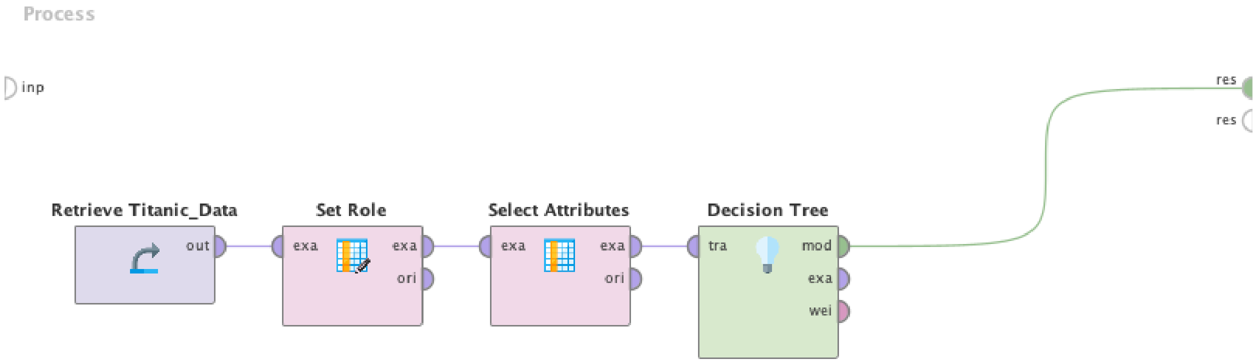
Assignment 4

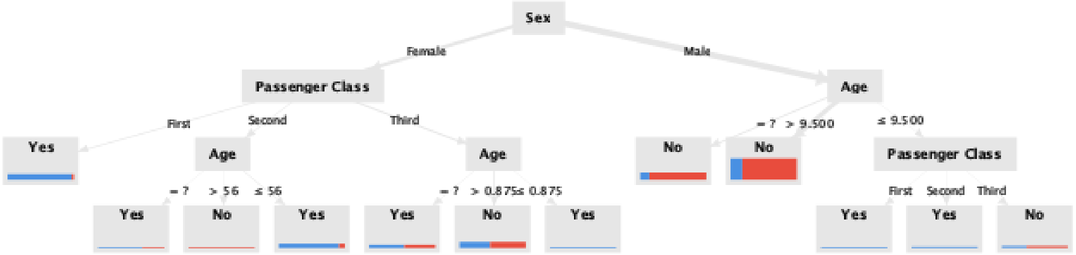
Name: Priyanshi Deliwala

Task 1: Decision Tree

Process:



Result:

