

|  |  |
| --- | --- |
| **Data Science Applications (ICT583)**  Technical Report | A Technical Report On  Mammographic Dataset  Rajesh Jyothi  Murdoch ID: 33669079 |

# Overview of the Project

This project is carried out on Mammographic dataset, taken from UCI repository. It is a medical data set dealing with the Breast Cancer diagnosis. In this report, we discuss the Machine Learning algorithms used for Breast Cancer Diagnosis, and performance of the model is evaluated.

The structure of the report is as follows:

1. **Understanding the Data set**
   1. Problem statement
   2. Questions to be analysed during the exploratory analysis.
   3. Finding the missing values in the data set.
2. **Exploratory Data Analysis**
3. **Predictive Modelling**
   1. Model 1: Decision Tree
   2. Model 2: Random Forest
   3. Model 3: Neural Networks
4. **Interpretation of the Results:**
5. **Conclusion**
6. **References**

# Dataset:

The widely used Breast Cancer screening method is Mammography. It is a manual process of diagnosing the Breast Cancer, based on its size, shape and mass. In the Breast screening process, an X-ray of the low dose is passed through the breast. It helps in diagnosis the cancer cells before it shows any symptoms. This process works effectively for women above 40 years and is less effective for women under 40 years. However, women under 40 years are less likely to get Breast Cancer. If they were still concerned, they can discuss with GP and use Iprevent Breast cancer tools to estimate the risk and get the personalised screening type that is appropriate to the patient (Wang, 2009).

For our analysis, we have collected the data from Mammography MASS data set from the UCI repository. It is collected at Radiology Institute, under University of Erlangen-Nuremberg of Germany, revising 2003 to 2006. This helps in better diagnosis of the Breast Cancer in the early stage. This data set has 961 observation and 6 attributes with a few missing values in it. It is a balanced data set with 516 benign observations and 445 malignant observations. Among the six attributes in the data set, four attributes are predictive, one is a response variable and other is a non-predictive. The response variable is categorical with a binary class, Malignant (M) or Benign (B), and the predictive variables are categorical with more than two classes. On the hand, the non-predictive variable is ordinal.

## Attributes of the dataset:

Table 1: Attributes description

|  |  |  |  |
| --- | --- | --- | --- |
| **Attributes** | **Description** | **Range of values** | **Types of data** |
| 1. BI-RADS assessment | Breast Imaging Reporting and Data System | 1 to 5 | Ordinal |
| 2. Age | Patient's age in years | Min = 18, Max = 96 | Integer |
| 3. Shape | The shape of the tissue | round=1 oval=2 lobular=3 irregular=4 | Nominal |
| 4. Margin | The margin of the tissue | Circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 | Nominal |
| 5. Density | The density of the tissue | high=1 iso=2 low=3 fat-containing=4 | Ordinal |
| 6. Severity | The severity of the tissue | benign=0 or malignant=1 | Categorical |

## Problem statement:

The Mammographic data is balanced data set with 516 benign and 445 malignant observations. It has 5 independent variables and one dependent variable. All the columns have a clear description of the data, hence it can be classified as supervised learning. The output variables or the dependent variable is of binary class, either M or B, Therefore is a classification problem.

By understanding the data set, two questions were raised for analysis.

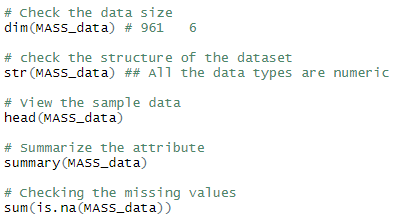
## Analysis Questions:

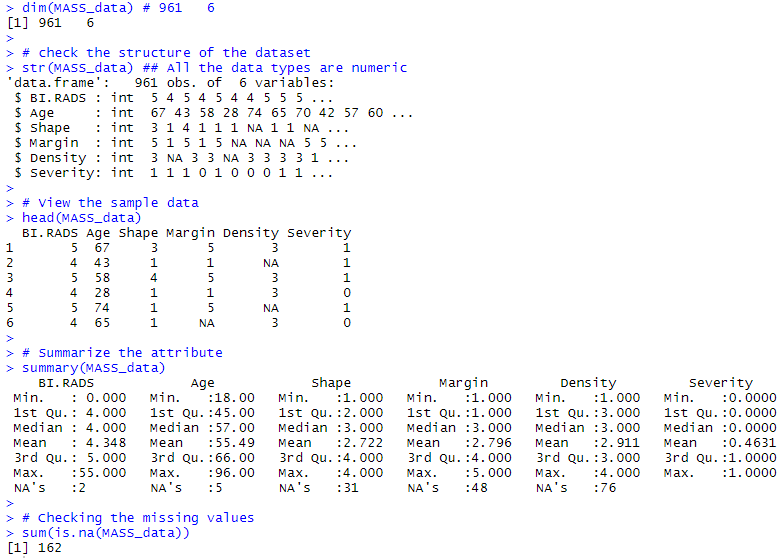
1. ***What shape of the Cancer tumour is most common in malignant class?***
2. ***What is the common density of breast mass in case of malignant patients?***

These questions can be analysed and answered in later stages of exploratory analysis.

