EXERCISE -9

(i) Implement Simple Linear Regression on an 'Employee' dataset.
(ii) Demonstrate the simple k-Means clustering algorithm on 'iris' dataset.

9.1.1 Problem Statement:

Implement Simple Linear Regression on an 'Employee' dataset.

9.1.2 Description:

<u>About Dataset used:</u>

Is Employee dataset with eight attributes

MarriedID

MaritalStatusID

GenderID

EmpStatusID DeptID

PerfScoreID

FromDiversityJobFairID

Salary

Termd

<u>About RapidMiner:</u>

Rapidminer is a comprehensive data science platform with visual workflow design and full automation. It means that we don’t have to do the coding for data mining tasks. Rapidminer is one of the most popular data science tools.

This is the graphical user interface of the blank process in rapidminer. It has the repository that holds our dataset. We can import our own datasets. It also offers many public datasets that we can try. We can also work with a database connection.

RapidMiner is an awesome visual workflow designer. The way they present visually is so unique. It helps in speeding and automating the creation of visual models. It helps in creating models in only 5 clicks by automated machine learning.

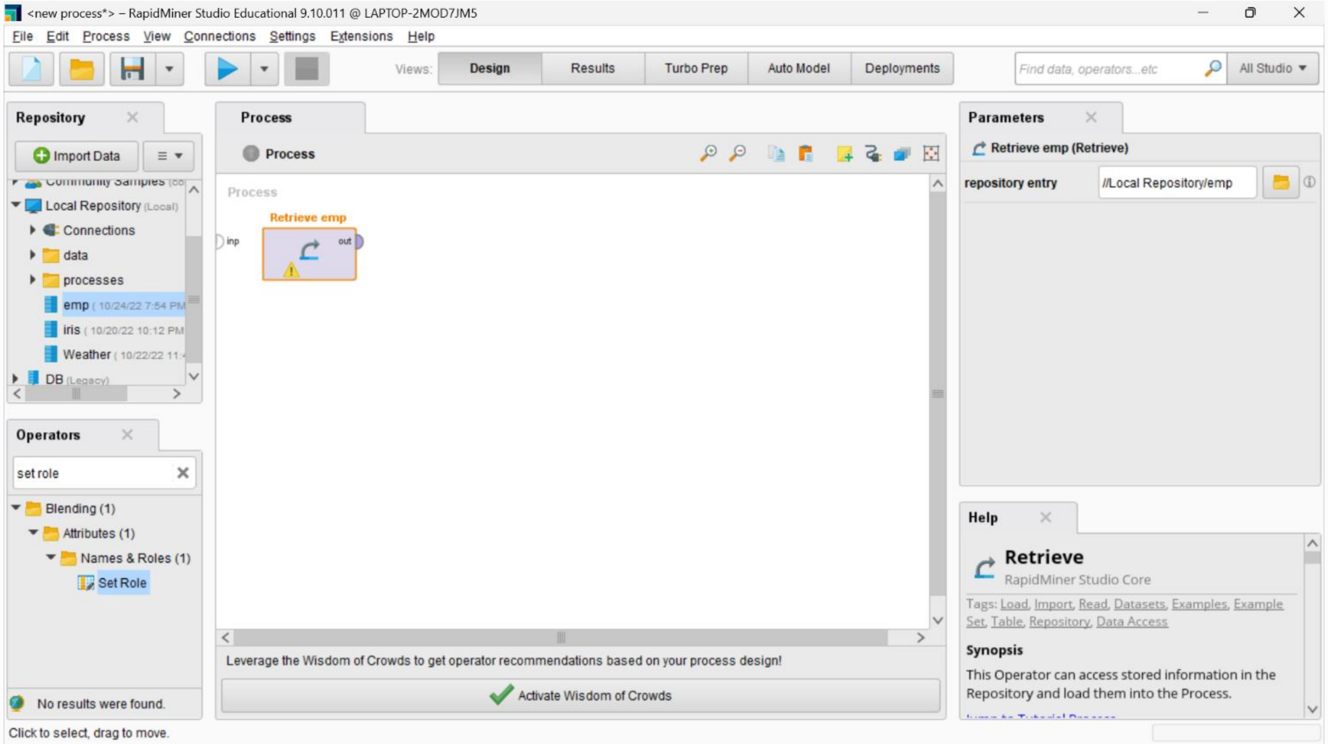
<u>About Linear Regression:</u>

In the simplest words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

Simple Linear Regression: A linear regression model with one independent and one dependent variable. Multiple Linear Regression: A linear regression model with more than one independent variable and one dependent variable

9.1.3 Steps to implement simple Linear Regression:

1.To implement the linear regression on Employee dataset using Rapidminer first import the dataset and drop the dataset into the design screen.


<new process*> – RapidMiner Studio Educational 9.10.011 @ LAPTOP-2MOD7JM5
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Community Samples
Local Repository (Local)
Connections
data
processes
emp (10/24/22 7:54 PM
iris (10/20/22 10:12 PM
Weather (10/22/22 11:
DB (Legacy)
Operators
set role
Blending (1)
Attributes (1)
Names & Roles (1)
Set Role
No results were found.
Click to select, drag to move.
Process
Retrieve emp
inp
out
Activate Wisdom of Crowds
Parameters
Help
Retrieve
RapidMiner Studio Core
Tags: Load, Import, Read, Datasets, Examples, Example Set, Table, Repository, Data Access
Synopsis

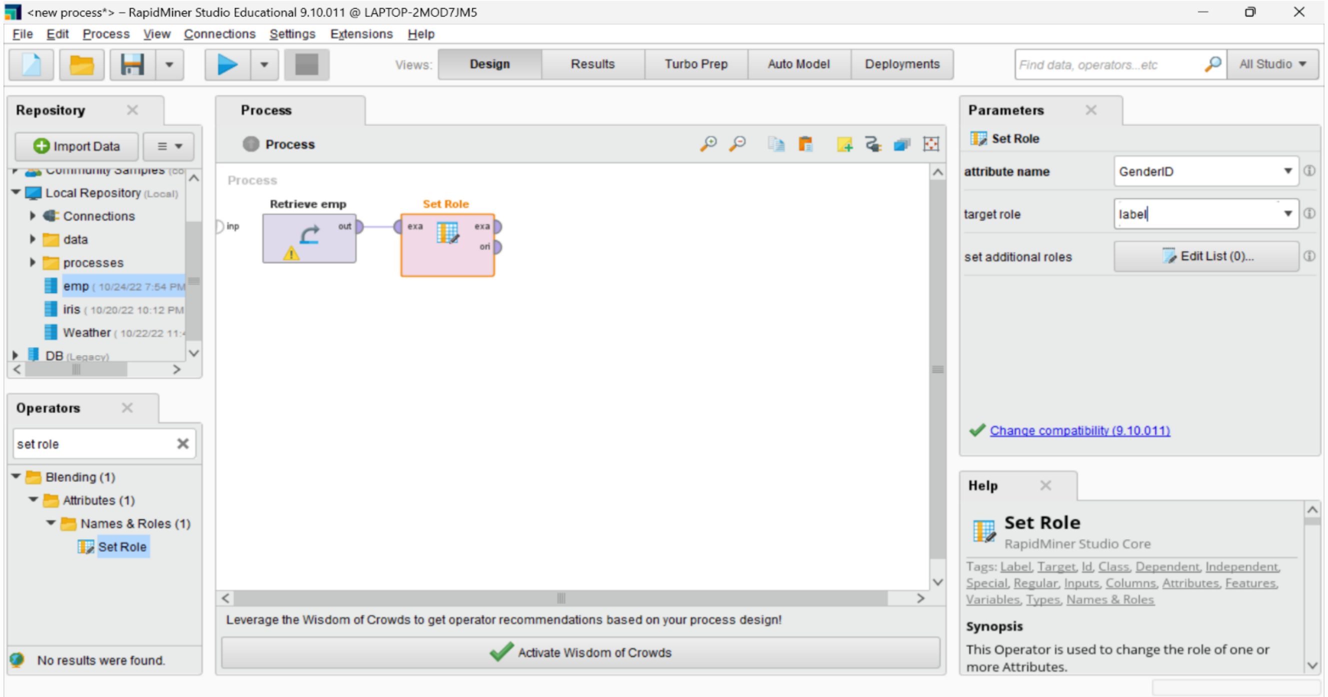
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Retrieve emp (Retrieve)

repository entry: //Local Repository/emp

This Operator can access stored information in the Repository and load them into the Process.