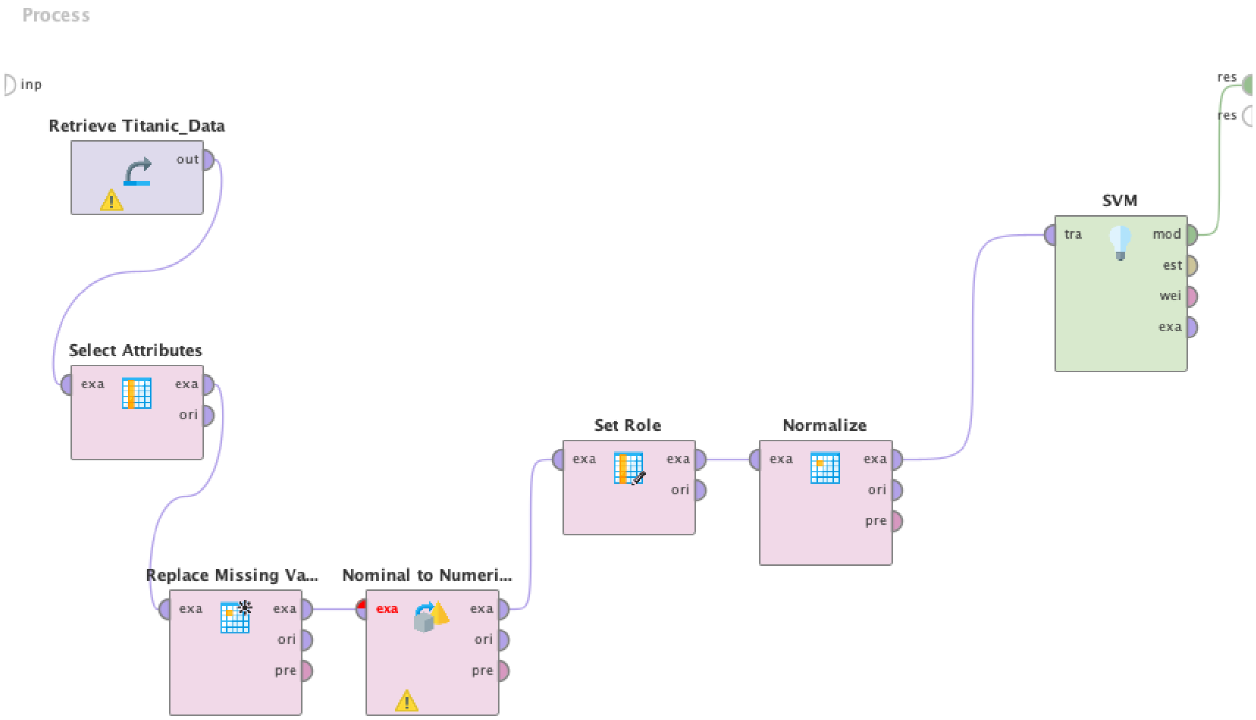


Interpretation:

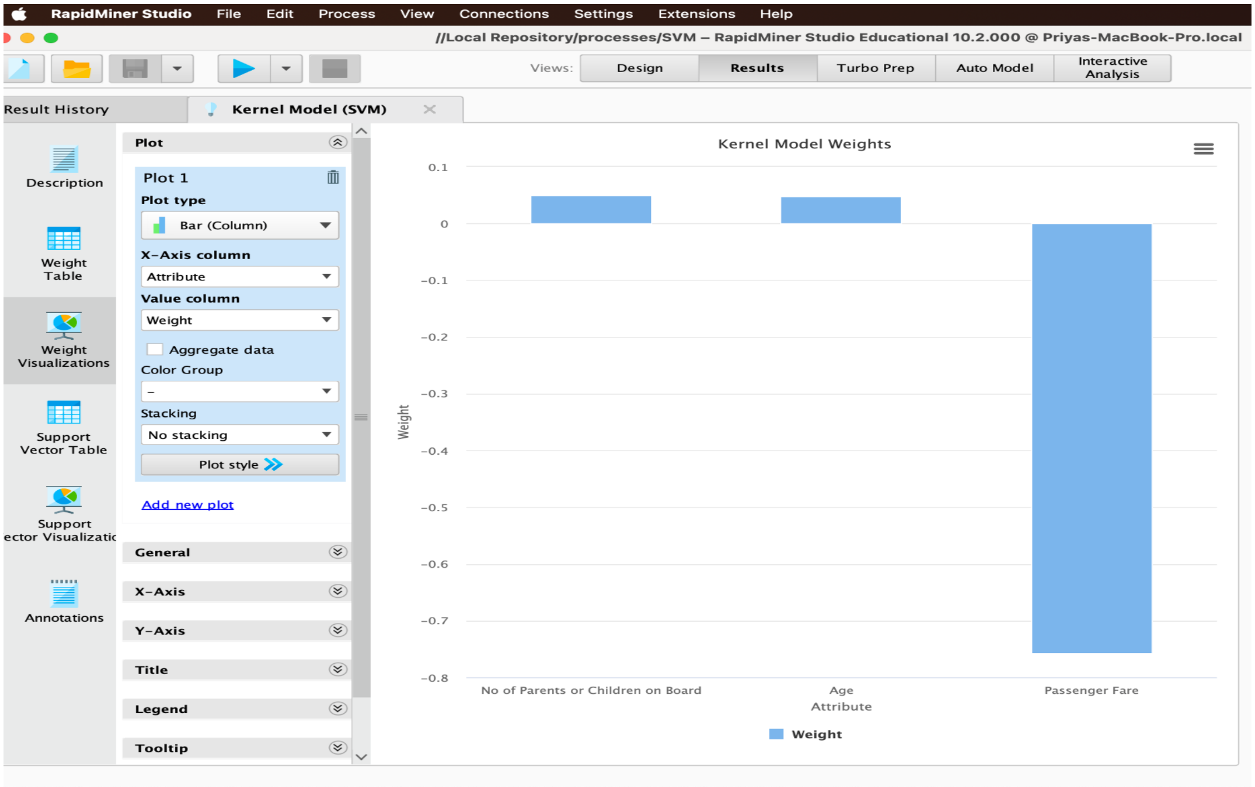
The decision tree with a depth of 4 was likely chosen to balance model complexity with the need to capture meaningful patterns in the data. In this case, deeper branches represent more specific conditions for making predictions, which can be valuable for understanding and explaining the decision-making process. Going deeper than 4 might lead to overfitting, making the model less generalizable to new data, while a shallower tree might not capture enough nuance in the dataset. Thus, a depth of 4 strikes a balance by providing a reasonably detailed decision tree without overly complex branching, making it a practical choice for the given dataset and problem.