Additionally, I would like to predict the diagnosis of the cancer patients using machine learning algorithms such as logistic Regression, Decision Tree, Random Forest, Neural Networks etc., and compare the model accuracy.

The dataset can be better understood by using the following code:



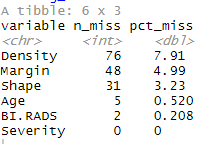


## Finding missing values:

From the above code be found we have some missing values in our data set. Few visualization methods are used to plot the missing values.

### Method 1 – finding missing values using simple colSums ()





### Method 2 – Finding missing values using gg\_miss\_var () method



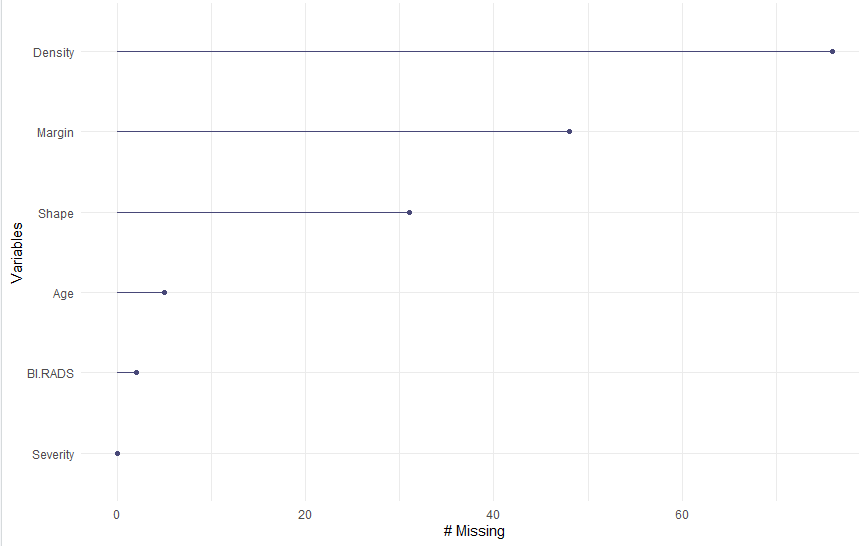
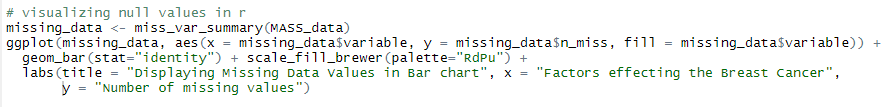


