BAN 210 – FINAL ASSESSMENT

ANALYSIS ON THE BREAST CANCER DATASET

BY-

STUDENT NAME: Poornima Singh

STUDENT ID: 125638213

SUBJECT: Predictive Analytics (BAN210ZBB.103325.2221)

PROFESSOR: Uzair Ahmad

DATE: 13th April 2022

INTRODUCTION:

The Final Assessment of 210 is an analysis where I have used predictive modelling to predict the class of the target variable of the Breast Cancer data. In the below assessment I have used Logistic Regression and Decision Tree models to predict the class of the data point of whether the Target is a recurrence event or non-recurrence. I have also conducted an analysis to understand which model is the best to run the prediction.

OBJECTIVE OF THE ANALYSIS:

The first step of any analysis is to understand the reason and the purpose of the analysis. In our cases, we will be able to answer the following two questions from the results we will obtain:

- What is the class of the Target Variable, whether the value of the Response Variable is a “Recurrence Event” or a “Non-Recurrence Event”
- Which model is performing better and by how much accuracy

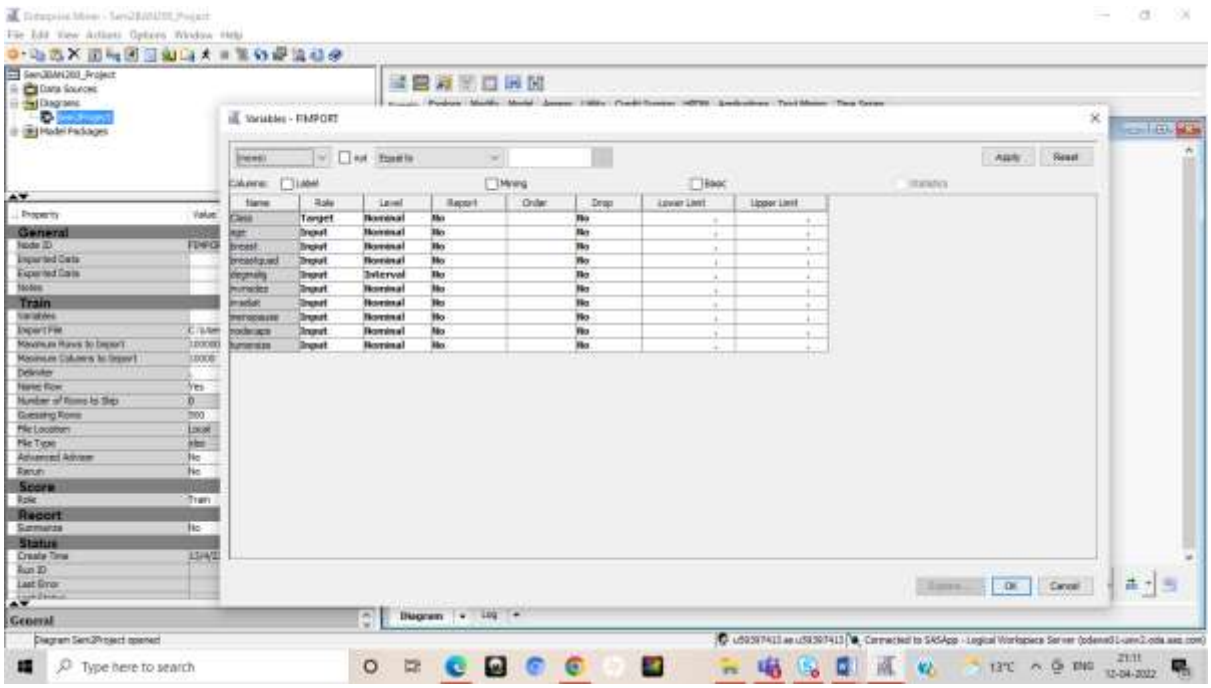
METHODOLOGY AND INFERENCES:

Below are the steps I followed using SAS Miner to analyse the dataset:

Step 1: File Import

In the first step, File Import node is used to import and read the dataset on the user system. The file can be read by adding the link to the path on the “Import File” option in the Properties of the node.

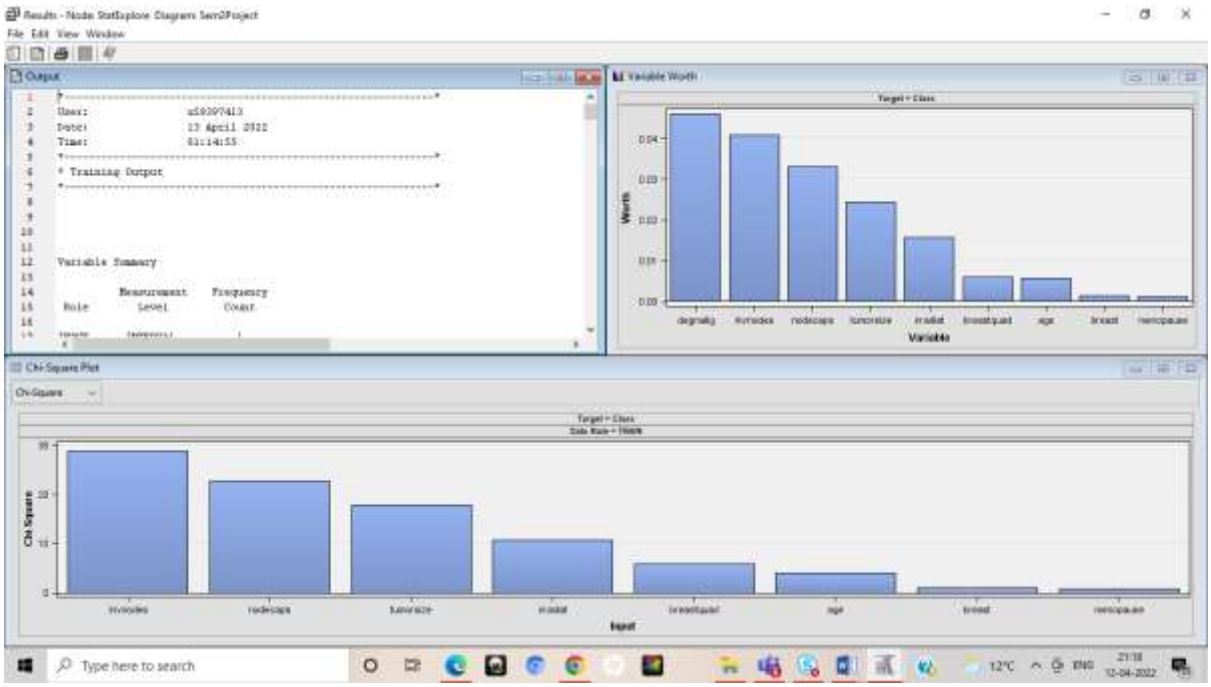
Using the “Variable” under Properties, assign Role as “Target” to the Class variable. The remaining features are identified as “Input” variables as they are the independent variables.