2. Now under operators drag and drop the set role operator and edit the parameters like attribute name and target role.


<new process*> – RapidMiner Studio Educational 9.10.011 @ LAPTOP-2MOD7JM5
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Community Samples
Local Repository (Local)
Connections
data
processes
emp (10/24/22 7:54 PM
iris (10/20/22 10:12 PM
Weather (10/22/22 11:
DB (Legacy)
Operators
set role
Blending (1)
Attributes (1)
Names & Roles (1)
Set Role
No results were found.
Process
Retrieve emp
inp
out
Set Role
exa
exa
ori
Activate Wisdom of Crowds
Parameters
Change compatibility (9.10.011)
Help
Set Role
RapidMiner Studio Core
Tags: Label, Target, Id, Class, Dependent, Independent, Special, Regular, Inputs, Columns, Attributes, Features, Variables, Types, Names & Roles
Synopsis

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Set Role

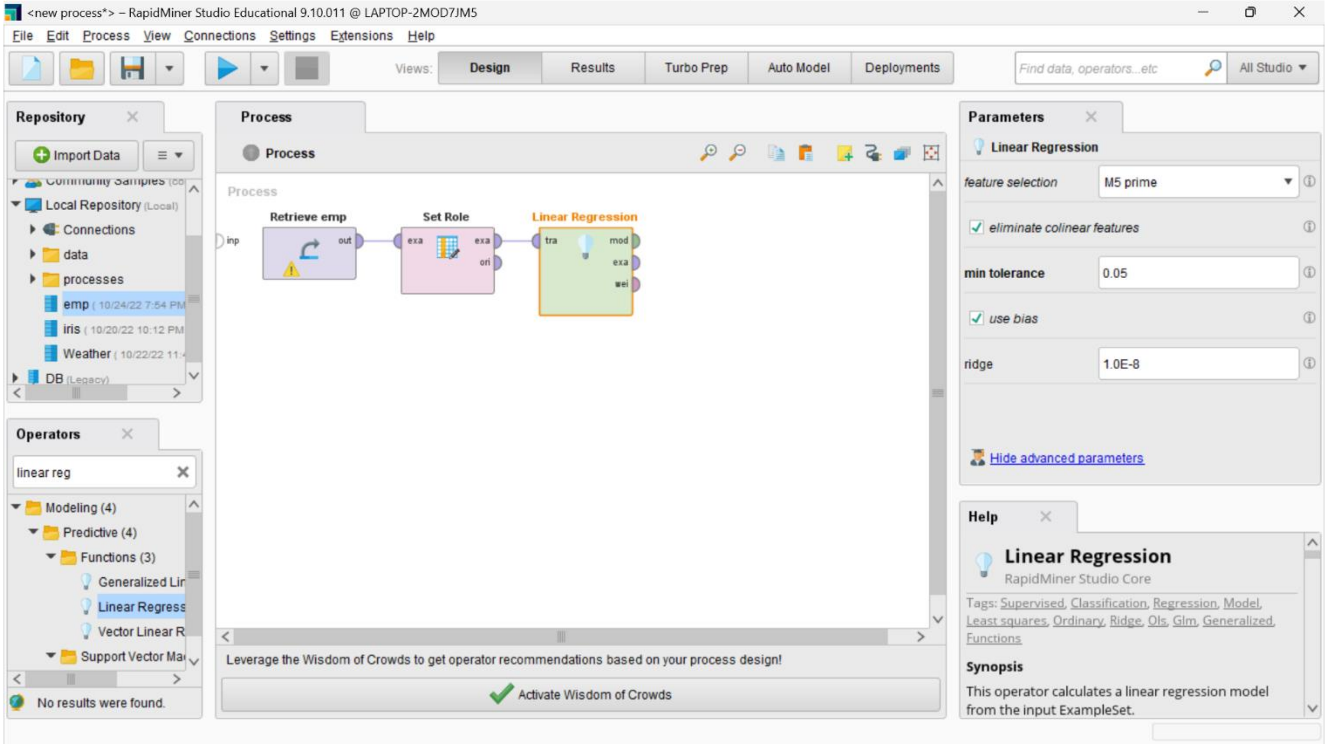
attribute name: GenderID

target role: label

set additional roles: Edit List (0)...

This Operator is used to change the role of one or more Attributes.