Task 2: Support Vector Machine

Process:



Result:



Plot type: Bar (Column)

X-Axis column: Attribute

Value column: Weight

☐ Aggregate data

Color Group: –

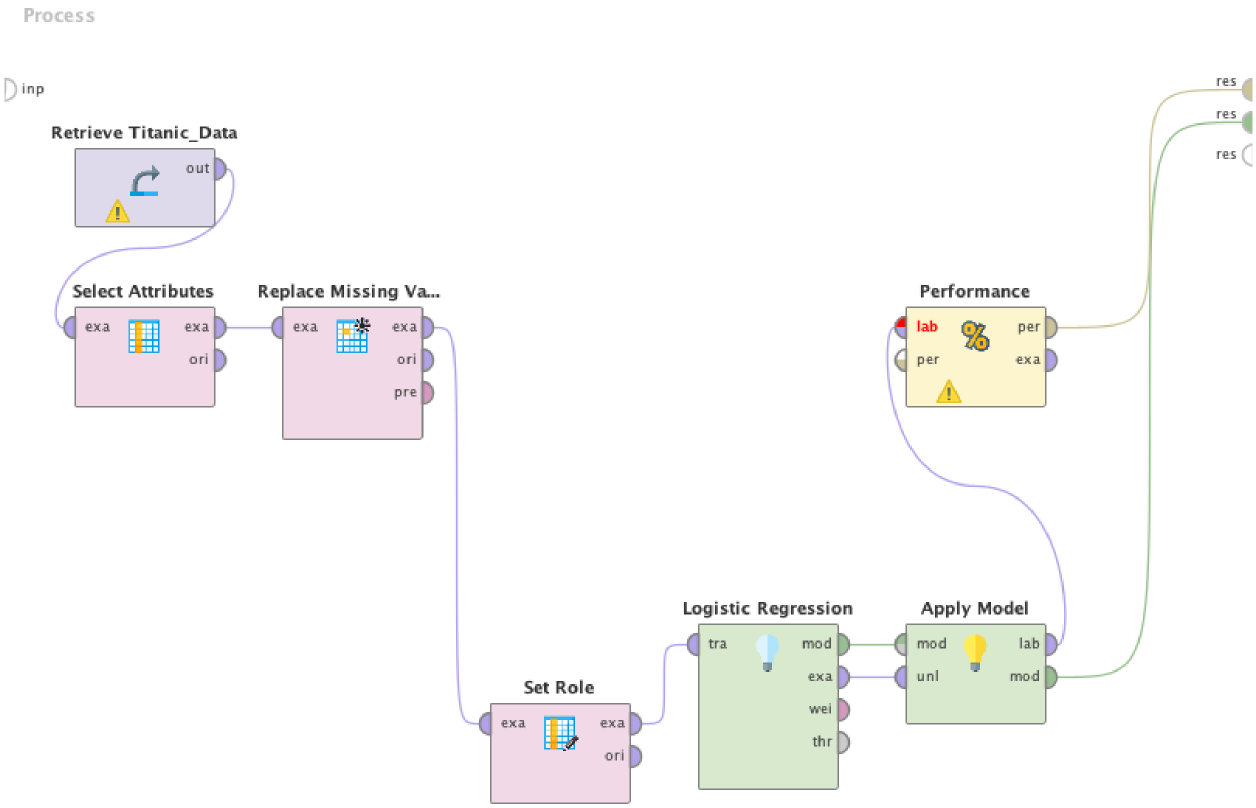
Stacking: No stacking

Interpretation:

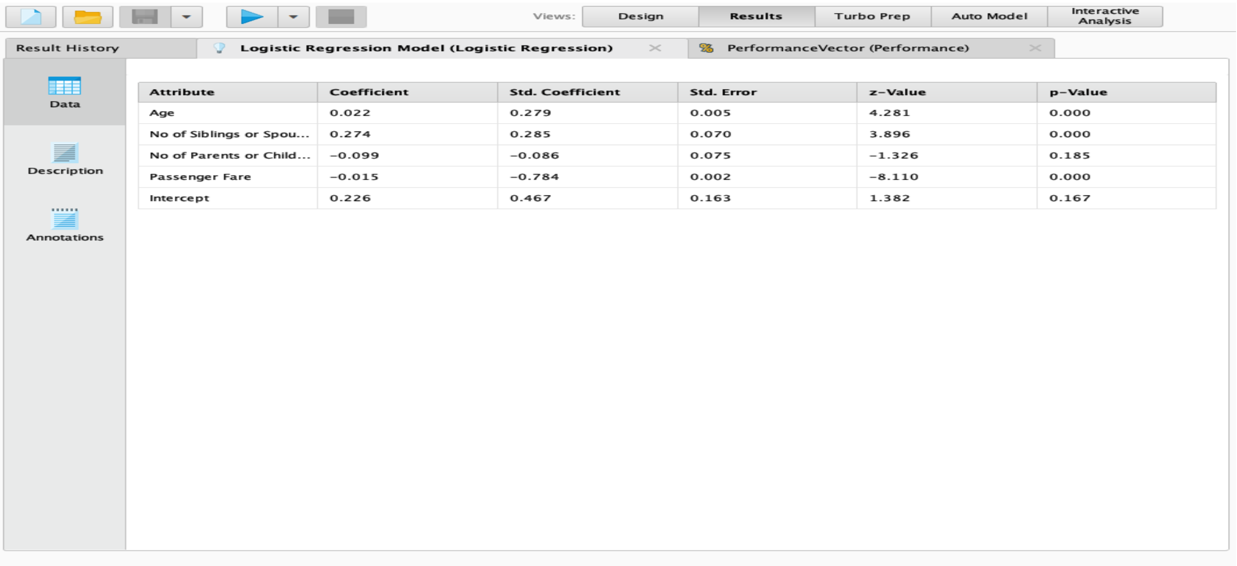
Whereas a positive weight for "No of Parents or Children on Board" indicates that passengers traveling with more parents or children have a higher possibility of being categorized positively, a positive weight for "Age" indicates that older passengers are more likely to be classed positively. On the other hand, the "Passenger Fare" negative weight suggests that a greater fare has a negative impact on the predictions, with higher-paying passengers having a lower likelihood of being categorized favorably. The size of these weights indicates how important the feature is; interestingly, "Passenger Fare" has the biggest impact on the decision boundary of the model.