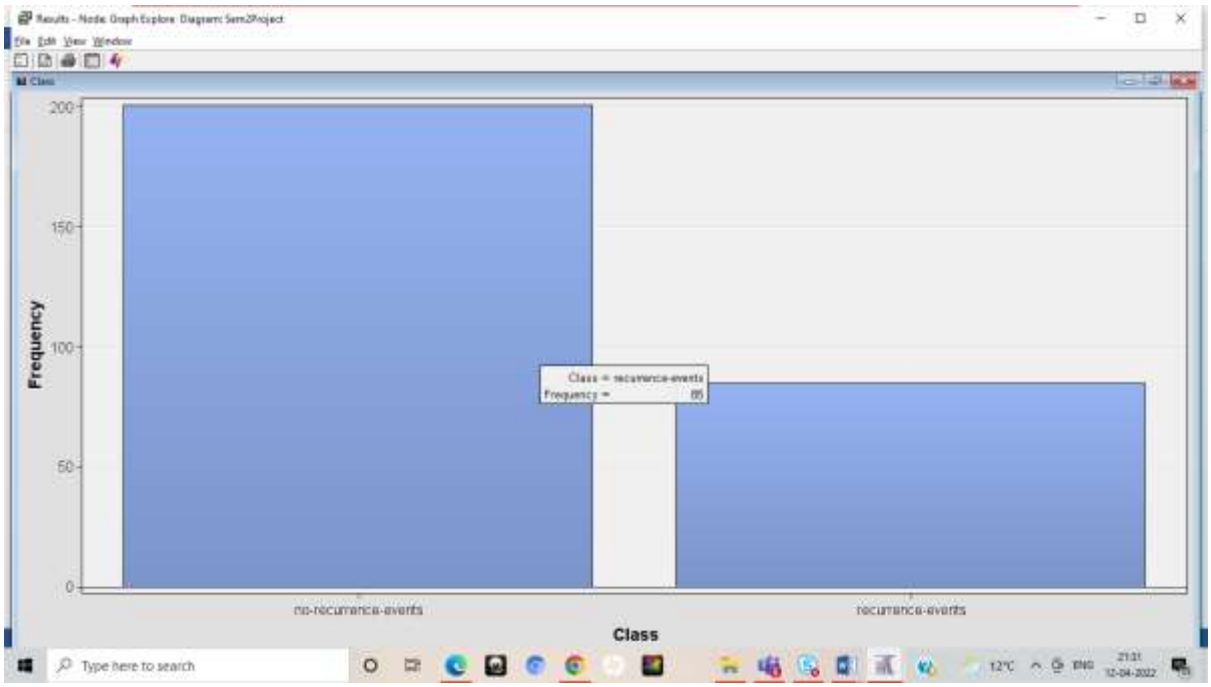


Step 2: Stat Explore

After the File Import, we add the Stat Explore node to study the class variable. Below is the screenshot of the result obtained from the Stat Explore node. The output has been shared on the GitHub link provided below.

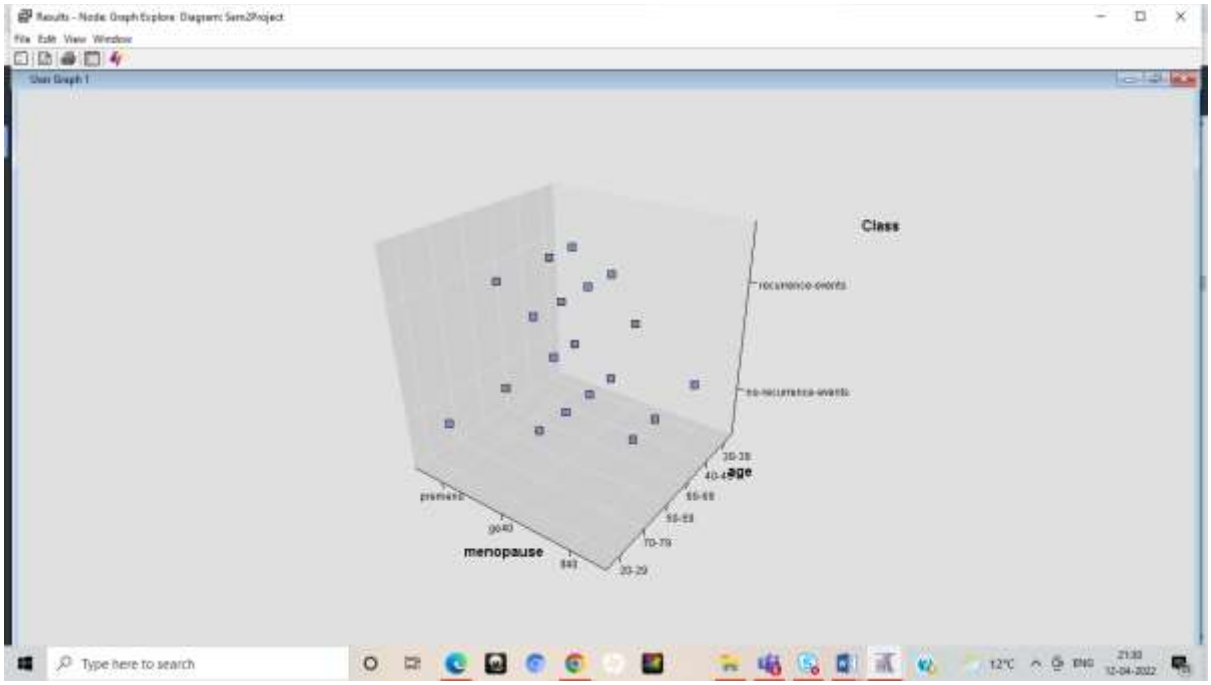
The below frequency graph shows the distribution of the Target variable with 86 recurrence event.



