

# **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: 13<sup>th</sup> 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

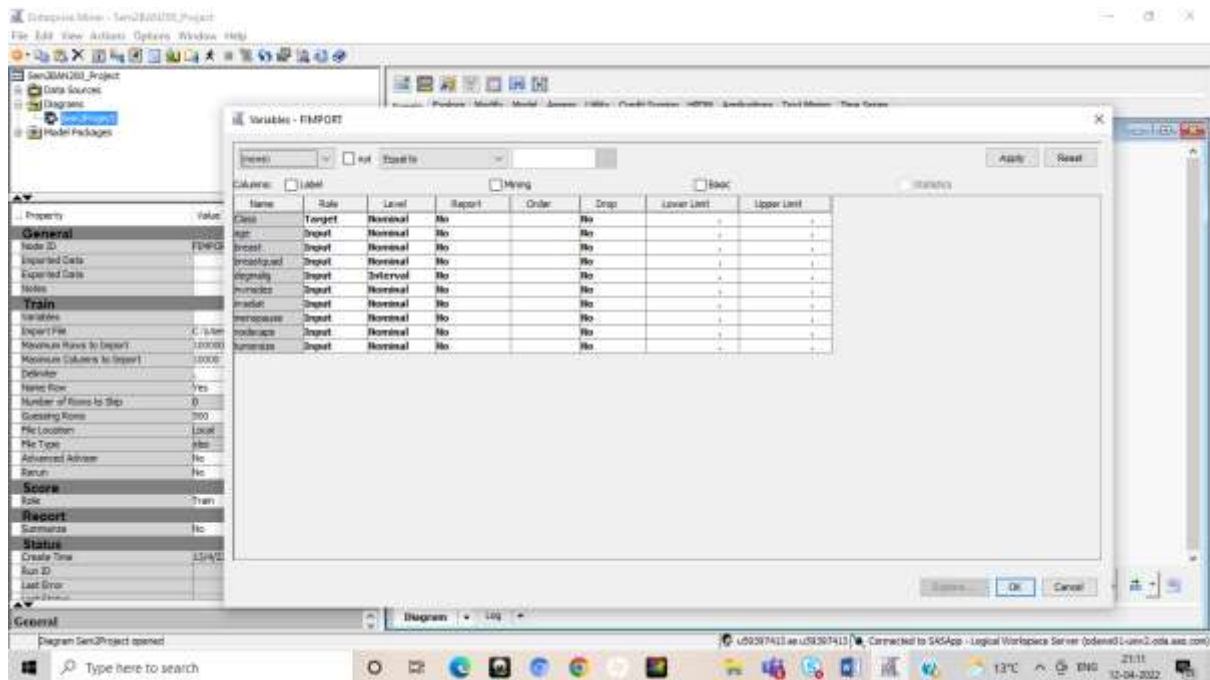
## METHODOLOGY AND INFERENCES:

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.