3. Drag and Drop Linear regression Operator into the design screen.



Linear Regression

feature selection: M5 prime

☑ eliminate colinear features

min tolerance: 0.05

☑ use bias

ridge: 1.0E-8

Hide advanced parameters

Linear Regression

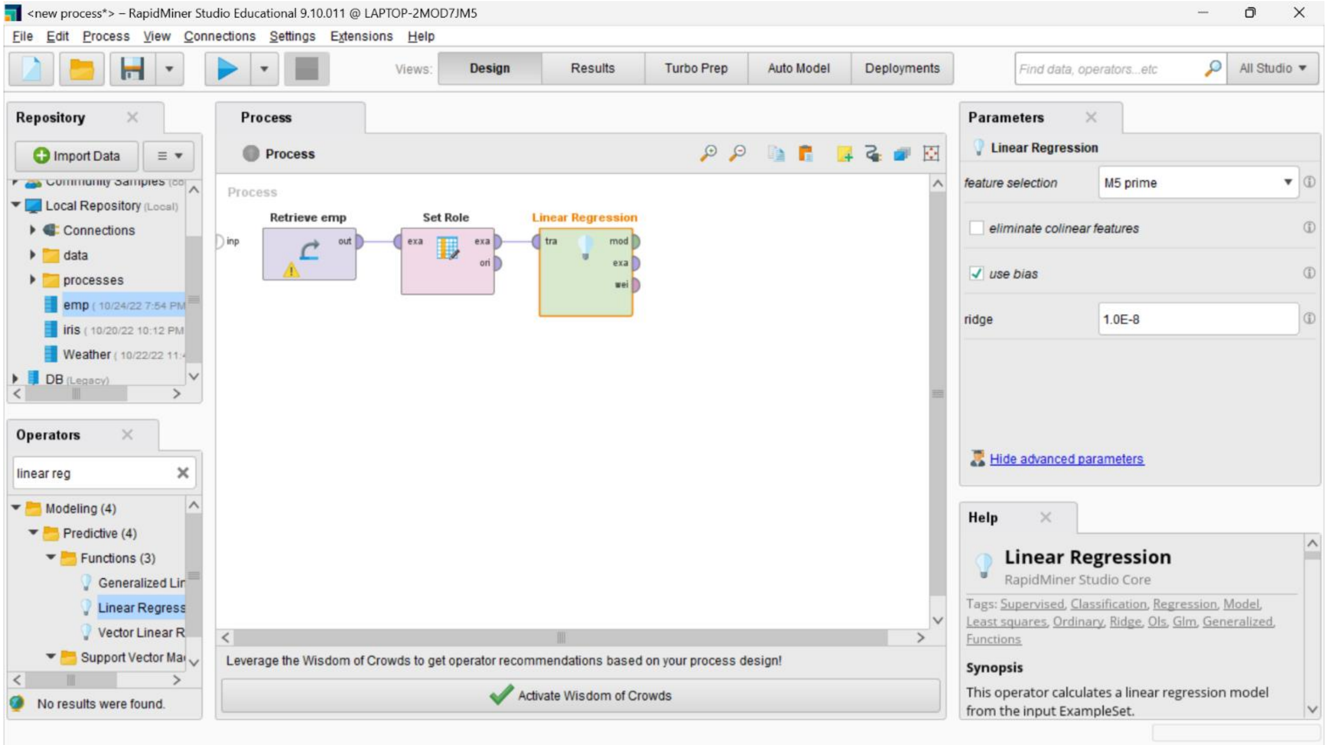
RapidMiner Studio Core

Tags: Supervised, Classification, Regression, Model, Least squares, Ordinary, Ridge, Ols, Glm, Generalized, Functions

Synopsis

This operator calculates a linear regression model from the input ExampleSet.

Now in the parameters of Linear regression enable the eliminate colinear features.



Linear Regression

feature selection: M5 prime

☐ eliminate colinear features

☑ use bias

ridge: 1.0E-8

Hide advanced parameters

Linear Regression

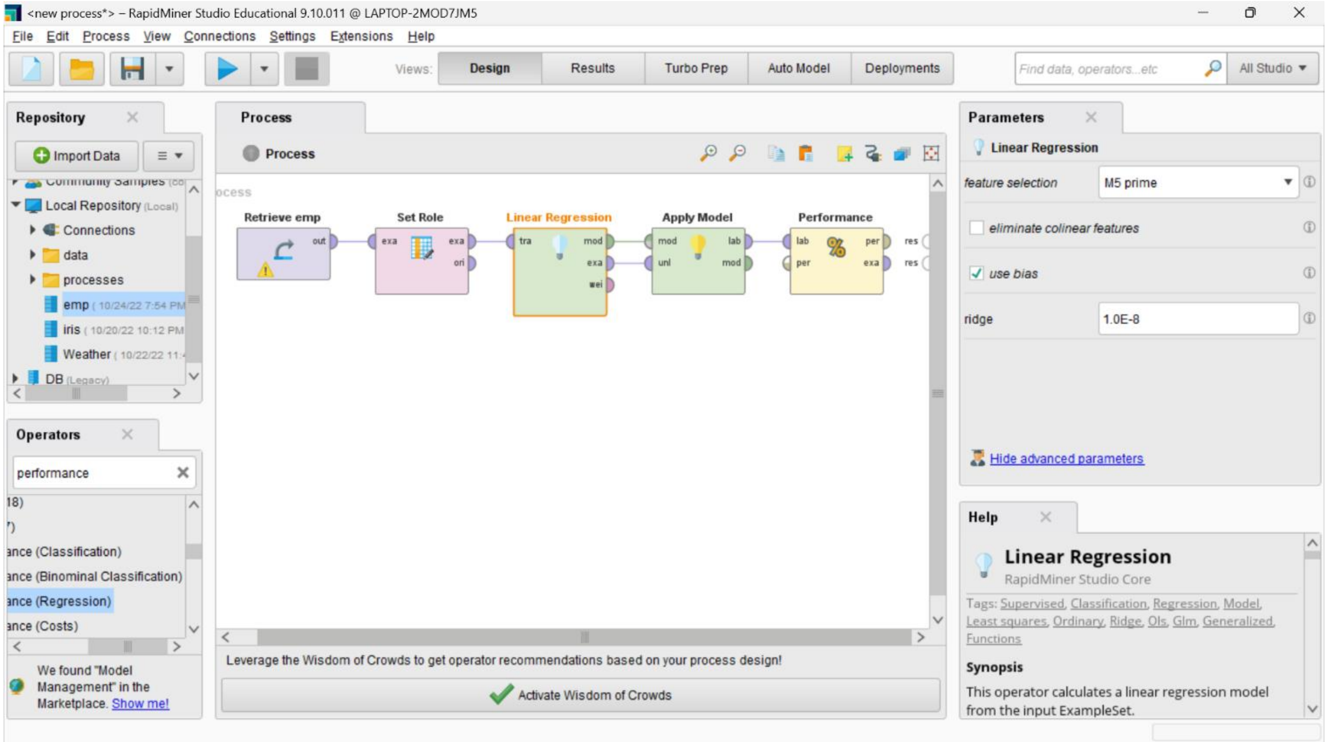
RapidMiner Studio Core

Tags: Supervised, Classification, Regression, Model, Least squares, Ordinary, Ridge, Ols, Glm, Generalized, Functions

Synopsis

This operator calculates a linear regression model from the input ExampleSet.

4. Drag and drop Apply Model and performance (Regression) operator into the design screen.


<new process*> – RapidMiner Studio Educational 9.10.011 @ LAPTOP-2MOD7JM5
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Community Samples
Local Repository (Local)
Connections
data
processes
emp (10/24/22 7:54 PM
iris (10/20/22 10:12 PM
Weather (10/22/22 11:
DB (Legacy)
Operators
performance
ance (Classification)
ance (Binominal Classification)
ance (Regression)
ance (Costs)
We found "Model Management" in the Marketplace. Show me!
Process
Retrieve emp
Set Role
Linear Regression
Apply Model
Performance
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Hide advanced parameters
Help

Linear Regression

feature selection	M5 prime
☐ eliminate colinear features	
☑ use bias	
ridge	1.0E-8

