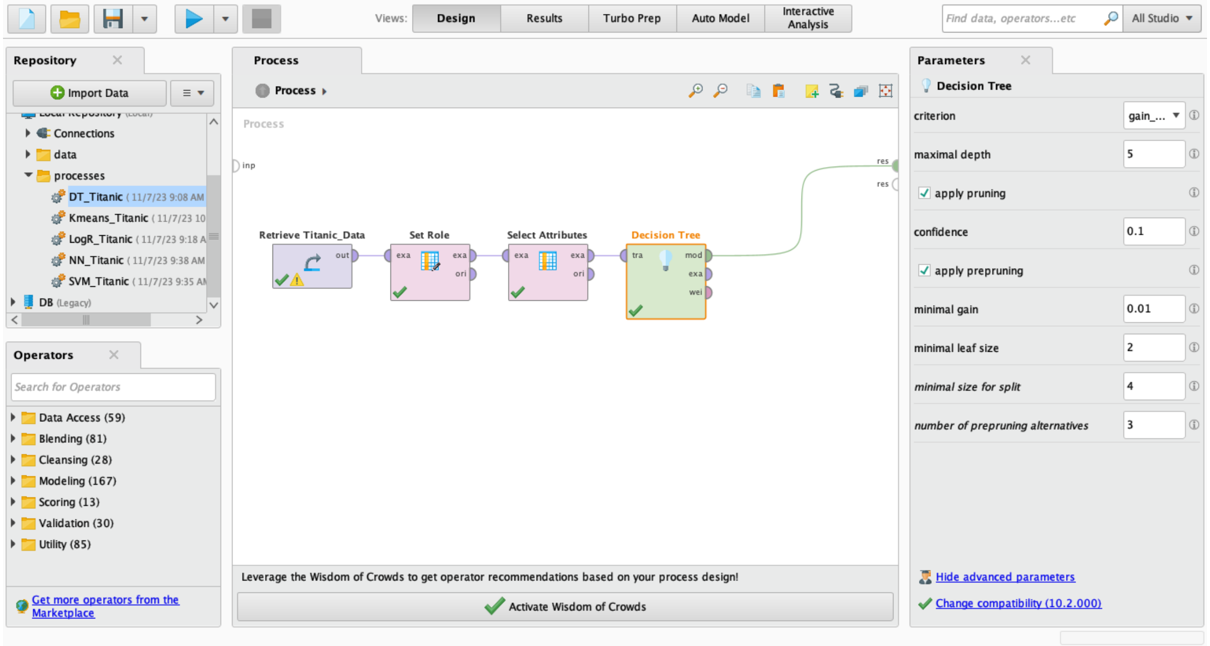
Assignment 4
Foundations of AI

Name: Komal Patil

Task 1: Decision Tree

a) Build a decision tree model



Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Repository
Import Data
Local Repository (Local)
Connections
data
processes
DT_Titanic (11/7/23 9:08 AM
Kmeans_Titanic (11/7/23 10
LogR_Titanic (11/7/23 9:18 A
NN_Titanic (11/7/23 9:38 AM
SVM_Titanic (11/7/23 9:35 AM
DB (Legacy)
Operators
Search for Operators
Data Access (59)
Blending (81)
Cleansing (28)
Modeling (167)
Scoring (13)
Validation (30)
Utility (85)
Get more operators from the Marketplace
Process
Retrieve Titanic_Data
Set Role
Select Attributes
Decision Tree
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Hide advanced parameters
Change compatibility (10.2.000)

Decision Tree

criterion: gain_...

