BAN 210 – Final Assessment

Breast Cancer Dataset

Submitted By: Amish Gupta

Student ID: 157547217

Predictive Analytics(BAN210ZAA)

**Introduction**

In this assessment I am going to implement predictive modelling and I am going to predict this by using class as a target variable. I am going to perform this analysis on the dataset that has been provided and it consists of 286 observations and 10 different parameters which can be considered significant in carrying out our research.

**Exploratory Data Analytics**

So, now we are going to perform EDA which basically means exploratory data analytics which is a approach which is used to summarise all the major factors which can be proved significant in carrying out of the research. It basically involves graphical representations to summarise the data so that useful information can be extracted from them.

**Step 1**: In order to start the first step to explore the data that is provided we need to import the data into the SAS EM. This can be done by clicking on the file import node and then selecting the provided dataset. After adding the dataset right click on the file import node then we need to click on Edit Variables and then we are assigning the target to the class.

Graphical user interface, text

Description automatically generated

**Step 2: Stat Explore**

Stat explore node is a node which is used to discover statistics and the distributions in our data. It can select the most important variables automatically which are significant in carrying out the research.

Now I have added stat explore node just after the file import node to get an understanding about the target variables.

Graphical user interface, application

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

**Step 3: Graph Explore**

After the stat explore graph explore is attached to the file import which can help us to understand trends in the datasets like in our case it’s telling us about the frequency of the no recurrence events and recurrence events

The scatter plot which is shown is 3D representation of the variables where X is class Y is age and Z is menopause

Graphical user interface, application, table

Description automatically generated

Chart, scatter chart

Description automatically generated

**Step 4: Data Partition**

Data Partition is done on the data in order to split the data in training and validation data which I have taken as 70% and 30% respectively. Data partition is required as it helps to deal with the problem of overfitting.

A screenshot of a computer

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

**Step 5: Impute**

Impute node is used for removing the missing values that are present in the dataset.

Graphical user interface, application, Word

Description automatically generated

**Step 6: Data Transformation**

It is used for transforming the variables

Graphical user interface, application, Word

Description automatically generated

**Predictive Models**

For making the predictive models I choose to make decision tree and neural network. I connected the decision tree and the neural network node to the transform variables node. So, after running the decision tree node we are getting the following output.

Decision Tree

Decision trees are basically supervised learning models which use classification and regression. It is basically used to predict the value of the target variable by learning about the data.

Graphical user interface, application

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Dataset | Misclassification Rate | Average Squared Error |
| Training | 0.241206 | 0.169919 |
| Validation | 0.241379 | 0.201799 |

Classification table for Validation Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Target Variable | 0 | 1 | Total |
| 0 | Tn=55 | Fp=6 | 61 |
| 1 | Fn=15 | Tp=11 | 26 |
| Total | 70 | 17 | 87 |

Where 0 means No-Recurrence-Events

and 1 means Recurrence Events

Recall = Tp/(Tp+Fn) = 11/(11+15)=11/26=0.423

Precision = Tp/(Tp+Fp) =11/17=0.647

F1= 2 \* (Precision \* Recall) / (Precision + Recall) = 0.547/1.07 = 0.511

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Neural Network

Neural network usually works as human brain. There are different types of neural networks. The major components of neural network are input layer, processing layer and the output layer.

Graphical user interface

Description automatically generated

**Neural Network**

|  |  |  |
| --- | --- | --- |
| Dataset | Misclassification Rate | Mean Squared Error |
| Training | 0.145729 | 0.241133 |
| Validation | 0.218391 | 0.187572 |

Classification table for Validation Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Target Variable | 0 | 1 | Total |
| 0 | Tn=57 | Fp=4 | 61 |
| 1 | Fn=15 | Tp=11 | 26 |
| Total | 72 | 15 | 87 |

Where 0 means No-Recurrence-Events

and 1 means Recurrence Events

Recall = Tp/(Tp+Fn) = 11/(11+15)=11/26=0.423

Precision = Tp/(Tp+Fp) =11/15=0.73

F1= 2 \* (Precision \* Recall) / (Precision + Recall) = 0.61/1.153 = 0.535

Graphical user interface, application, table

Description automatically generated

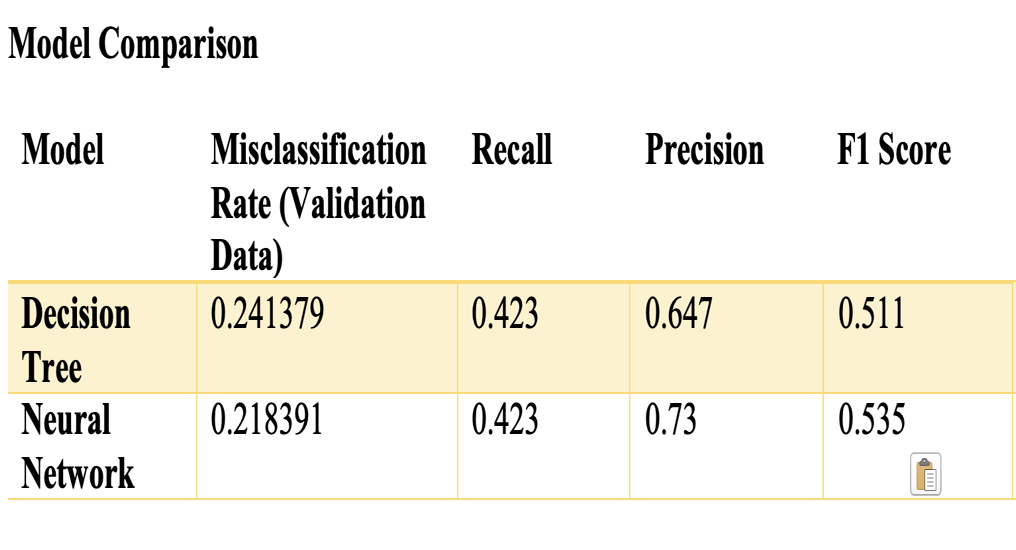
Graphical user interface, text, application, Word