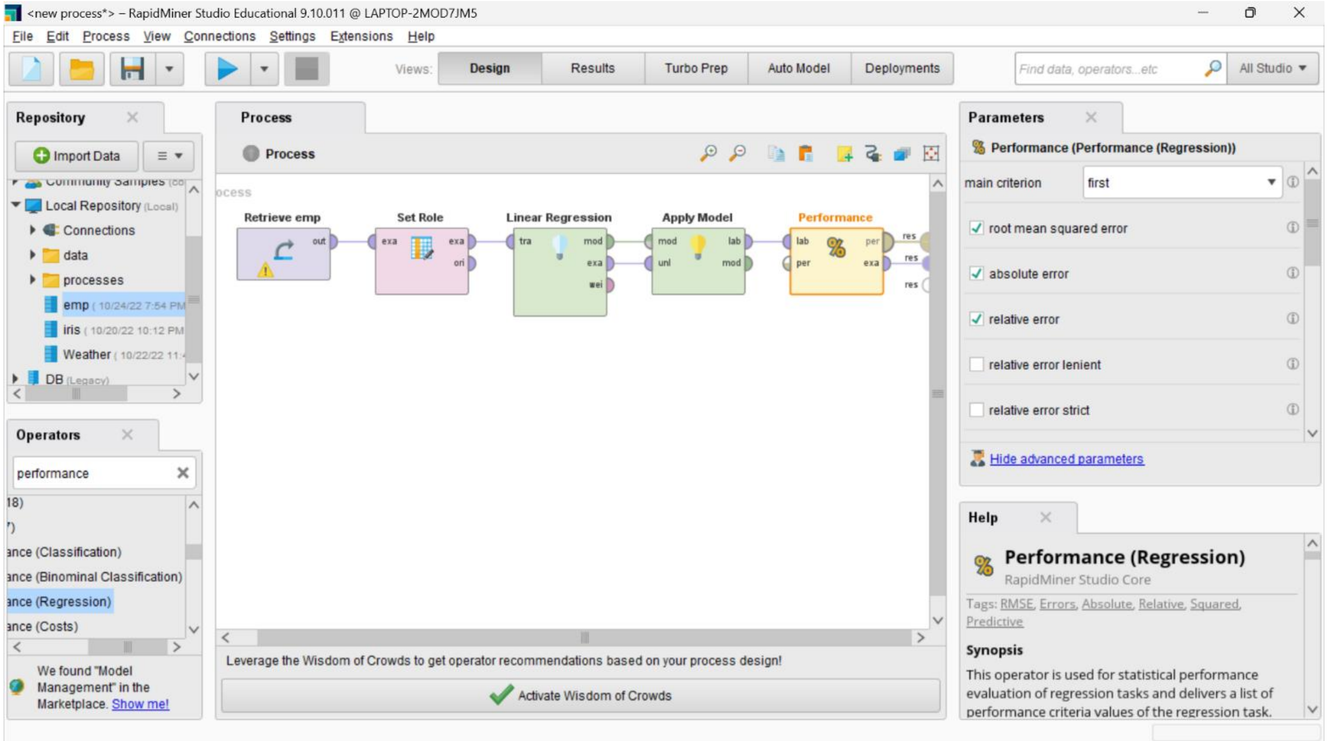
Linear Regression
RapidMiner Studio Core

Tags: Supervised, Classification, Regression, Model, Least squares, Ordinary, Ridge, Ols, Glm, Generalized, Functions

Synopsis

This operator calculates a linear regression model from the input ExampleSet.

5. In Performance select the parameters that has to display result when the design is executed and give the flow of connections as shown in the below figure, and click on run button to see the result.


<new process*> – RapidMiner Studio Educational 9.10.011 @ LAPTOP-2MOD7JM5
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Community Samples
Local Repository (Local)
Connections
data
processes
emp (10/24/22 7:54 PM
iris (10/20/22 10:12 PM
Weather (10/22/22 11:
DB (Legacy)
Operators
performance
ance (Classification)
ance (Binominal Classification)
ance (Regression)
ance (Costs)
We found "Model Management" in the Marketplace. Show me!
Process
Retrieve emp
Set Role
Linear Regression
Apply Model
Performance
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Hide advanced parameters
Help

Performance (Performance (Regression))

main criterion	first
☑ root mean squared error	
☑ absolute error	
☑ relative error	
☐ relative error lenient	
☐ relative error strict	

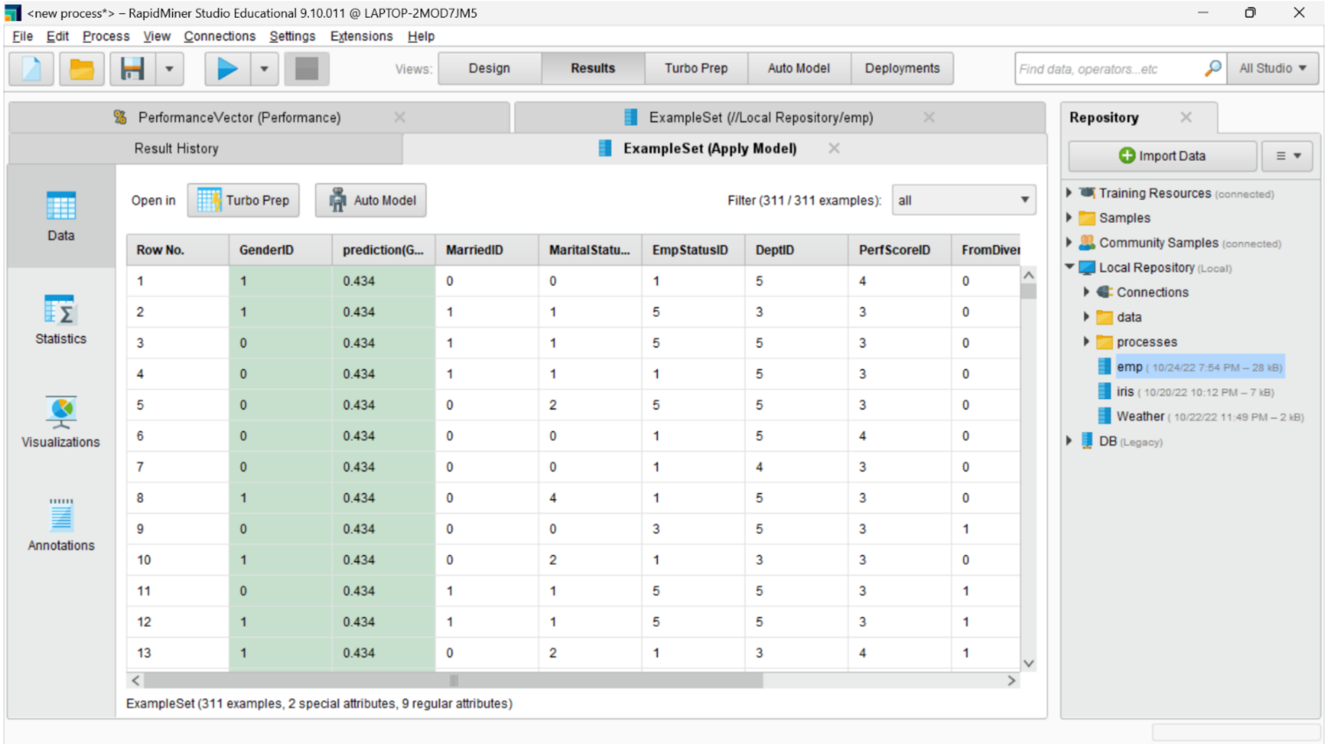
Performance (Regression)
RapidMiner Studio Core

Tags: RMSE, Errors, Absolute, Relative, Squared, Predictive

Synopsis

This operator is used for statistical performance evaluation of regression tasks and delivers a list of performance criteria values of the regression task.