Figure 1: Missing values by columns

### Method 3 – using ggplot2



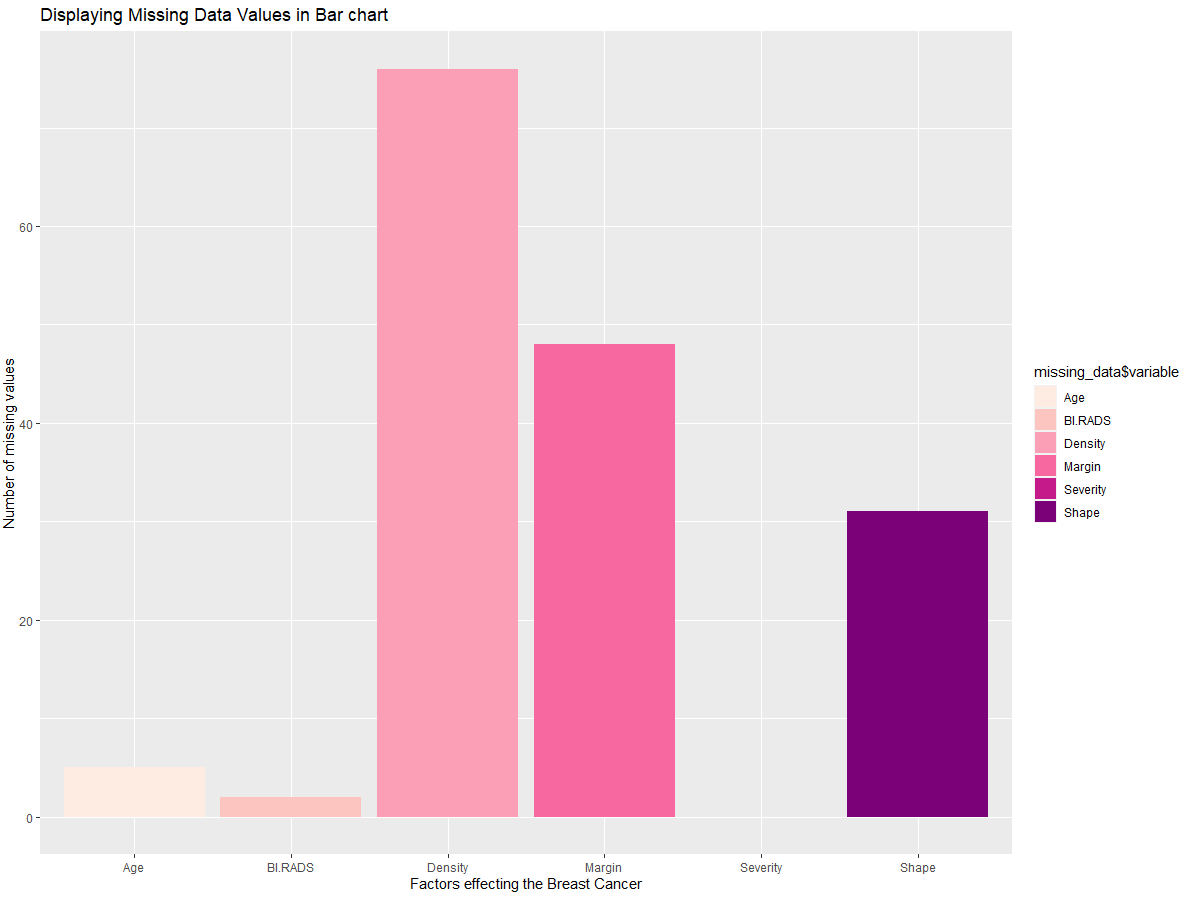


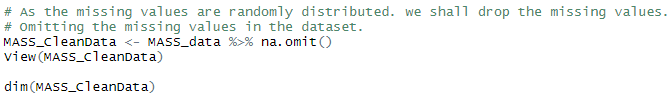
Figure 2: Missing values visualizations

From the above analysis, we found that there are missing values randomly distributed. The visualization plots above show the number of missing values in each attribute.

# Data Manipulation and Cleaning:

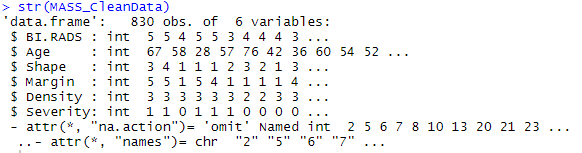
Data manipulation a process in which, the data is transformed to required form that makes easier for operation. As a part of Data manipulation

1. Removing the missing values in the data set. As they are categorical variables, the imputation of the missing values is not the right choice. Moreover, we have an ordinal variable that gives priority to the data imputed. In these cases, omitting the missing values we are the best option.

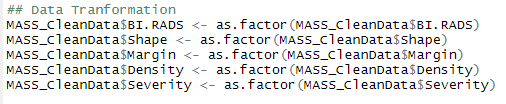


1. Changing the variable type from integer to factor. All the independent variables are of categorical type, if not converted into factor, during any arithmetic operation the value will be summed rather than finding the frequency of occurrence.

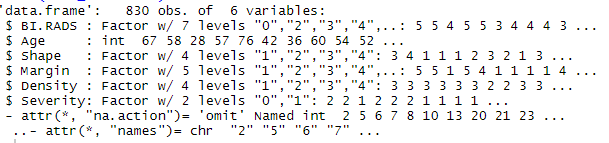




Code to convert the integer data type to factors:



Now all the attributes are converted into Factor variables.



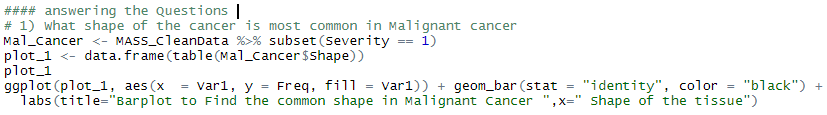
# Exploratory Data Analysis:

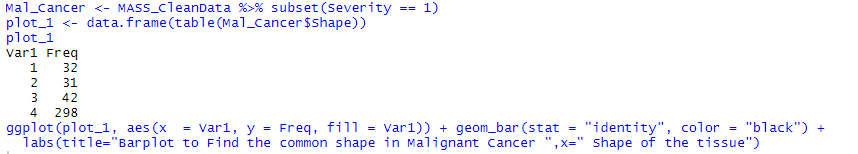
Exploratory Data Analysis is used to analyse the data based on visual plots and graphs. This helps to understand the hidden patterns in the data.