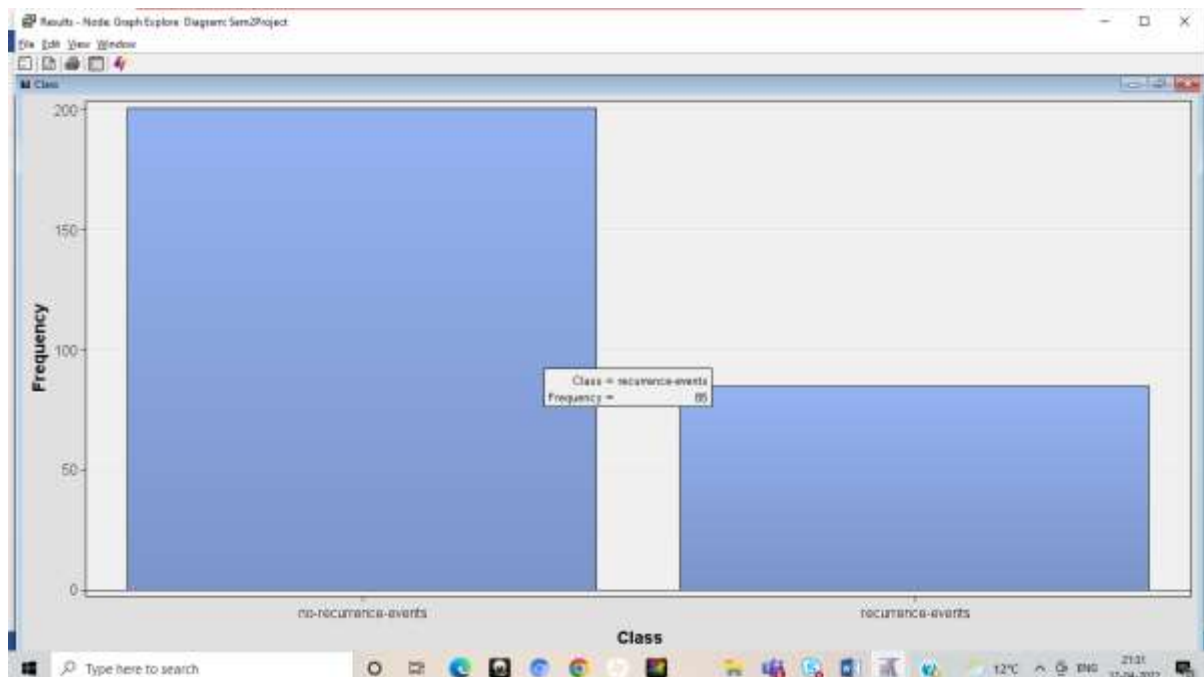
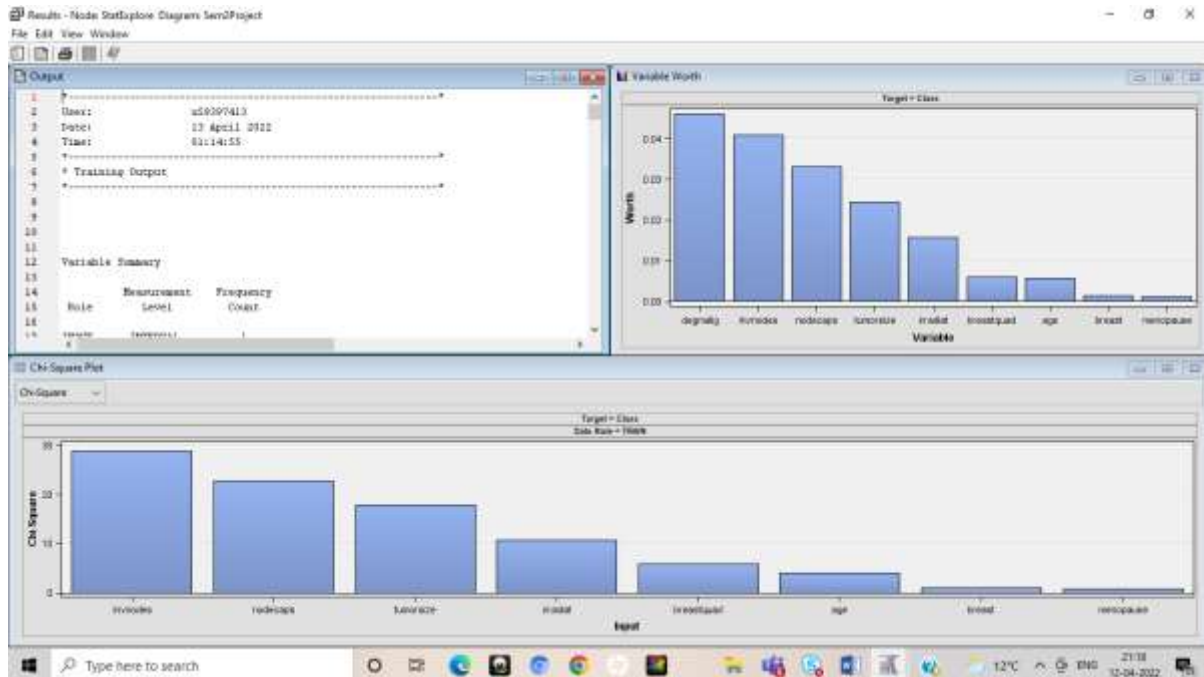
Using the “Variable” under Properties, assign Role as “Target” to the Class variable. The remaining features are identified as “Input” variables as they are the independent variables.