6. Now result of the design is shown in many ways like Data, Statistics, Visualizations, Performance vector.



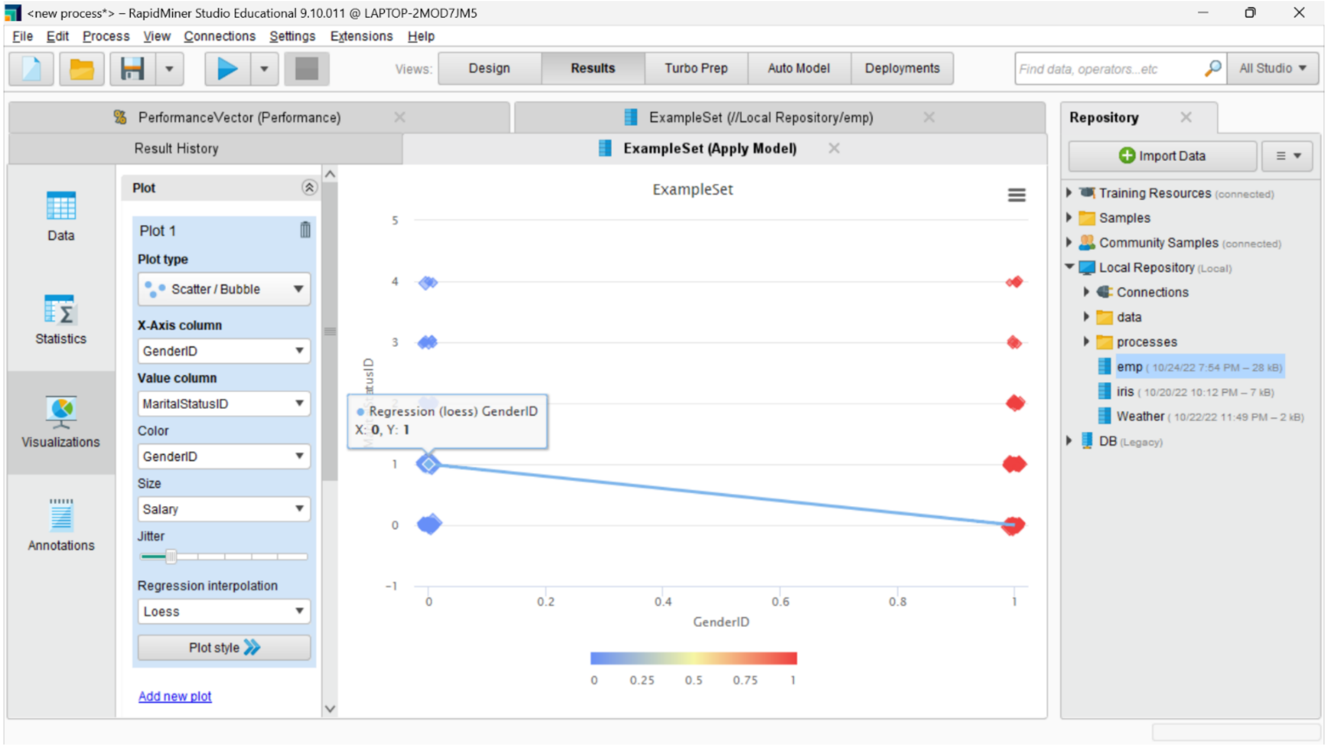
<new process*> – RapidMiner Studio Educational 9.10.011 @ LAPTOP-2MOD7JM5
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
PerformanceVector (Performance)
ExampleSet (//Local Repository/emp)
Result History
ExampleSet (Apply Model)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
emp (10/24/22 7:54 PM – 28 kB)
iris (10/20/22 10:12 PM – 7 kB)
Weather (10/22/22 11:49 PM – 2 kB)
DB (Legacy)

Filter (311 / 311 examples): all

Row No.	GenderID	prediction(G...	MarriedID	MaritalStatu...	EmpStatusID	DeptID	PerfScoreID	FromDiver
1	1	0.434	0	0	1	5	4	0
2	1	0.434	1	1	5	3	3	0
3	0	0.434	1	1	5	5	3	0
4	0	0.434	1	1	1	5	3	0
5	0	0.434	0	2	5	5	3	0
6	0	0.434	0	0	1	5	4	0
7	0	0.434	0	0	1	4	3	0
8	1	0.434	0	4	1	5	3	0
9	0	0.434	0	0	3	5	3	1
10	1	0.434	0	2	1	3	3	0
11	0	0.434	1	1	5	5	3	1
12	1	0.434	1	1	5	5	3	1
13	1	0.434	0	2	1	3	4	1

ExampleSet (311 examples, 2 special attributes, 9 regular attributes)

7. In Visualization we can plot the graph according to our modifications.



<new process*> – RapidMiner Studio Educational 9.10.011 @ LAPTOP-2MOD7JM5
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
PerformanceVector (Performance)
ExampleSet (//Local Repository/emp)
Result History
ExampleSet (Apply Model)
Data
Statistics
Visualizations
Annotations
Plot
Plot 1
Plot style
Add new plot
ExampleSet
5
4
3
1
0
-1
0
0.2
0.4
0.6
0.8
1
GenderID
MaritalStatusID
Regression (loess) GenderID
X: 0, Y: 1
0
0.25
0.5
0.75
1
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
emp (10/24/22 7:54 PM – 28 kB)
iris (10/20/22 10:12 PM – 7 kB)
Weather (10/22/22 11:49 PM – 2 kB)
DB (Legacy)

Plot type: Scatter / Bubble

X-Axis column: GenderID

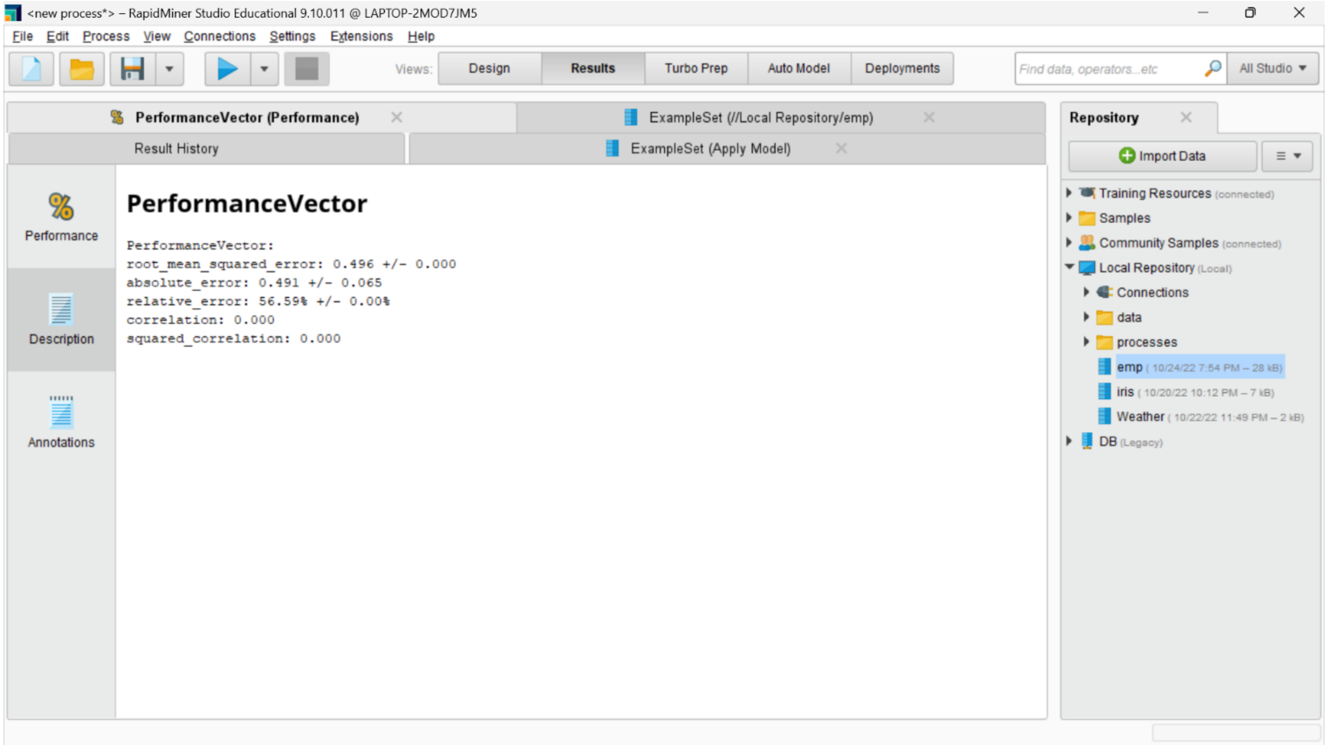
Value column: MaritalStatusID

Color: GenderID

Size: Salary

Jitter

Regression interpolation: Loess


<new process*> – RapidMiner Studio Educational 9.10.011 @ LAPTOP-2MOD7JM5
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
PerformanceVector (Performance)
ExampleSet (//Local Repository/emp)
Result History
ExampleSet (Apply Model)
Performance
Description
Annotations
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
emp (10/24/22 7:54 PM – 28 kB)
iris (10/20/22 10:12 PM – 7 kB)
Weather (10/22/22 11:49 PM – 2 kB)
DB (Legacy)