maximal depth: 5

☑ apply pruning

confidence: 0.1

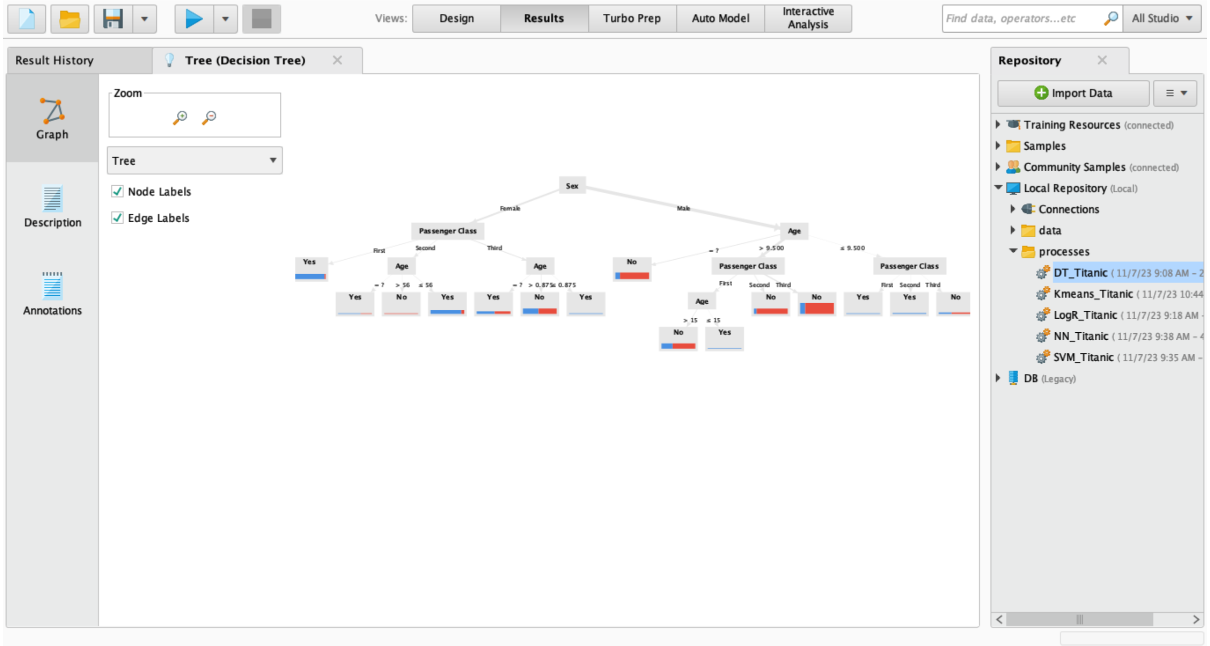
☑ apply prepruning

minimal gain: 0.01

minimal leaf size: 2

minimal size for split: 4

number of prepruning alternatives: 3



Views: Design Results Turbo Prep Auto Model Interactive Analysis
Result History
Tree (Decision Tree)
Graph
Description
Annotations
Zoom
Tree
☑ Node Labels
☑ Edge Labels
Sex
Female
Male
Passenger Class
Age
First
Second
Third
Yes
No
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
DT_Titanic (11/7/23 9:08 AM
Kmeans_Titanic (11/7/23 10:44
LogR_Titanic (11/7/23 9:18 AM
NN_Titanic (11/7/23 9:38 AM
SVM_Titanic (11/7/23 9:35 AM
DB (Legacy)

Considering gender and age in particular, the decision tree offers insightful information on the variables affecting survival on the Titanic. Females have a far higher chance of surviving, particularly those in the first and second passenger class. Male age is a crucial factor, with younger men having a better probability of surviving, especially those in the first and second passenger classes. The intricate links between gender, age, and passenger class are well-represented by the hierarchical structure of the tree, which enhances the interpretation of survival patterns.

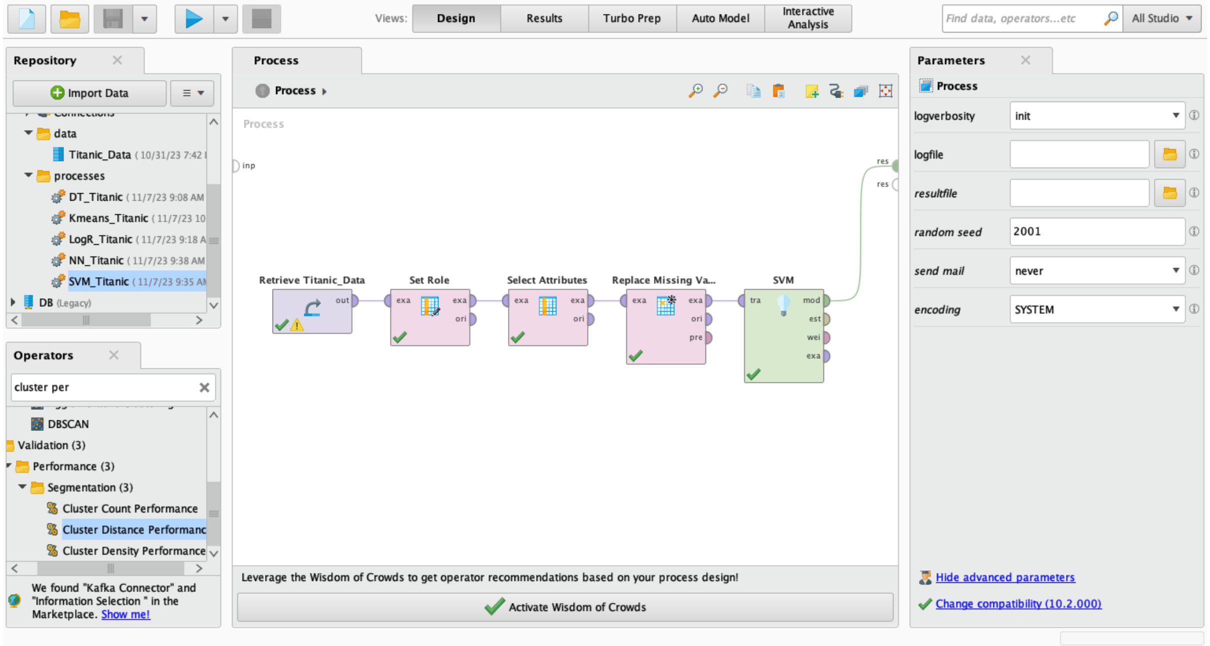
b) What is your target variable?
Ans: Target variable is survived and target role is Label.

c) What are your key attributes to estimate your target variable?
Ans: Age, Passenger class and sex

d) Choose your best model complexity (i.e., depth of decision tree) and provide your logic
Ans: The decision to make the decision tree depth 5 seems like a good choice. This depth helps the tree understand specific conditions for making predictions without getting too complicated. If we went deeper then 5, it might focus too much on our current data and not work well with new information. On the other hand, if the depth is less then 5, it might miss important details. So, keeping it at a depth of 5 strikes a nice balance between being detailed and practical for our dataset and problem.

