0.28	0.36	1307	
		I	0.64	0.82	0.72	1342	
		М	0.49	0.56	0.52	1528	
		micro avg	0.56	0.56	0.56	4177	
		macro avg	0.55	0.55	0.53	4177	
ŀ		weighted avg	0.55	0.56	0.54	4177	
		Accuracy:-					
		0.55566195834	33086				
				End Of	Summary		

## 14.1.8.2.2. Algorithm Type Regression

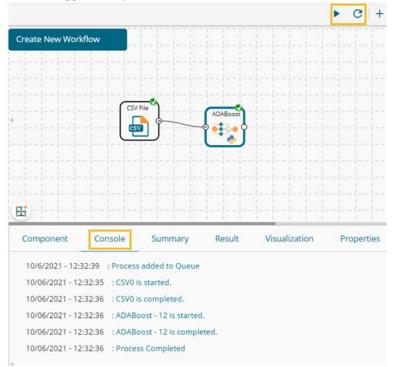
- i) Drag the ADABoost algorithm component to the workspace and connect it with a configured data source.
- ii) Configure the '**Dynamic Fields**' of the dragged algorithm component.
- iii) Independent Columns: Select the required input columns against which the analysis gets applied to the target column.
- iv) Dependent Column: Select the target column on which the analysis gets applied.
- v) **Types of Algorithms:** The selected algorithm type is **Regression.**
- vi) Loss Function: Select an option out linear, square, or exponential
- vii) Algorithms: Select an option out of SAMME.R and SAMME (the default option for this field is SAMME.R)
- viii) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	. <u>↑</u>
General	Dynam	nic Fields						
Custom Group		Independent Colum	nns	8 checked	-			
		Dependent Column	1	length	~			
		Type of Algorithm		Regression	~			
•		Loss Function		linear	~			
		Algorithm		SAMME	~			
							A	pply

ix) Run the workflow after getting the success message.



x) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- xi) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the **'Result**' tab.

Com	ponent	Console	Summary	Result	Visualization	Properties Dat	alnsight		÷ ±
Show	10 ¥ e	ntries					S	earch:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	0.479
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	0.385
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	0.524
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	0.485
1	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	0.385
1	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	0.4
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	0.524
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	0.524
М	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	0.485
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	0.552

xii) Click the 'Summary' tab to open the model summary.



omponent Co	nsole	Summary	Result	Visualizatio
Su	mmary Of T	he ADABoost Reg	ressor	
Call:-				
AdaBoostRegressor( n_estimat		ator=None, lear ndom_state=None		ð, loss='linear
Feature Importance	:-			
	Importanc			
0 diameter				
	0.01965			
<pre>2 weight_whole</pre>				
3 weight_shucked	0.16879	7		
4 weight_viscera	0.11648	19		
5 weight_shell	0.06584	0		
6 rings	0.05665	5		
7 sex	0.00000	0		
Mean Absolute Erro	n (MAE):-			
0.018				
Mean Squared Error	(MSE):-			
0.001				
Root Mean Squared	Error (RMS	E):-		
0.0234				
	F-1	4.05.5		

Note: The Regression algorithm together with the '**square**' and '**exponential**' Loss Function options bring slightly changed processed data under the result and summary tabs. Please look into the following image for the same:

# 14.1.8.2.2.1. Regression Algorithm Type with 'Square' as Loss Function

Summary Component Console Result Visualization Properties DataInsight <u>i</u> T General Dynamic Fields Custom Group Independent Columns 8 checked • Dependent Column length ~ Type of Algorithm Regression Loss Function square SAMME Algorithm AD

**Dynamic Fields:** 

Result Tab:



Com	ponent	Console	Summary	Result	Visualization P	Properties Dat	alnsight		÷ 4
Show	10 <b>v</b> er	ntries					S	earch:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	0.483
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	0.395
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	0.495
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	0.486
1	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	0.386
I.	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	0.415
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	0.52
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	0.538
Μ	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	0.486
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	0.562

## Summary:

Component	Console	Summary	Result	Visualization
	Summary Of	The ADABoost Reg	ressor	
6-11-				
Call:-				
AdaBoostBogna	sson/basa asti	mator=None, lear	ning nato-1 0	loss-'squape'
-	· _	andom_state=None		, 1055- Square ,
	cimacor 5-50, 11		/	
Feature Impor	tance:-			
Fea	ture Importan	ce		
0 diam	eter 0.2920	36		
1 he	ight 0.0015	50		
2 weight_w	hole 0.2840	19		
	cked 0.1424			
4 weight_vis				
5 weight_s				
	ings 0.1019			
7	sex 0.0000	66		
Mean Absolute	Error (MAE):-			
0.02				
Mean Squared	Error (MSE):-			
0.001				
	ared Error (RM	SE):-		
0.0253				
	-	nd Of Summany		
	E	nd Of Summary		

# 14.1.8.2.2.2. Regression Algorithm with 'exponential' as Loss Function

Dynamic Fields:



Component	Cons	sole	Summary	Result	Visual	lization	Properties	Data	Insig	ht	<u>+</u>	Ť
General												
Custom Group		Dynam	nic Fields									
			Independent Col	umns		8 checked			*			
			Dependent Colu	mn		length			~			
			Type of Algorithr	n		Regression			~			
>			Loss Function			exponential			~			
			Algorithm			SAMME			~			
												-
											Appl	y I

### Result Tab:

Comp	onent	Console	Summary	Result	Visualization F	Properties Dat	alnsight		÷ -
how	10 <b>v</b> en	tries					5	earch:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	0.471
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	0.389
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	0.525
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	0.482
1	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	0.388
1	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	0.399
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	0.528
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	0.525
М	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	0.482
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	0.569

## Summary Tab:

Component Console Summary Result Visualization
Summary Of The ADABoost Regressor
Call:-
AdaBoostRegressor(base_estimator=None, learning_rate=1.0, loss='exponential',
n_estimators=50, random_state=None)
Feature Importance:-
reactive importance
Feature Importance
0 diameter 0.271227
1 height 0.019628
2 weight_whole 0.311047
3 weight_shucked 0.161855
4 weight_viscera 0.125040
5 weight_shell 0.076101
6 rings 0.035102
7 sex 0.000000
Mean Absolute Error (MAE):-
0.019
Neer Coursed Coper (NCC) -
Mean Squared Error (MSE):-
0.001
Root Mean Squared Error (RMSE):-
0.0235
End Of Summary

# 14.1.8.3. Extremely Randomized Trees



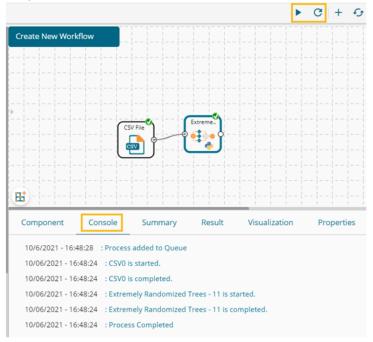
- i) Drag the Extremely Randomized Trees algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
- iv) Click the '**Apply**' option.

Component	Console		Result	Visualization	Properties	DataInsight			ł	L	Ť
General											
Custom Group		Dynamic Fields									
		Indepen	ident Columns		8 checked		-				
		Depend	ent Column		sex		~				
		Type Of	Algorithm		Classification		~				
		Number	r Of Estimators		10						
		Maximu	im Features		auto		~				
>		Maximu	im Depth		2						
		Minimu	m Sample Split	(0.0, 1.0]	1.0						
		Minimu	m Sample Leaf		1						
									A	pply	

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.



how	10 👻 ent	tries						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValue
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	м
N	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	1
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	1
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	М
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	м
N	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	м

Showing 1 to 10 of 4,177 entries

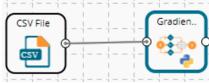
#### viii) Click the 'Summary' tab to open the model summary.

	Component	Console	Sum	nmary	Result	Visualization	Properties	DataInsight	÷	<u>+</u>
		Summary O			assifier					
	Call:-									
•	ExtraTreesCla ma mi mi mi	<pre>ix_depth=2, m in_impurity_d in_samples_le in_weight_fra</pre>	ax_featur ecrease=0 af=1, min ction_lea	es='auto', .0, min_im _samples_s f=0.0, n_e	<pre>max_leaf_no purity_split plit=1.0, stimators=10</pre>		1			
	Prediction Pr 		36769974	0.2720	6034 0.27206	034 0.21027872]				
	Confusion Mat [[ 0 221 1 [ 0 1048 [ 0 334 1	.086] 294]								
	Classificatio	n Report:-								
				<i>a</i>						
		precision	recall	f1-score	support					
	F	0.00	0.00	0.00	1307					
	I	0.65	0.78	0.71	1342					
		0.46	0.78	0.58	1528					
	micro avg	0.54	0.54	0.54	4177					
	macro avg	0.37	0.52	0.43	4177					
1	weighted avg	0.38	0.54	0.44	4177					
	Accuracy:-  0.53674886282	102058								
			End Of	Summary						

Note: The Extremely Randomized Algorithm component can be configured with the selection of the Regression as algorithm type as well.

## 14.1.8.4. Gradient Boosting Model

- i) Drag the Gradient Boosting Model algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to get the configuration fields.



- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a. Independent Column: Select the required input columns against which the analysis gets applied to the target column.



- b. Dependent Column: Select the target column on which the regression analysis gets applied.
- c. Type of Algorithms (Select an algorithm type out of Classification or Regression)
- d. Learning Rate:
- e. Number of Estimators (the default value for this field is 100)
- f. Maximum Depth (the default value for this field is 3)
- g. Minimum Samples Split (0. infinity) (the default value for this field is 2)
- h. Minimum Sample Leaf (0. infinity) (the default value for this field is 1)
- i. Sub Sample (0,1]- (the default value for this field is 1.0)
- j. Maximum Features: Select an option out of auto, sqrt, log2
- k. Random Seed: Provide a number (or vector) used to initialize a pseudorandom number generator (the default value for this field is 10).
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataIn
General						
Custom Group	Dynar	nic Fields				
		Independent Co	lumns	8 checked		•
		Dependent Colu	imn	sex		~
		Type Of Algorith	m	Classificat		~
		Learning Rate		Classificat Regression		
		Number Of Estir	nators	100		
		Maximum Depti	n n	3		
		Minimum Samp	les Split	2		
		[0,Infinity]				
		Minimum Samp	les Leaf	1		
		[0,Infinity]				
		Sub Sample (0,1	]	1.0		
		Maximum Featu	ires	auto		~
		Random Seed		10		

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

how	10 👻 entr	ies						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
м	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	м
м	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	1
4	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	м
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	м
	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	М
N	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1
	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	м



omponent	Console		nmary	Result	Visualization
	Summary C	of The GBP	Classifie	r	
1000					
Call:-					
GradientBoos	tingClassifie				
	max feature			nce', max_de	eptn=3,
				impurity sp	lit-None
	min samples				vii c-isone,
				n estimator:	s=100.
					ndom state=10,
				idation_frac	
	verbose=0,				
Confusion Ma	trix:-				
[[ 601 151 [ 50 1124 [ 167 236	168]				
Classificatio	on Report:-				
	precision	recall	fl-score	support	
F		0.46		1307	
I		0.84	0.79	1342	
м	0.61	0.74	0.67	1528	
micro avg	0.68	0.68	0.68	4177	
macro avg		0.68	0.67	4177	
weighted avg	0.69	0.68	0.67	4177	
Accuracy:-					
0.6823078764	663634				

Note: The users can run the model using the Regression Algorithm Type from the component configuration fields. The processed data under the Result tab will slightly get changed and the Summary will get displayed for the Regressor as given below:

Show 10 v entries Search:									
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValue
м	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	0.464
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	0.361
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	0.526
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	0.473
	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	0.346
1	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	0.412
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	0.536
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	0.532
м	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	0.474
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	0.554

a. Result Tab with Regression Algorithm Type:

b. Summary of the model created choosing the Regression Algorithm type:

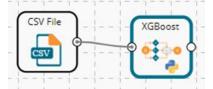


omponent	Cons	ole	Summary	Result	Visualization
	Summ	ary Of	The GBM Regresso	r	
Call:-					
GradientBoos	<pre>learning max_fea min_imp min_sam min_weig n_iter_u</pre>	g_rate= tures=' urity_d ples_le ght_fra no_chan le=1.0,	0.1, loss='ls', sqrt', max_leaf_ decrease=0.0, min saf=1, min_sample sction_leaf=0.0, nge=None, presort tol=0.0001, val	<pre>max_depth=3, nodes=None, _impurity_spl s_split=2, n_estimators= ='auto', rand</pre>	100,
Feature Impo			,		
Fe	ature In	nportan	ice		
0	sex	0.5227	/33		
1 dia	meter	0.1132	16		
2 h	eight	0.1377	193		
3 weight_					
4 weight_sh	ucked	0.0514	402		
5 weight_vi		0.0646	668		
6 weight_		0.0042			
7	rings	0.0002	189		
Mean Absolut	e Error	(MAE):-			
0.012					
	····· //				
Mean Squared	Error (	152):-			
0.0					
Root Mean Sq	uared Er	ror (RM	ISE):-		
0.0165					
		е	nd Of Summary		

# 14.1.8.5. XGBoost

The XGBoost is an algorithm that uses a gradient boosting framework designed for speed and performance.

- i) Drag the XGBoosting algorithm component to the workspace and connect it with a configured data source.
- ii) Click on the algorithm component to open the configuration fields.