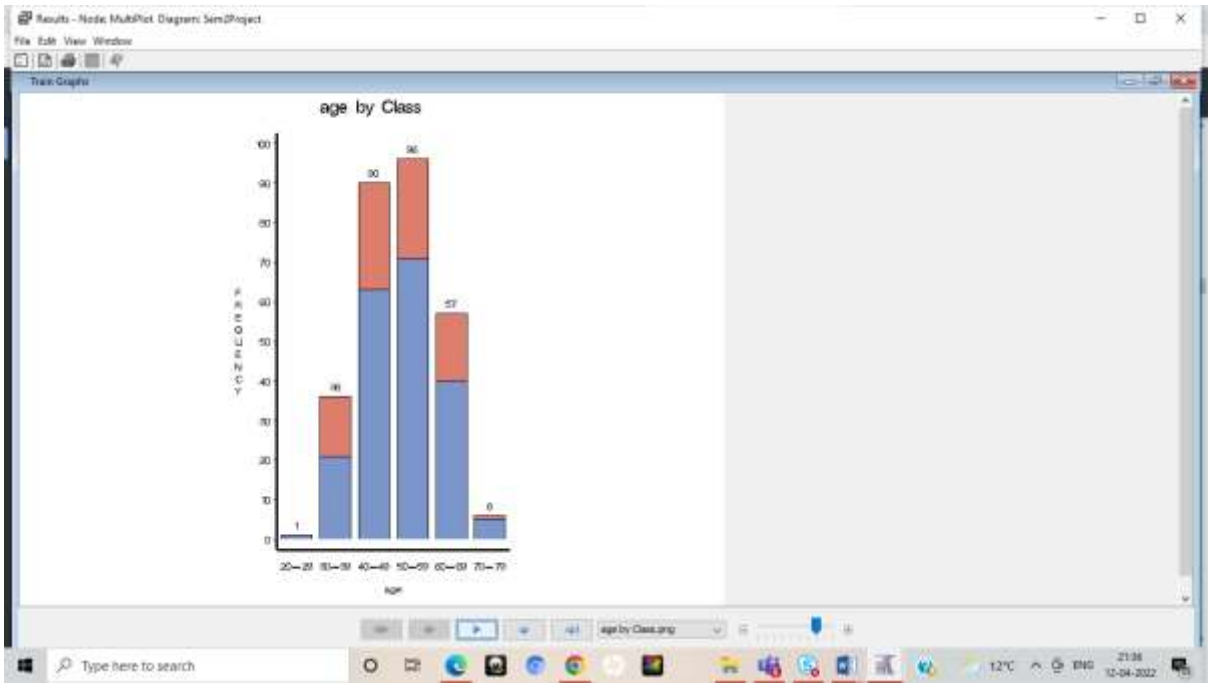


Step 3: Graph Explore

Next to the Diagram, Graph Explore node is added connecting it to the Stat Explore. In the below graph we have a 3D – Scatter Plot among the Target Variable, Age and Menopause.



