BAN 210 Final Project

Breast Cancer Data Exploratory Analysis

SUBMITTED BY:

MITALI MAHESHWARI (157554213)

SUBMITTED BY:

DR. ADEEL JAVED

Introduction

I have used Breast Cancer Data Set for the exploratory data analysis and predictive modeling. To determine whether the class of the data point represents a recurrence event or a non-recurrence of the Target variable. I have compared decision tree node and logistic regression node to identify best performing regression model.

Data Preparation

We must first prepare the data. In this process, some variables are eliminated, missing values are imputed, remodelled, and transformed. It is also important to investigate if there is an inaccurate correlation between the input and the target variables.

Dataset

There are 85 occurrences of one class and 201 instances of another in this data set. There are nine qualities – some linear and some nominal – are used to characterize the instances.

Deriving Target Variable

I have selected Class as the target variable because the objective is to identify weather or not the response variable is a recurrence event or non-recurrence event. There are total 10 attributes most of them being nominal. All other variables have input roles as they are independent variables.


Variables - FIMPORT

| Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Nominal	No		No	.	.
Class	Target	Nominal	No		No	.	.
breast	Input	Nominal	No		No	.	.
breast_quad	Input	Nominal	No		No	.	.
deg_maliq	Input	Nominal	No		No	.	.
inv_nodes	Input	Nominal	No		No	.	.
irradiat	Input	Nominal	No		No	.	.
menopause	Input	Nominal	No		No	.	.
node_caps	Input	Nominal	No		No	.	.
tumor_size	Input	Nominal	No		No	.	.