## Step 2: Stat Explore

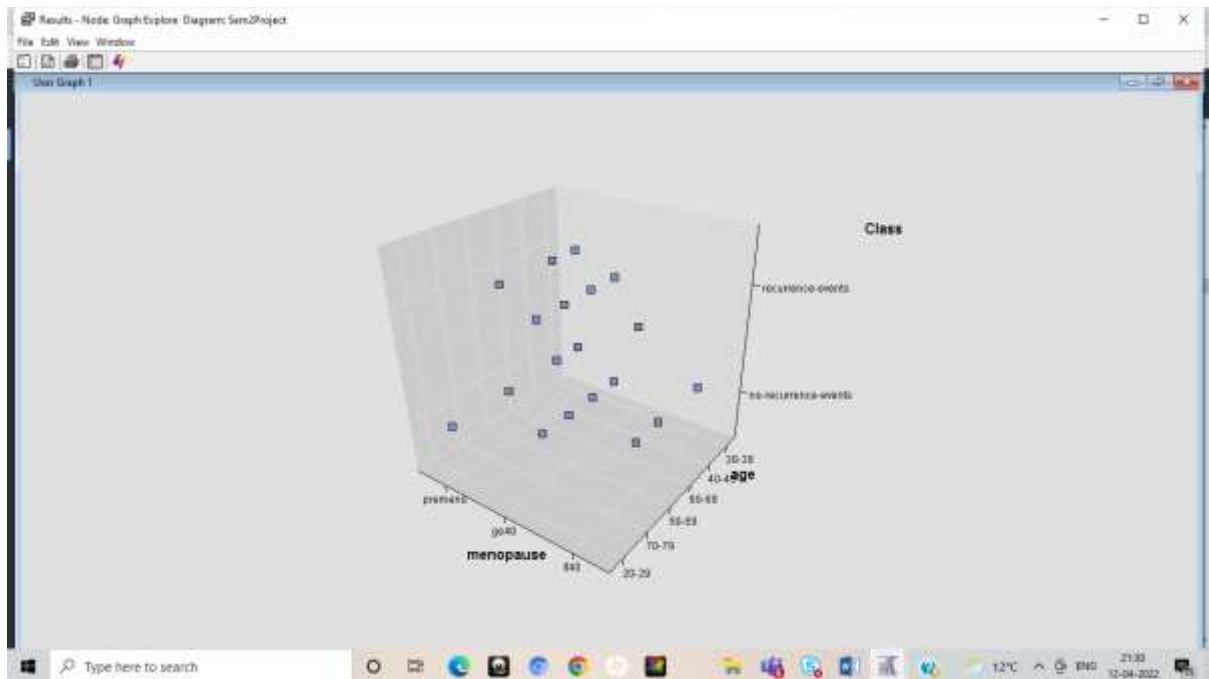
After the File Import, we add the Stat Explore node to study the class variable. Below is the screenshot of the result obtained from the Stat Explore node. The output has been shared on the GitHub link provided below.

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.