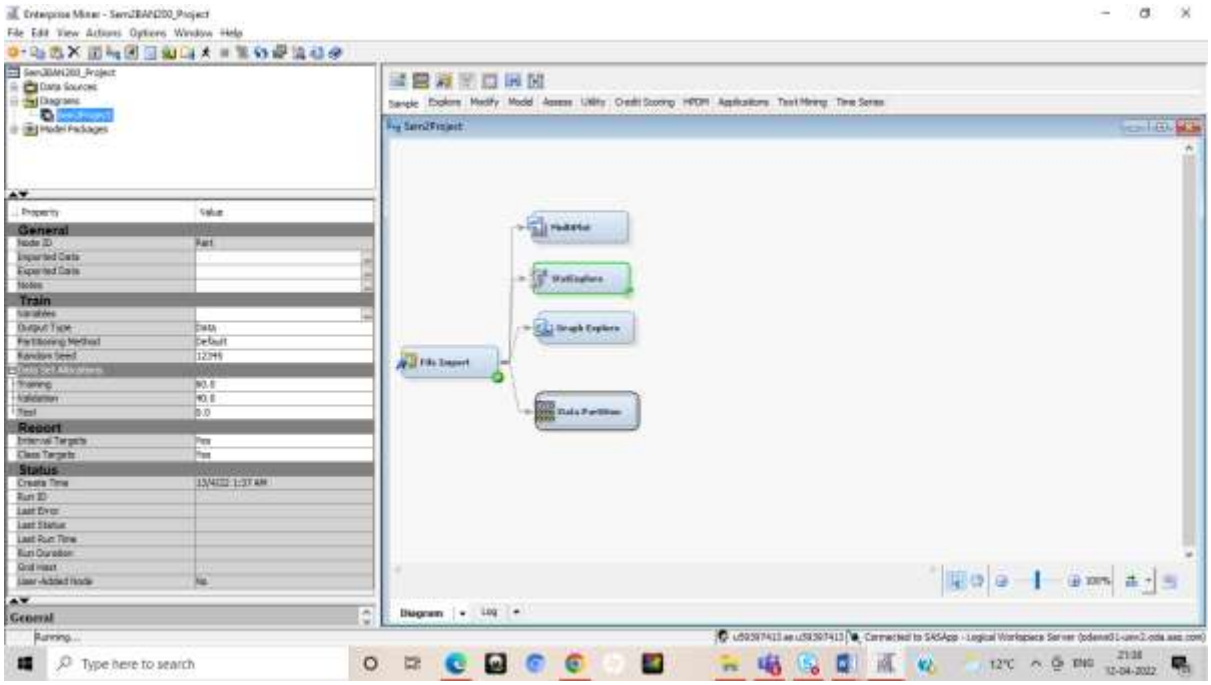
Step 4: Multiplot

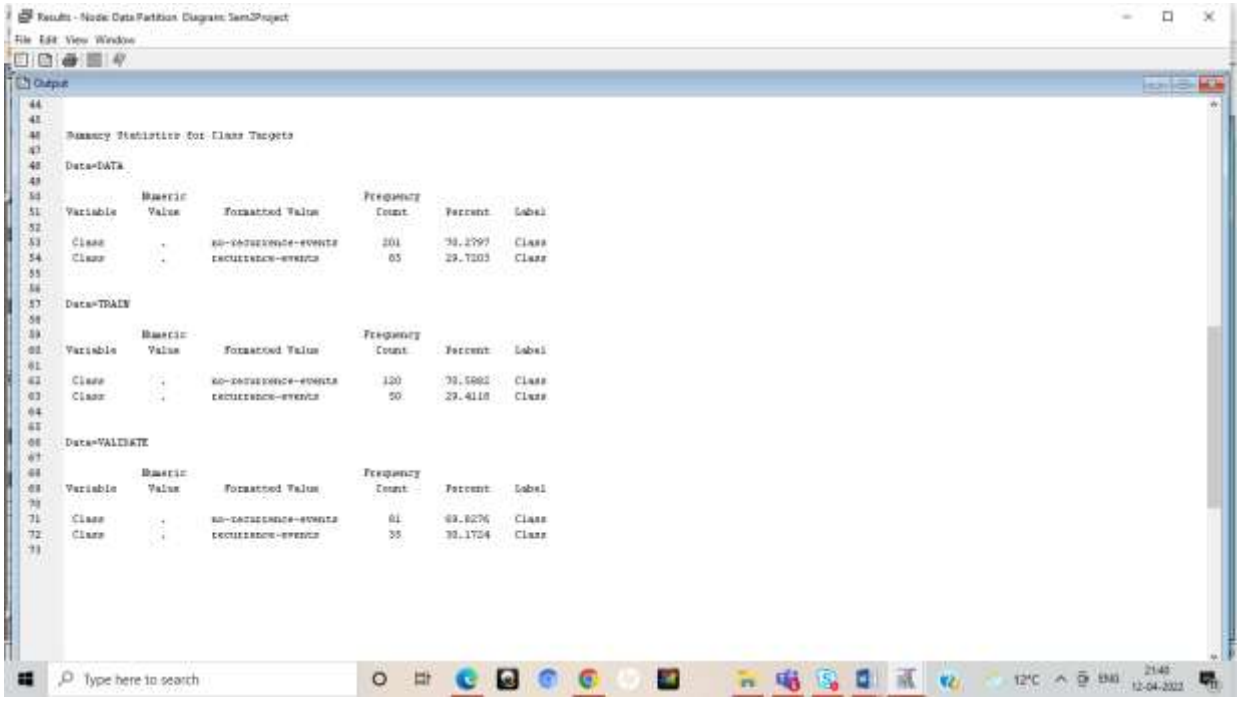


Step 5: Data Partition

To avoid any overfitting and under fitting, by partioning the data as 40% Validation dataset and 60% Train dataset.

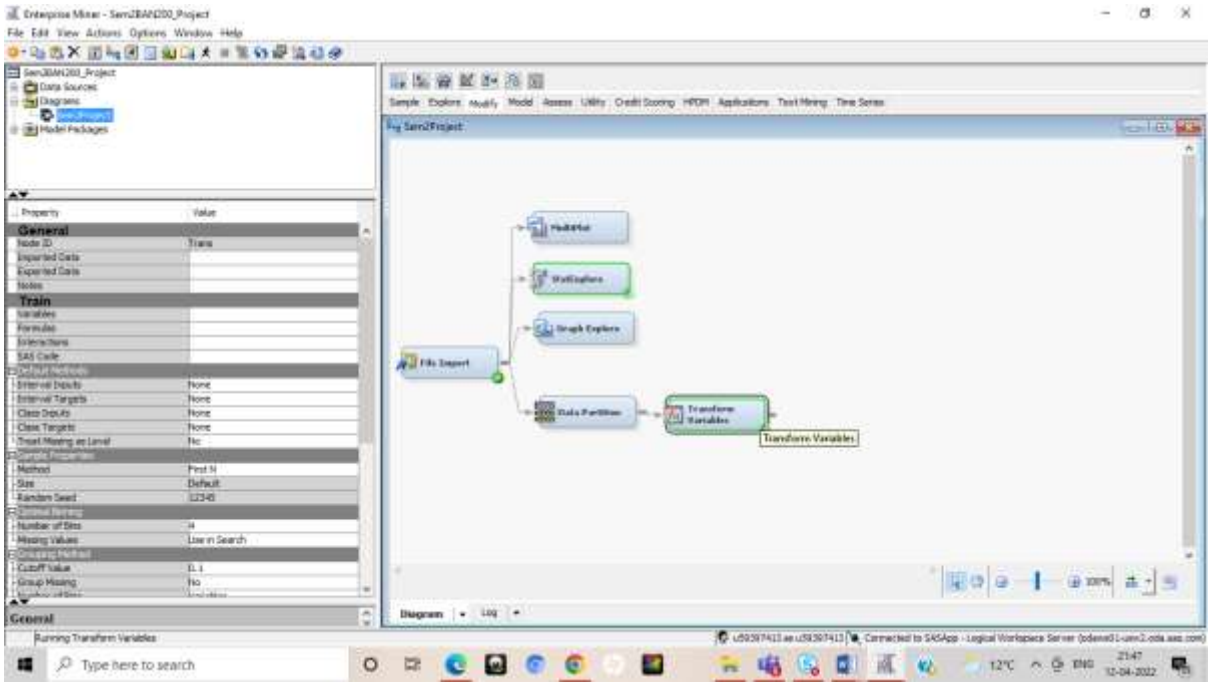
The below screenshot of the results window shows the distribution of the population.