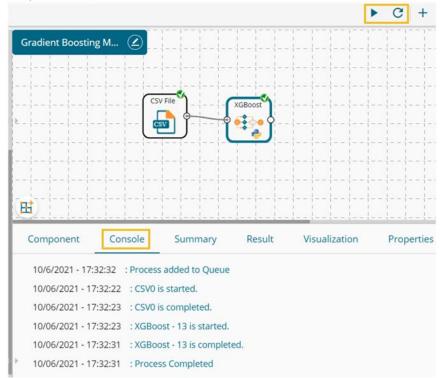


- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
- iv) Click the 'Apply' option.



ustom Group	Dynamic Fields			
ustom Group	Independent Columns	8 checked	•	
	Dependent Column	sex	~	
	Type Of Algorithm	Classification	~	
	ETA Value [0,1]	0.3		
	Gamma [0,Infinity]	0		
	Maximum Depth [0,Infinity]	6		
	Minimum Child Weight	1		
	[0,Infinity]			
	Maximum Delta Step [0,Infinity]	0		
	Sub Sample (0,Infinity]	1		
	ColSample By Tree (0,1]	1		

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.



Com	ponent	Console	Summary	Result	Visualization P	roperties Dat	alnsight		÷ <u>1</u>
Show	10 v er	ntries						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValues
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	М
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	1
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	м
1	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
I.	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	F
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	F
М	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	F

viii) Click the 'Summary' tab to open the model summary.

Component		Console	Sun	nmary	Result	Visualization
		Summary C	f The XGB	oost Class:	ifier	
Call:-						
co ma n_ ob	lsample x_delta estimat jective	_bytree=1 _step=0, cors=100, ='multi:s	<pre>, eta=0.3 max_depth n_jobs=1, oftprob',</pre>	, gamma=0, =6, min_ch: nthread=Ne random_sta	<pre>learning_ra ild_weight=1 one, ate=0, reg_a</pre>	, missing=None,
Predictio	n Proba	bility:-				
[0.238769	04 0.74	18545 0.	48815355	0.0107	5355 0.07452	959 0.00150557]
Confusion	Matrix					
[[ 913 [ 36 12 [ 109 1	17 89	ĵ.				
Classific	ation R	eport: -				
	pr	ecision	recall	f1-score	support	
	F	0.86	0.70	0.77	1307	
	I	0.83	0.91	0.86	1342	
	М	0.76	0.82	0.79	1528	
micro	avg	0.81	0.81	0.81	4177	
macro	avg	0.82	0.81	0.81	4177	
weighted	avg	0.81	0.81	0.81	4177	
Accuracy:	2					
0.8084749						
	8204452	96				
0.8084749						

# 14.1.9. Instance-based Algorithm

The Instance-based algorithms also known as memory-based algorithms compare new problem instances with the instances already seen in training and have been stored in memory. These algorithms are sometimes referred to as 'Lazy' computation gets postponed until a new instance is observed.

The Data Science Workbench provides the K-Nearest Neighbor as displayed in the below-given image:



## 14.1.9.1. K-Nearest Neighbor (KNN)

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to understand and use, but it becomes significantly slow while dealing with a large amount of data.



# ☆ k-Nearest Neighbors

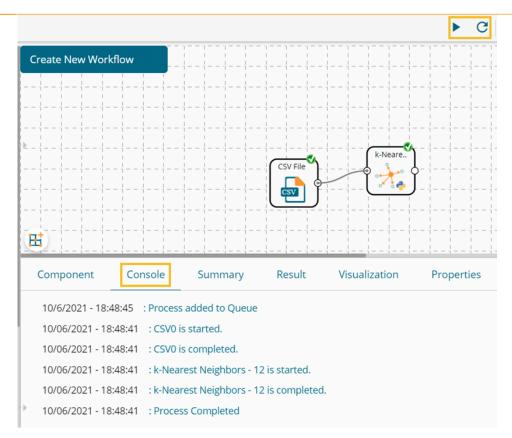
- i) Drag the K-Nearest Neighbor (KNN) algorithm component to the workspace and connect it with a configured data source.
  - CSV File
- ii) Click the K-Nearest Neighbor (KNN) algorithm component.

- iii) Configure the 'Dynamic Fields' of the dragged algorithm component.
  - a) Independent Columns: Select the required input columns against which the analysis gets applied to the target column.
  - b) Dependent Column: Select the target column on which the analysis gets applied.
  - c) Algorithm Type: Select an algorithm type out of Classification or Regression.
  - d) N Neighbors: Set a value for the N Neighbors (the default value for this field is 5).
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General	_							
Custom Group	Dynar	mic Fields						
		Independent Col	umns	8 checked		•		
		Dependent Colur	nn	sex		~		
>		Type of Algorithm	n	Classification		~		
		N Neighbors		5				
								_
							Apply	y

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a) Click the dragged algorithm component onto the workspace.
  - b) Click the '**Result**' tab.

how	10 🗸 ent	tries						Search:	
sex	length	diameter	height	weight_whole	weight_shucked	weight_viscera	weight_shell	rings	PredictedValue
М	0.455	0.365	0.095	0.514	0.224	0.101	0.15	15	1
М	0.35	0.265	0.09	0.226	0.1	0.048	0.07	7	1
F	0.53	0.42	0.135	0.677	0.256	0.142	0.21	9	1
М	0.44	0.365	0.125	0.516	0.216	0.114	0.155	10	м
I	0.33	0.255	0.08	0.205	0.09	0.04	0.055	7	1
I	0.425	0.3	0.095	0.352	0.141	0.078	0.12	8	1
F	0.53	0.415	0.15	0.778	0.237	0.142	0.33	20	F
F	0.545	0.425	0.125	0.768	0.294	0.15	0.26	16	F
М	0.475	0.37	0.125	0.509	0.216	0.112	0.165	9	1
F	0.55	0.44	0.15	0.894	0.314	0.151	0.32	19	F

viii) Click the 'Summary' tab to open the model summary.



Component	Console	Sun	nmary	Result	Visualization
	Summary O	f The K-N	N Classifi	er	
Call:-					
		None, n_j		size=30, met n_neighbors=	
Confusion Matri:					
Confusion Macri	·				
[[ 869 116 32 [ 166 1094 8 [ 399 219 910	2]				
Classification	Report:-				
р	recision	recall	f1-score	support	
F	0.61	0.66		1307	
I	0.77	0.82		1342	
м	0.69	0.60	0.64	1528	
micro avg	0.69	0.69	0.69	4177	
macro avg	0.69	0.69	0.69	4177	
weighted avg	0.69	0.69	0.69	4177	
Accuracy:-					
0.6878142207325	332				
		~			

# 14.1.10. NLP

The following algorithms are included under the NLP algorithm tree node.

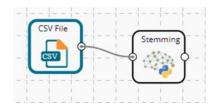
A 🦚 NLP

- Stemming
- Document Summarization
- Remove Stopwords
- 🕸 Lemmatization
- 🆓 Topic Model
- Sentence Parsing
- 🆓 Named Entity Recognition
- Part of Speech Tagging
- Document Keyword Extraction
- 🆓 N-Gram
- Word Tokenization
- Sentence Tokenization



## 14.1.10.1. Stemming

- i) Drag the Stemming algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.



- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Stemming Algorithms
  - b. Input Column Name
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General Custom Group	Dynami	c Fields						
custom Group	S	itemming Algorithm		Porter	~			
	h	nput Column Name		words				
•								
							Apply	y

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





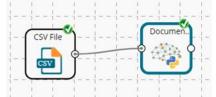
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.

Component Console Summery Result Vasaltation Properties Detailinight	+
Pow 11 w enter	Search.
lan -	
	$w,w,t,w,w',\cdot,w,v,w',\cdot, \mathbf{y},w,w,w,\tau,\cdot,t,\cdot,w,w,w,\tau,\cdot,w,w,w,\tau,\tau,\tau,w,w,v,\tau,\tau,\tau,w,w,\tau,\tau,\tau,\tau,w,w,t,\tau,\tau,\tau,w,w,w,\tau,\tau,\tau,\tau$
	的现在分词 化化化化化化化化化化化化化化化化化化化化化化化化化化化
N 1911 De la construction de la const La construction de la construction de	$(\mathbf{x},\mathbf{x},\mathbf{w},\mathbf{w}) \in [0,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty,\infty$
	法法法法法法法 化化化学 化化化物化化物化物化物化物化物化物化物化物化物
	алы ты тыкалар акуылы жалар улар. Кайтын калар тарын калар жалар жалар кайтын калар тарын калар кайтын каралар Кайтын калар

#### 14.1.10.2. Document Summarization

b.

- i) Drag the Document Summarization algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.

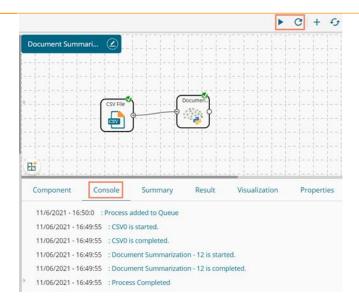


- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Input Column Name
  - b. Summarize by Ratio (value between 0 to 1)
  - c. Summarize by Word Count
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visu	alization	Properties	DataInsig	ht	<u>+</u>	Ŧ
General	Dyna	mic Fields								
Custom Group		Input Column I	Name		content					
		Summarize by	Ratio (value		0.0					
•		between 0 & 1)								
		Summarize by	Word Count		0					
									Appl	y

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

Compone	t Console	Summary	Result	Visualization	Properties	DataInsight	+	<u>+</u>
Show 10	✓ entries					Search:		
summary								
You'll save	wo dollars AND get a	medium drink inste	ead of a small AN	D a cookie".				
Today inste	d of getting 10 pieces	s my bucket had 8 a	and they forgot to	o include the chocolate	e chip cookies.			
Food is goo	but every single time	e I stop there to or	der pot pies, they	always only seem to h	nave 1 leftFrustra	ting!!!		
About the s	ime as other location:	s.						
after hangi	g up and calling back	we were told a new	v pizza was on th	e way.				
Ordered piz	za tonight at 8:41 with	n an expected 60-7	0 minute delivery	time.				
Showing 1 to	10 of 100 entries			Previou	is 1 2	3 4 5 1	10 N	ext

### 14.1.10.3. Remove Stopwords

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration tab.

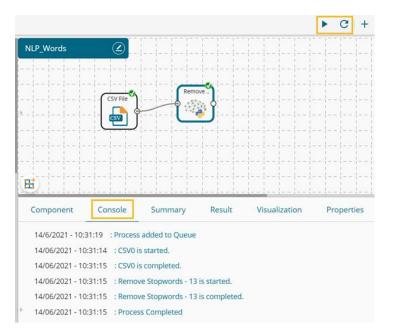


- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Input Column Name
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General	Dynami	c Fields						
Custom Group		input Column Name		words				
Þ								
							Appl	ly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

iow 10 V entries	Search:
words	
d new pressing the sense of the sense of the match in the sense within this in this within a within the sense w With reaching grade within the match in a sense think appendix visite in within this sense grade book in the within with new presses grade within the sense of the sense of the sense within the sense grade book in the sense of	
	'u', 'n', W', 'n', 'b', 'c', 'k', 'u', 'g', 'e', 'h', 'e', 'r', 'e', 'f', 'u', 'r', 'r', 'e', 'r', T, '
b, e, v, h, c, k, e, n, h, v, e, e, v, e, r, h, h, g, r, e, f, t, v, r, s, e, w, h, h, e, e, v, e, h, h, e, r, v, k, e, r, e, g, t, v, h, h, e, b, e, e, f, f, r, e, v,	'r, 'n', 'h', 'V, 'e', 'e', 'V, 'e', 'r', 'b', 'e', 'e', 'n']
1. g. b. v. e. v. e. n. n. g. t. e. e. n. p. n. e. n. e. n. e. n. p. p. e. n. e. t. w. n. t. e. n. v. e. t. t. t. t. v. n. n. g]	
n bunn nen mer trom y u gen winn uten prei ein nich ein och ein u protein gift nic groen men mer prei night in Due un einer meunen viel riertig prei eine un einer einer wurp der einer men mit holl der trom einer un ein tro einer nei blick mit einer einer einer men einer mit mit trom viel einer wurd auf einer men mit werden.	
w.acch.ch.g.g.g.h.g.g.t.c.w.cc.w.t.w.t.w.n.ec.t.f.e.w.cc.n.t.k.m.h.w.c.g.w.cc.n.g.h.e.c.c.w.m.g.f.c.n.e.t.v.ec.t.c.t c.t.t.e.g.n.ch.ec.t.c.t.e.c.t.w.n.t.e.c.ex.t.t.e.c.t.w.c.t.w.c.t.w.c.e.c.g.m.g.t.t.t.t.g.t.t.t.g.t.t.g.t.t.g.t.	k, 'e', 'n', 'n', 'r', 'v', 'e', 'r', 'e', 'l', 'v', 'e', 'r', 'w', 'h', 'h', 'e', 'w', 'h', 'e', 'n', '
N, C, T, T, C, C, T, V, C, V, C, C, D, T, T, T, T, T, D, N, C, C, C, V, C, T, T, C, N, C, C, T, N, Y, C, T, C, C, D, C,	, h, e, n, g, e, r, n, c, r, e, n, n, f, f, e, r, f, x, h, e, u,
. Notified in the content of the second to prove the second of the second	
a restricte preziency with note or prezience and the entry term with the view of the view of the prezience of the restricted prezience of the	'r', 'b', 'u', 'n', 'f', 'r', 'u', 'n', 'e', T', T', T', T', 'b', 'e', 'h', 'e', T', T', 'b', 'e', 'u',



## 14.1.10.4. Lemmatization

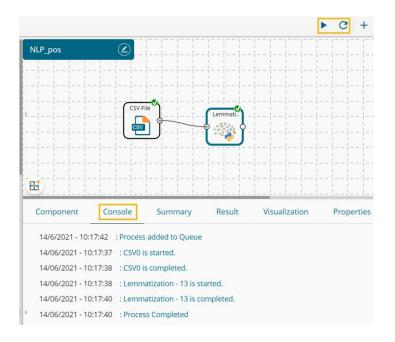
- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.



- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Data is POS Tagged: Select either the 'True' or 'False' option from the given menu.
  - b. Input Column Name: Provide the name for the input column.
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General	Dynam	ic Fields						
Custom Group	_	Data is POS Tagged		True	~			
		Input Column Name		pos				
÷								
							Appl	y

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



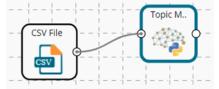


- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

que la 🗸 encies	Search.
imma	
	<pre>cy = x = x = y = z = z = x = z = x = z = z = z = z = z</pre>
n en ante en la construction de la La construction de la construction d	(A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,
	e an an in Allen North an an an an an an an in the an an a
NARAN'I NY TANÀNA MANTRANA MANJARANA MININA MININA MININA MANJARANA MININA MININA MININA MININA MININA MININA M Noro-Promonometry designa dia manjara dia manjara minina minina minina minina minina minina minina minina minina	$w_i \in \mathcal{N}, \ i \in \mathbf{A}, \ i \in \mathbf{A}, \ \mathbf{A}, $
	n general an ann an Anna an Anna an Anna an Anna An Anna Anna
	このものにも本にのためにものもにものものます。
	······································

#### 14.1.10.5. Topic Model

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.



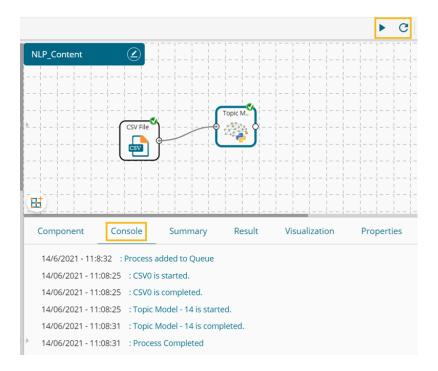
- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Number of Topics
  - b. Input Column Name
  - c. Number of Passes (epochs)
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	Ŧ	Ť
General	Dynar	mic Fields						
Custom Group	_	Number of Topic	5	2				
		Input Column Na	me	content				
•		Number of Passe	s (epochs)	50				
							Apply	y

v) Run the workflow after getting the success message.



vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



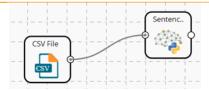
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.

now 10 🗸 entries			Search:
Document_num	Dominant_Topic	Topic_Perc_Contrib	Keywords
0	1	0.988	order, get, food, n't, time, wait, minute, go, 's, place
1	1	0.986	order, get, food, n't, time, wait, minute, go, 's, place
2	0	0.94	pizza, order, n't, domino, 's, call, delivery, get, bad, say
3	1	0.978	order, get, food, n't, time, wait, minute, go, 's, place
4	1	0.962	order, get, food, n't, time, wait, minute, go, 's, place
5	0	0.984	pizza, order, n't, domino, 's, call, delivery, get, bad, say
6	1	0.954	order, get, food, n't, time, wait, minute, go, 's, place
7	1	0.581	order, get, food, n't, time, wait, minute, go, 's, place
8	1	0.786	order, get, food, n't, time, wait, minute, go, 's, place
9	0	0.986	pizza, order, n't, domino, 's, call, delivery, get, bad, say

### 14.1.10.6. Sentence Parsing

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click on the algorithm component to open the configuration fields.

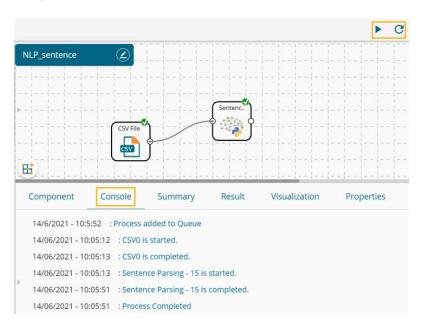




- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Input Column Name
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ŧ
General	Dynam	nic Fields						
Custom Group	bynan	Input Column Nar	ne	sentence				
•								
							Appl	y .

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.



Component	Console	Summary	Result Visualization	Properties Datalr	nsight 🕂 🕹
Show 10 🗸	entries				Search:
Row_id	Text	Тад	Dependency	Head_text	Head_tag
•	Do	VB	aux	expect	VB
0	n't	RB	neg	expect	VB
0	expect	VB	ROOT	expect	VB
0	the	DT	det	order	NN
0	order	NN	nsubj	taker	VB
0	taker	VB	nsubj	try	VB
0	to	то	aux	try	VB
0	try	VB	ccomp	expect	VB
0	to	то	aux	save	VB
0	save	VB	xcomp	try	VB
Showing 1 to 10 of	f 8,334 entries			Previous 1 2 3	4 5 834 Next

# 14.1.10.7. Named Entity Recognition

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.

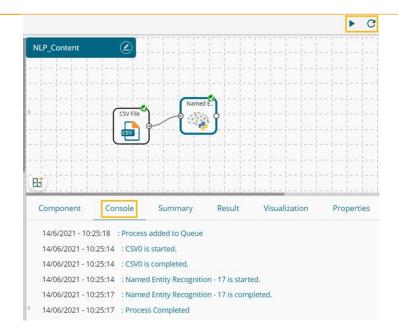


- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Input Column Name
- iv) Click the '**Apply**' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>1</u>	Ť
General	Dynar	nic Fields						
Custom Group	_	Input Column Na	ime	content				
							Арр	ly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





#### vii) Follow the below-given steps to display the Result view:

a. Click the dragged algorithm component onto the workspace.

#### b. Click the '**Result**' tab.

Component	Con	sole	Summary	Resu	lt	visualization	Properti	es DataInsight	t								÷
now 10	✓ entries														S	earch:	
PERSON	NORP	FAC	ORG	GPE	LOC	PRODUCT	EVENT	WORK_OF_ART	LAW	LANGUAGE	DATE	TIME	PERCENT	MONEY	QUANTITY	ORDINAL	CARDINAL
														['two dollars']			
('Learn']			['KFC' 'Brunswick']								['Today']						['10' '8']
																	['16' '10']
																	[11]
												('2 hours')					['40']
['Adam']												['an hour and a half ago']					['7' '35']
			[WTP]	['530p' '630p']													['3' '30']
												['tonight' '8:41' '60- 70 minute' '11:00' 'the past half hour']				('first')	

### 14.1.10.8. Part of Speech Tagging

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.

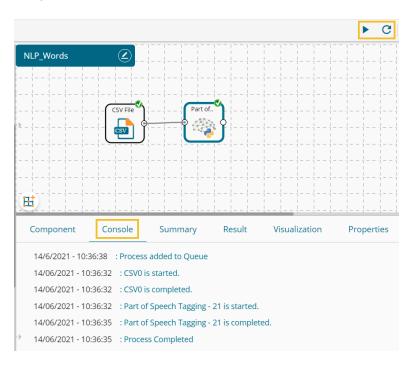
CSV File	<u>-</u> + +	 Part of.	<b>`</b>
csv			<b>0</b>   .
	<b>-</b> +	 	

- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Input Column Name
- iv) Click the 'Apply' option.



Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	T
General		a eta tata						
Custom Group		Input Column Name		words				
		input column Name		WOIUS				
•								
							App	ly

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.



Com	ponent	Console	Summary	Result	Visualization	Properties	DataInsight		÷
w	10	← entries						Search:	
oos	î:								
5,NN /8', ' NNP' (,NN (,NN (,NN (,NN (,NN (,NN) (,NN) (,NN)	(, W.NN a.DT, J VBZ, V. O.NN P. O.VB (, O.NN a.DT, J V. J, T I.NN, d (, NNP (, NNP (, NNP (, NNP (, NNP (, NNP (, NNP (, NNP) (, NNP (, NNP) (, NNP) (, NNP (, NNP) (,	LNN, V62, a.C. , r.NN, IANN, J. , m.N, IANN, J. , m.J. (e.NN, J.N. , m.J. (e.NN, J.N. , m.Y. (POS), a.C. , s.NN, m.VBD, , s.NN, m.VBD, , s.NN, m.N, b.N, , m.N, NN, b.N, , a.D. , s.NN, fa.N, a.D.	DT, hgg, dJ.N.Y., JN, N.N.Y. (VHO), 'a, DT, N.Y. (NNN, 'W, NNY, N.Y. (NNN, 'W, NNY, 'MNP, 'D, NNY, 'W, NNY, 'MNP, 'D, NNY, 'NNY, 'MNP, 'U, 'NNY, 'NNY, 'NNY, 'NNY, 'W, 'NNY, 'NNY, 'NNY, 'W, 'NNY, 'C, 'H, NNY, 'W, 'NNY, 'C, 'H, NNY, 'W, 'NNY, 'MNY, 'NNY,	PP: s.VB2; m.VB; JJJ; r.YNR; a.DT; NNP; m.NN; eh N;NNP; m.VB] VB2; a.DT; VMN v, D; NN, a.NN; v, D; NN, a.NN; v, J;NN; v, NNP; m.FW; J; NN;NNP; m.FW; J;NN; a.DT; J;NN; a.DT; J;NN; a.DT; J;NN; a.DT; J;NN; a.DT; J;NN; a.DT; J;NN; a.DT; J;NN; a.DT; J;NN; a.DT; J;NN;NN;	"	NNP, d.NN, T.NNT, NN, D.NN, C.NN, T.NN, INV, BNP, S.NE, T. INVEP, g.NN, INN, d.NN, T.NN, UNN, DT, J., C.NN, D.NE, S.NN, S.NN, S.NN, V. NN, T.NP, SNN, S.NN, V. NN, T.NP, SNN, S.NN, N, NN, D.NE, S.NN, S.NN, N, NNP, INVES, B N, UNN, NNP, INVES, B N, UNN, S.NP, SNN, SNN, ADT, INN, S.NP, SNN, SNN, SNN, SNN, SNN, SNN, SNN, S	I.NN. N. 1997, Y.NN, C.C. NANN, N.NN. EWN, T.NN, NNP, WANNE O, D.D.T. VANN, & NNN, W.NNE O, D.D.T. VANN, & NNN, W.NNE Q, MARK, Y.	N. LAN, EANN GAN, YAR, ADT, N. YAN, SAN, GAN, YAN, AN, YAN, YAN, YAN, YAN, YAN, YA	"UNN, 'y,NN, '' 'a,DT, 'y,NN, '' 'o,NN, '', 'ND 'D,NNP, 'NNP, 'NNP, 'NNP, '0,VB2, 'd,I' 'NNV, 'NNN, '0,NN 'NNV, 'NNN, '0,NN 'N, 'NNP, 'g,'' 'NVBP, 'g, 't,NN ', 'NVB2, 'e,J', ' P, ''NN, 'NNN, ' NN, 'NNP, 'g,NN
JI, Y ENN ENN ENN ENN ENN ENN ENN ENN ENN EN	0,NN, 'a, , ',NN, , 'e,NN, , ',NNP, NNP, 'h, , 'NNP, , 'NNP,	DT, TANE, TANE, 'eNN, TANE, 'NN, 'BLD', 'NN, 'ANN 'BLD', 'NN, 'ANN 'BLD', 'NN, 'ANN 'NN, 'ENN, 'NN, 'ENN 'NN, 'ENN, 'NN, 'NN, 'NN, 'SNN, 'NN, 'NN, 'SNN, 'NN, 'SNN, 'SNN, 'NN, 'NN, 'SNN, 'NN, 'NN, 'SNN, 'NN, 'NN, 'NN, 'NN, 'NN, 'NN,	1.1.N. CVBP, ULT, NNF, 25, N, CVBP, 25, NNF, 25, T, NJ, GUN, NNF, CNN, T, NN, NNF, CNN, CNN, T, NN, NNF, NNF, CNN, T, NN, NNF, NNN, NNF, NN, NNF, SNN, INNF, 2, T, NN, NNF, CNN, T, 2, T, NN, NNF, CNZ, 2, T, NN, NNF, CNZ, 3, T, NNF, CNZ, 1, T, NN, SNN, SNN, NN, NNF, CAR, NN, NNF, CAR, 1, T, NN, SNN, SNN, NN, SNN, SNN, SNN, SNN, SNN, SNN, 1, T, SNN, SNN, SNN, 1, T, SNN, SNN, SNN, SNN, 1, T, SNN, SNN, SNN, SNN, SNN, SNN, SNN,	INN (ADT, 'INN /8P, 'INNN, 'INN /8P, 'INNN, 'INN 1, CO, 'OCD, ' JJ ', INN, 'OJ, 'INN, 'INN B2, 'ADT, 'SJ, 'I P, 'CNN, 'INN B2, 'ADT, 'SJ, 'I NN, 'OJN, 'INN 'J, 'INN, 'INN 'J, 'INN, 'INN 'INN, 'INN, 'INN INN, 'INN, 'INN 'INN, 'INN, 'INN 'INN, 'INN, 'INN 'INN, 'INN, 'INN 'INN, 'INN, 'INN 'INN', 'INN, 'INN 'INN', 'INN, 'INN', 'I	1. MPP, 2MPF - EM 	NY, CANP, J. ANP, J. ANP, J. ANP, J. ANP, J. ANP, J. AND, J. ANN, J. A	TIMINY, JANE, ANDE, ANDE, ANDE, JANE, J	<sup>1</sup> ANP, TAN, GAN, TANP, TANE, TANE, T GAN, BAN, TANE, TANE, TANE, SA GAN, BAN, TANE, TANE, SA MAN, SA MAN, SA MAN, TANE, TANE, SA MAN, SA MAN, TANE, TANE, TANE, SA MAN, TANE, SA MAN, SA MAN, SA MAN, SA MAN, SA MAN, JANE, SA MAN, SA MA	BP, 1, J. 10, J. 1, 10, 20, 20, 20, 20, 20, 20, 20, 20, 20, 2

# 14.1.10.9. Document Keywords Extraction

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.



- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Input Column Name
- iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u>	Ť
General	Dynami	c Fields						
Custom Group	-	nput Column Name		content				
>								
							Аррі	by .

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



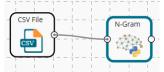
					▶ C +
NLP_Content		1 - 1 - 1 - 1 - 1 - 1 1 - 1 - 1 - 1 - 1   1 1 1 1	- 1 1 1		
				$  _{} =   _{} =   _{} =   _{}$	
te =   = =   = =   = =   =	CSV File	)		<u> </u>	
$ \frac{1}{1}  = -\frac{1}{1} = -\frac{1}{$	csv		Docu	umen)	
$  _{1}^{2} _{1}^{2} _{1}^{2} _{1}^{2}- _{1}^{2}$				🦥 J	$ \frac{1}{1} \frac{1}{1} \frac{1}{1} \frac{1}{1} \frac{1}{1}$
				$\frac{1}{1} = -\frac{1}{1} =$	
E	_				
Component	Console	Summary	Result	Visualization	Properties
14/6/2021 - 10:4	48:21 : Process	added to Queue			
14/06/2021 - 10	):48:16 : CSV0 i	s started.			
14/06/2021 - 10	:48:17 : CSV0 i	s completed.			
14/06/2021 - 10	):48:17 : Docun	nent Keyword Exti	raction - 19 is s	started.	
14/06/2021 - 10	):48:17 : Docun	nent Keyword Ext	raction - 19 is o	completed.	
14/06/2021 - 10	):48:17 : Proces	ss Completed			

- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

Component Console Summary	Result	Visualization	Properties	s Dat	alnsight		÷	4
how 10 🗸 entries				Se	arch:			
keywords								
['drink', 'employees']								
['pieces', 'chip', 'fried']								
[best]								
[lemonade', 'terrible', 'kept']								
[pot]								
['entertainment', 'good', 'mashed potatoes']								
['like', 'understaffed']								
['terrible']								
['pizza', 'separate', 'ordered', 'order']								
['pizza', 'delivery', 'half']								
howing 1 to 10 of 100 entries		Pre	vious 1	2 3	4 5	10	Ne	×t

#### 14.1.10.10.N-Gram

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click on the algorithm to open the configuration fields.



- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Number of words (0)



b. Input Column Name

#### iv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	Ŧ	Ť
General Custom Group	Dynam	nic Fields Number of words Input Column Na		3 words				
>							Appl	y

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.

NLP_Words			$\gamma = -\gamma = -\gamma = -$		
_		· · · · - · · - · · - · · - · · - · · - · · - · · - · · - · · - · · · - ·		· · · - · · - · · - · · - · · - · · - · · - · · - · · - · · - · · - ·	
	iii	· · · · · · · · · · · · · · · · · · ·		· · · - · · - · · - · · - · · - · · - · · ·	
		· · · - · · - · · - · · · · · · · · · ·	N-Gram	·	
	CSV File	);	🚓 oʻ	· · · - · · - · · - · · - · · - · · - · · - · · - · · · - ·	$-\frac{1}{2}$ $-\frac{1}{2}$ $-\frac{1}{2}$ $-\frac{1}{2}$ $-\frac{1}{2}$ $-\frac{1}{2}$
	csv				$- \left  \right  \left  \right $
		· · · · · · · · · · · · · · · · · · ·	1 1 1		
	ii		i i i		
Ħ					
				Visualization	Properties
Component	Console	Summary	Result	VISUAIIZAUUTI	
		Summary 	Result	VISUAIIZAUOTI	
	:5:12 : Process		Result	VISUAIIZALIOIT	
14/6/2021 - 11	:5:12 : Process 1:05:07 : CSV0	added to Queue is started.	Result	VISUAIIZAUON	
14/6/2021 - 11 14/06/2021 - 1 14/06/2021 - 1	:5:12 : Process 1:05:07 : CSV0 1:05:07 : CSV0	added to Queue	Result	VISUAIIZAUUT	
14/6/2021 - 11 14/06/2021 - 1	:5:12 : Process 1:05:07 : CSV0 1:05:07 : CSV0 1:05:07 : N-Gra	added to Queue is started. is completed.		visualization	

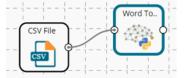
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

Component Console Summary Result	Visualization	Properties	DataInsight		÷
how 10 v entries				Search:	
ngram					
$\begin{array}{c} (1,1,1,1) \\$	10,				1 th of [th or ,



## 14.1.10.11. Word Tokenization

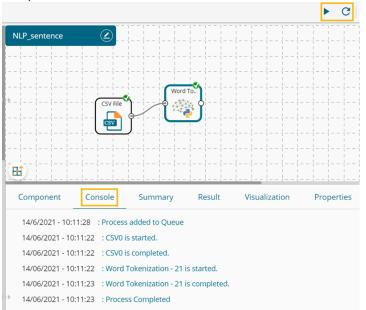
- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.



- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
- a. Input Column Nameiv) Click the 'Apply' option.

Component	Console	Summary	Result	Visualization	Properties	DataInsight	± T
General	Dynan	nic Fields					
Custom Group	Dynan	Input Column Nat	ne	sentence			

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.



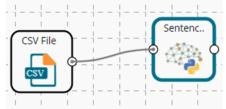
- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the 'Result' tab.



Con	nponent	Console	Summary	Result	Visualization	Properties	DataInsight	÷	<u>1</u>	
w	ords									
DOI: 1011 Seaded: Table Vote: Tabler, too top, top Votew, journ Woney, and With Seaders 7. To Voted Head Votew, Vote Woney, Votewater, Head Votew, Yourn Woney, Votewater, Head Votew, Yourn Woney, Your										
Elementer: These "goes" to "bis" particular XKC "by "order" is" neer "light". Today "solard" of getting, To yuess" by "best" and "by least" by "bust" to "bust" is "bust" of getting, To yuess" by "bust", bust" of the "bust" is									'get',	
['Be '1]	(Best: chicken; % have; tever; had; %; has; great; flavor; %; Same; with; the; sides; even; their; tookies; are; good; %; %; must; admit; this; is; the; best; semi-fast; food; restaurant; %; have; even; been; to; 1]									
Phosolutely: temble: Senice: 1: The' gen', taking: four: order: kept: Spnoring: what: we' taked: for: 1: and: would: waik: wey', while' we' vere' taking: 1: Then' we' were' supposed: to: gen', in' rec' temonade: 1: and: the' gay: refused: to: give', in: to: vere', 'bough', we', 'ordered' w' 10: piece' meat, and: the 'temonade'; comes, with' w'10; piece', or 'arger'. Never', gang', bad();										
['Fo	od", "is", "good".	'but', 'every', 'sir	ngle', 'time', 'T, 'stop	p', 'there', 'to', 'or	der', 'pot', 'pies', '/, 'th	ey', 'always', 'only', 's	eem', 'to', 'have', '1', 'left', '', '.Frustrating', 'T, 'T, 'T]			
1.17	eot', 'as', 'a', 'go	od', 'value', 'as',	'it', 'used', 'to', 'be'	'They', 'used',	'to', 'have', 'really', 'go	od', 'specials', 'but',	High Tai, "Bittle", "greasy, "1, "The", "mashed: "potatoest, "and", "gravy, "are", "fake", ", "pri- they, "are", "rare", "now," 1, "UPDATE, "6-9-16, "1, "They", "doi, "not, "honor", the", "coup "said", "they," 'doi, "not, "know," 1, "1, "only", "work", "here, "1, "WRONG", "ANSWER", "[]		pî.	
							.) 'sitting', 'there', 'waiting', 'for', 'delivery', 'for', 'like', '40', 'mins', '', 'no', 'drivers', 'to ger', 'f', 'you', 'reading', 'this', 'got', 'ta', 'do', 'something', 'buddy', '.']	/, 'deliver', 'is', 'what', 'they',		
							ndl, 'asked', 'to', 'talk', 'to', 'the', 'manager', ' 'after', '7', 'mins', 'of', 'being', 'on', 'hold ancer', 'the', 'order', 'or', 'wAit', 'another', '35', 'mins', 'for', 'a', 'delivery', ' 'Terrible', '			
'tim	e', 'in', 'the', 'sa	me', 'phone', 'ca	ill', '.', 'after', 'hangi	ing', 'up', 'and', 'ca	alling', 'back', 'we', 'we	re', 'told', 'a', 'new', 'p	et 'hourt, 'thet' 'ordert' wast, 'wrong' 10 'when', 'wet' 'called', 'to', 'complain', 10, wast Jizza', 'wast' 'on', 'thet' 'way', 17, '800p', 'rolls', 'around', 'no', 'pizza', 11, 'took', 'another' ihen', 'lim', 'spending', '20bucks', 'of', 'my', 'hard', 'earned', 'cash', 't', 'want', 'what', 't',	'phone', 'call', 'and', '30',	te'.	
'halt 'unf	f, 'hour', 'and',	'have', 'not', 'bee	en', 'able', 'to', 'read	ch', 'anyone', 'as',	'of', 'yet', '.', 'Not', 'the	, 'first', 'time', 'this',	1:00; \l. \and; \l. have: been, 'on; the' phone; 'on; hold; 'trying; to; 'cancel; 'm 'has; 'happened'.\l. \t. \was; forced; to; 'order, 'from'; a; 'more! reliable: 'piza; driver, 'as; 'it; 'rs', 'not; 'his; 'fault; \l. but; will; 'mot; 'be; 'paying; 'for; the' piza;	'place', 'nearby',		
Show	ing 1 to 10 of	100 entries					Previous 1 2 3	4 5 10 N	lext	

### 14.1.10.12. Sentence Tokenization

- i) Drag the algorithm component to the workspace and connect it with a configured data source.
- ii) Click the algorithm component to open the configuration fields.

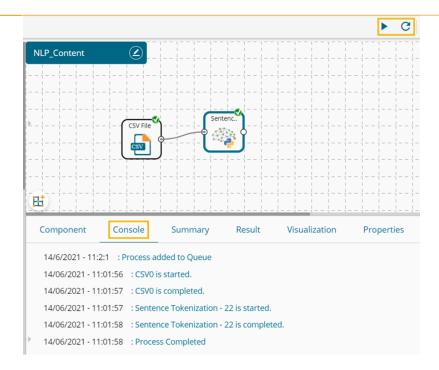


- iii) Configure the '**Dynamic Fields**' of the dragged algorithm component by providing the following details:
  - a. Input Column Name
- iv) Click the 'Apply' option.

General Custom Group Input Column Name content	Component	ion Properties DataInsight	<u>1</u>	Ŧ
,		t		
	>		Apply	

- v) Run the workflow after getting the success message.
- vi) The Console tab opens displaying the step-by-step completion of the process. The completion of the console process gets marked by the green checkmarks on the top of the dragged components.





- vii) Follow the below-given steps to display the Result view:
  - a. Click the dragged algorithm component onto the workspace.
  - b. Click the '**Result**' tab.

now 10	entries Search:							
row_id	sentence							
0	Don't expect the order taker to try to save you money at this location!							
0	I ordered a bowl and small drink.							
0	Normally, a restaurant worker would say "Allow me to save you some money by making this a combo.							
0	You'll save two dollars AND get a medium drink instead of a small AND a cookie".							
0	This is the type of service that would bring me back to this restaurant!							
0	When I brought this to the employees attention, he made no attempt to make it right.							
0	He could have at least given me a cookie, which would have dramatically changed the tone of this review.							
1	Everytime I have gone to this particular KFC my order is never right.							
1	Today instead of getting 10 pieces my bucket had 8 and they forgot to include the chocolate chip cookies.							
1	There was no napkins or condiments included in the bag it's very frustrating when you get home to discover your order is not correct and you do not want to back out to get the rest of your order.							

# 14.2. Custom Scripts (Python Scripts)

The users can create and add customized algorithm components using the '**Custom Python Script**' component. The created scripts will be stored in the '**Saved Scripts**' module provided for the Python Workspace.

### 14.2.1. Creating a New Python Script

- i) Click the 'Custom Scripts' tree-node on the Predictive Analysis home page.
- ii) Click the **'Create New Script'** option.