Task 3: Logistic Regression

Process:



Result:

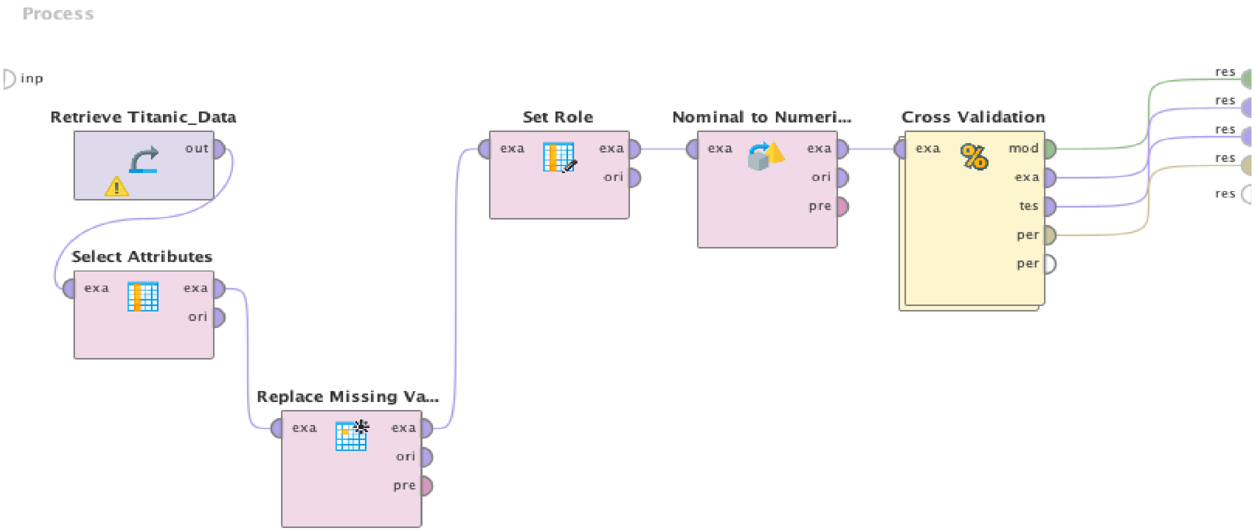


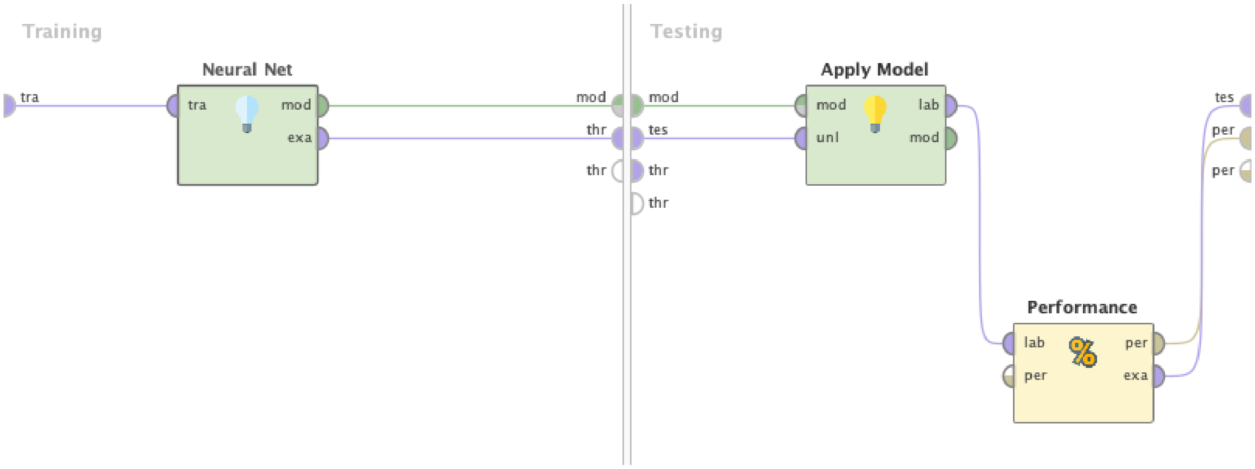
Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Age	0.022	0.279	0.005	4.281	0.000
No of Siblings or Spou...	0.274	0.285	0.070	3.896	0.000
No of Parents or Child...	-0.099	-0.086	0.075	-1.326	0.185
Passenger Fare	-0.015	-0.784	0.002	-8.110	0.000
Intercept	0.226	0.467	0.163	1.382	0.167