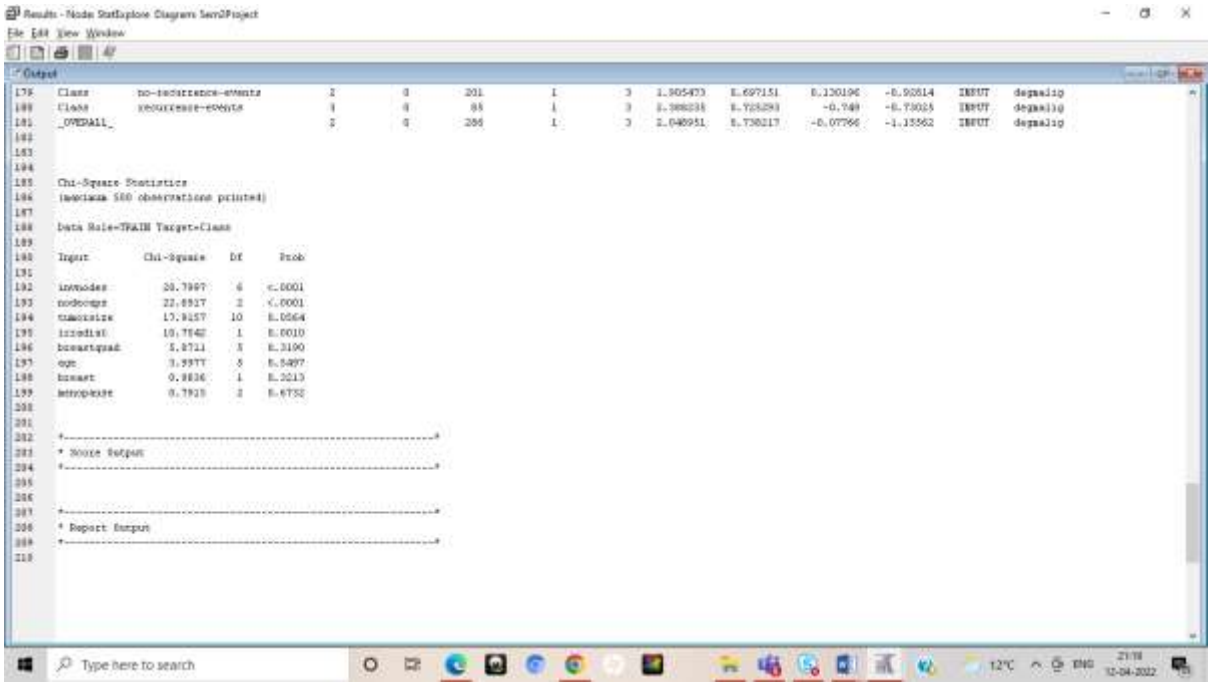




Step 6: Data Transformation

The data transformation node has been used to transform the variables for the model.





Step 7: Regression Model

Because we are running prediction on classification variable, Logistic Regression model has been used. The Regression node has been connected to the data Transformation node and Logit has been selected under Properties.

The following has been obtained from the Regression output:

DATASET	MISCLASSIFICATION RATE	MEAN SQUARE ERROR
Train Data	0.258824	0.201877
Validation Data	0.258621	0.195585

Event Classification Table for Validation dataset:

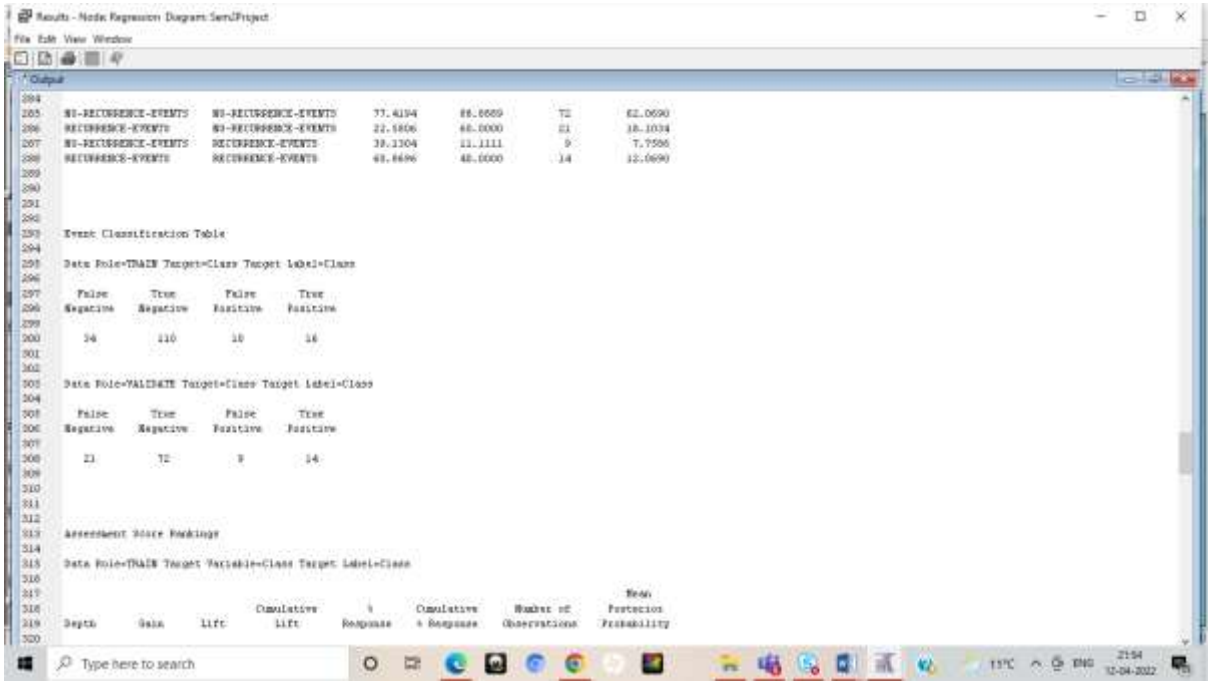
Target Variable	Target = 0	Target = 1	Total
Target = 0	TN = 72	FP = 9	81
Target = 1	FN = 21	TP = 14	35
Total	93	23	116

