

final Assessment

Ban210BAN210NAA Predictive Analytics



Submitted to: Dr. Gulsun Demirezen

By: SAi MOhan tirunagari

ID: 147150213 T

**Breast cancer Data Set**

**Introduction**

Breast cancer is defined as uncontrolled growth of breast cells. In human body when cancer cells are formed when the gene undergoes abnormal mutations. Normal cells in human body dies everyday and are replaced by new ones simultaneously as part of the immune system. When the cells which underwent mutation grow in normal cell loose the ability to replace with healthy tells. These mutated cells grow in number over time forming tumors. A tumor can be benign or malignant. Malignant tumors pose potential treat to the body and cause cancer. When these mutated cells accumulate in the breast leads to breast cancer. “Over time, cancer cells can invade nearby healthy breast tissue and make their way into the underarm lymph nodes, small organs that filter out foreign substances in the body. If cancer cells get into the lymph nodes, they then have a pathway into other parts of the body. As per the statistics 1 out of 8 women are prone to get breast cancer. With advancement in technology, in medical field, machine learning and AI ar being implemented to predict the potential of getting a breast cancer. Based on medical data, diagnosis of breast cancer in very early stage becomes easier and can be cured. In few instances there are chances of the cancer returning, such recurrences can also be predicted using machine learning models.

This project aims to study the Breast cancer data, which is available in UCI machine learning repository, and build prediction models further comparing which model best fits to predict accurate outcomes. “This data is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary tumor.) This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear, and some are nominal.

**Data dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Measurement** | **Description in SAS Enterprise** |
| Age | Numeric | Ordinal | Target Variable |
| Menopause | Numeric | Nominal | Input |
| Tumor size | Numeric | Ordinal | Input |
| Inv-nodes | Numeric | Nominal | Input |
| Node caps | Numeric | Binary | Input |
| Degree of malignancy | Numeric | Interval | Input |
| Breast | Numeric | Nominal | Input |
| Breast quadrant | Numeric | Nominal | Input |
| Irradiation | Numeric | Binary | Input |
| Class | Char | Nominal | Target Variable |

**Exploratory Data analysis**

Table

Description automatically generated

After Importing the file, the variables and its type can be viewed. The class variable which is recurrence and no-recurrence is the target variable.

**Stat explorer node**

In this node we can visualize the variable and se the relationship between the variables.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

We can see that, the variables deg\_malig, inv\_nodes, node\_capes, tumor\_size have hiher significance compared to other variables.

**Explorer node**

Chart, histogram

Description automatically generated

Visual representation of frequency of recurrence and non-recurrence. This shows that no-recurrence number is higher meaning the caner did not return.

Chart, schematic, radar chart

Description automatically generated

The above chart depicts measure of class variable with age. The results gives the impression that, people aged between 20-29 have minimal chances of the cancer coming back.

Chart, bar chart

Description automatically generated

The above chart is a measure of class against menopause variable. The results depicts the percentages of recurrence based on the menopause stage.

Chart, bar chart

Description automatically generated

The above chart is the measure of radiation therapy and class. The percentage of no-recurrence is high with radiation therapy.

**Data Partition**

The data set is divided as follows:

Table

Description automatically generated

**Replacement Node**

With the help of this node, we can find out the missing values and analyse them.

Table

Description automatically generated

Table

Description automatically generated

Treating the data and declaring the missing values and correcting the errors in data.

Result of replacement node

Table

Description automatically generated

**Model building**

The target variable is a binary variable hence we can build neural network and logistic regression model and compare them to check which fits to give best results.

**Impute node**

The impute node is used to create new variables replacing the missing values. This will be useful to transform the data. Data transforms is done to have normal distributions.

Graphical user interface, application

Description automatically generated

The above shows the imputation results.

**Variable node**

Table

Description automatically generated

The variable node helps us to reduce the number od input variable with less significance based on the R-square value. The lowest R-square value variable will be rejected. Also, the variable deg\_malig has been transformed using log 10 transformation in the **transformation node.**

**Model comparison**

Now the models’ neural network and logistic regression model.

Graphical user interface

Description automatically generated with low confidence

From the above snap it is observed that the misclassification of both the models are same and average square error of neural networks is less compared to regression model.

A screenshot of a computer

Description automatically generated with medium confidence

Also form the graphs the lift percentage of neural networks is better compared to that of regression model. Hence, looking at he graphs and comparison results we can say that Neural networks model fits for building a predictive model for breast cancer data set.

Diagram

Description automatically generated with medium confidence

Image of the nodes created to build the model and compare.

**Declaration**

I, Sai Mohan Tirunagari 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.