The results of the **StatExplore** node are shown below with the statistics of Class variable. There is total **85 recurrence** events. Deg_malg and inv_nodes have the highest contribution in decision making whereas the menopause and breast are the least.
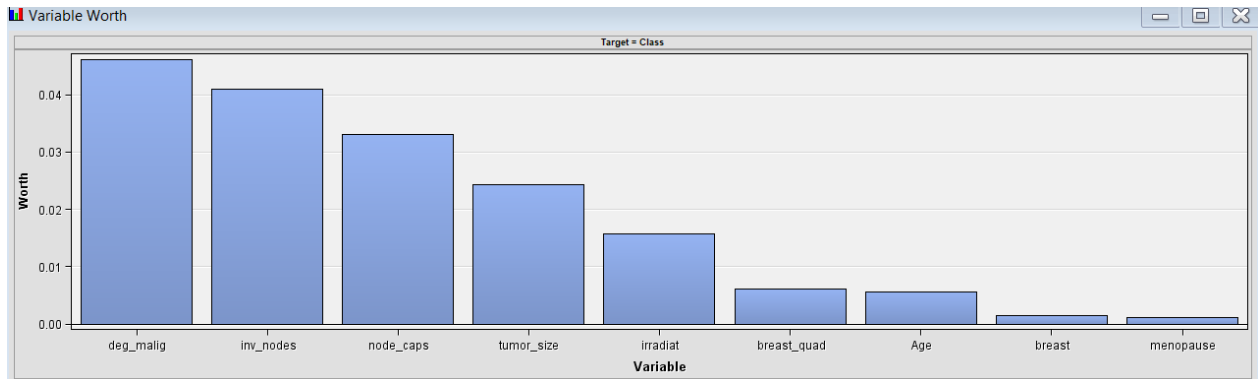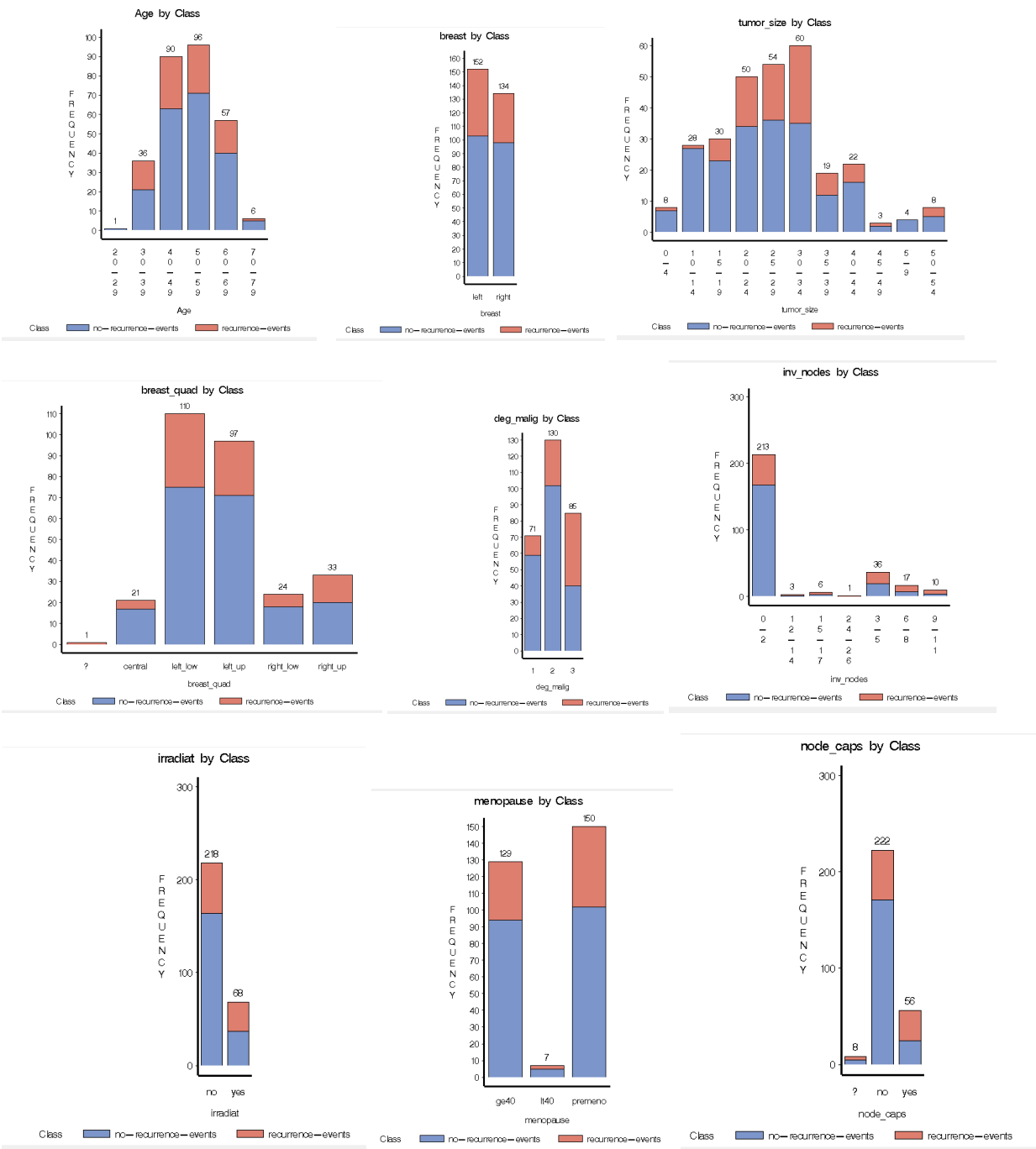