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.496 +/- 0.000
absolute_error: 0.491 +/- 0.065
relative_error: 56.59% +/- 0.00%
correlation: 0.000
squared_correlation: 0.000
```

9.1.4 Results and Discussion:

Therefore, Implementation of Simple Linear Regression on an 'Employee' dataset using RapidMiner is executed successfully.

9.2.1 Problem Statement:

Demonstrate the simple k-Means clustering algorithm on ‘iris’ dataset.

9.2.2 Description:

About Dataset used:

The attributes are:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class
6. Number of samples of each species of iris flowers
7. Predicted attribute: class of iris plant
8. Missing Attribute Values: None

The Iris Dataset contains information of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Data Set Characteristics: Multivariate

Area: Life Sciences

Number of samples (or instances) in the dataset: 150

Number of attributes (or features): 05

Attribute Information:

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica Class Distribution: 33.3% for each of 3 classes

About RapidMiner:

Rapidminer is a comprehensive data science platform with visual workflow design and full automation. It means that we don’t have to do the coding for data mining tasks. Rapidminer is one of the most popular data science tools.

This is the graphical user interface of the blank process in rapidminer. It has the repository that holds our dataset. We can import our own datasets. It also offers many public datasets that we can try. We can also work with a database connection.

RapidMiner is an awesome visual workflow designer. The way they present visually is so unique. It helps in speeding and automating the creation of visual models. It helps in creating models in only 5 clicks by automated machine learning.

About k-Means Clustering algorithm:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

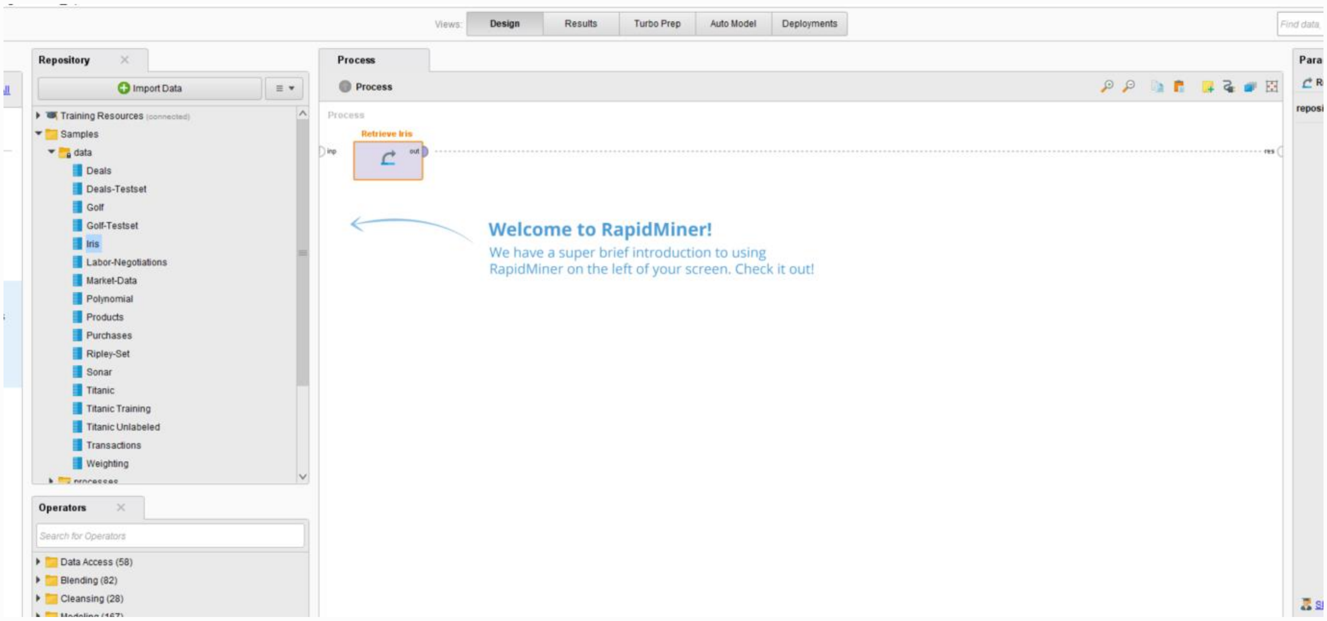
The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

9.2.3 Steps to implement simple K-Means clustering algorithm:

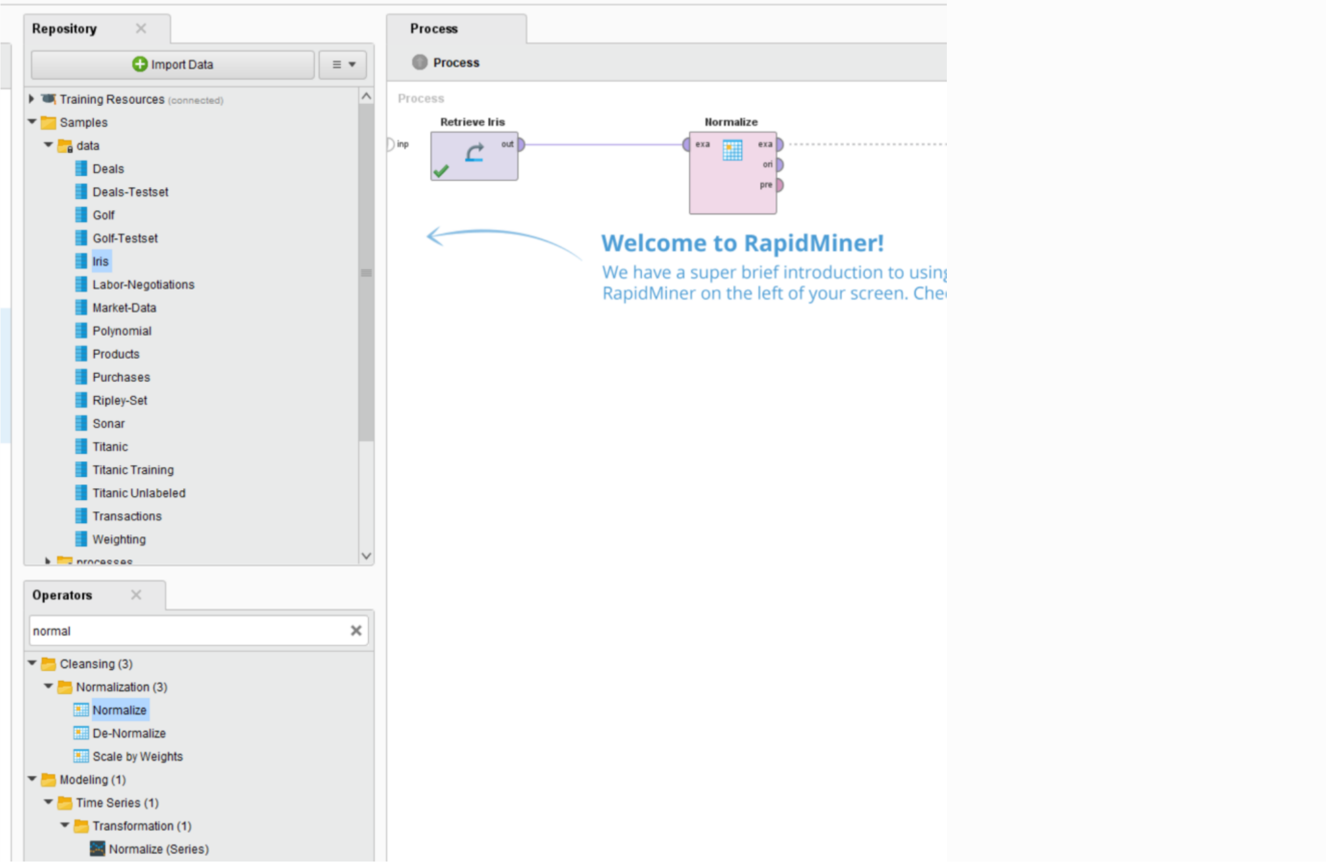
1. To implement the K-Means clustering on Iris dataset using Rapidminer, first import the dataset and drop the dataset into the design screen.



Welcome to RapidMiner!

We have a super brief introduction to using RapidMiner on the left of your screen. Check it out!

2. Now under operators drag and drop the Normalize operator and connect the dataset to it.

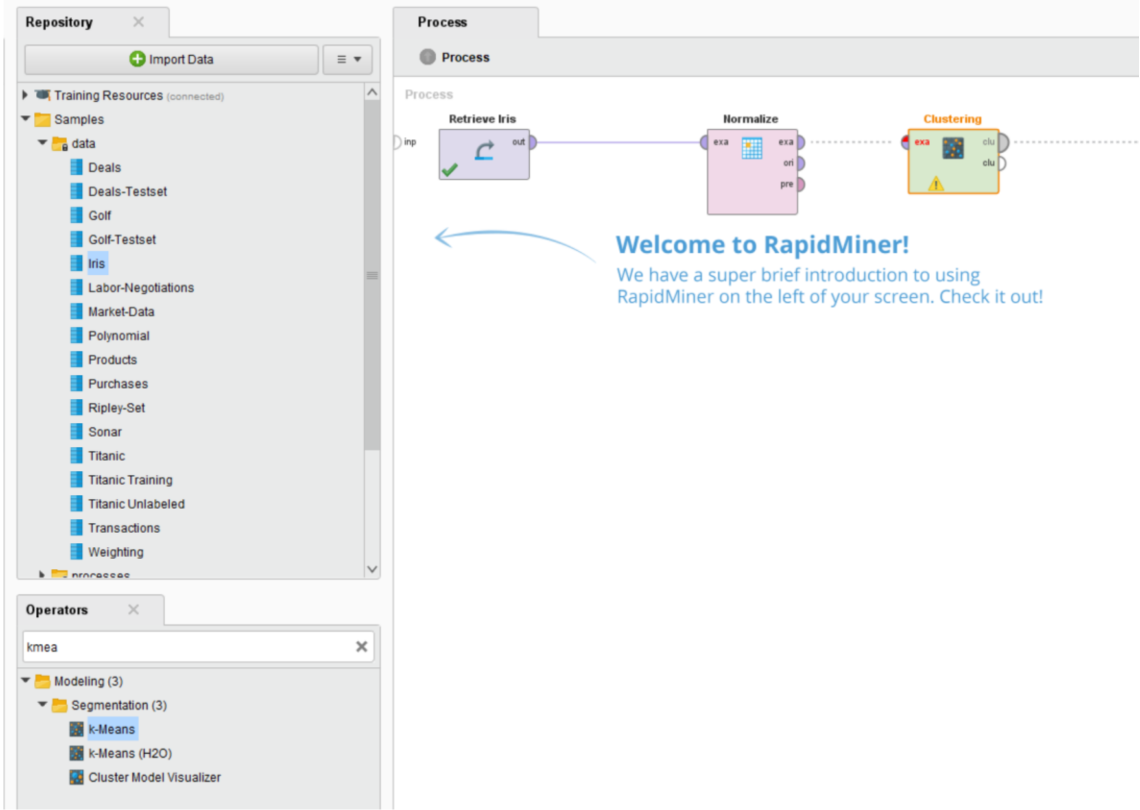


normal

Welcome to RapidMiner!

We have a super brief introduction to usin
RapidMiner on the left of your screen. Che

3. Drag and Drop k-means operator and connect it.

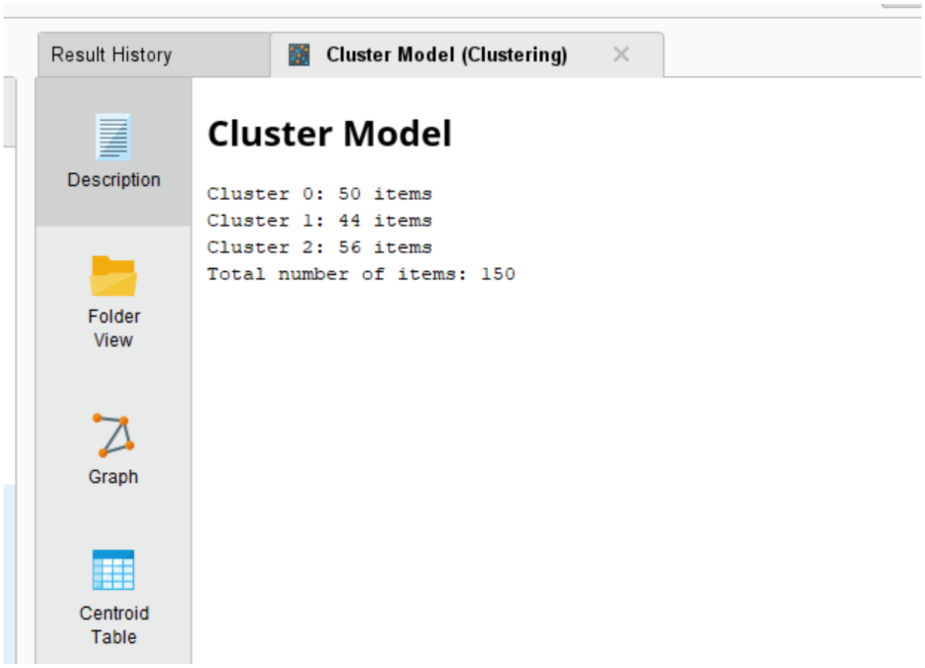


kmea

Welcome to RapidMiner!

We have a super brief introduction to using
RapidMiner on the left of your screen. Check it out!

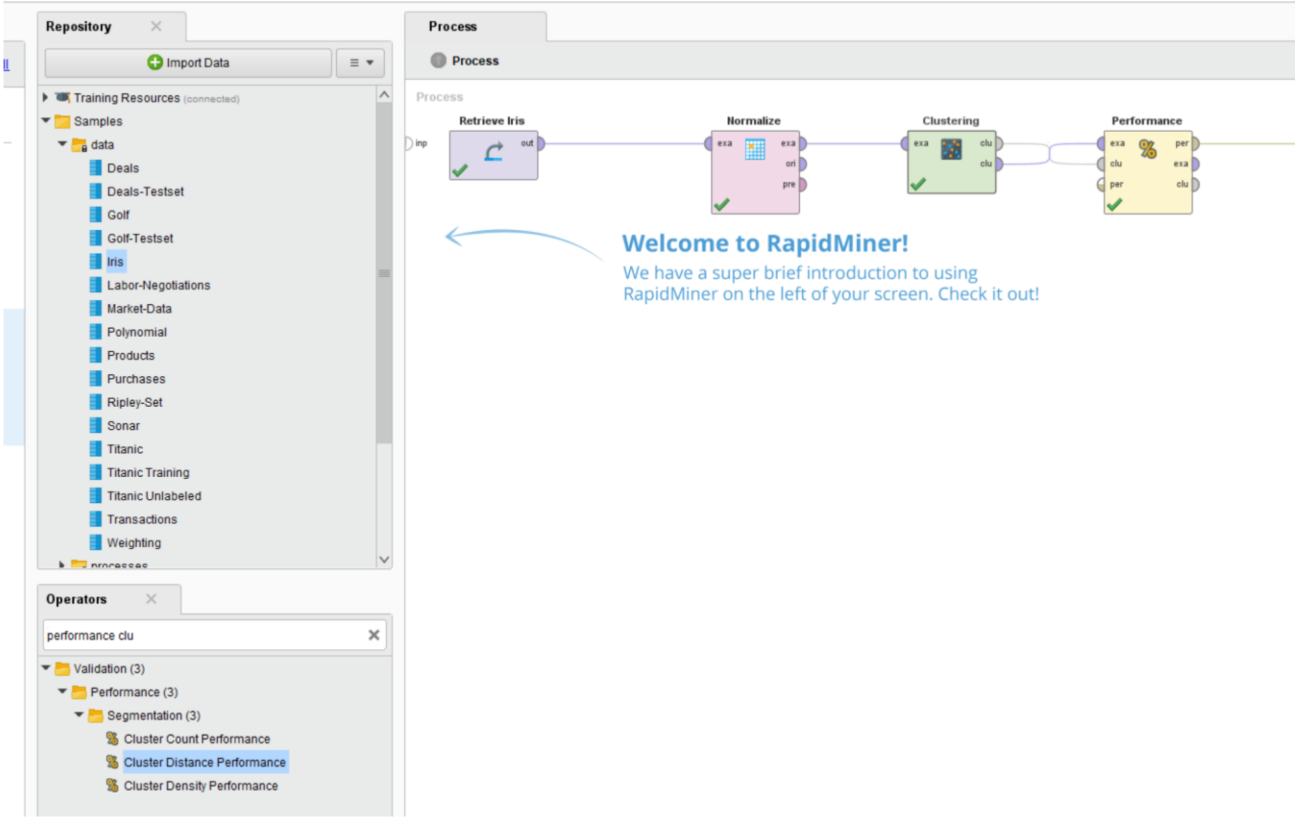
4. Set the k value to 3 to create three clusters and run it and see the clusters.



Cluster Model