Initial questions can be easily answered using this exploratory analysis:

## Analysis Questions Answered

1. **What shape of the Cancer tumour is most common in malignant class?**





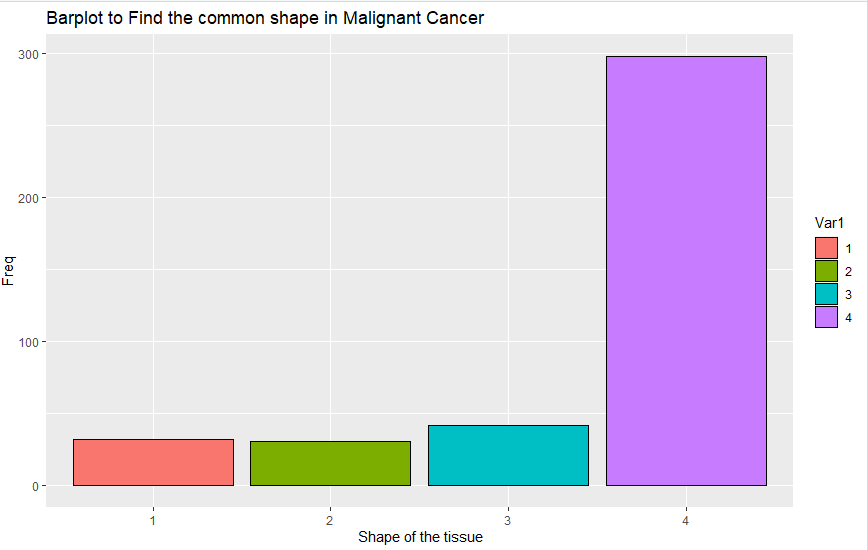
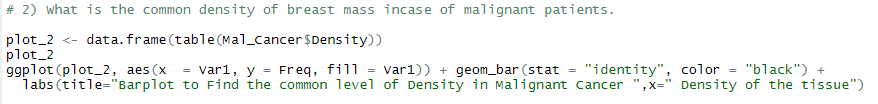
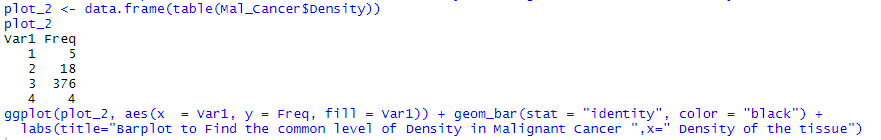


Figure 3: Types of Tumour in malignant

From the above visualization, the tissue type is found to be irregular, this shows that for a Breast Cancer patient most commonly the tissue is irregular.

1. **What is the common density of breast mass in case of malignant patients?**





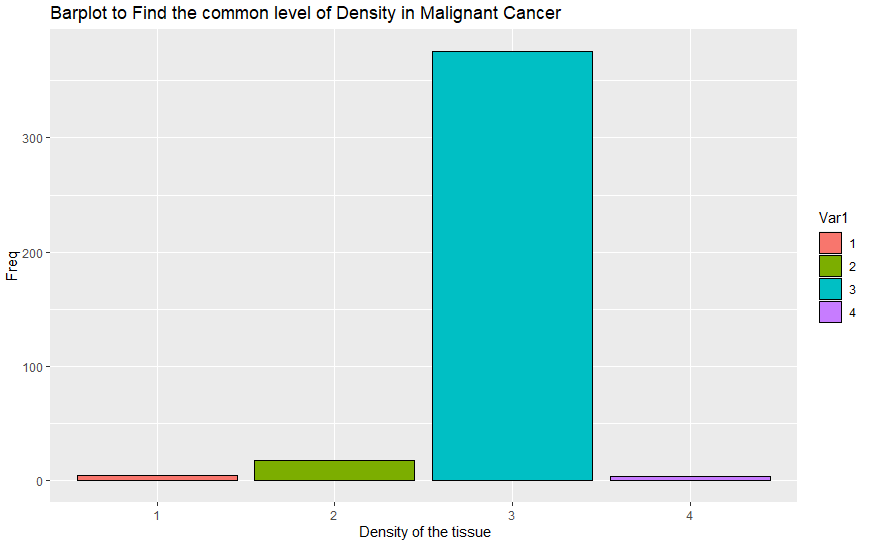


Figure 4: Density of the tissue

The most common density type in Density in malignant patients is **Low.** This is more dominating than any other Density type.

## Additional Histograms and Bar charts

Here are a few other visualizations to understand the data.