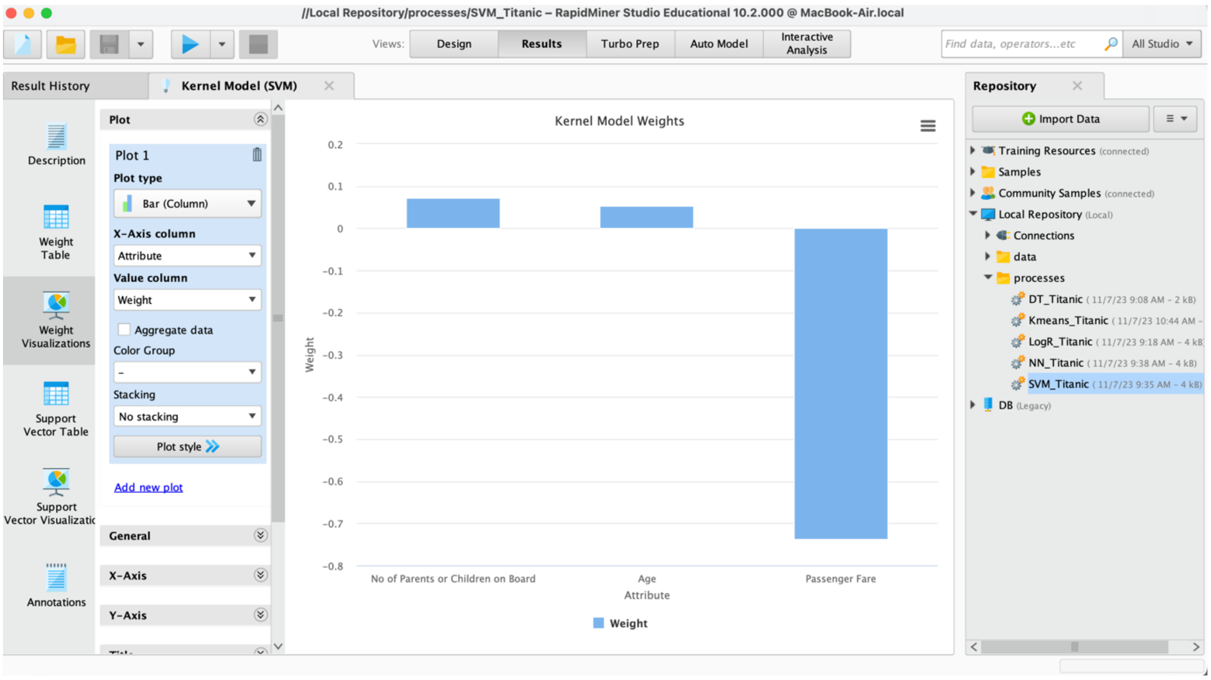
Task 2: Support Vector Machine (SVM)

a) Build your SVM model



Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Repository
Import Data
data
Titanic_Data (10/31/23 7:42
processes
DT_Titanic (11/7/23 9:08 AM
Kmeans_Titanic (11/7/23 10
LogR_Titanic (11/7/23 9:18 A
NN_Titanic (11/7/23 9:38 AM
SVM_Titanic (11/7/23 9:35 AM
DB (Legacy)
Operators
cluster per
DBSCAN
Validation (3)
Performance (3)
Segmentation (3)
Cluster Count Performance
Cluster Distance Performanc
Cluster Density Performance
We found "Kafka Connector" and "Information Selection " in the Marketplace. Show me!
Process
Process
inp
res
res
Retrieve Titanic_Data
out
Set Role
exa
exa
ori
Select Attributes
exa
exa
ori
Replace Missing Va...
exa
exa
ori
pre
SVM
tra
mod
est
wei
exa
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Process
Hide advanced parameters
Change compatibility (10.2.000)

logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM


//Local Repository/processes/SVM_Titanic – RapidMiner Studio Educational 10.2.000 @ MacBook-Air.local
Views:
Design
Results
Turbo Prep
Auto Model
Interactive Analysis
Find data, operators...etc
All Studio
Result History
Kernel Model (SVM)
Description
Weight Table
Weight Visualizations
Support Vector Table
Support Vector Visualizatic
Annotations
Plot
Plot 1
Add new plot
General
X-Axis
Y-Axis
Kernel Model Weights
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
Weight
No of Parents or Children on Board
Age
Passenger Fare
Attribute
Weight
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
DT_Titanic (11/7/23 9:08 AM – 2 kB)
Kmeans_Titanic (11/7/23 10:44 AM –
LogR_Titanic (11/7/23 9:18 AM – 4 kB
NN_Titanic (11/7/23 9:38 AM – 4 kB)
SVM_Titanic (11/7/23 9:35 AM – 4 kB)
DB (Legacy)

Plot type: Bar (Column)