```
Cluster 0: 50 items
Cluster 1: 44 items
Cluster 2: 56 items
Total number of items: 150
```

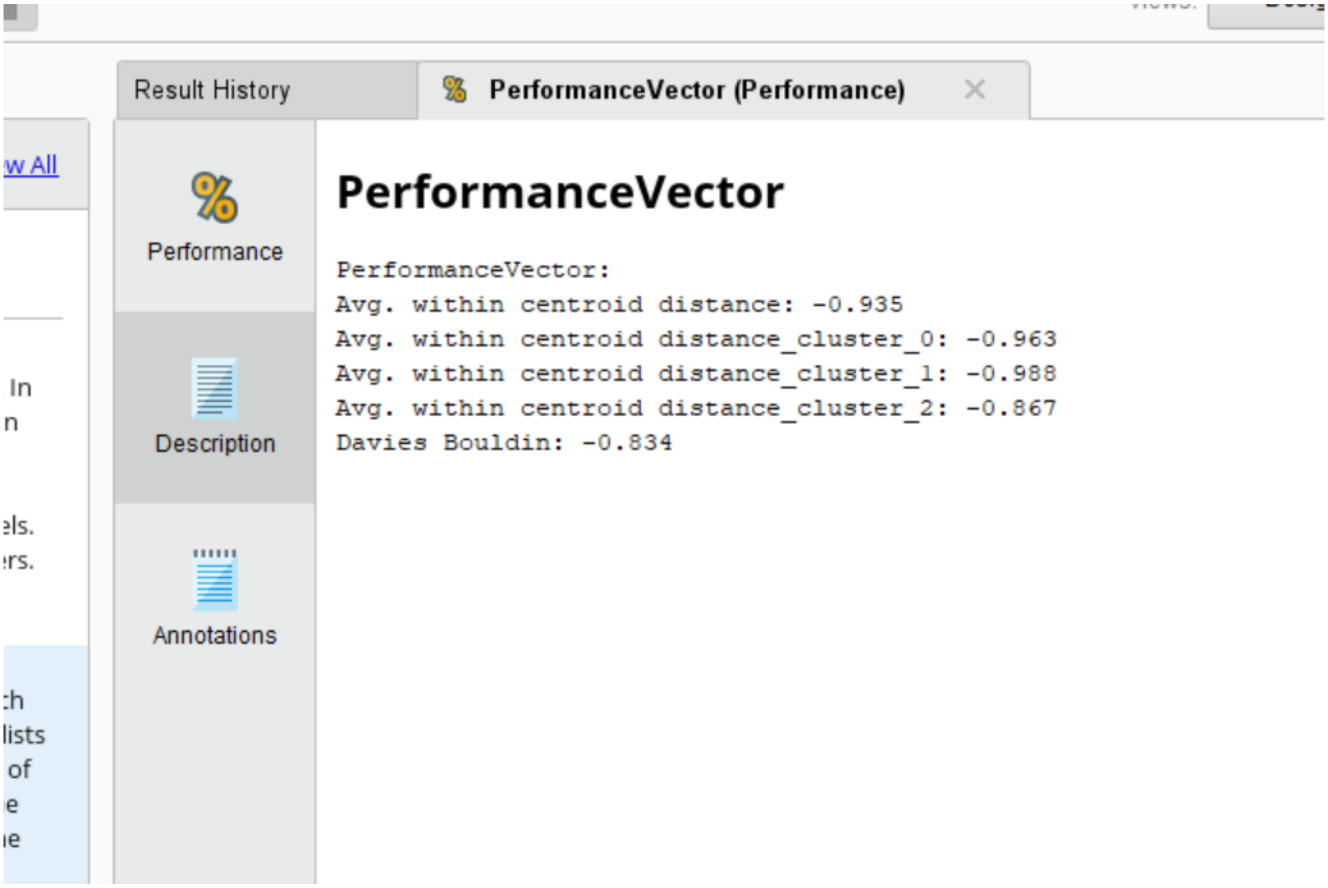
5. Drag and drop the Cluster Distance Performance from the operators and see the performance of our k-means algorithm.



Welcome to RapidMiner!

We have a super brief introduction to using RapidMiner on the left of your screen. Check it out!

6. Now the performance vector is seen.



PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.935
Avg. within centroid distance_cluster_0: -0.963
Avg. within centroid distance_cluster_1: -0.988
Avg. within centroid distance_cluster_2: -0.867
Davies Bouldin: -0.834
```

9.2.4 Results and Discussion:

Therefore, Implementation of Simple k-means clustering algorithm on ‘iris’ dataset using RapidMiner is executed successfully.