Interpretation:
The most crucial variable to estimate the target variable, according to the findings of the logistic regression model, seems to be "No of Siblings or Spouses on Board." This judgment is supported by several factors: First, compared to the other attributes, it has the largest positive coefficient (0.2740622543993109), suggesting a higher positive influence on the target variable. Second, when taking into account the standardization of variables, the standard coefficient (Std. Coefficient) is likewise rather high (0.2854792468407258), indicating that its impact is considerable. Third, this variable's p-value (9.769401908805583E-5) is rather low, suggesting that it has a statistically significant effect on the target variable's prediction. In the logistic regression model, "No of Siblings or Spouses on Board" stands out as the most significant feature overall due to its statistical significance, substantial positive coefficient, and standardization.

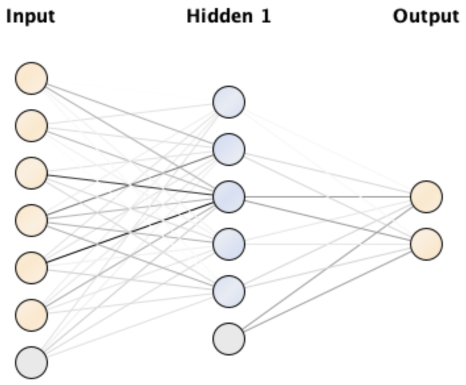
Task 4: ANN

Process:





Result:

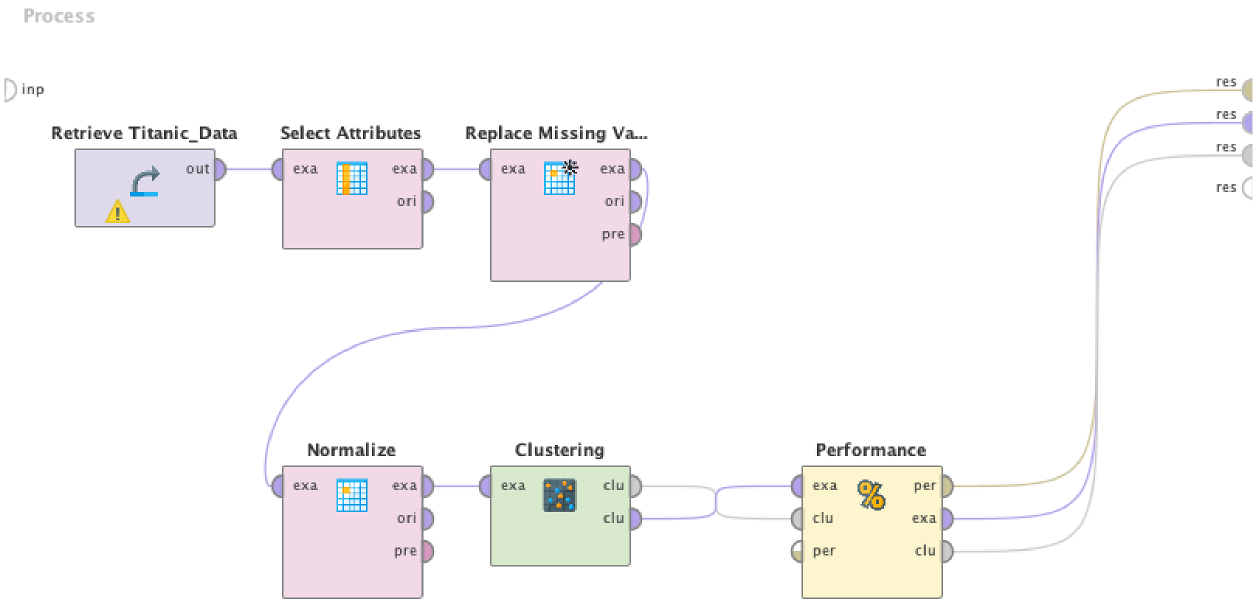


Interpretation:

The Artificial Neural Network (ANN) model demonstrated a consistent performance with an accuracy of 69.82% and a standard deviation of 2.75%. The confusion matrix showed that while it accurately predicted 223 instances of "Yes" and 691 instances of "No," it incorrectly identified 118 instances of "Yes" as "No" and 277 instances of "No" as "Yes." The total accuracy of 69.82% is further supported by the micro average accuracy. To summarize, the ANN model exhibits a moderate degree of accuracy, albeit with some misclassifications. The low standard deviation suggests that the performance is consistent. However, additional study may be required to address the particular consequences of these misclassifications.