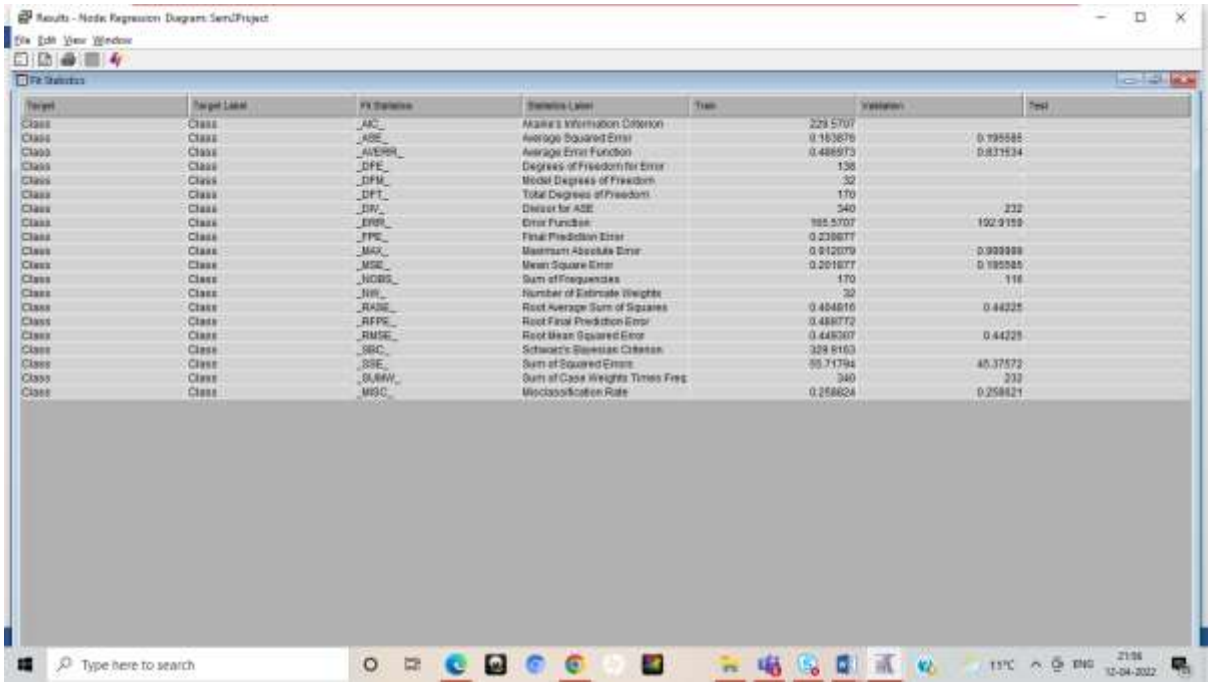
Recall = TP/(TP+FN) = 14/(14+21) = 14/35 = 0.40

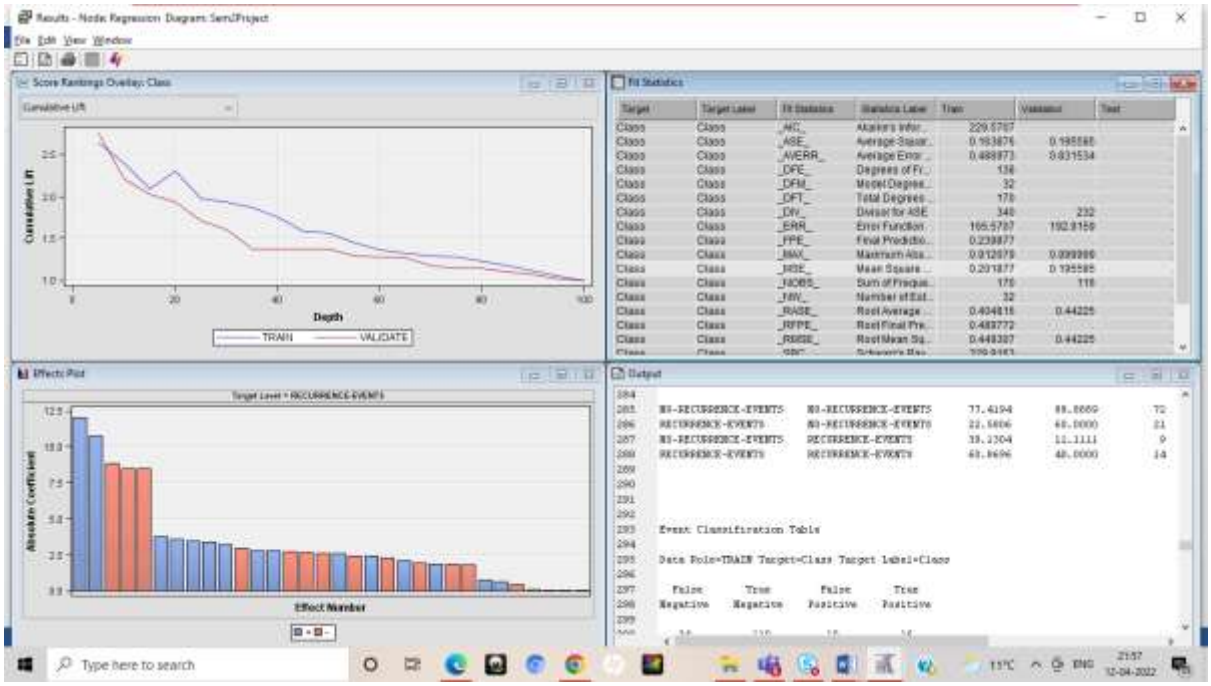
Precision = TP/(TP+FP) = 14/(14+9) = 14/23 = 0.6086

F1 = 2.P.R/(P+R) = 0.48688/1.0086 = 0.4827

Below are the screenshots of the SAS Miner outputs:







Step 8: Decision Tree

To run a comparison, I connected the Decision Tree to run the second model. Below are the results obtained:

DATASET	MISCLASSIFICATION RATE	MEAN SQUARE ERROR
Train Data	0.241176	0.806723
Validation Data	0.232759	0.806723

Event Classification for Validation Dataset:

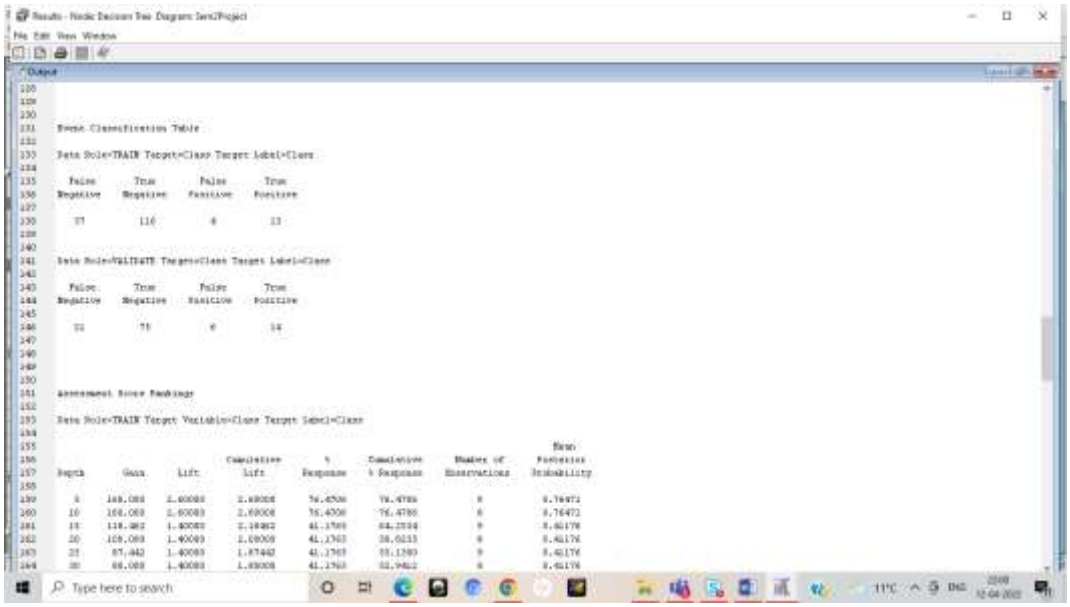
Target Variable	Target = 0	Target = 1	Total
Target = 0	TN = 75	FP = 6	81
Target = 1	FN = 23	TP = 14	37
Total	98	20	118