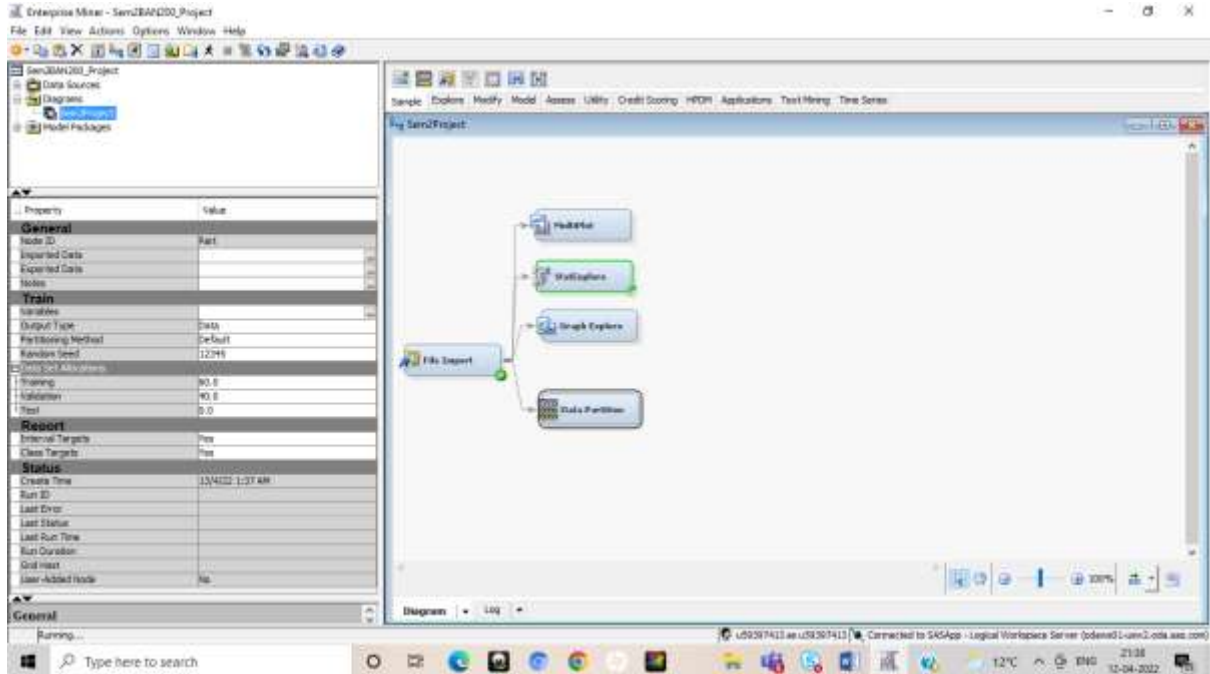
### Step 4: Multiplot



## Step 5: Data Partition

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

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



The screenshot displays the 'Results - Node Data Partition: Diagram: Sem2Project' window. The 'Output' pane shows the results of the data partitioning process. The results are organized into three sections: 'Summary Statistics for Class Targets', 'Data=DATA', and 'Data=TRAIN'. Each section contains a table with columns for 'Variable', 'Value', 'Formatted Value', 'Frequency Count', 'Percent', and 'Label'.

Variable	Value	Formatted Value	Frequency Count	Percent	Label
Class	.	no-outcome-events	201	70.2997	Class
Class	.	outcome-events	85	29.7003	Class

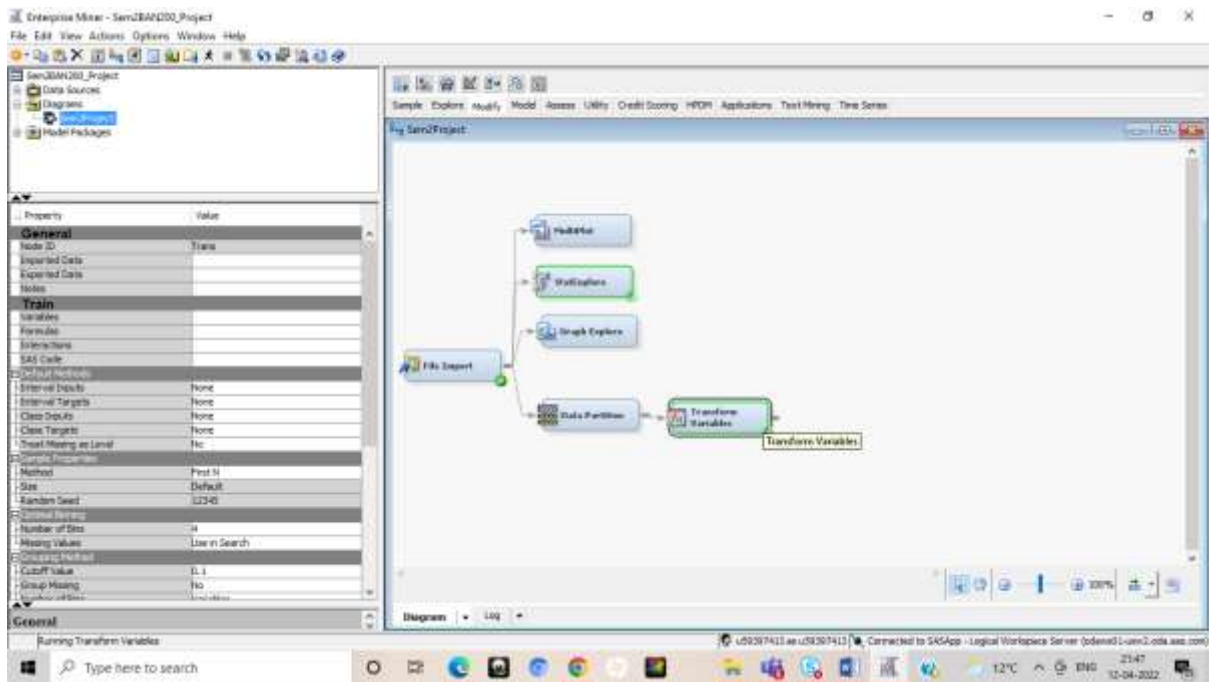
Variable	Value	Formatted Value	Frequency Count	Percent	Label
Class	.	no-outcome-events	120	70.5882	Class
Class	.	outcome-events	50	29.4118	Class