🔺 🚳 Custom Scripts	
Create New Script	
▶ 🝯 Saved Scripts	
🕵 Utility Module Scripts	5

- iii) The users get the 'Component' tab.
- iv) Configure the following fields in the 'General' tab:
  - a. Basic
    - i. **Component Name:** Enter a name or title that you wish to give a saved Python Script.
    - ii. **Component Type:** Default Component type will be displayed in this field.
    - iii. **Description:** Describe the Component (It is an optional field).
    - iv. Script Type: Select one option out of 'Custom Python Script' or 'Utility Module Script' for the script to get saved under the selected script type.
- v) Click the 'Next' option.

Component Console	Summary Result	Visualization	Properties	DataInsight	<u>+</u> <u>↓</u>
Basic					
Component Name	Custom Python Script				
Component Type	Algorithms				
Description	Optional				
Script Type	Custom Python Script	O Utility Module	∮ Script		
General Script	Settings				Next

- vi) The users get redirected to the 'Script' tab.
- vii) Provide the following information:

#### a. Script Editor

- i. Write the required python script in the given space under the 'Script Editor.'
- ii. Click the 'Validate' option.
- iii. A success message should appear after the validation (as shown in the below image).
- iv. Click the 'Next' option.

	Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u> <u>+</u>
	Script Editor							
1		Python script has be	een validated succe	ssfully!				θ
	19 20	lt.bar(index + bar_wid alpha=opacity, color='g', abel='data_2')	ith, data_2, bar_widtl	١,				
	23 plt.xlabel( 24 plt.ylabel( 25 plt.title('So	'Scores') cores by person') index + bar_width, ('A'	, 'B', 'C', 'D'))					_
	32	(SAVE_IMG_PATH)	df deseribe()					
**[		itput":df, "summary": d						
	General	Script	Setting	5			Previous	Next



- **b.** Configure the required **'Primary Function Details'** to embed the customized Python script into a function.
  - i. **Primary Function Name:** Select the name of the created function from the drop-down menu.
  - ii. Input Data Frame: Select a dataset (that has been used above) from a drop-down menu.

(The 'Output Data Frame' option and the 'Model Variable Name' are pre-selected for the Primary Function Details)

viii) Click the **'Next'** option (The users can click the **'Previous**' option if wish to open the previous page).

Component	Console Sur	mmary Result	Visualization Prope	rties DataInsight		*	<u>+</u>
ne	rtion Details Primary Function Name w_fun • Show Visualization	Input DataFrame	Output DataFrame Output Data Frame Show Summary	Summary Variable Name Summary			
	ameter Definition	Property D	isplay Name	Control Type TextBox	• 0		
General	Script	Settings			Previous Ar	oply	

- ix) The users get directed to the 'Settings' tab.
- x) Configure the following fields:

#### a. Primary Function Details

This option configures the following details:

- i. Primary Function Name: Select an option from the drop-down menu.
- ii. Input Data Frame: Select an option from the drop-down menu.
- iii. Output Data Frame: Provide a name for the Output Data Frame.
- iv. Summary Variable Name: Provide a name for the Summary Variable Name.
- v. The user can select the 'Show Visualization' and 'Show Summary' options from this section.

#### b. Function Parameter Definition

- i. Function Parameters: Actual names of parameters configured in the script.
- ii. **Property Display Name**: Parameter name to be displayed while configuring the saved script as a component.
- iii. Control Type: User can select out of the following options:
  - 1. Text box,
  - 2. Drop-down menu,
  - 3. Column Selector (single),
  - 4. Column Selector (multiple).
- iv. Settings option To set the display for mandatory fields and validate the datatype for the input column. This field is associated with function parameters.



¢
<ul> <li>Mandatory Field</li> </ul>
🔲 Show Default
Select Datatype

xi) Click the 'Apply' option.

Componer	nt Console	Summary Re	esult Visualization	Properties	DataInsight		+	Ŧ
Primary F	unction Details		0.1.1015					
	Primary Function Name		ne Output DataFrai		ny Variable Name mary			
	Show Visualization		🗹 Show Summa	iry				
Function	Parameter Definition							
	Function Parameters		Property Display Name	Control		v Ö		
	^		^	Text	504	· •		
•								
		_	-					
General	Script	Settings				Previous A	pply	

xii) A message appears to notify that the newly created Python script has been saved successfully.xiii) The newly created Python Script gets saved in the **'Saved Scripts'** list.

= C 👪 Search Tree Q	Component Console Summary Result Visualization Properties DataInsight	÷ 1
📲 Performance	Primary Function Details	
> 🆀 Data Writer 💡	Primary Function Name Input DataFrame Output DataFrame Summary Variable Name	
🛚 💁 Custom Scripts	new_fun v df v Output Data Frame Summary	
Create New Script	Show Visualization     Show Summary	
4 🝯 Saved Scripts		
💆 Custom Python Script		
😼 script	Function Parameter Definition Function Parameters Property Display Name Control Type	
🕵 datatype_checking	x     x     TextBox	
🕵 trial		
🕵 custom_python_script		
🕵 Utility Module Scripts		
🛛 🗾 Jupyter Notebooks	General         Script         Settings         The Python Script is saved.	

#### **Guidelines for Writing a Python Script**

- 1. The first argument of the function should be a data frame.
- 2. The Python script needs to be written inside a valid Python function. E.g., the entire code body should be inside the proper indentation of the function (Use 4 spaces per indentation level.)
- 3. The Python script should have at least one primary function. Multiple functions are acceptable, and one function can call another function, but it should be written above the calling function body (if the called function is an outer function) or above the calling statement (if the called function is an inner function).



- 4. Continuation lines should align wrapped elements either vertically using Python's implicit line joining inside parentheses, brackets, and braces, or using a hanging indent. When using a hanging indent, the following should be considered; there should be no arguments on the first line, and further indentation should be used to distinguish itself as a continuation line clearly.
- 5. Spaces are the preferred indentation method.
- 6. Limit all lines to a maximum of 79 characters. The Python standard library is conservative and requires limiting lines to 79 characters (and doctrines/comments to 72).
- 7. Do not use "type" as the function argument, as it is a predefined keyword.
- 8. In Python, single-quoted strings and double-quoted strings are the same.
- 9. All the packages used in the function need to import explicitly before writing the function.
- 10. The Python script should return data in the form of a data Frame only and should define while writing the function.
- 11. The column names should remain the same while creating new columns in the Output Table Definition.
- 12. If users need to define column selector (Multiple), then in the definition ': List[String]' should be used and the body of the function should be in '.to Array'.
- 13. If users need to define a column selector (Single), then 'String' must be used in the definition.

#### Note:

- **a.** Click the **'Information**' button <sup>1</sup> to get the rules to write a valid Python script.
- **b.** All the supported date data types are listed in date formats in the data type definition, all other date formats are considered string data types.
- c. Mssql data types are considered string data types.

### 14.2.2. Saved Python Scripts

#### 14.2.2.1. Viewing a Saved Python Script

- i) Select a Script from the 'Saved Scripts' list.
- ii) Use right-click on the selected Script.
- iii) A context menu opens.
- iv) Select the 'View' option.
- v) The users get redirected to the 'Component' tab.

Data Science Workbench 🗬				▶ C   +	छ।∎ ⊠ ७ €	0   Q
E C 🕰 Search Tree 🛛	Create New Workflow					
-						
4 🝯 Custom Scripts						
Create New Script						
🔺 🕵 Saved Scripts						
ダ Custom Python Script	₩					
View	Component Console	Summary Result	Visualization	Properties	DataInsight	⊥ ⊺
Edit	Basic					
Share	Component Name	Custom Python Script				
Delete	Component Type	Algorithms				
Pull from VCS script	Component Type	Algonithms				
Push into VCS	Description	Optional				
Jupyter Notebooks	General Script	Settings				

# 14.2.2.2. Editing a Saved Python Script



- i) Select a Script from the list of 'Saved Scripts' list.
- ii) Use a right-click on the selected script.
- iii) A context menu opens.
- iv) Select the 'Edit' option.
- v) The users get redirected to the '**Component**' tab.
- vi) The users can edit the required fields provided under the **General**, **Script**, and **Settings** tabs.

E C Search Tree Q	Component Cons	ole Summa	ry Result	Visualization	Properties	DataInsight	<u>+</u> ⊥
🔺 🕵 Custom Scripts	Basic						
Create New Script	Component Name	Custo	m Python Script				
4 💰 Saved Scripts	Component Type	Algori	ithms				
🕵 Custom Python Script	Description	Optio	nal				
View Edit 1g Share Delete Pull from VCS script Push into VCS 5	Script Type		stom Python ript	O Utility Module	Script		
Jupyter Notebooks	General Sci	ript	Settings				Next

### 14.2.2.3. Sharing a Saved Python Script

The users can share a custom Python script with other users and groups using the Share option. The following options are available to share a custom Python script:

- 1. **Share With**: This option allows the user to share a custom Python script with selected users or user groups. Any changes made to the custom Python script will be transferred to all the users with whom the custom Python script has been shared.
  - i) Select a Python script from the list of **Saved Scripts**.
  - ii) Select the 'Share' option from the context menu.
  - iii) The 'Share With' option gets displayed (by default).
  - iv) Select either the 'Group' or 'Users' option.
    - a. By selecting a group, all group members inside the group will be listed. Users can be excluded by not selecting them from the group when the '**Group**' option has been selected.
    - b. The users can be excluded by not selecting a username from the list when the 'Users' option

has been selected.

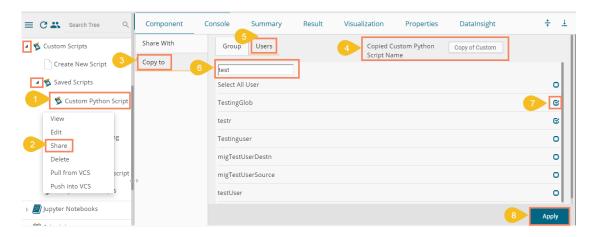
- v) Search for specific users or groups by using the Search space.
- vi) Select a specific user or group from the list by check-marking the box.
- vii) Click the 'Apply' option.



E C Search Tree Q	Component	Console	Summary	Result	Visualization	Properties	DataInsight	*	<u>+</u>
Custom Scripts 3	Share With	4 Group	Users						
Create New Script	Copy to	5 Search							
🔺 🕵 Saved Scripts		Select All	Group						0
1 Scustom Python Script		Admin Ro	ble					6	C
View		Viewer R	ole					*	0
Edit		UserGrou	qı					*	0
2 Share Delete		Kpmg Gr	oup					*	0
Pull from VCS script		testUser	Group					*	0
Push into VCS	•	Powerus	er					*	0
Jupyter Notebooks									
								7 Арр	ly

viii) The selected Python script gets shared with the chosen user(s)/group(s).

- 2. **Copy To**: This option creates a copy and shares a copy of the custom Scala script with the selected users and user groups. Any changes to the original custom Scala script after sharing will not show up for the users that received the shared file via the '**Copy To**' option.
  - i) Select a Python script from the list of 'Saved Scripts'.
  - ii) Select the 'Share' option from the context menu.
  - iii) Select the 'Copy To' option.
  - iv) The copied custom Python script name will be displayed in a box.
  - v) Select either the 'Group' or 'Users' tab.
    - a. By selecting a group, all group members inside the group will be listed. Users can be excluded by not selecting them from the group when the '**Group**' option has been selected.
    - b. Users can be excluded by not selecting a username from the list when the 'Users' option has been selected.
  - vi) Search for a user or group by using the search space.
  - vii) Select a specific user or group from the list by check-marking the box.
  - viii) Click the 'Apply' option.



# 14.2.2.4. Deleting a Saved Python Script

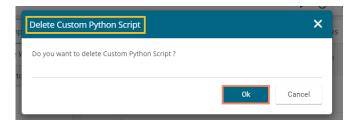
- i) Select a Python Script from the 'Saved Scripts' list.
- ii) Right-click on the selected Scala Script.



- iii) A context menu opens.
- iv) Select the '**Delete**' option.

💶 蓫 Custom Scripts	
Create New Scrip	ot
💶 ≶ Saved Scripts	
🝯 Custom Pyth	on Script
View	
Edit	ng
Share	0
Delete	
Pull from VCS	script
Push into VCS	5

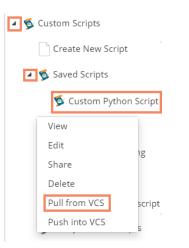
- v) The Delete Custom Python Script window opens to assure the deletion.
- vi) Click the '**Ok**' option.



vii) The selected script gets deleted.

### 14.2.2.5. Pull from VCS

- i) Select a saved script.
- ii) Select the 'Pull from VCS' option from the context menu.



- iii) The 'Pull from Version Controlling System' window opens.
- iv) Select the version(s) of the script that you wish to pull.
- v) Click the '**Pull**' option.



Pull fr	om Version	Controlling System	×
Branchr VCS_b	name for Pull * pranch		
	Version	Commit Date	Committed by
0	v1	19 Mar, 2020 06:13:40 GMT	William Martin
Commit I Python	Message Saved Script		
Warnii	<b>ng</b> : Deployed	workflow will undeploy after Pull o	peration.
			Pull Cancel

vi) A message appears to confirm that the data is pulled.



# 14.2.2.6. Push into VCS

- vii) Select a saved script.
- viii) Select the 'Push into VCS' option from the context menu.