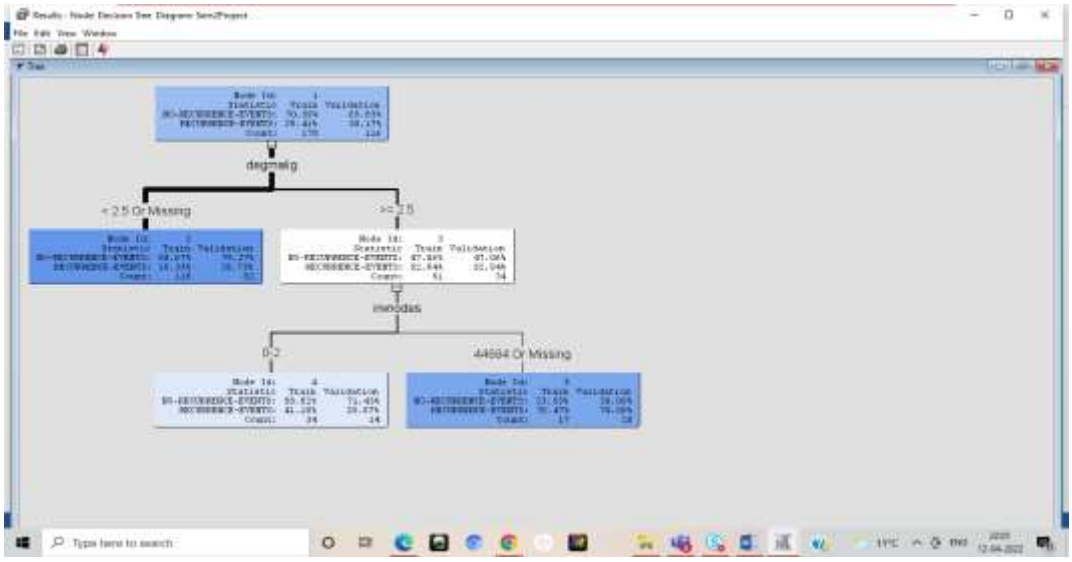
Recall = TP/(TP+FN) = 14/23 = 0.6086

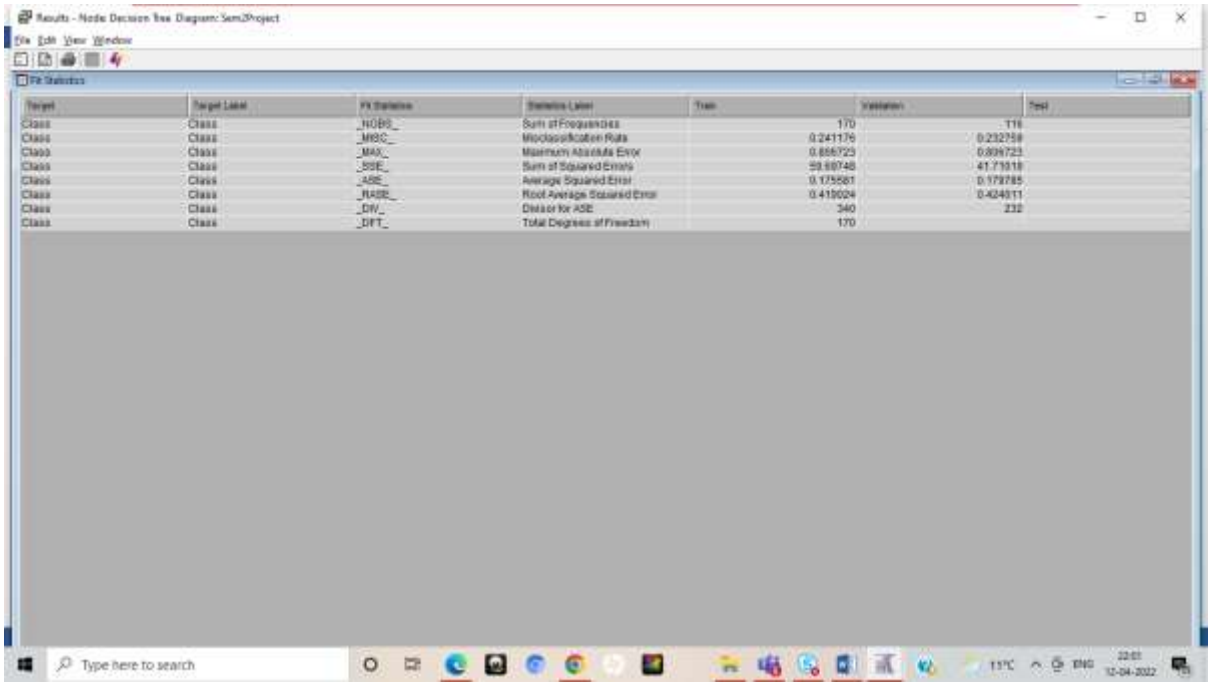
Precision = TP/(TP+FP) = 14/20 = 0.7

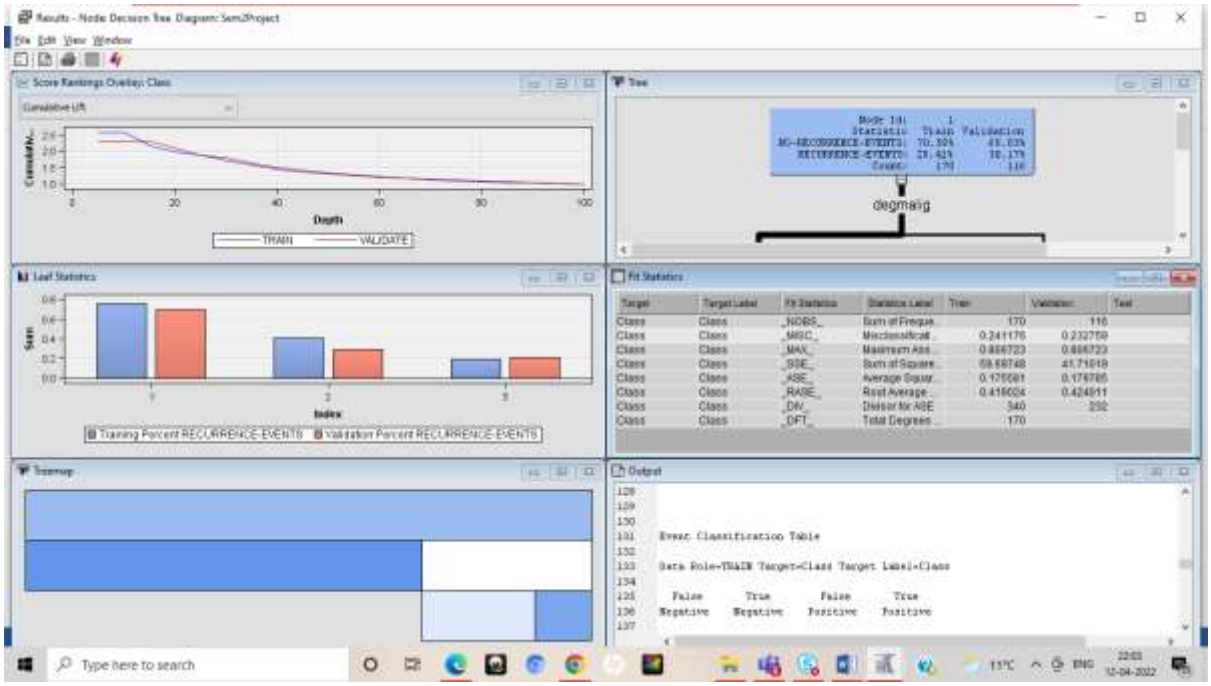
F1 = 2.P.R/(P+R) = 0.8520/1.3086 = 0.6510

Below are the screenshots of the SAS Miner outputs:



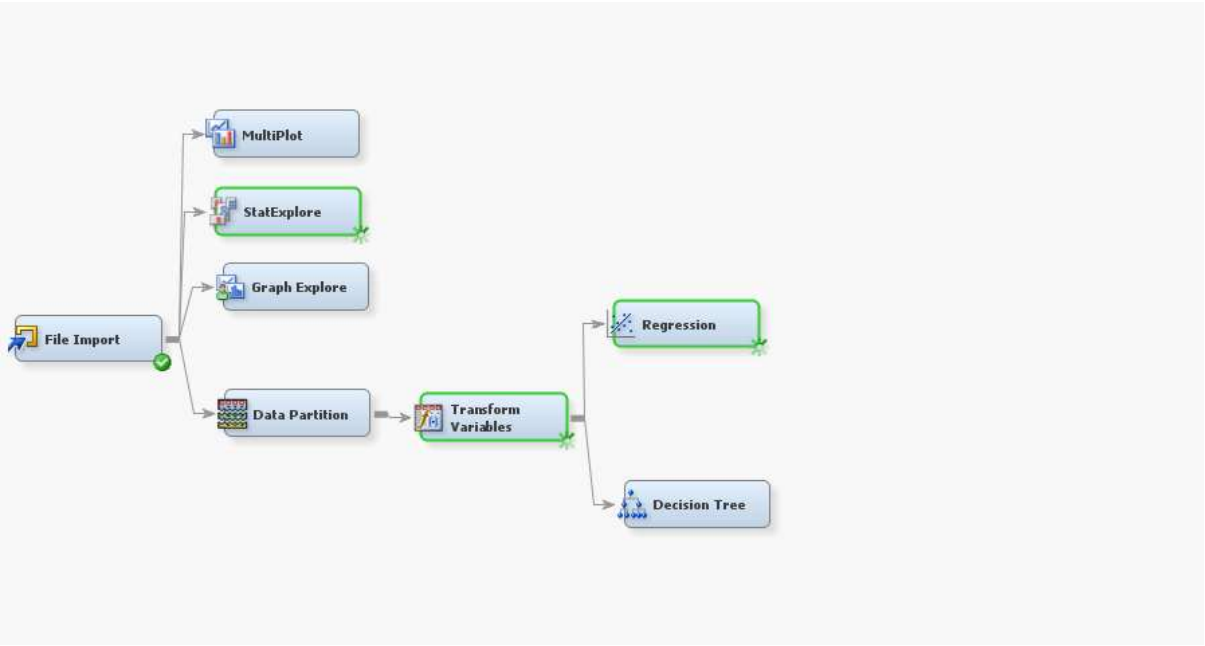






Step 9: Final Diagram

I have obtained the final diagram as below:



CONCLUSION:

I compare the two models inferring MSE, Misclassification Rate, Recall, Precision and F1 Score. Logistic Regression model is better than the Decision Tree model, as the Misclassification Rate is higher as compared to the DT model and also has a lower MSE.

Model	Misclassification Rate (Validation Dataset)	MSE	Recall	Precision	F1 Score
Logistic Model	0.258621	0.201877	0.4000	0.6086	0.4827
Decision Tree	0.232759	0.806723	0.6086	0.7000	0.6510

GITHUB LINK:

Please find below the GitHub link for the final assessment:

psing361/Breast-Cancer-Prediction-SAS-Miner-Project: This project has been conducted on SAS miner and is being used to detect breast cancer basis the available data. The project is part of my Predictive Analytics subject during the Business Analytics course at Seneca College (github.com)

DECLARATION:

I, **Poornima Singh**, declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. I have not distributed my work to other students.