Variable	Value	Formatted Value	Frequency Count	Percent	Label
Class	.	no-outcome-events	61	69.8276	Class
Class	.	outcome-events	26	30.1724	Class

## Step 6: Data Transformation

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



Results - Node StatExplore Diagram SendProject

Output

Input	Chi-Square	Df	Prob
innodes	20.7997	4	<.0001
nodegroup	23.8917	2	<.0001
nodevalue	17.9157	10	0.0564
innodeid	10.7142	1	0.0010
innodepath	5.0711	3	0.1900
out	3.9977	3	0.0497
innode	0.8836	1	0.3413
innodepath	0.7923	2	0.6732

Chi-Square Statistics  
(overall: 500 observations printed)

Data Role=TRAIN Target=Class

Model Output

Report Output

## 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:

The screenshot displays the SAS Miner interface with the following outputs:

**Model Performance Metrics:**

Model	Recall	Precision	F1 Score	Accuracy
Model 1	0.40	0.6086	0.4827	0.48688

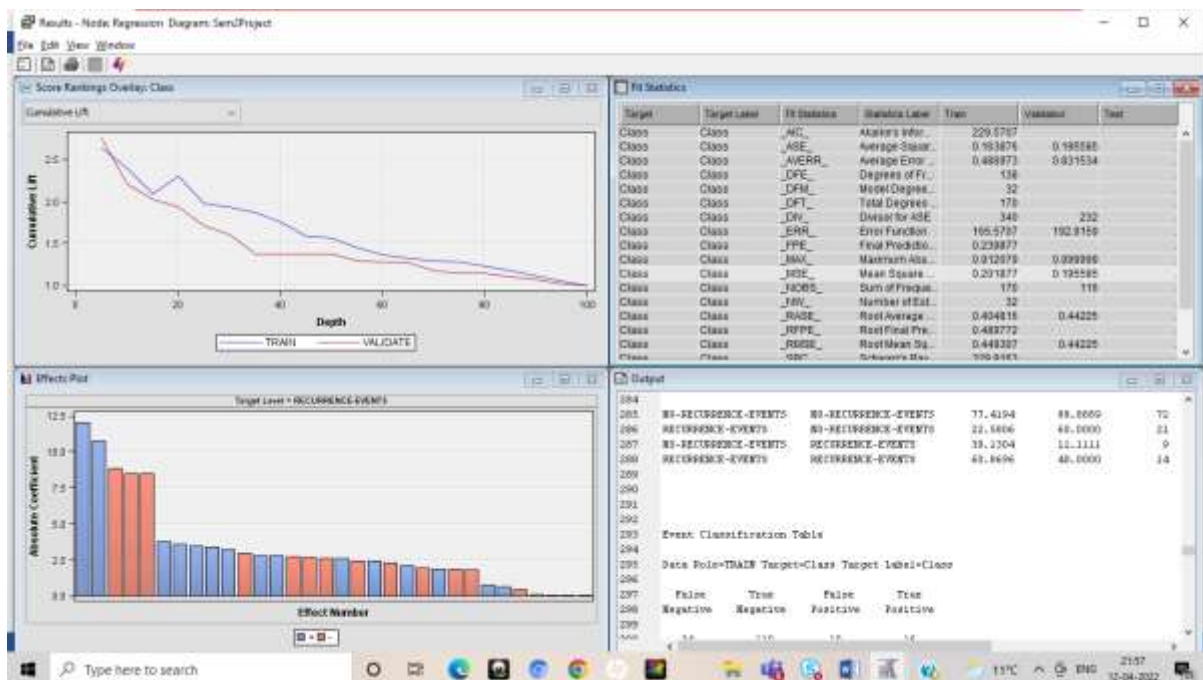
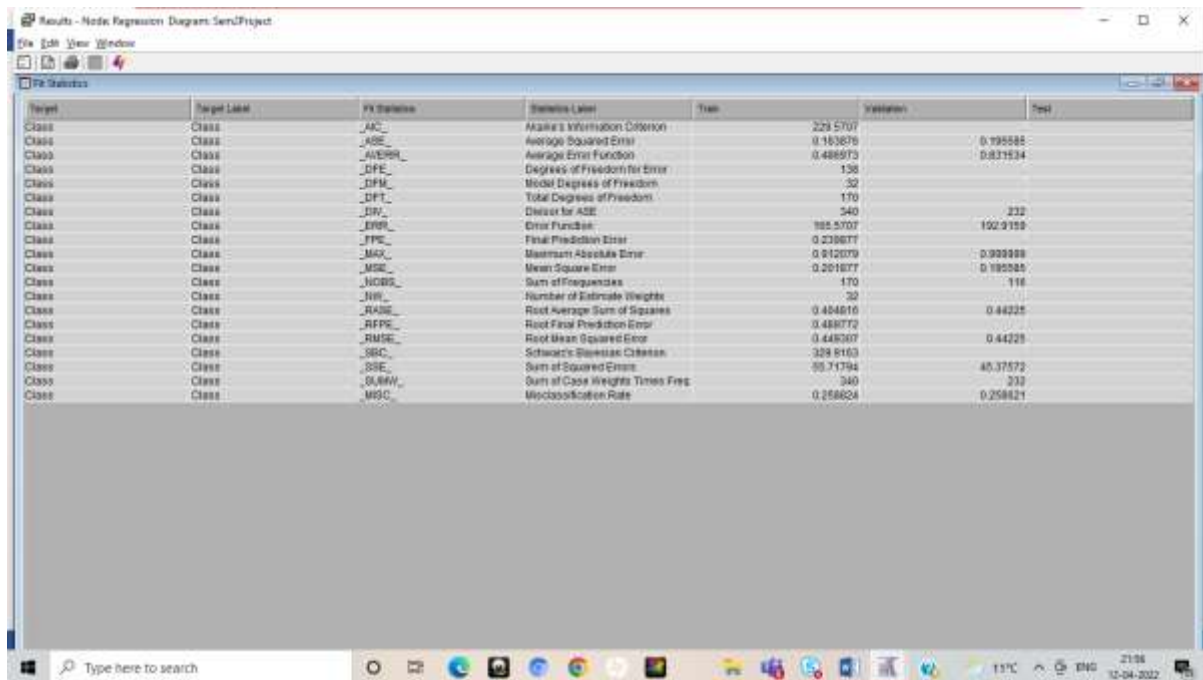
**Event Classification Table (Validation):**