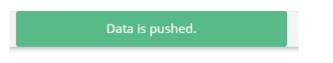
4 🕸	Custom Scripts	
	Create New Scr	ipt
4	🚳 Saved Scripts	
	ダ Custom Pyt	hon Script
	View	
	Edit	ıg
	Share	0
	Delete	
	Pull from VCS	script
	Push into VCS	5

- ix) The 'Push into Version Controlling System' window opens.
- x) Select a branch name for the push.
- xi) Provide the commit message.
- xii) Click the '**Push**' option.



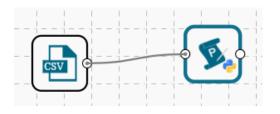
Push into Version Controlling System
Branchname for Push * VCS_branch
Commit Message (required) * Python Saved Script
<b>Warning</b> : Please save data if selected entity has been updated to prevent any data lose. In case if updated version exists on configured repository, data will be updated.
Push Cancel

xiii) A success message appears to confirm that the data has been pushed.



# 14.2.2.7. Connecting Saved Python Script with a Data Source

- i) Click the 'Custom Python Script' tree node.
- ii) Select and drag a saved Python script to the workspace.
- iii) Connect the Python Script to a configured data source.
- iv) Click the dragged 'Python Script' component.



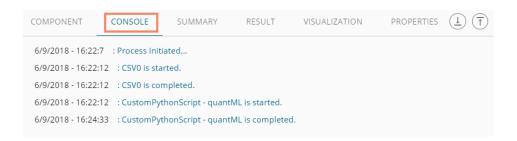
- v) Configure the required fields in the 'Custom Group' tab.
- vi) Click 'APPLY'

COMPONENT	CONSOLE	SUMMARY	RESULT	VISUALIZATION	PROPERTIES	$(\underline{1})(\overline{\uparrow})$
General	Dynamic Field	S				
Custom Group	х		ello			
>						
						APPLY

vii) After getting the success message run the workflow



#### viii) Users will get the process status under the 'CONSOLE' tab

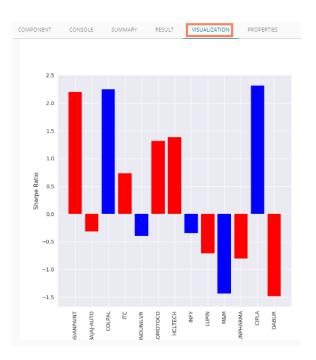


#### ix) Follow the below-given steps to display the result view:

- a. Click the dragged Python component on the workspace.
- b. Click the 'RESULT' tab.

Show 10 • entries Search:									
Category	Sharpe	Mean	Risk	Skew	%ир	%Down	Suggestion		
ASIANPAINT	2.2030408166105375	0.14000661722622762	0.22014896192869232	-0.06900642087301212	0.75	0.25	3		
AJAJ-AUTO	-0.3177065940151844	-0.013857152174100246	0.15109092518619893	0.11717177808347531	0.5	0.5	3		
OLPAL	2.251838714300893	0.07889388828628727	0.12136590604885886	0.9535998577259107	0.75	0.25	-3		
rc	0.7331135544309868	0.06519084746374554	0.30803920978740906	1.473192027990805	0.5	0.5	3		
IINDUNILVR	-0.4002884334177015	-0.011890271063565994	0.10289856952410058	-0.09109831006676725	0.5	0.5	-3		
EROMOTOCO	1.3202203304714948	0.05652638362336265	0.14831852857292047	0.03267872250176619	0.6666666666666666	0.3333333333333333333	3		
CLTECH	1.3869160530891287	0.03971886370384778	0.0992058456612971	-0.4683947882728144	0.6666666666666666	0.25	3		
IFY	-0.3437118922664428	-0.01835622747553245	0.1850033085167015	0.5903718468849175	0.416666666666666	0.58333333333333334	-3		
UPIN	-0.7128405424741218	-0.037619918477645675	0.18281679084561048	-0.1086621290968751	0.416666666666666	0.58333333333333334	3		
18:M	-1.4382216587471626	-0.06983137833970447	0.1681959029212423	0.32982346399266066	0.33333333333333333333	0.6666666666666666	-3		

x) Click the 'VISUALIZATION' tab to display the result data through a column chart.





xi)	Click the 'SUMMARY'	tab to view a summary of the process.	
-----	---------------------	---------------------------------------	--

	(*												
	CIPLA	SUNPHARMA	M&M	LUPIN	INFY	HCLTECH	HEROMOTOCO	HINDUNILVR	ITC	COLPAL	BAJAJ-AUTO		DABUR
	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	7.000000	count 7.00000
	0.182414	0.536085	-0.430005	0.474813	-0.226670	0.710588	0.793963	-0.342911	0.939934	0.200814	0.562386	0.927741 5	mean 0.53568
	1.596073	1.183617	1.321089	1.189576	1.268560	1.163812	1.073325	1.214917	1.010162	1.599156	1.112336	1.192077 5	std 1.36738
-	-3.000000 -	-0.808388	-3.000000	-0.712841	-3.000000	-0.468395	0.032679	-3.000000	0.065191	-3.000000		9	1.48425
			01101021			01000102	01202.22	-0.245693	01101020	01200200	01002007	-	0.02800
	0.250000	0.333333	0.168196	0.182817	0.185003	0.250000	0.333333	-0.011890	0.500000	0.250000	0.151091	0.250000 0	50% 0.25000
	01725002	0.000.200	01002070			11020/02		0.301449		01002000		-	0.96402
	2.316329	3.000000	0.666667	3.000000	0.590372	3.000000	3.000000	0.500000	3.000000	2.251839	3.000000		max 3.00000

### **14.3.** Jupyter Notebooks

The integrated Jupyter Notebook tree-node allows the users to create and share documents that contain live code, equations, visualizations, and narrative text. It can be used in numerical simulation, statistical modeling, data visualization, machine learning. The key motive is to introduce live coding inside the Data Science workbench and more efficiently use it as a component.

- i) Upload a Data Source and run it.
- ii) Connect the Jupyter Notebook component to it.



- iii) Configure the General tab for the Jupyter Notebook component.
- iv) Provide the Component name.
- v) Click the '**Next**' option to load the Jupyter Notebook Kernel. This will trigger the Notebook Kernel in the backend and start it.



Component	Console	Summary	Result	Visualization	Properties	DataInsight 🛓	Ť
General	Basic						
Properties	Component	Name	Notebook_sani	ty_02			
•	Component	Туре	Jupyter Notebo	ook			
	Description		Optional				
						Next	t

vi) After loading Notebook Kernel a new page gets listed in the footer tab as shown below:

Component Console	Summary Result Visualization Properties DataInsight	4	ŧ	<u>+</u>
General	Notebook Kernel			
Properties		Kernel 🌒		
>	E3 + 3< 2 K + ↓ HRun ■ Code ▼			_
>	Notebook Kernel Variables NOTEBOOK_INPUT_DATA - To access the input Data of Notebook Kernel NOTEBOOK_SAVE_DATA(output_dataframe)-Call method to write Output Data of Notebook Kernel NOTEBOOK_SAVE_VIZ() - Call method to write Visualisation(s) of Notebook Kernel ZIP_FILE_LOCATION(),ZIP_FILE_OUTPUT_LOCATION()- Use for ZIp operations			
			Save	

- vii) Provide the script with proper Input, Output, and Save functions. The user must follow the instructions given for the Notebook Kernel Variables to move further.
  - 1. To load the input data using NOTEBOOK\_INPUT\_DATA

eneral	Notebook	Kernel														
perties															🖌   Kernel	0
	8	+ *	08	<b>↑ ↓</b>	N Run	Code	•									
			Notes	book Kernel	variables											
			NOTE	BOOK_INP	UT_DATA	- To acc	ess the Input Data of	Notebo	ok Ke	rnel						
			NOTE	BOOK SAN	E DATA	output o	iataframe)-Call metho	d to wr	ite Ou	tput Da	ta of N	otebook k	ernel			
										4						
			NOTE	BOOK SA	E VIZO -	Call met	hod to write Visualis.	ation(s)	of No	tebook	Kernel					
			NOTE	BOOK_SA	E_VIZ() -	Call met	hod to write Visualis	ation(s)	of No	tebook	Kernel					
				1997 - 1997 - 1997 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 19												
				1997 - 1997 - 1997 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 19			hod to write Visualis UTPUT_LOCATION()									
		In [5]:	ZIP_F	1997 - 1997 - 1997 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 19	ION(), ZIP											
			ZIP_F	FILE_LOCAT	ION(), ZIP											
		In [5]: Out[5]:	ZIP_F	FILE_LOCAT	DATA	FILE_O		Use fo	r Zip o		ins	Ticket	Fare	Cabin	Embarked	
			ZIP_F	FILE_LOCAT	DATA	FILE_O	UTPUT_LOCATION()	Use fo	r Zip o Age	operatio	Parch		Fare 7.250	Cabin C92	Embarked	
			ZIP_F	FILE_LOCAT	DATA Survived	FILE_O	UTPUT_LOCATION()- Name Braund, Mr. Owen Harris	Use foi	r Zip o Age	sibSp	Parch	Ticket				
			ZIP_F	FILE_LOCAT	DATA Survived	FILE_O	UTPUT_LOCATION() Name Braund, Mr. Oven Harris Cumings, Mrs. John Bradley (Porence	Use foi Sex male	Age	sibSp	Parch 0	Ticket	7.250			
			ZIP_F NOTER	FILE_LOCAT	DATA Survived	FILE_O Pclass 3	UTPUT_LOCATION()- Name Braund, Mr. Owen Harris Cumings, Mrs. John	Use foi Sex male	Age	sibSp 1	Parch 0	Ticket A/5 21171	7.250	C92	S	
			ZIP_F NOTER	FILE_LOCAT	DATA Survived 0	FILE_O Pclass 3	UTPUT_LOCATION() Name Braund, Mr. Oven Harris Cumings, Krs. John Bradley (Forence Briggs Th	Use foi Sex male	Age 22.0 38.0	sibSp 1	Parch 0 0	Ticket A/5 21171 PC 17599 STON/02	7.250	C92	S	
			ZIP_F : NOTEE : 0 1 2	FILE_LOCAT	DATA Survived 0 1	FILE_O Pclass 3 1 3	Name Braund, Mr. Oven Harris Cumings, Mrs. John Bradley (Forence Briggs Th Heikkinen, Miss. Laina	Sex Male female	<b>Age</b> 22.0 38.0 26.0	sibsp 1 1	Parch 0 0	Ticket A/5 21171 PC 17599 STON/02 3101282	7.250 71.283 7.925	C92 C85 None	S C S	
			ZIP_F NOTEE 0 1	FILE_LOCAT	DATA Survived 0	FILE_O Pclass 3	UTPUT_LOCATION() Name Braund, Mr. Oven Harris Cumings, Krs. John Bradley (Forence Briggs Th	Sex Male female	Age 22.0 38.0	SibSp 1	Parch 0 0	Ticket A/5 21171 PC 17599 STON/02 3101282	7 250 71 283	C92 C85	s	



2. To save the output data use NOTEBOOK\_SAVE\_DATA(output\_dataframe)

eneral																	
neral	Noteb	ook Ke	rnel														
operties																🖋 Kernel C	>
		8 +	30 0	2 6 1	• • • NR	un 🔳	Code										
				3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100	C123	S	
				4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050	None	s	
				886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.000	None	s	
				887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.000	B42	s	
				888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.450	None	s	
				889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.000	C148	С	
				890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	None	Q	
				891 rows	× 12 column	s											
			In [8]:	NOTEBOOK	_SAVE_DATA	(outpu	t)										
			Out[8]:	Output	has been S	aved S	uccessf	ully'									
			In [ ]:	I													

3. To call method to write Visualizations use NOTEBOOK\_SAVE\_VIZ()

Co	omponent Co	onsole S	ummary Res	ult Visualization Properties DataInsight	+	<u>+</u>
Ger	neral	Notebo	ok Kernel			
Pro	operties					
Þ			8 + % 4	🗄 🚯 🛧 🖌 Markdown 🔻		
			In [6]:	<pre>import pandas as pd import numpy as np import numpy as np import seaborn as sns import matplotlib.pyplot as plt  def correlation(df, columns, method):     #df = df[columns]     corrmat = columns.corr(method= method)     result = pd.DatErname(corrmat).reset_index()     result = result.rename(columns-{'index':'category'})     f, ax = plt.subplots(figsize = (8, 7))     plt.title("Correlation Matrix")     sns_plot= sns.heatmap(corrmat, ax = ax, cmap ="magma", linewidths = 0.1)     plt.savefig(NOTEBOOK_SAVE_VII())     return {'opt': df, 'summary' : result.describe(include="all")}</pre>		
>				<pre>df = df columns = df[['Pclass','Age', 'Fare', 'Survived', 'Parch', 'PassengerId']] method = 'pearson' df_out = correlation(df, columns, method) df_out {'opt': PassengerId Survived Pclass \</pre>		
					Sav	e

4. To load input data from a zip reader use **ZIP\_FILE\_LOCATION()** 



Component	Console	Summary	Result	Visualizatio	n F	Properties	DataInsight	+	Ŧ
General		Notebook Kernel							
Properties								Python 3 O	>
		8 + % 4	8 10 1 4	N Run	Code				
		In [11]:	NOTEBOOK_S NOTEBOOK_S ZIP_FILE_LOO	NPUT_DATA SAVE_DATA( SAVE_VIZ() - CATION(),ZIP_ s as pd	- To acce output_di Call meth _FILE_OU	ataframe)-Call i nod to write Vis	nta of Notebook Kernel method to write Output Data of Notebook Kernel ualisation(s) of Notebook Kernel ION()- Use for Zip operations		
		In [12]:	df						
		Out[12]:		r_number	Date	itemDescription			
			0	1808 21-	07-2015	tropical frui	t		
			1	2552 05-	01-2015	whole mill	¢.		
			2	2300 19-	09-2015	pip frui	t		
			3	1187 12-	12-2015	other vegetables	3		
			4	3037 01-	02-2015	whole mili	¢.		
								Sa	ave

5. To save the output data to a zip file location use **ZIP\_FILE\_OUTPUT\_LOCATION()** 

neral	Notebook H	Kernel					
operties							
	8 +	* 2		↑ ↓ N Run			
				Member_number	Date	itemDescription	
			0		21-07-2015	tropical fruit	
			1		05-01-2015	whole milk	
			2		19-09-2015 12-12-2015	pip fruit	
			4		01-02-2015	other vegetables whole milk	
			38760	4471	08-10-2014	sliced cheese	
			38761		23-02-2014	candy	
			38762		16-04-2014	cake bar	
			38763			fruit/vegetable juice	
			38764	1521	26-12-2014	cat food	
			38765 r	ows × 3 columns			
	I	n [15]:	df.to_	csv( <mark>ZIP_FILE_0</mark>	UTPUT_LOCA	TION() + "df.cs	v")
		In [ ]:					

viii) Once you have saved the output of the Jupiter Notebook, click the 'Save' option.



