**BAN 210 – FINAL ASSESSMENT**

**TOPIC: BREAST CANCER DATASET**

**Submitted by: Nishalben krutay joshi**

**Student id: 158865212**

**Subject: Predictive Analytics (BAN210ZAA)**

**INTRODUCTION:**

For the analysis of breast cancer data, I have used predictive modeling. It is done by using the target variable as a class. I have used Decision Tree and Neural network. The analysis determines which model is the best to use for the prediction.

**GOAL:**

This analysis of the breast cancer data set aims to determine which characteristics are most useful in the prediction. It also aims to find which model is performing better with good accuracy.

**Exploratory Data Analytics**

EDA is exploratory data analytics. This method is used to summarize all the key elements that may be demonstrated to be relevant in the execution of the research. In order to summarize the data and extract relevant information from it, graphical representations are primarily used.

**Step 1:**

We must import the data into the SAS EM before we can begin the first stage, which involves exploring the given data. You may accomplish this by selecting the specified dataset after clicking the file import node. Right-click on the file import node after importing the dataset, then select Edit Variables before assigning the target to the class.

Graphical user interface

Description automatically generated

**Step 2: Stat Explore**

Identifying statistics and distributions in our data is done using the stat explore node. It may autonomously choose the factors that are most crucial for doing the investigation. To better comprehend the target variables, I have now added a stat explore node right after the file import node.

Graphical user interface

Description automatically generated with medium confidence

Table

Description automatically generated

**Step 3: Graph Explore**

In our scenario, the stat explore graph explore is telling us about the frequency of the no recurrence events and recurrence events once the stat explore graph is linked to the file import. This can help us analyze trends in the datasets.

Graphical user interface

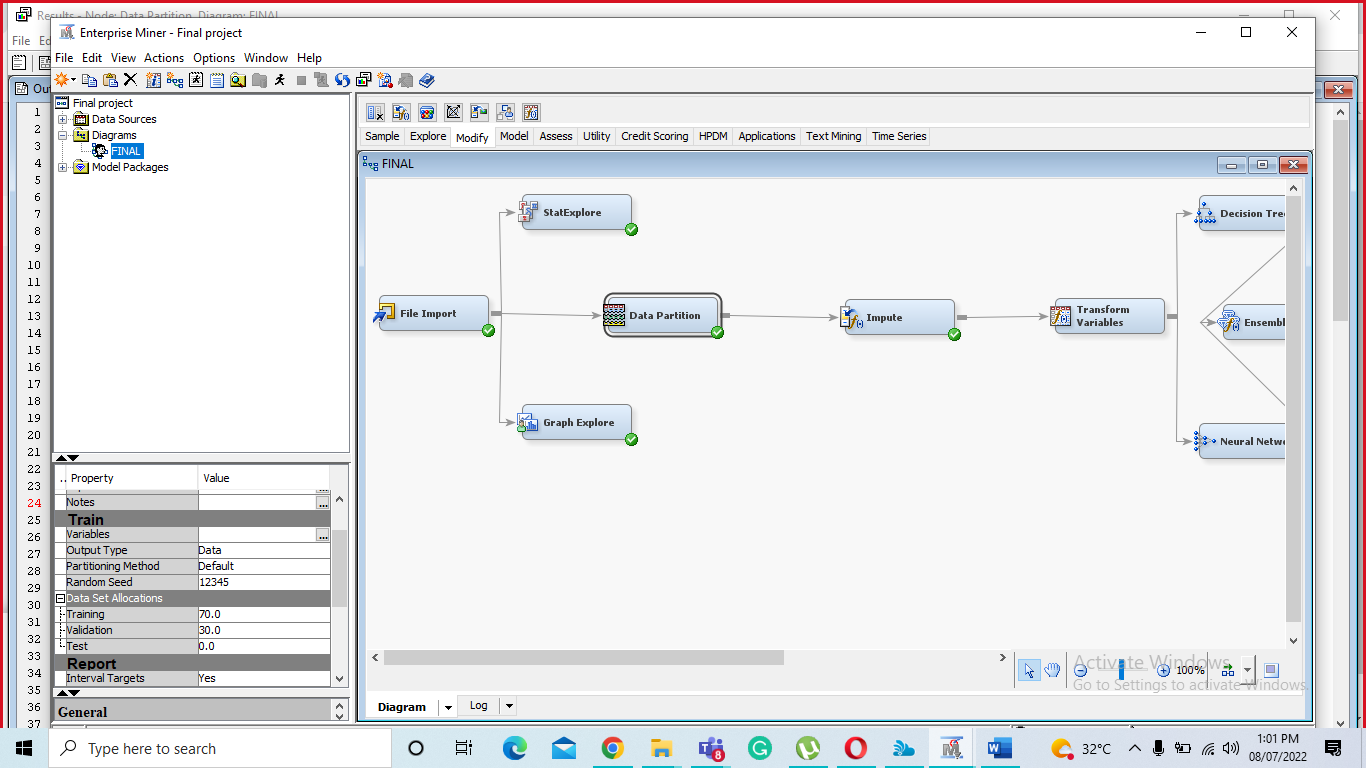
Description automatically generated

**Chart, scatter chart

Description automatically generated**

**Step 4: Data Partition:**

The data is divided into training and validation data, which I have determined to be 70% and 30% of the total data, respectively. Data division is necessary since it aids in addressing the overfitting issue.



Graphical user interface, text, application

Description automatically generated

**Step 5: Impute**

The dataset's missing values may be eliminated using the impute node.

Graphical user interface, application, table, Word

Description automatically generated

**Step 6: Data Transformation:**

The variables for the model have been transformed using the data transformation node.

Graphical user interface, application, table, Word

Description automatically generated

**Predictive Models:**

I decided to create a decision tree and neural network for the prediction models. I linked the transform variables node to the neural network and decision tree nodes. So, following the decision tree node's execution, the result is as follows.

Step 7 : Decision Tree: The decision tree is connected to the data transformation node. Following is the output. In the decision tree, I have checked out the Misclassification rate, recall, precision, and F1 score, and based on the f1 score we can predict a decision about the model accuracy.

Graphical user interface, application

Description automatically generated

Diagram

Description automatically generated with medium confidence

Table

Description automatically generated

A picture containing text

Description automatically generated

**Decision tree:**

|  |  |  |
| --- | --- | --- |
| 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

Step 8: Neural Network:

To compare, I connected the Neural network to run the second model. Below are the output:

Graphical user interface

Description automatically generated

Table

Description automatically generated

A picture containing chart

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

**Step 9: Model Comparison:**

Below is the comparison model of the Decision tree and Neural network, It compares both the model and help us finding best model with good accuracy.

Graphical user interface

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**Step 10: Final Diagram**

Graphical user interface, application, Word

Description automatically generated

**Conclusion:**

When the validating set of data is sorted and structured in descending order, neural networks perform more effectively in terms of score ranking results. Additionally, for this reason, we can assert that the neural network is a superior model to the decision tree.

**Declaration:**

I Nishalben krutay joshi, 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 websites, unless specified as references. I have not distributed my work to other students.