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

**Assessment Score Rankings:**

Depth	Gain	Loss	Cumulative Loss	Response	Cumulative Response	Number of Observations	Mean Posterior Probability
1	0.0000	0.0000	0.0000	0.0000	0.0000	116	0.0000





## 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

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) = 14/23 = 0.6086$$

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) = 14/20 = 0.7$$

$$\text{F1} = 2 \cdot \text{P} \cdot \text{R}/(\text{P}+\text{R}) = 0.8520/1.3086 = 0.6510$$

Below are the screenshots of the SAS Miner outputs:

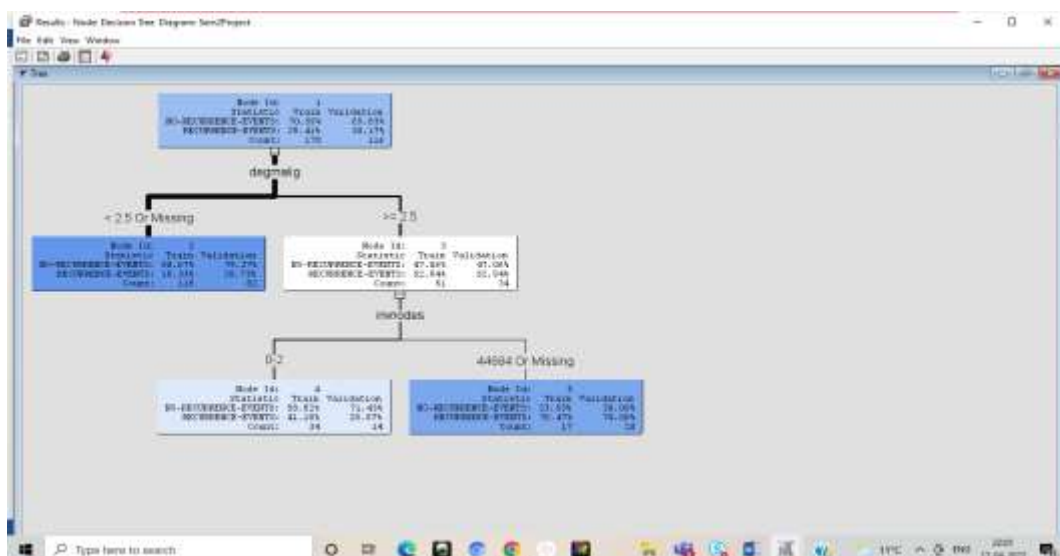
The screenshot shows the SAS Miner Output window with the following content:

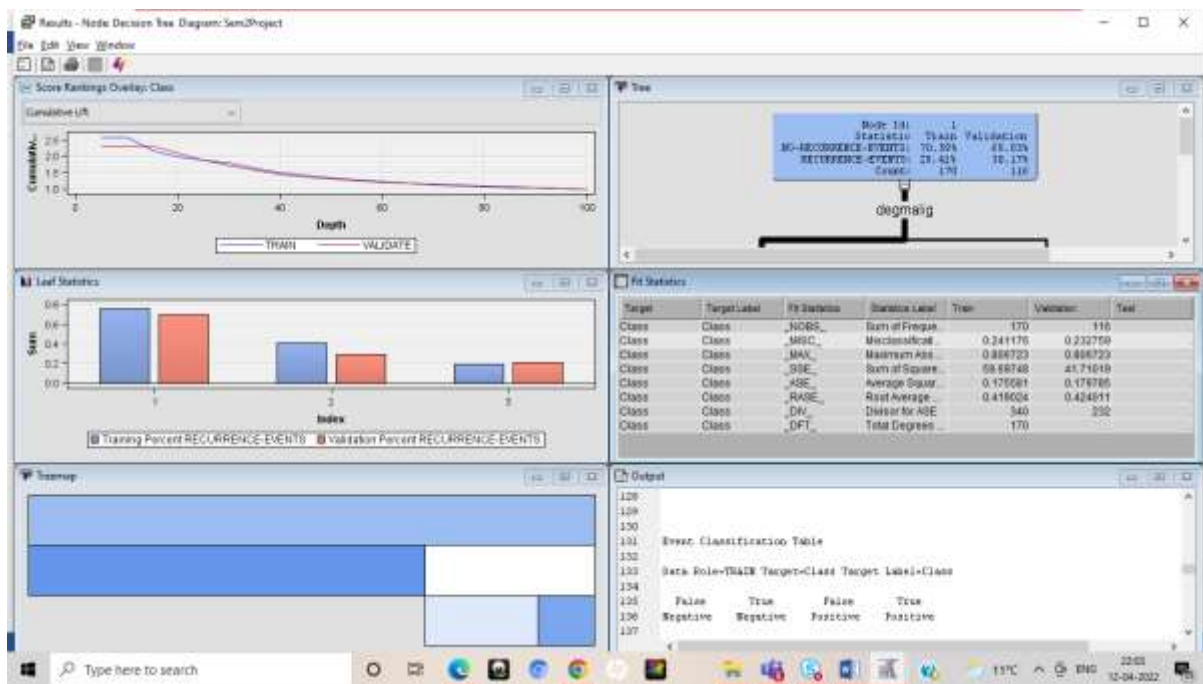
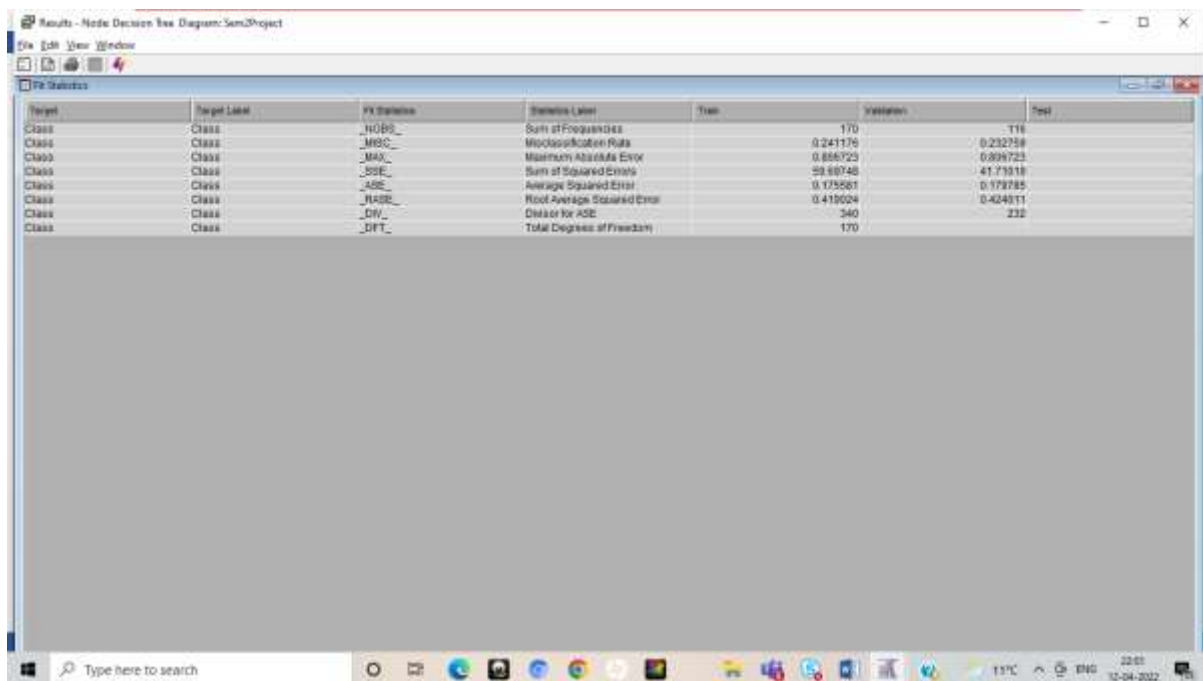
**Event Classification Table:**