X-Axis column: Attribute

Value column: Weight

□ Aggregate data

Color Group: –

Stacking: No stacking

Plot style

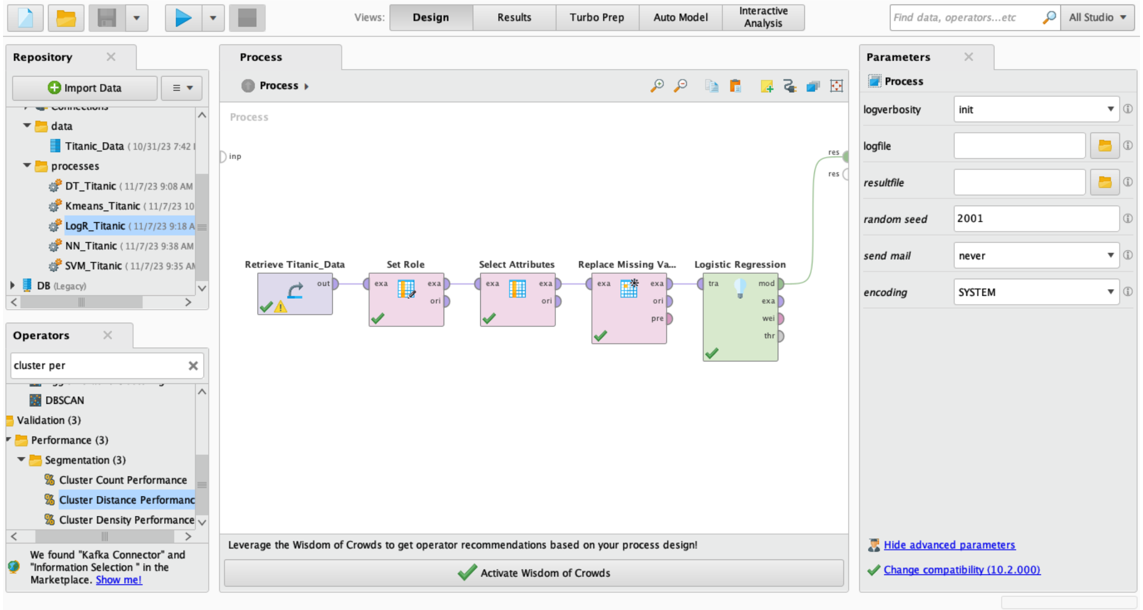
b) Once you run your model, try to interpret your results (i.e., Kernel Model (SVM)). Use 'Weight Visualization' or 'Support Vector Visualization,' and provide your interpretations.

Ans: The weight associated with the passenger fare (w[Passenger Fare] = -0.738) is the most significant among the three features. This indicates that passenger fare has the strongest influence on the model's predictions. A negative weight suggests that a higher passenger fare is associated with a lower probability of survival.

While age and the number of parents or children on board also have some influence on the model's predictions, their weights are much smaller than the weight of passenger fare. This suggests that these factors are less important in predicting survival outcomes.