Description automatically generated

**Model Comparison**

On comparing both the models it is evident that the neural network is a better model than the decision tree as the F1 Score of the Neural network is higher in comparison to the decision tree.

Moreover, the results that were obtained from the Model comparison node is as shown, here also neural network is selected as a model with better accuracy

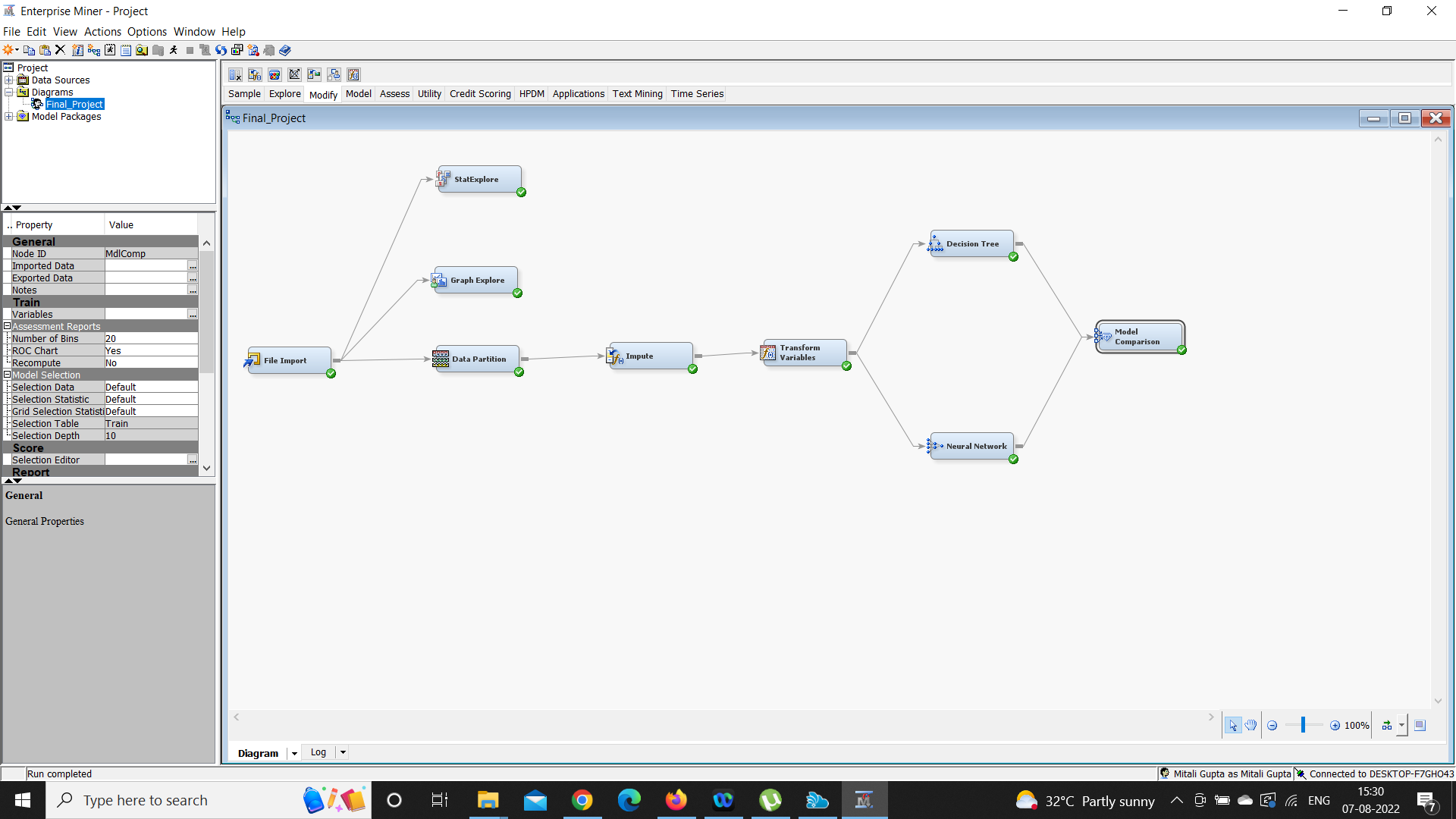
****

Graphical user interface, application

Description automatically generatedGraphical user interface, application, Word

Description automatically generated

**Final Diagram**



**Conclusion**

Neural network has a better output in score ranking performance in which the validation data set is sorted and arranged in the format of descending order. Due to this reason also, we can say that the Neural network can be considered as a better model in comparison to the decision tree

I, Amish Gupta , 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.