1. The below plot represents the age distribution on severity Boxplot. It is overserved that, the malignant cases are more

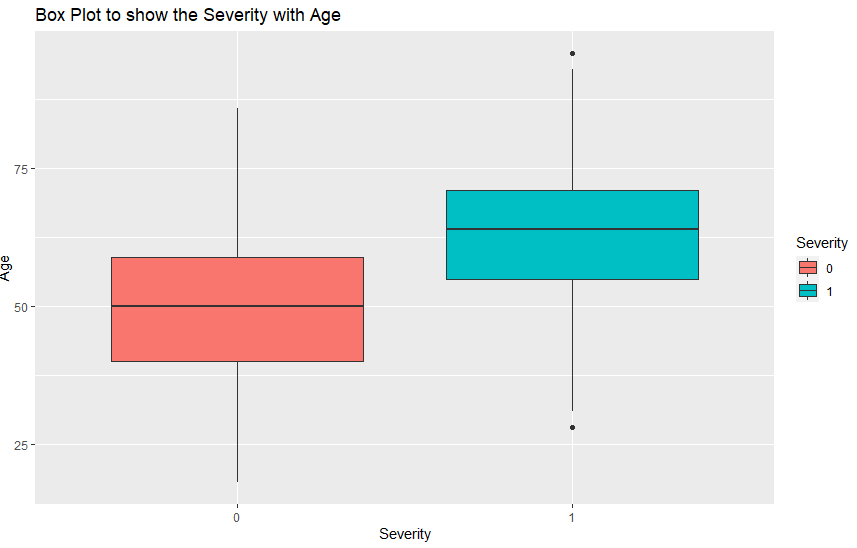


Figure 5: Boxplot to show the Severity

1. Histogram to show the age distribution based on BI.RADS

## 

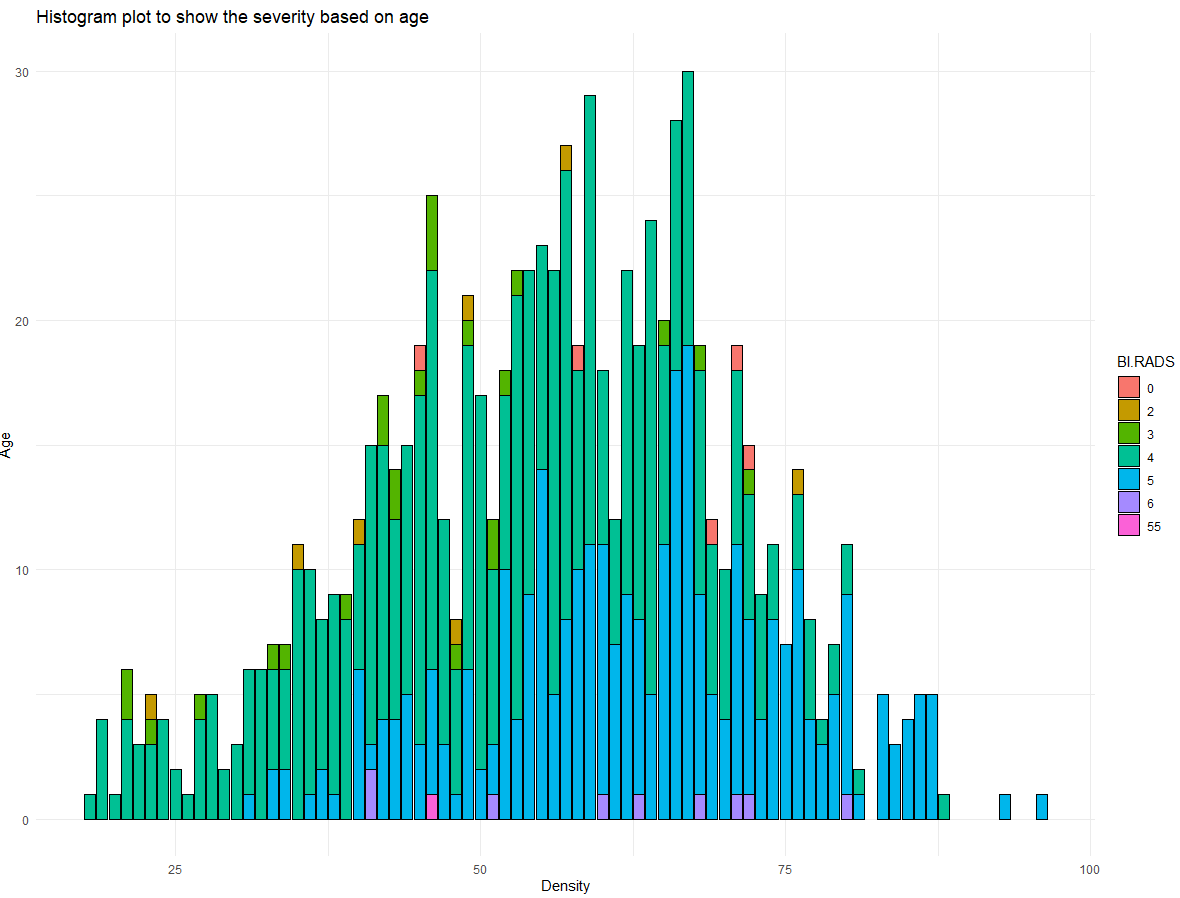


Figure 6: Histogram plot for Bi-RADS

1. Histogram to plot the Severity with age are shown below

## 

## 

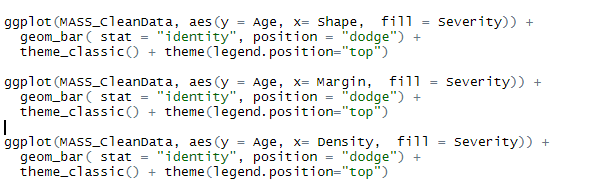
Figure 7: Histogram based in severity

1. Bar plots for Severity, Margin and Density are shown below.

## 

|  |  |
| --- | --- |
| Figure 8: Bar chart Severity Vs age | Figure 9: Severity vs Margin |
| Figure 10: Severity vs Density |  |