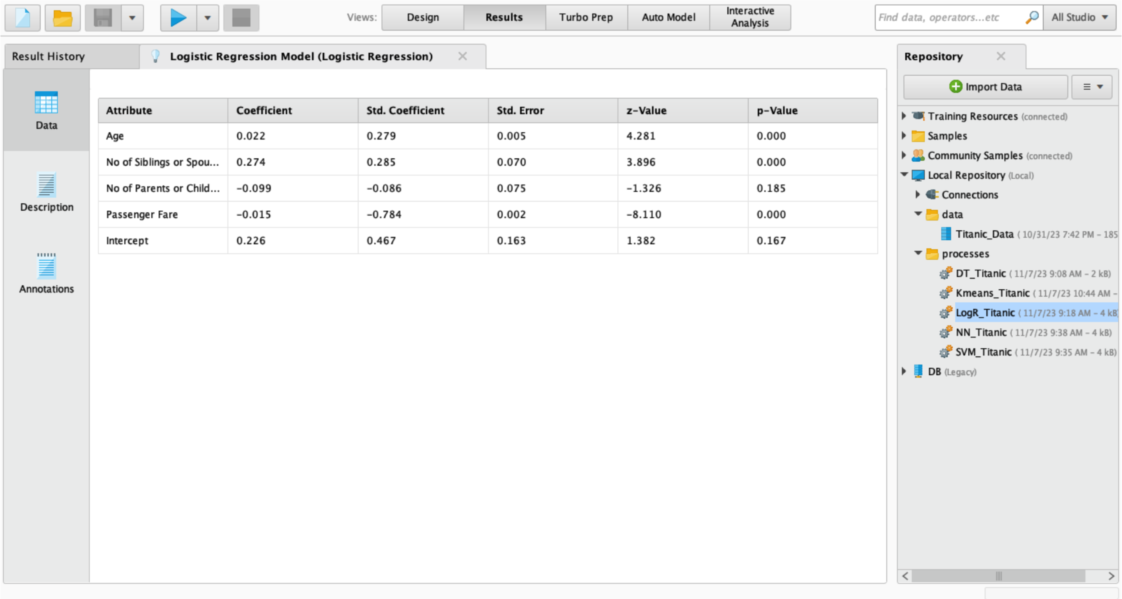
Task 3: Logistic Regression

a) Build your logistic regression model


Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Repository
Import Data
data
Titanic_Data (10/31/23 7:42
processes
DT_Titanic (11/7/23 9:08 AM
Kmeans_Titanic (11/7/23 10
LogR_Titanic (11/7/23 9:18 A
NN_Titanic (11/7/23 9:38 AM
SVM_Titanic (11/7/23 9:35 AM
DB (Legacy)
Operators
cluster per
DBSCAN
Validation (3)
Performance (3)
Segmentation (3)
Cluster Count Performance
Cluster Distance Performanc
Cluster Density Performance
We found "Kafka Connector" and "Information Selection " in the Marketplace. Show me!
Process
Process
inp
res
res
Retrieve Titanic_Data
out
Set Role
exa
exa
ori
Select Attributes
exa
exa
ori
Replace Missing Va...
exa
exa
ori
pre
Logistic Regression
tra
mod
exa
wei
thr
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Process
Hide advanced parameters
Change compatibility (10.2.000)

logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

b) Specify your model in the equation (note: you can refer ‘results’ tab > data > Coefficient)


Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Result History
Logistic Regression Model (Logistic Regression)
Data
Description
Annotations
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
Titanic_Data (10/31/23 7:42 PM – 185
processes
DT_Titanic (11/7/23 9:08 AM – 2 kB)
Kmeans_Titanic (11/7/23 10:44 AM –
LogR_Titanic (11/7/23 9:18 AM – 4 kB
NN_Titanic (11/7/23 9:38 AM – 4 kB)
SVM_Titanic (11/7/23 9:35 AM – 4 kB)
DB (Legacy)

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Age	0.022	0.279	0.005	4.281	0.000
No of Siblings or Spou...	0.274	0.285	0.070	3.896	0.000
No of Parents or Child...	-0.099	-0.086	0.075	-1.326	0.185
Passenger Fare	-0.015	-0.784	0.002	-8.110	0.000
Intercept	0.226	0.467	0.163	1.382	0.167

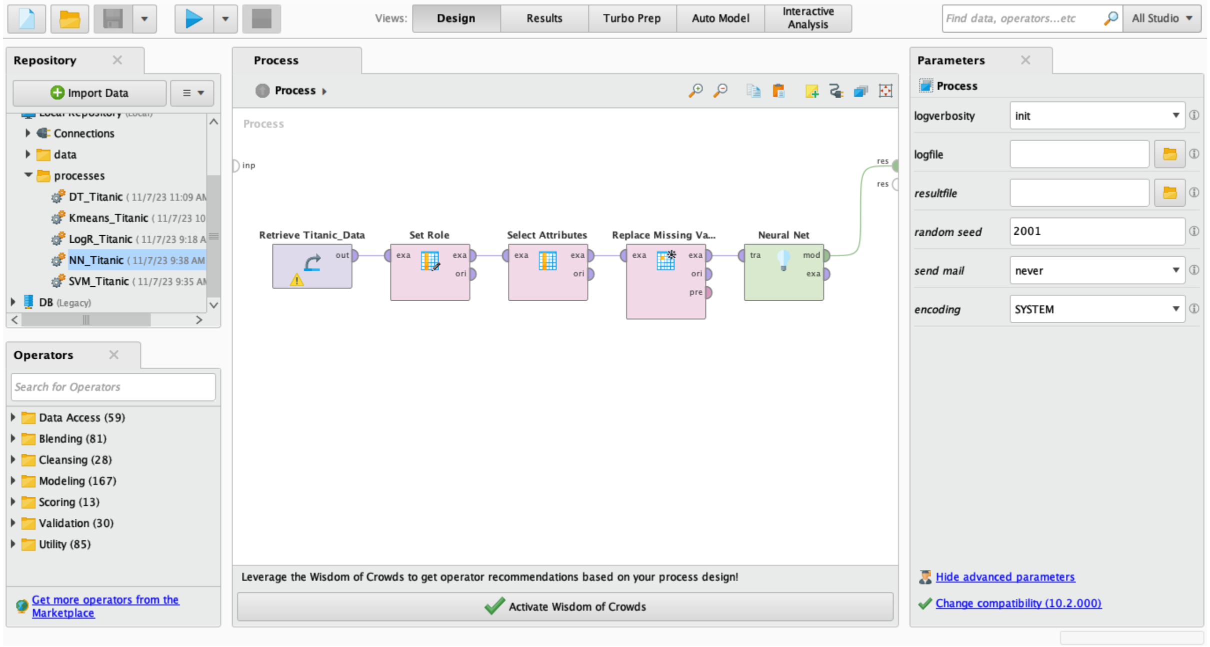
c) Based on your results, what is the most important variable to estimate your target variable? Can you provide your logic on why?

Ans: The **most significant** attribute is No of siblings or spouses on board because it has the highest coefficient i.e. 0.274. and the p value is 0.