In the **Cluster** node, we can see that the variables are well distributed.
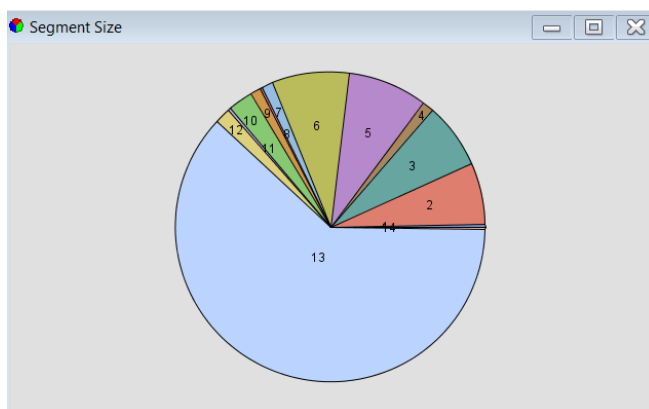
The results of the **GraphExplore** node are as follows. We can observe that the recurrence events are less than the no-recurrence events. The 3D scatter plot of Class, Breast and Age variable is also shown below.
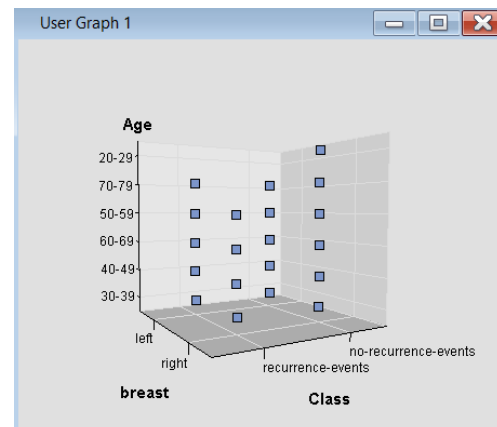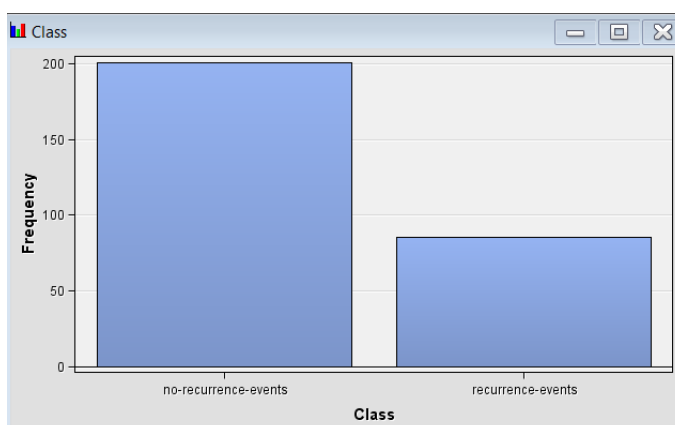
The **MultiPlot** node graphical inferences are shown as below.

Data Transformation and Partition

In the **Data Partition** node, discriminant function approach was used in the as the imputation technique for input variables, while regression method was used to impute the other variable. The data was divided into 60% for training, 30% for validation, and 10% for testing using the Partition node.

.. Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
⊟ Data Set Allocations	
Training	60.0
Validation	30.0
Test	10.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	04/08/22 7:39 PM
Run ID	a663b1df-7905-4f27-a24
Last Error	
Last Status	Complete
Last Run Time	07/08/22 6:36 PM
Run Duration	0 Hr. 0 Min. 2.93 Sec.
Grid Host	
User-Added Node	No

```
Partition Summary

                                          Number of
Type              Data Set               Observations

DATA       EMWS2.FIMPORT_train              286
TRAIN      EMWS2.Part_TRAIN                 170
VALIDATE   EMWS2.Part_VALIDATE               86
TEST       EMWS2.Part_TEST                   30
```