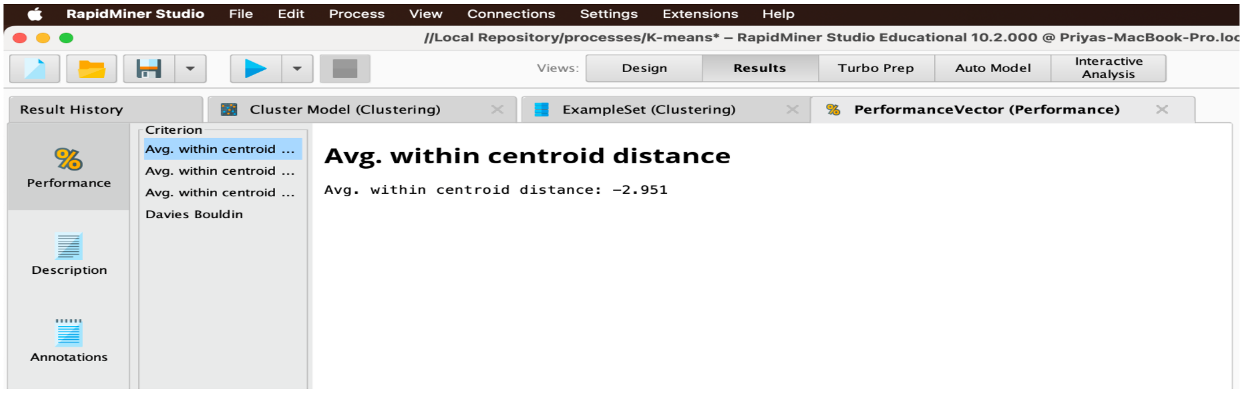
Task 5: K-Means Clustering

Process:



Result:

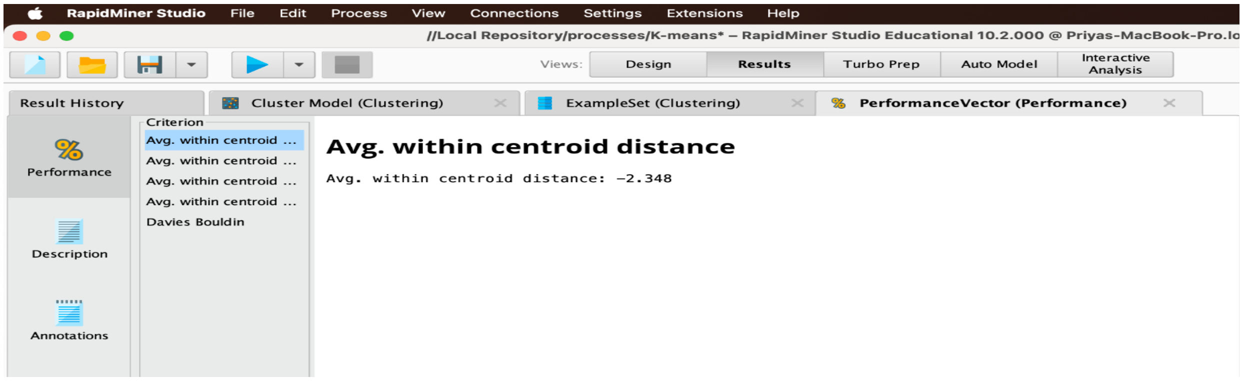
K=2



Avg. within centroid distance

```
Avg. within centroid distance: -2.951
```

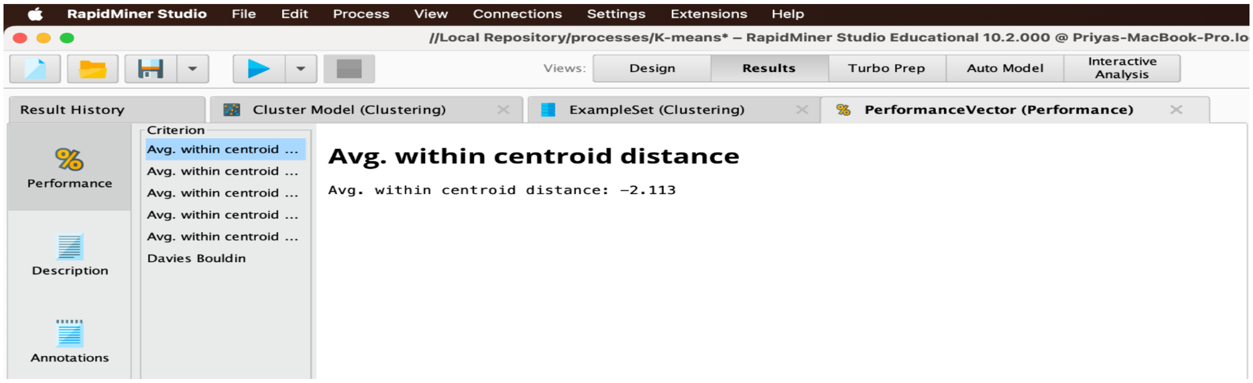
K=3



Avg. within centroid distance

```
Avg. within centroid distance: -2.348
```

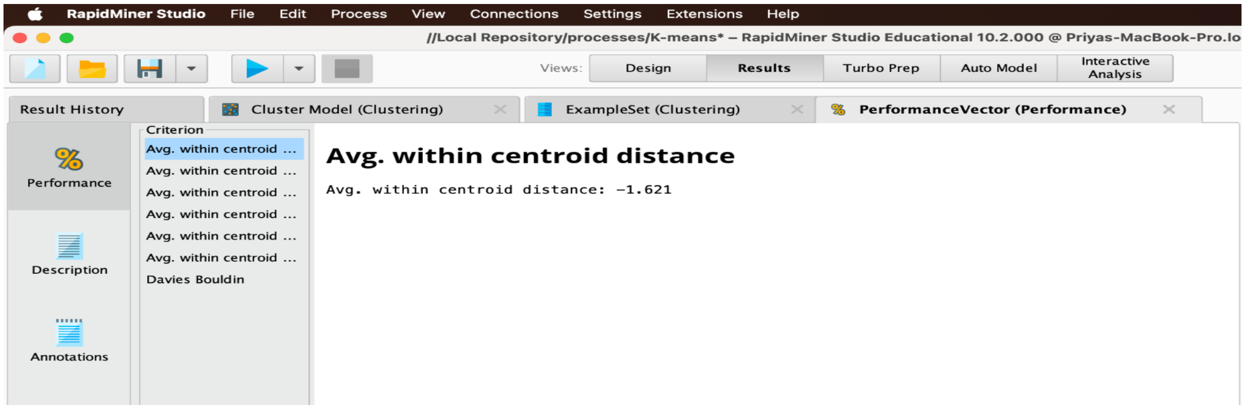
K=4



Avg. within centroid distance

```
Avg. within centroid distance: -2.113
```

K=5



Avg. within centroid distance

```
Avg. within centroid distance: -1.621
```

Interpretation:

Finding the optimal K value in the average within-cluster distance (negative values) for a clustering method. The number of clusters at which increasing K further does not significantly reduce the within-cluster distance is indicated by this point. It seems that K=2 might be the best option in this situation. Between K=2 and K=3, the average within-centroid distance dramatically drops; it then drops less sharply for K=4 and K=5. The greatest within-cluster distance decrease is observed at K=2, suggesting a more effective partitioning of data points into clusters.