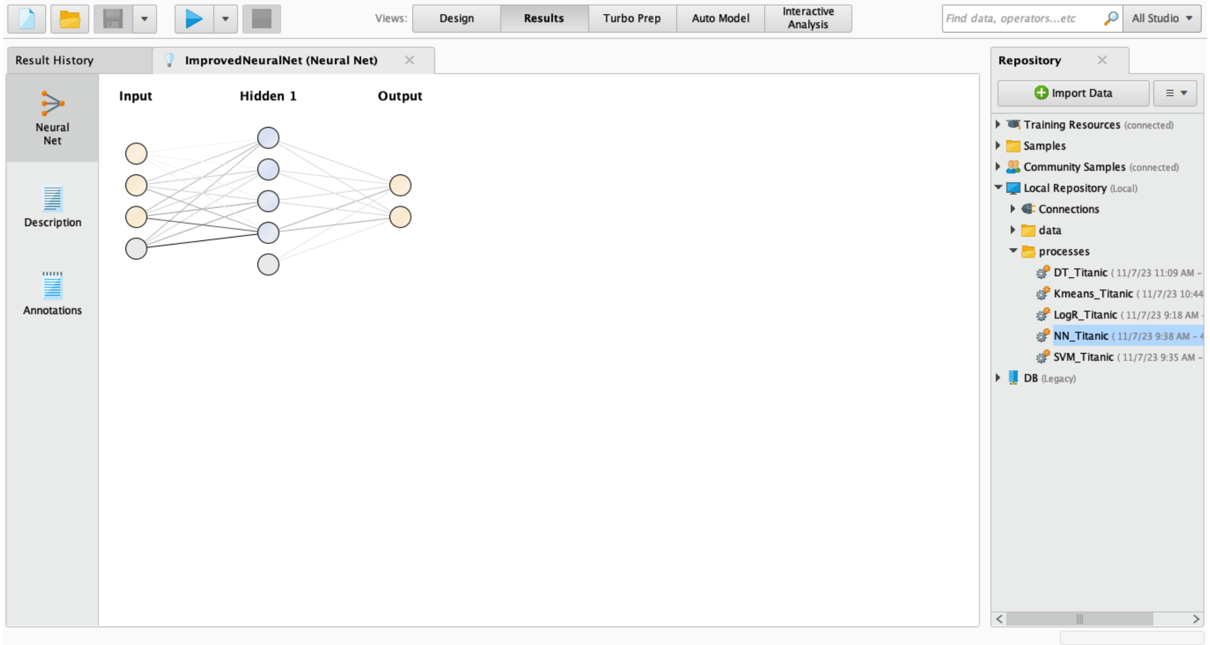
Age, No of siblings or spouses on board and Passenger Fare are the significant variable to the target variable survived because the p values of these attributes are 0 or less than 0(p value <= 0) but the p value for No of parents and children’s on board is more than zero i.e. 0.185 which is not the significant attribute for survival.

Task 4: Artificial Neural Network (ANN)

a) Build your ANN model


Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Repository
Import Data
Connections
data
processes
DT_Titanic (11/7/23 11:09 AM
Kmeans_Titanic (11/7/23 10
LogR_Titanic (11/7/23 9:18 A
NN_Titanic (11/7/23 9:38 AM
SVM_Titanic (11/7/23 9:35 AM
DB (Legacy)
Operators
Search for Operators
Data Access (59)
Blending (81)
Cleansing (28)
Modeling (167)
Scoring (13)
Validation (30)
Utility (85)
Get more operators from the Marketplace
Process
Process
inp
res
res
Retrieve Titanic_Data
out
Set Role
exa
exa
ori
Select Attributes
exa
exa
ori
Replace Missing Va...
exa
exa
ori
pre
Neural Net
tra
mod
exa
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Process
Hide advanced parameters
Change compatibility (10.2.000)

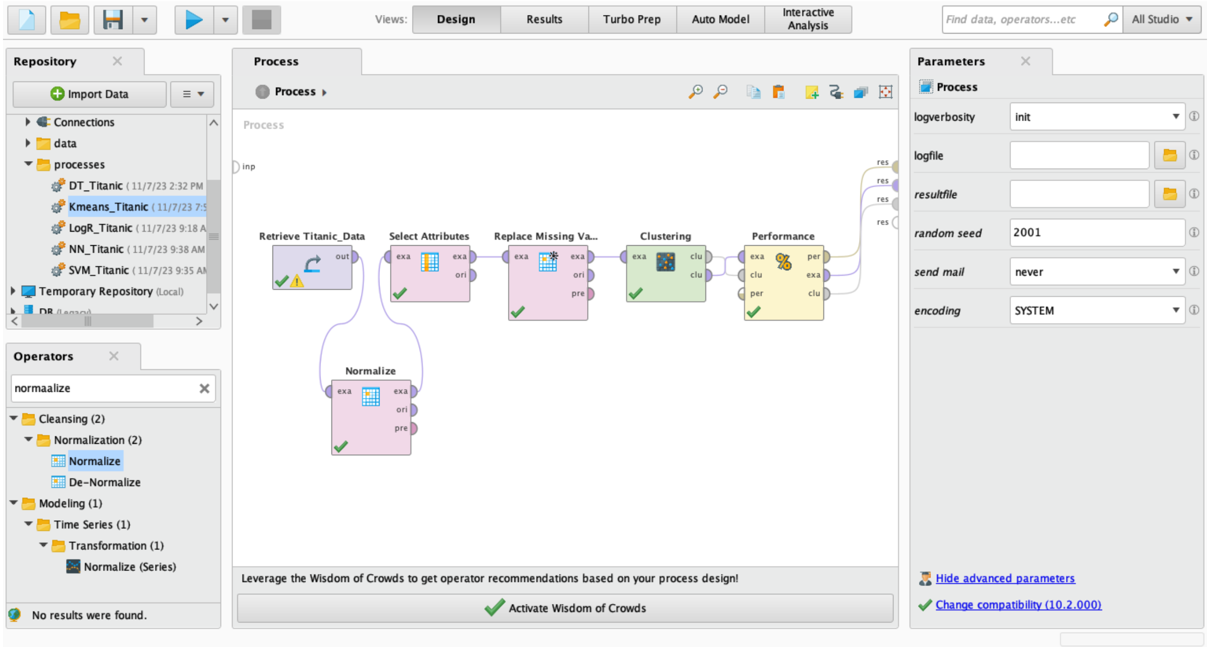
logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never
encoding: SYSTEM


Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Result History
ImprovedNeuralNet (Neural Net)
Neural Net
Description
Annotations
Input
Hidden 1
Output
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
DT_Titanic (11/7/23 11:09 AM -
Kmeans_Titanic (11/7/23 10:44
LogR_Titanic (11/7/23 9:18 AM
NN_Titanic (11/7/23 9:38 AM -
SVM_Titanic (11/7/23 9:35 AM -
DB (Legacy)

b) Provide your neural network including input, hidden layer, and output
Ans: Neural Net model suggests that passenger fare is the most significant factor in predicting survival. The relationship between age and survival is more complex, while the number of parents or children on board may increase the probability of survival. The output layer consists of two sigmoid functions, one for class 'Yes' (survival) and one for class 'No' (non-survival). The thresholds for these functions are both 1.078, indicating that the model is more likely to predict survival for individuals with higher activation values from the hidden layer nodes.

Task 5: Clustering (K-means)

a) Build your K-means


Views: Design | Results | Turbo Prep | Auto Model | Interactive Analysis
Find data, operators...etc
All Studio
Repository
Import Data
Connections
data
processes
DT_Titanic (11/7/23 2:32 PM
Kmeans_Titanic (11/7/23 7:5
LogR_Titanic (11/7/23 9:18 A
NN_Titanic (11/7/23 9:38 AM
SVM_Titanic (11/7/23 9:35 AM
Temporary Repository (Local)
DB (Legacy)
Operators
normaalize
Cleansing (2)
Normalization (2)
Normalize
De-Normalize
Modeling (1)
Time Series (1)
Transformation (1)
Normalize (Series)
No results were found.
Process
Retrieve Titanic_Data
Select Attributes
Replace Missing Va...
Clustering
Performance
Normalize
inp
out
exa
ori
pre
clu
per
res
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Hide advanced parameters
Change compatibility (10.2.000)

Process

logverbosity: init

logfile:

resultfile:

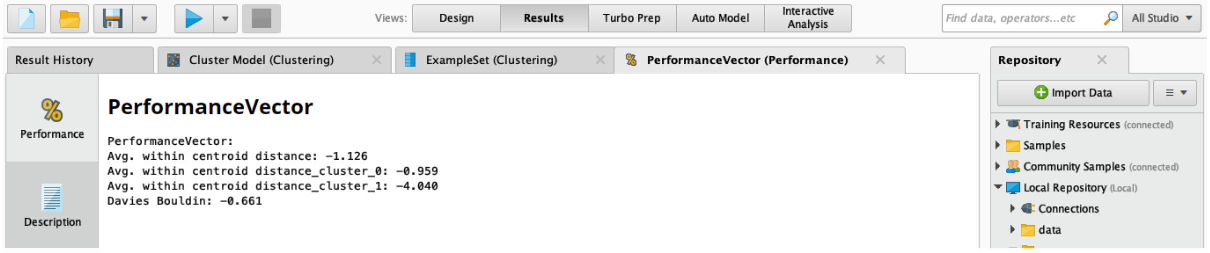
random seed: 2001

send mail: never

encoding: SYSTEM

b) Try to different k (i.e., number of clustering) ranging from 2-5, find your best k, and provide your logic.