Model Building

The model was developed using SAS® Enterprise Miner™ 14.1. The process flow diagram is shown as below. In the model, the **File Import node** was added to import the Breast Cancer Dataset. Then the **StatExplore**, **GraphExplore** and **MultiPlot** node were added to understand the statistics of the dataset. Then the **Data Partition** node was connected to **File Import** node to split the original dataset into training and validation dataset. Thereafter, **Transform Variables** node was connected to the Data Partition to transform variables. This was followed by regression node. Logistic Regression is used as we are doing prediction for classification variable. The second model comprises of Decision Tree. The results for both models were then evaluated by connecting them to the **Model Comparison** node.
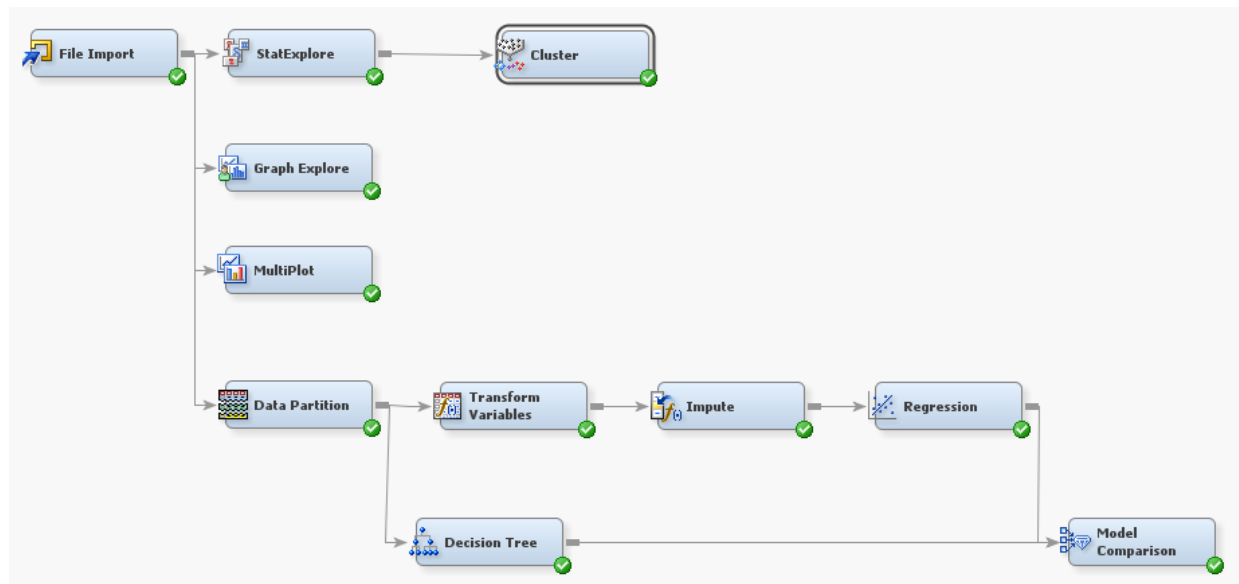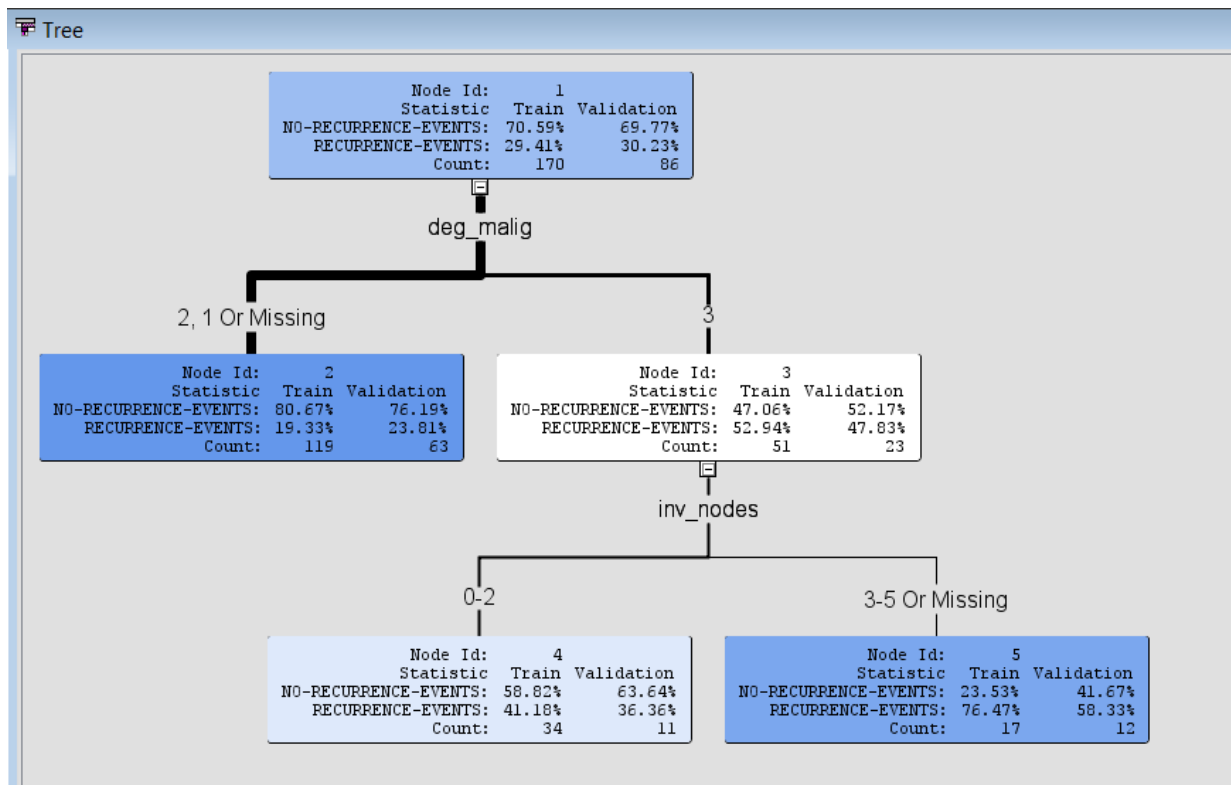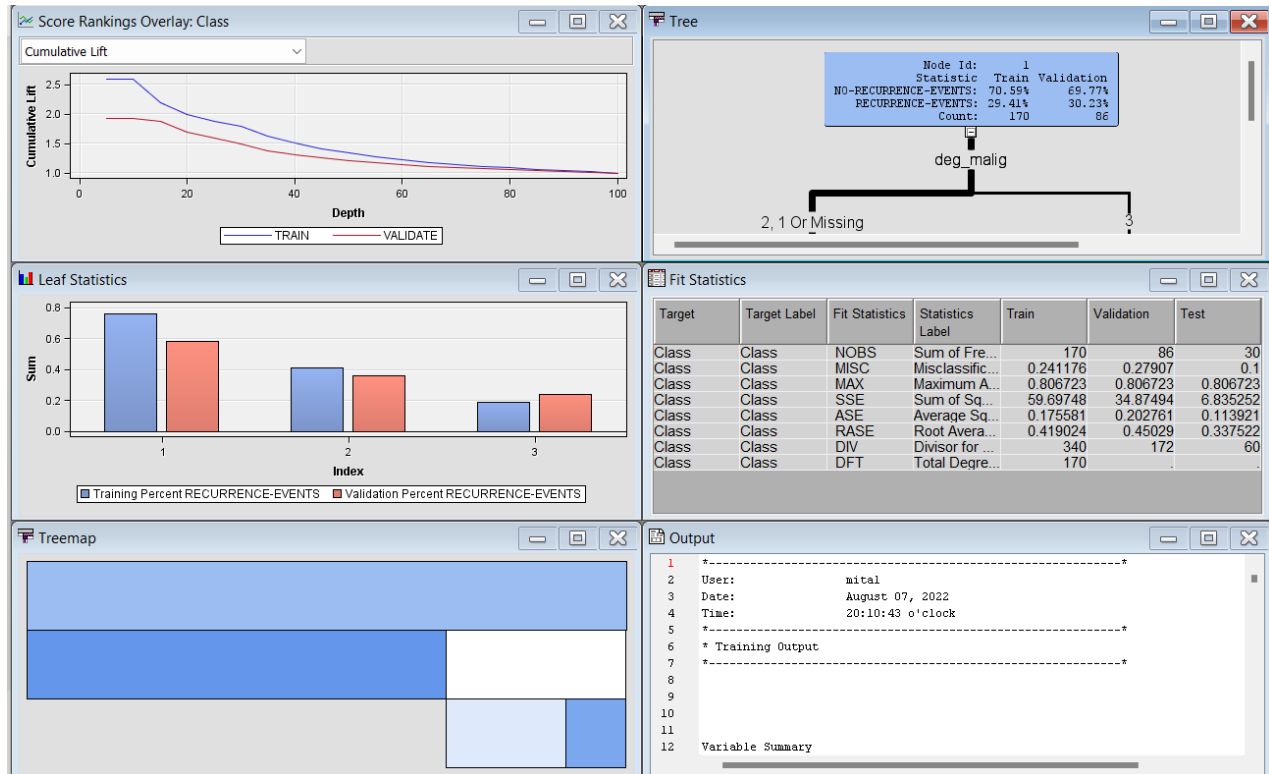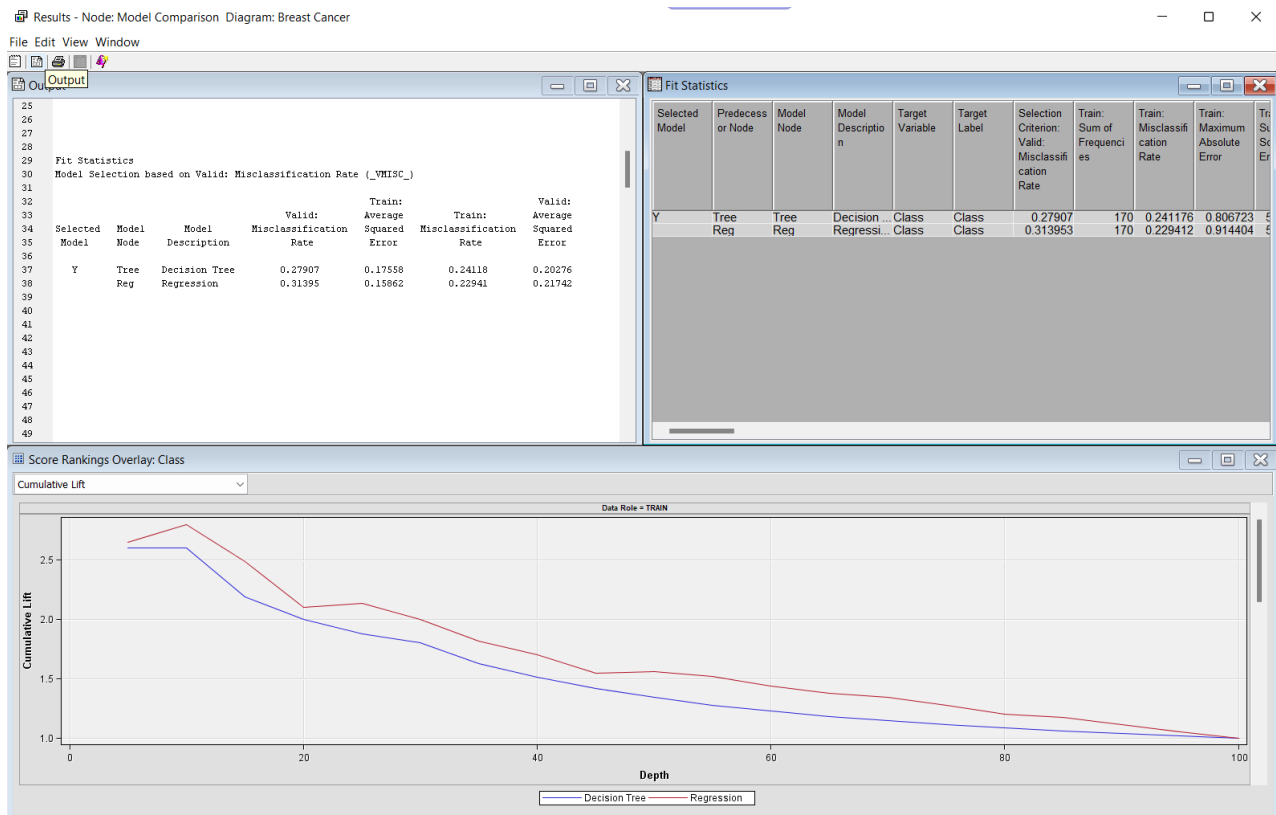
Results