Component	Console	Summary	Result	Visualization	Properties	DataInsight	<u>+</u> ↑	<u>+</u>
General	Notebo	ook Kernel						
Properties	F	ile					Python 3 O	
	8	+ % 4		Run Markdown	• PASILIT (IASCP	ipe(iucinde= 911 )}		
		#0 m0	f = df plumns = df[ columns = df. ethod = 'pear	['Pclass','Age', ' .loc[1:3]	Fare', 'Survived	', 'Parch', 'Passenge	erId']]	
		In [ ]: NO	DTEBOOK_SAVE	_DATA(df_out['opt'	])			-
· P								
							Sav	/e

- ix) A message appears to inform that the Jupyter Notebook has been saved.
- x) The Jupyter Notebook component gets added to the **Saved Notebooks** section.

= C 🔹 Search Tree Q	Component	Console Sum	mary Result	Visualization	Properties	DataInsight	+	<u>+</u>
💵 Performance	General	Notebook Kerr	nel					
🕨 🖀 Data Writer	Properties	File					Python 3	0
Scustom Scripts		B + %		un Markdown v ur, summary : r	.ezotr.oszcutos(	IUCIOO6= 911 )}		
Jupyter Notebooks      Jupyter Notebook      Saved Notebooks      Notebook, sanity, 02		In [ ]:	<pre>df = df columns = df[['Pcl #columns = df.Loc[ method = 'pearson' df_out = correlati df_out</pre>	1:3]		Parch', 'Passenger	.Id.]]	
Notebook_sanity_02	÷	In [ ]: In [ ]:	NOTEBOOK_SAVE_DATA	(df_out['opt'])				l
🗾 titanic_dataset				Save	d.			

- xi) Run the workflow.
- xii) Once the Workflow runs successfully the user can see Summary, Result, Visualization, and DataInsight for the newly saved Jupyter Notebook component.
  - a. Click the 'Result' tab to see the processed data.

now 10 🔻 e	ntries								Se	arch:	
Passengerid	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	C92	s
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.283	C85	С
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/02. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1	C123	s
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05		s
5	0	3	Moran, Mr. James	male		0	0	330877	8.458		Q
7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.862	E46	s
8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.075		s
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.133		s
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.071		с

b. Click the 'Summary' tab to see the model summary.



C	Compone	nt Cor	isole S	ummary	Resu	lt	Visualization	Properties	DataInsight	+	<u>+</u>
•		Summar	y of the mode	21							
		PassengerId	Survived	Pclass			Name				
	count	891.000000	891.000000	891.000000			891				
	unique	NaN	NaN	NaN			891				
	top	NaN	NaN	NaN	Panula	, Master	. Eino Viljami				
	freq	NaN	NaN	NaN			- 1				
	mean	446.000000	0.383838	2.308642			NaN				
	std	257.353842	0.486592	0.836071			NaN				
	min	1.000000	0.00000	1.000000			NaN				
	25%	223.500000	0.00000	2.000000			NaN				
	50%	446.000000	0.00000	3.000000			NaN				
	75%	668.500000	1.000000	3.000000			NaN				
	max	891.000000	1.000000	3.000000			NaN				
		Sex	Age S	SibSp F	Parch	Ticket	Fare				
	count	891 714.0	00000 891.00	00000 891.00	0000	891	891.000000				
	unique	2	NaN	NaN	NaN	681	NaN				
	top	male	NaN	NaN	NaN C	A. 2343	NaN				

c. Click the 'Visualization' tab to see the visual presentation of the data.



d. Click the 'DataInsight' tab to see the data insights.

Component Console	Summary	Result	Visualization	Propertie	Datalnsi	ight 🗧
Profiling Report		Overview	Variables	Correlations	Missing values	Sample
Overview	/					
Dataset inf	o		Variable	es types		
Number of variables	12		Numeric		5	
Number of observati	ons 891		Categorical		5	
Missing cells	865 (8.1%)		Boolean		1	
Duplicate rows	0 (0.0%)		Date		0	
Total size in memory	1 1		URL		0	
Average record size			Text (Unique)		1	
memory	III 90.1 D		Rejected		0	
			Unsupported		0	

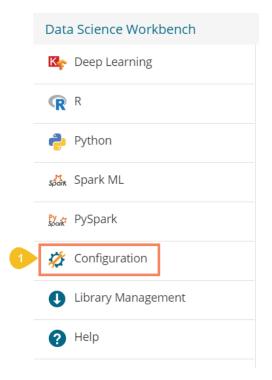


# **15. Configuration**

The user gets redirected to the Admin module containing the server configuration option for the Data Science plugin.

## 15.1. Configuring Python Server

i) Click the 'Configuration' option from the Data Science Workbench homepage.



- ii) By default, the Python Server details open under the Admin Module.
- iii) The user can select another Python server from the available server list by selecting the radio button icon

or	

Click the 'Add new server' + icon to configure a new Python server.

=	Admin Module			
2 Pytho	n Server			3 + Test Save
	Server name	Default	Actions	
	172.31.42.225	0	× 1	
	Prod Python 3.6	•	× =	
	dl-server	•	/ 1	



- iv) The 'Create Python Server' page opens by clicking the 'Add new server' option.
- v) Provide the following information:
  - i. Host: Host address of the Spark server
  - ii. Port: Spark server's port number
  - iii. Username: Enter a username to log in to the Spark server
  - iv. Password: Enter the password for the above username
  - v. Python Server Name: Provide Python Server Address
  - vi. Elastic Search Port: Provide the elastic search port number
  - vii. Server API URL: Provide the server API URL link
  - viii. Tensor Board Visualization URL: Provide the Tensor Board Visualization URL link
  - ix. Python Server Protocol: Select a protocol option by using the radio option out of HTTP and HTTPS
- vi) Click the 'Test' option to verify the connection.

5 Server Information	
Host *	Port *
User name *	Password *
	Elastic search port *
Python server name *	9200
Server API URL *	TensorBoard visualization URL *

- vii) A success message appears to assure about the Python Server connection.
- viii) Click the enabled 'Save' option to save the verified Python server information.



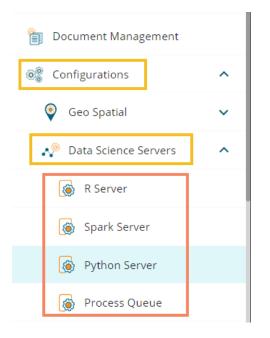
- ix) A success message appears to ensure that the predictive settings got updated.
- x) The newly configured Python Server gets added to the 'Python Server' window.



			III Ļ <sup>11</sup> (	? W
Python Server			+ Tes	st Save
Server name	Default	Actions		
172.31.42.225	۲	× •		
Prod Python 3.6	۲	× +		
dl-server	•	× =		
10 Sample Python Server	۲	× •		

#### Note:

- a. Click the 'Edit' 🖍 icon to modify an existing python server configuration
- b. Click the '**Delete**'<sup>**I**</sup> icon to remove the selected Python server details from the list.
- xi) To access the other Data Science Servers, the user can click on the '**Configurations**' option provided under the Admin Module.
- xii) Choose the 'Data Science Servers' option.
- xiii) All the available options appear as displayed below:



## **15.2.** Configuring R Server

i) Select the **R Server** option from the Data Science server list.



Configurations	^
Geo Spatial	~
Nata Science Servers	^
1 🛞 R Server	
Spark Server	
Python Server	
Process Queue	

- ii) The R Server page opens.
- iii) The user can select another R server from the available server list by selecting the radio button icon
   or

Click the 'Add new server' T icon to configure	ure a new K server.	
--	---------------------	--

	Admin Module					
2 R Serv	ver				3 <b>+</b> Te	est Save
	Server name	Cores	Workflow	Scheduler	Actions	
	34.209.28.33	1	0	0	× =	
	R server	1 3	۲	۲	× =	

- iv) The 'Create R Server' page opens by clicking the 'Add new server' option.
- v) Provide the following information to configure a new R server:
  - i. IP Address: IP address of the R-server
  - ii. Port: R-Server's port number
  - iii. Username: Enter a username to log in to the R- server
  - iv. Password: Enter the password for the above username
  - v. R Server Name: Provide the R- Server address
  - vi. Provide HTTP URL for R-Bokeh: Provide R Visualization URL
  - vii. Elastic Search Port: Provide an elastic search port number
  - viii. R Visualization URL: Provide HTTP URL for R-Bokeh
  - ix. Enable Parallel Processing: Avail this option by using the enable/disable button1. By enabling the Parallel Processing, it asks to configure 'Number of Cores'
  - x. Set as Default: Select this option by using a checkmark in the box
  - xi. The user gets further options for the Parallel enabled Processing:
    - 1. Utilize for Workflow and Scheduler



- 2. Utilize for the only Workflow
- 3. Utilize for the only Scheduler
- vi) Click the 'Test' option to verify the R-Server connection.

5 Server Information		
IP address *	Port *	
		Can contain only numbers. Maximum
User name *	Password *	
R server name *	Elastic search port	
Sample R Server	9200	
HTTP URL for R-Bokeh * https:// <ip:port domainname="" rviz=""></ip:port>		
Enable parallel processing		
Number of cores *		
2		

- vii) A success message appears to assure about the R Server connection.
- viii) Click the enabled 'Save' option to save the verified R server information.

		¢"	?	W
Successfully connected to R Server		0		
	Test	Sav	/e	←

- ix) A success message appears to ensure that the predictive settings got updated.
- x) The newly configured R Server gets added to the '**R Server**' window.

BBB@		9 Prec					¢"	0	W
≡ Admir	n Module	Prec		dated successfully					
R Server						+	Test		
	Server name	Cores	Workflow	Scheduler	Actions				
	34.209.28.33	1	0	0	× +				
	R server	2	•	0	Z 1				
10	Sample R Server	2	۲	۲	Z #				

Note:

- a. Click the 'Edit' 🖍 icon to modify an existing R server configuration
- b. Click the '**Delete'** icon to remove the selected R server details from the list.

### 15.3. Configuring Spark Server

i) Select the Spark Server option from the Data Science server list.



Configurations	^
Seo Spatial	~
🎺 Data Science Servers	^
R Server	
1 🛞 Spark Server	
Python Server	
Process Queue	

- ii) The Spark Server page opens.
- iii) The user can select another Spark server from the available server list by selecting the radio button icon. **or**

Click the 'Add new server'  $\pm$  icon to configure a new Spark server.

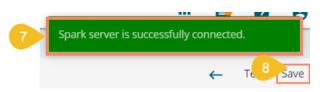
Admin Module			
2 Spark Server			3 + Test Save
Server name	Default	Actions	
172.31.41.103	0	× =	
Spark Server	3	2.1	

- iv) The 'Create Spark Server' page opens by clicking the 'Add new server' option.
- v) Provide the following information to configure a new Spark server:
  - i. Host: Host address of the Spark server
  - ii. Port: Spark server's port number
  - iii. Username: Enter a username to log in to the Spark server
  - iv. Password: Enter the password for the above username
  - v. Spark Server Name: Provide Spark Server Address
  - vi. Jetty Confirmation URL: Provide Jetty confirmation URL link
  - vii. Application: Provide the application name
  - viii. Spark Server Protocol: Select a protocol option by using the radio option
- vi) Click the 'Test' option to verify the Spark Server connection.

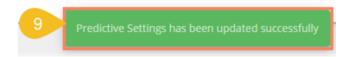


Create Spark Server	6 Te
5 Server Information	
Host *	
Port *	User name *
Password *	Spark server name *
Jetty configuartion URL *	Application *

- vii) A success message appears to assure about the Spark Server connection.
- viii) Click the enabled 'Save' option to save the verified Spark server information.



ix) A success message appears to ensure that the predictive settings got updated.



x) The newly configured Spark Server gets added to the '**Spark Server**' window.