1. Side-Bar charts for Malignant and Benign cases on shape, margin and Density.



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

## Outliers:

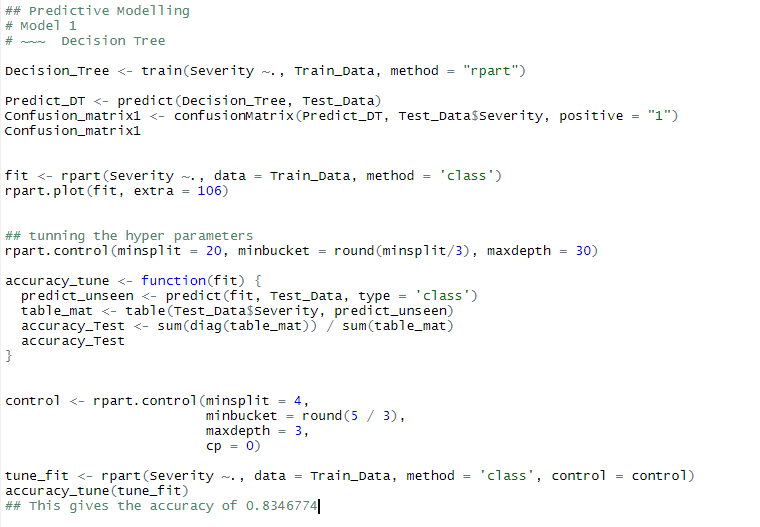
There is no outlier observed in this data set.

# Predictive Modelling:

The MASS data set is partitioned in to train and test subsets using a method “**createDataPartition**()". Using a random sample method, we partition 70 % of data into training and 30 % of data into testing. The training data is used for training the model and testing subset is used to validate the trained model.

The classification models such as Decision Tree, Random Forest and Neural networks are used to predict the diagnosis of breast cancer.