Regression Output:
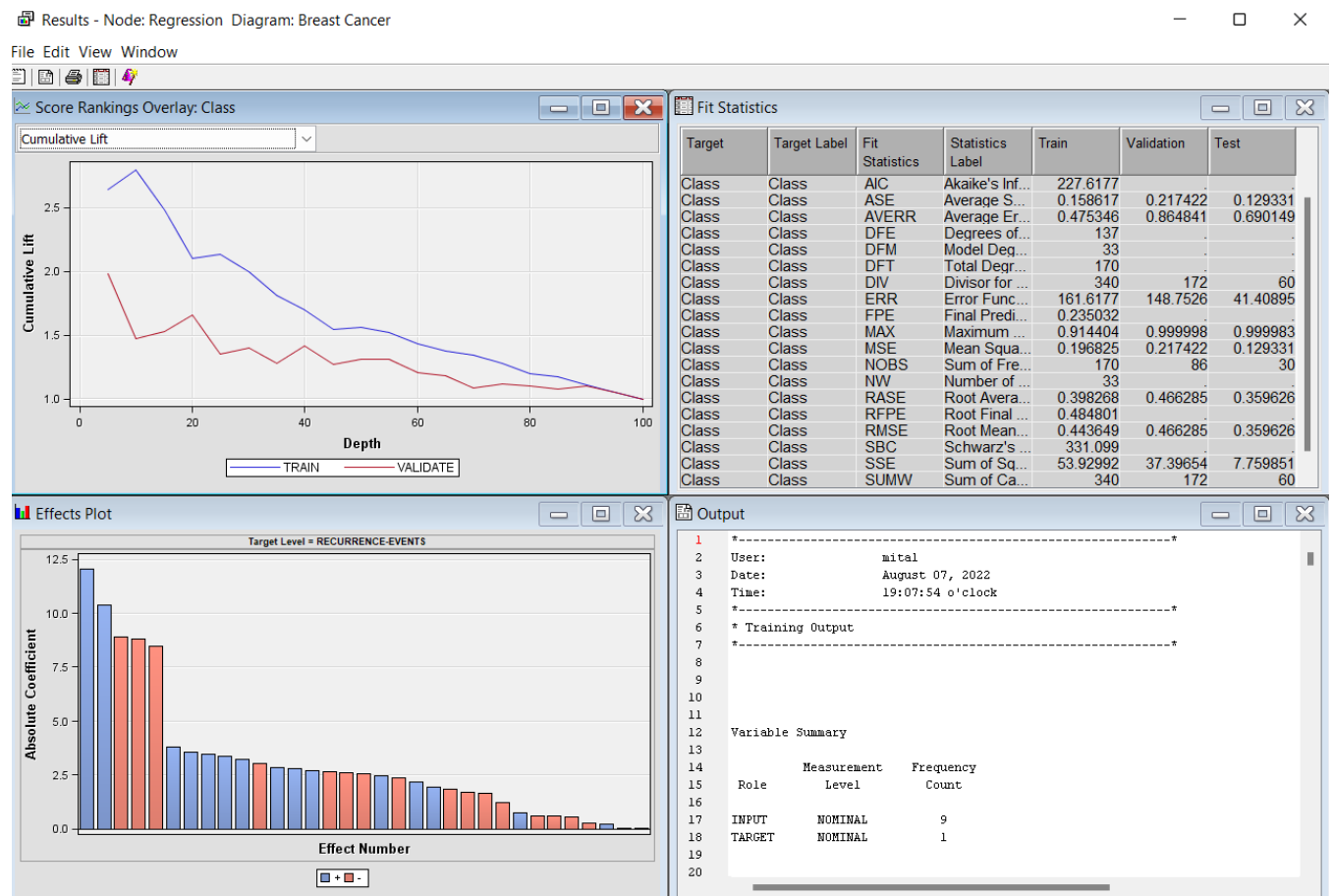
Décision Tree Output

Model Comparison:

The output of the model comparison node is as follows:

Conclusion:

As we can see, the decision tree node has been selected for the fact that the decision parameters are slightly better than the regression model. A model is better when the average squared error is low. Variables Breast and Menopause are of no significance. Whereas the seriousness of the cancer can be determined with the variables – deg-malig and inv-nodes.

GITHUB LINK:

https://github.com/Mitali027/Predicitve-Analytics-Final-Assessment

Declaration:

I, **Mitali Maheshwari**, 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.