🗮 Admin Module			
Spark Server			+ Test Save
Server name	Default	Actions	
172.31.41.103	۲	× 1	
Spark Server	۲	/ 1	
dl-server	0	× =	
10 Sample Spark Server	0	× =	

#### Note:

- a. Click the 'Edit' 🖍 icon to modify an existing Spark server configuration
- b. Click the '**Delete**' <sup>I</sup> icon to remove the selected Spark server details from the list.



# 15.4. Configuring Process Queue

The user can reset the Predictive process queue through this Predictive Settings option.

i) Click the 'Process Queue' option from the Data Science Servers configuration options.

Nota Science Servers	^
🔞 R Server	
Spark Server	
Python Server	
Process Queue	

- ii) The Process Queue page opens.
- iii) Click the 'Reset Queue' option.

Admin Module Process Queue	
	Click the 'Reset Queue' option to set the count of currently running processes to zero.
Java R	
	Reset Queue

- iv) A warning message appears, asking whether the user wants to flush the queue.
- v) Click the '**YES**' option to set the count of currently running processes to Zero for the Data Science Workbench.





# **16. Library Management**

The Library Management option facilitates the user to install R and Python libraries from an online repository or the source code uploaded on the server to the R and Python Data science servers.

Click the Library Management option from the Data Science Workbench. Details to open R and Python Packages get displayed.

Data Science Workbench	1	n
🔀 Deep Learning	R Packages	
R R	Install New Package	Install New Package It's an user interface for end users to install required libraries quickly in currently selected server in Predictive
🄁 Python	Python Packages	admin module.
🕰 Spark ML	Install New Package	In containerization deployment, libraries installed through this predictive will be lost after restart of server, as container will revert back to its base version.
🗶 PySpark	Installed Packages	container win revert back to its base version.
Library Management		Installed Packages
Help	-	It's an user interface for listing the libraries installed in currently selected server in Predictive admin module.
		Library name and it's version is visible

### i) R Packages

- a) Navigate to the Library Management page.
- b) The R Packages option displays.
- c) Click the 'Install New Package' option to open the R-Install New Packages screen.
- d) Select an option from where you want to install the package using the drop-down list.
- e) Provide the package names in the given box. Use a comma to separate multiple packages.
- f) Click the 'INSTALL' option to install the packages.

Data Science Workbench			🏚 Q
Rep Learning 2	R Packages	R - Install New Packages	
<b>R</b> R 3	Install New Package	4 Install From	
🥐 Python	Python Packages	Local R Server - Package Archieve(.tar.gz) •	
star Spark ML	Install New Package	5 Packages to be Installed (separate multiple with comma)* 0	
🕵 PySpark	Installed Packages		1
🔅 Configuration			
1 Library Management			
Help			
			6
			CANCEL

g) Click the 'Installed Packages' to display all the installed packages.

(



Data Science Workbench			<b>n</b>
Ceep Learning	R Packages Install New Package	R - Installed Packages Search	۹
🦆 Python	Python Packages	abind (1.4-5) acepack (1.4.1) ada (2.0-5) adabag (4.2) AnomalyDetection (1.0) arules (1.5-1)	
🖧 Spark ML	Install New Package	arulesViz (1.3-3)         assertthat (0.2.1)         backports (1.1.4)         BAS (1.5.3)         base64enc (0.1-3)         BH (1.69.0-1)           bindr (0.1.1)         bindrcpp (0.2.2)         bitops (1.0-6)         BizVi2PredictiveAnalysis (3.6.0)	
2. PySpark	Installed Packages	BizVizPredictiveAnalysisLight (0.1.0) Boruta (6.0.0) broom (0.5.0) Cairo (1.5-9) callr (3.2.0) car (3.0-3)	
🔅 Configuration		carData (3.0-2)         caret (6.0-80)         caTools (1.17.1.1)         cellranger (1.1.0)         checkmate (1.8.5)         classint (0.2-3)           cli (11.0)         clipr (0.6.0)         clisymbols (1.2.0)         clustofVar (1.1)         coin (1.3-0)         colorspace	-
Uibrary Management		combinat (0.0-8) corrgram (1.13) corrplot (0.84) corrr (0.3.0) cowplot (0.9.4) crayon (1.3.4)	
Help		crosstalk (1.0.0)         curl (3.2)         customelasticsearchr (0.2.1)         CVST (0.2-2)         data.table (1.11.4)         DBI (1.0.0)           ddalpha (1.3.4)         dendextend (1.8.0)         DEoptimR (1.0-8)         desc (1.2.0)         devtools (2.0.2)         dichromat (2.0-0)	
		digest (0.6.19) dimRed (0.1.0) diptest (0.75-7) docopt (0.4.5) doParallel (1.0.11) dotCall64 (1.0-0)	
		dplyr (0.8.1)         DRR (0.0.3)         DT (0.4)         dummies (1.5.6)         e1071 (1.7-0)         elasticsearchr (0.2.2)         ellipse (0           EMCluster (0.2-12)         evaluate (0.11)         exactRankTests (0.8-30)         expss (0.8.11)         factoextra (1.0.5)	.4.1)
		FactoMineR (1.41) fansi (0.4.0) fastDummies (1.4.0) FastRWeb (1.1-1) fields (9.8-3) flashClust (1.01-	-21

### ii) Python Packages

- a) Select the 'Install New Package' option from the Python Packages.
- b) The 'Python- Install New Packages' fields open.
- c) Select an option from where you want to install.
- d) If the selected 'Install From' option is 'Repository,' it displays the Repository URL link.
- e) Mention the packages you want to install in the given box.
- f) Follow the below-given rules:
  - 1. Provide Package version as 'package name==version' (optional)
  - 2. Separate Multiple Packages with Comma (optional)
- g) Click the '**INSTALL**' option to install the new Python packages.

R Packages Install New Package	2 Python - Install New Packages	
Installed Packages	3 Install From	
Python Packages	Repository •	
Installed Packages	4 Repository URL*	
	5 Packages to be Installed	
		1
	6 * Provide Package version as package_name==version (optional) * Separate Nultiple Packages with Comma (optional)	7
		CANCEL

h) Click the 'Installed Packages' option to display all the installed Python Packages.



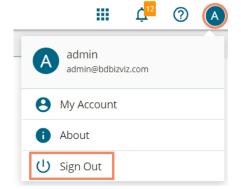
Data Science Workbench		♠ Q
🔀 Deep Learning	R Packages	Python - Installed Packages Search Q
R R	Installed Packages	
🥐 Python	Python Packages	absl-py (0.7.1)         altgraph (0.16.1)         amqp (2.4.2)         asgi-rabbitmq (0.5.5)         asgiref (1.1.2)         asn1crypto (0.24.0)           astor (0.7.1)         attrs (19.1.0)         autobahn (19.3.3)         automat (0.7.0)         backcall (0.1.0)         bcrypt (3.1.6)
्रद्धे <sub>स</sub> Spark ML	Install New Package	beautifulsoup4 (4.7.1)         belliard (3.5.0.5)         bils (0.2.4)         bokeh (1.0.0)         boto (2.49.0)         boto 3 (1.9.129)
Bar PySpark	Installed Packages	botcore (1.12.129)         bs4 (0.0.1)         bz2file (0.98)         cached-property (1.5.1)         catboost (0.11.2)         celery (4.2.0)           certifi (2019.3.9)         cffi (1.12.2)         channels (1.1.8)         chardet (3.0.4)         click (7.0)         cloudpickle (0.8.1)
💋 Configuration		Colorama (0.4.1)         Confluent-kafka (1.1.0)         Constantiy (15.1.0)         Cryptography (2.6.1)         csselect (1.0.3)
Library Management		cx-oracle (7.0.0)         cycler (0.10.0)         cymem (2.0.2)         daphne (1.4.2)         dask (1.1.5)         dbutlis (1.3)           decorator (4.4.0)         distributed (1.26.1)         django (2.1)         djangorestframework (3.9.0)         dlib (19.17.0)
Help		docopt (0.6.2) docutils (0.14) editdistance (0.5.2) elasticsearch (2.3.0) en-core-web-sm (2.1.0)
		et-xmlfile (1.0.1)         face-recognition (1.2.3)         face-recognition-models (0.3.0)         featuretools (0.5.1)         future (0.17.1)           fuzzywuzzy (0.17.0)         gast (0.2.2)         gensim (3.7.3)         geojson (2.3.0)         graphviz (0.10.1)         grocio (1.19.0)
		(1022)m022y (0.17.6)         gesc (0.2.2)         genation (3.7.5)         geographic (0.10.1)         graphic (0.10.1)         graphic (1.10.1)           gtts (2.0.3)         gtts-token (1.1.3)         h2o (3.22.0.3)         h5py (2.9.0)         hdfs (2.2.2)         heapdict (1.0.0)
		hosklearn (0.1.0) hvoerlink (18.0.0) hvoeroot (0.1.1) icc-rt (2019.0) idna (2.8) imaee (1.5.25)

Note: The containerized deployment does not support the libraries installed through this option as the container reverts to its base version.

# **17. Signing Out**

The following steps describe how to Sing-off from the BDB Platform.

- i) Click the 'User Profile' icon on the Platform homepage.
- ii) Click the 'Sign Out' option.



iii) The user successfully signs off from the BDB Platform.

### Note:

- a. By clicking the 'Sign Out' option, the user gets back to the Sign-in page of the BDB platform.
- b. Click the 'About' option to open the default homepage for the BDB Platform.



B.e.			ୟ 🏭 🗘 🍱 🕐
My Documents			Administration User
Welcome to BDB Decision platform	<i>2</i>		admin.user@bdb.ai
	Dashboard Designer	Business Story	1 About
BDB is a complete decision platform for all your business needs. Drive from data to dynamic visuals and derive an actionable insight into your business data. Avail 360° view of your business by assembling,	Design, save and publish a splendid governed dashboards. Display relevant KPIs through comprehensive and stunning visual reports to attain your business objectives.	Go beyond the classic BI with our ground-breaking self-service BI tool. Gain pertinent insights into your business by creating wide-ranging views on your own without external IT help.	Sign Out your next ousniess move: access an apply accurate and customizable Predictive models to maximize futu opportunities.
processing, and analyzing the acquired data. Access incomparable analytics at any time from anywhere on any	Data Center	© © ETL	
device.	Supports a wide range of Data sources	E I L A self-driven Data Wrangling tool to	Data Preparation Experience a secure yet self-driven
Version: 6.5.0	supports a white range of Data sources starting from the spreadsheets in your system to a cloud-based database. Establish connections to these data sources and build Data Sets or Data	extract data from diverse sources, including the merged data. Enforce data quality and consistency standards to deliver the output in a presentation-	experience a secure yet sen-orwern mode of data preparation. Streamlin the entire process of dealing with retrievable business data empowerin the business users to decide with
Released on: Sep 30, 03:28	Stores to enable rich business	ready format.	unprecedented agility.

### **17.1. Forgot Password Option**

The users are provided with a choice to change the password on the Login page of the platform.

i) Click the 'Forgot Password?' option from the Sign In page.

Decision Platform
Email / User Id *
Password *
Auth Type Enterprise
1 Forgot Password ?
Copyright © 2015-2020 BDB (BizViz Technologies Pvt Ltd)

- ii) The 'Forgot Password?' page opens.
- iii) Provide the email id that is registered with BDB to send the reset password link.
- iv) Click the 'Continue' option.

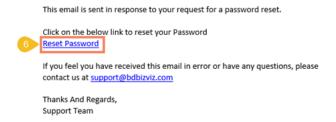
orgot Password?	
lease enter the registered emai eset your password.	laddre
mail	
dmin.user@bdb.al	
Concession of the local division of the loca	
Continue	
Continue	



v) The user may be redirected to select a space in case of multiple spaces under one server link (The user needs to select a space and click the 'Continue' option once again). If a user does not have multiple spaces then, a message appears to notify the user that the password reset link (The users receive the reset link via their registered email.)



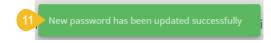
vi) Click the link from your registered email.



- vii) The user gets redirected to the 'Reset Password' page to set a new password.
- viii) Set a new password.
- ix) Confirm the newly set password.
- x) Click the 'Continue' option.

		ord	
	ave confirmed ov e reset your pass		
New Pa	ssword *		
Confirm	New Password *		

xi) The password for the selected BDB account gets reset and a message appears to inform the user.



Note: The user gets redirected back to the Sign In page after successfully resetting the password.

### 17.2. Force Login

The 'Force Login' functionality has been introduced to control the number of active sessions up to three. The users can access only 3 sessions at a time when they try to access the 4<sup>th</sup> session, a warning message displays

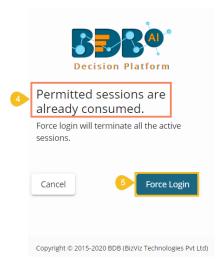


to inform that the user has consumed the permitted sessions and, a click on the 'Force Login' would kill all those active sessions.

- i) Navigate to the BDB Platform Login page.
- ii) Enter the valid credentials to log in.
- iii) Click the 'Sign In' option.



- iv) The user gets the following message if the permitted active sessions (3 sessions at a time) are consumed.
- v) Click the 'Force Login' option.



- vi) A warning message appears the currently active sessions get killed, and the user gets redirected to the BDB Platform Sign In page.
- vii) The user needs to provide valid credentials once again and click the '**Continue**' option to access the platform.