## Model 1: Decision Tree



**Decision Tree structure**

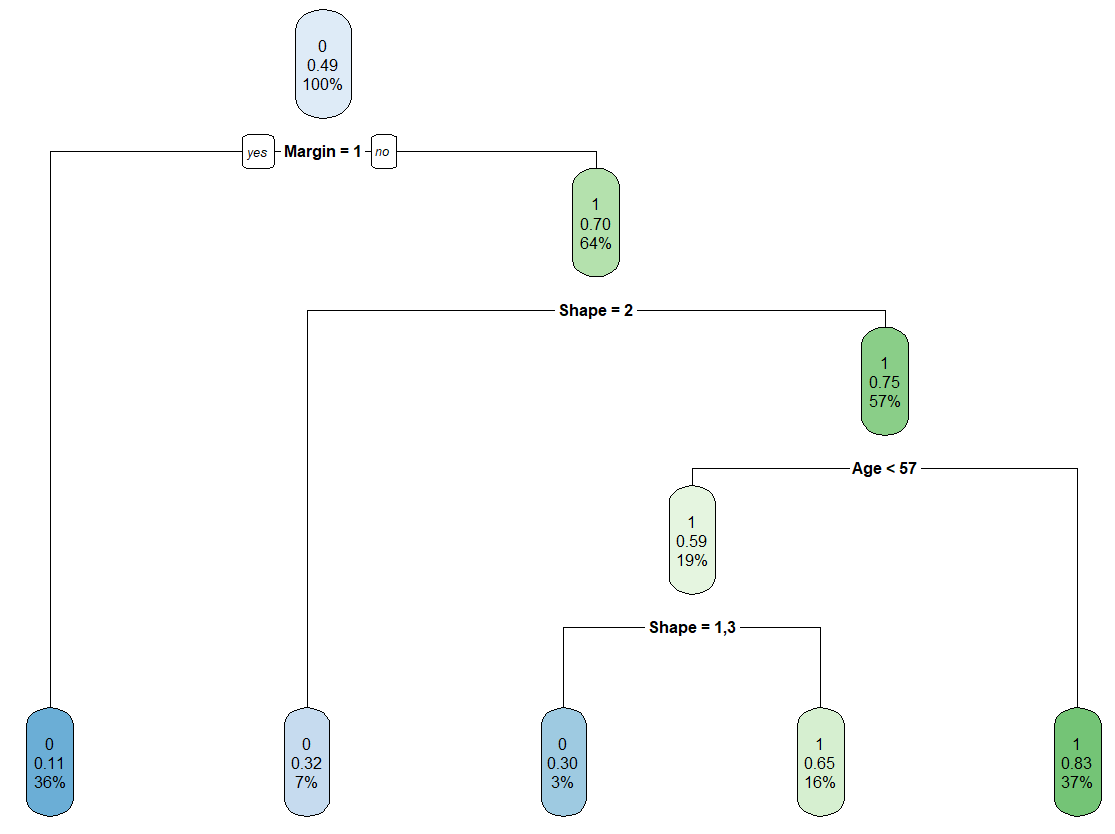
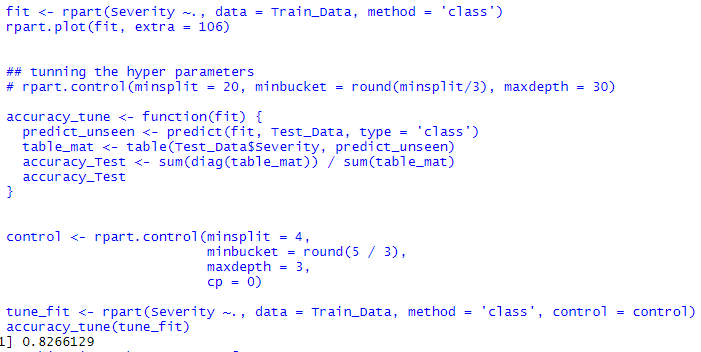
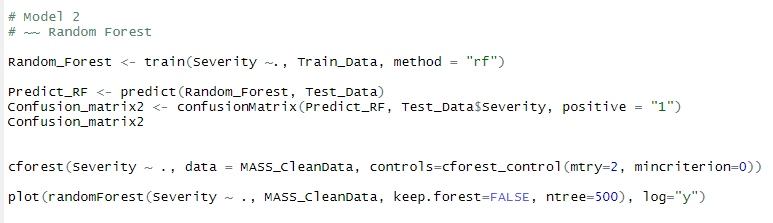


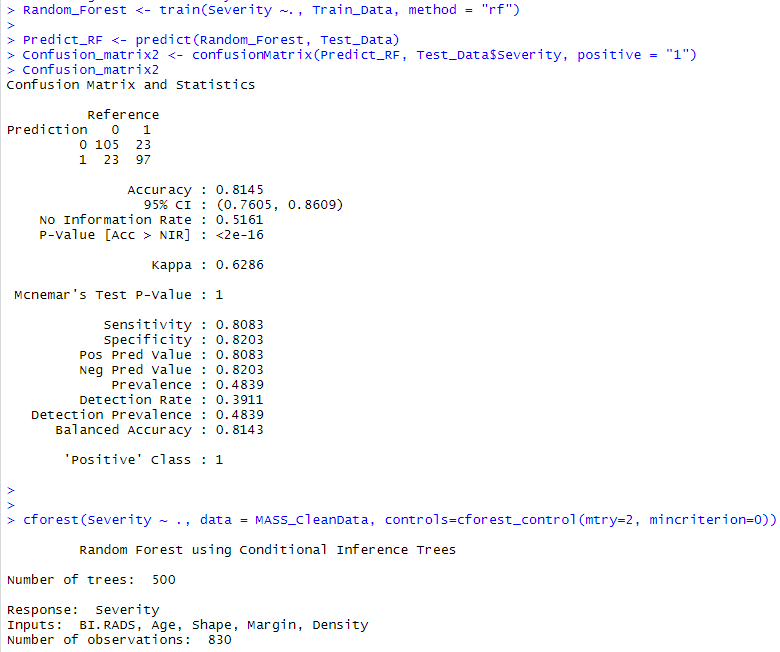
Figure 11: Decision tree for the MASS dataset