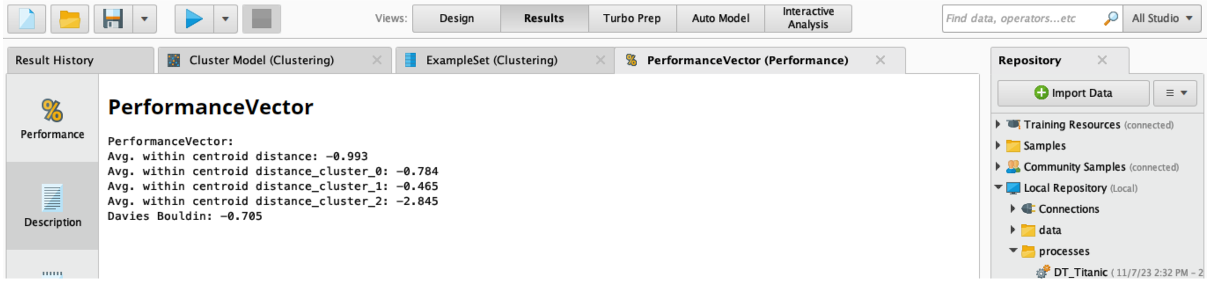
K= 2


Views: Design | Results | Turbo Prep | Auto Model | Interactive Analysis
Find data, operators...etc
All Studio
Result History
Cluster Model (Clustering)
ExampleSet (Clustering)
PerformanceVector (Performance)
Performance
Description
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -1.126
Avg. within centroid distance_cluster_0: -0.959
Avg. within centroid distance_cluster_1: -4.040
Davies Bouldin: -0.661
```

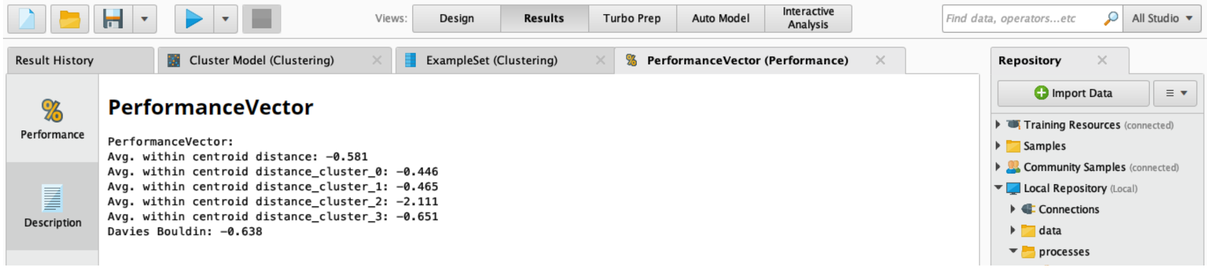
K= 3


Views: Design | Results | Turbo Prep | Auto Model | Interactive Analysis
Find data, operators...etc
All Studio
Result History
Cluster Model (Clustering)
ExampleSet (Clustering)
PerformanceVector (Performance)
Performance
Description
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
DT_Titanic (11/7/23 2:32 PM – 2

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.993
Avg. within centroid distance_cluster_0: -0.784
Avg. within centroid distance_cluster_1: -0.465
Avg. within centroid distance_cluster_2: -2.845
Davies Bouldin: -0.705
```

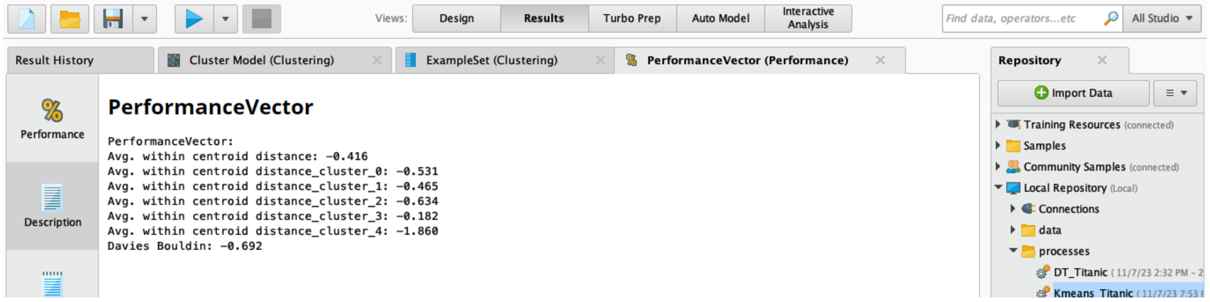
K=4


Views: Design | Results | Turbo Prep | Auto Model | Interactive Analysis
Find data, operators...etc
All Studio
Result History
Cluster Model (Clustering)
ExampleSet (Clustering)
PerformanceVector (Performance)
Performance
Description
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.581
Avg. within centroid distance_cluster_0: -0.446
Avg. within centroid distance_cluster_1: -0.465
Avg. within centroid distance_cluster_2: -2.111
Avg. within centroid distance_cluster_3: -0.651
Davies Bouldin: -0.638
```

K=5


Views:
Design
Results
Turbo Prep
Auto Model
Interactive Analysis
Find data, operators...etc
All Studio
Result History
Cluster Model (Clustering)
ExampleSet (Clustering)
PerformanceVector (Performance)
Performance
Description
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
DT_Titanic (11/7/23 2:32 PM – 2
Kmeans_Titanic (11/7/23 7:53 I

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -0.416
Avg. within centroid distance_cluster_0: -0.531
Avg. within centroid distance_cluster_1: -0.465
Avg. within centroid distance_cluster_2: -0.634
Avg. within centroid distance_cluster_3: -0.182
Avg. within centroid distance_cluster_4: -1.860
Davies Bouldin: -0.692
```

Insights:

Attributes are Age, Passenger Fare.

K=4 looks to be the optimal choice for clustering. This is because the data has now been effectively separated into four clusters, according to the Davies Bouldin Index, which shows a maximum value at K=4(-0.638). Higher values of the Davies Bouldin Index indicate stronger separation between clusters and within-cluster dispersion.

Although a higher average within centroid distance for k=2 (-1.126) may suggest more compact clusters, it is not always the best choice. A compromise between the meaningfulness and compactness of the clusters must be made to arrive at the optimal number of clusters.