Target	True Negative	False Positive	True Positive	False Negative
0	75	6	14	23

**Lift Chart:**

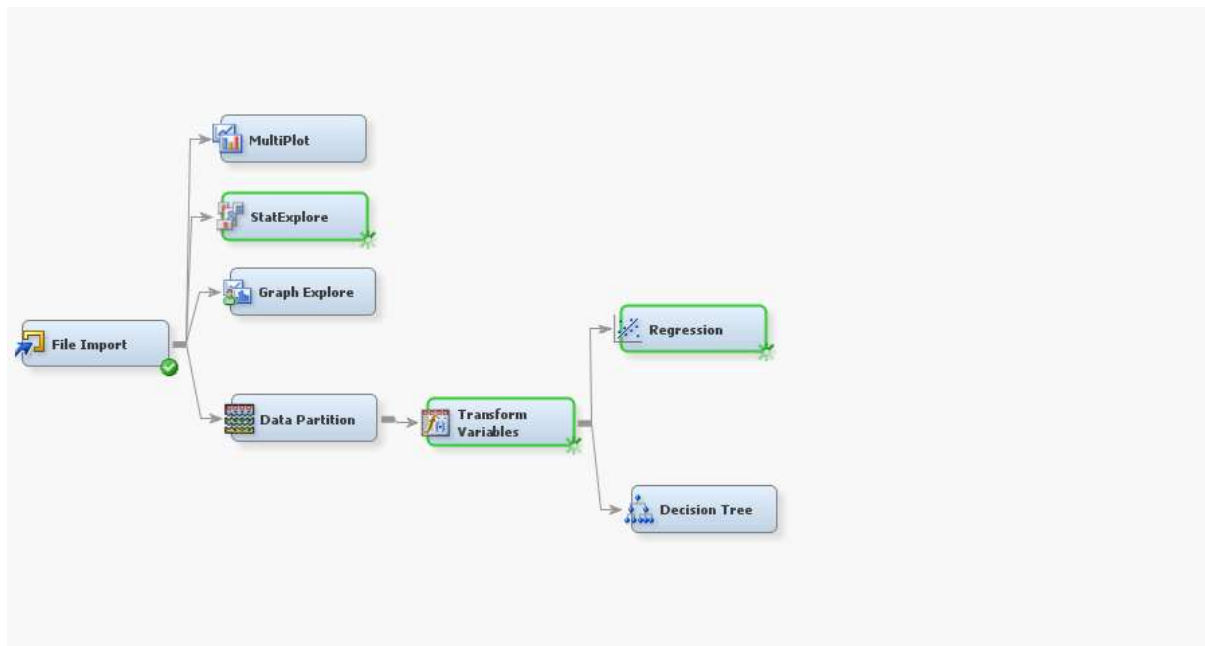
Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Event Probability
1	148.083	2.0000	2.0000	76.4706	76.4706	81	0.7647
2	128.083	1.0000	2.0000	41.1765	76.4706	37	0.4118
3	128.083	1.0000	2.0000	41.1765	76.4706	37	0.4118
4	128.083	1.0000	2.0000	41.1765	76.4706	37	0.4118





## 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](https://github.com/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](https://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.