## 



## Model 2: Random Forest





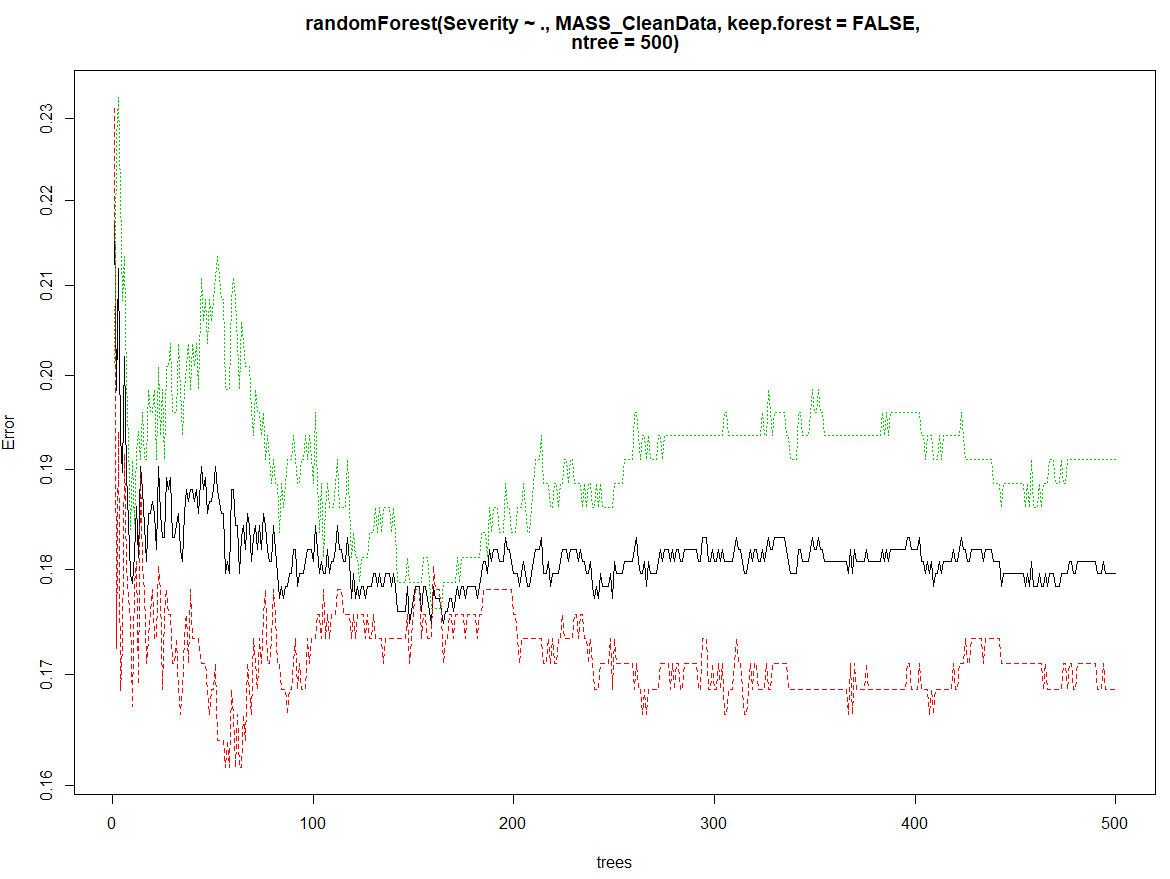
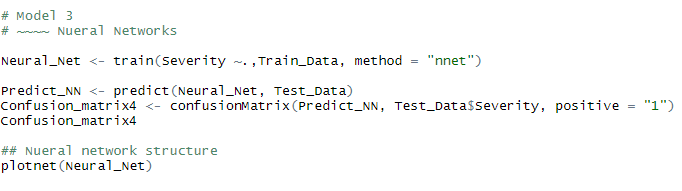
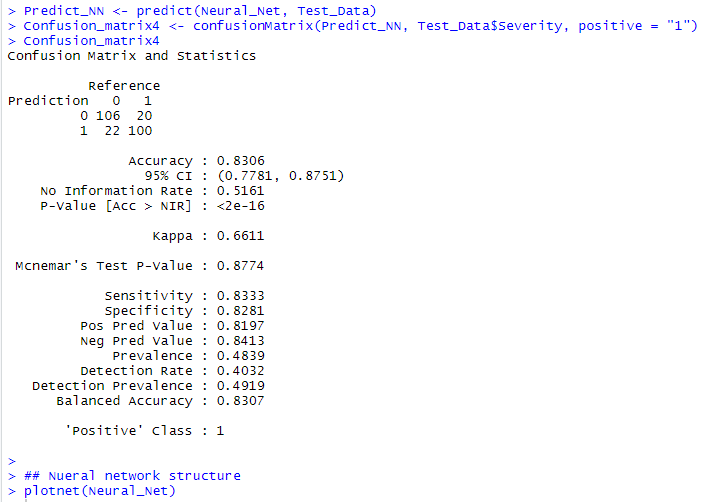


Figure 12: Number trees vs Error rate

From the above Random Forest Error vs Number of trees graph, we can observe that the error rate reduced from 1 to 150 trees. And then slightly increased from 150 trees to 300 and stabilized. As we considered random sampling methods, the performance of the model not evaluated at the best. To achieve better performance evaluation, Cross-Validation can be used.

## Model 3: Neural Networks





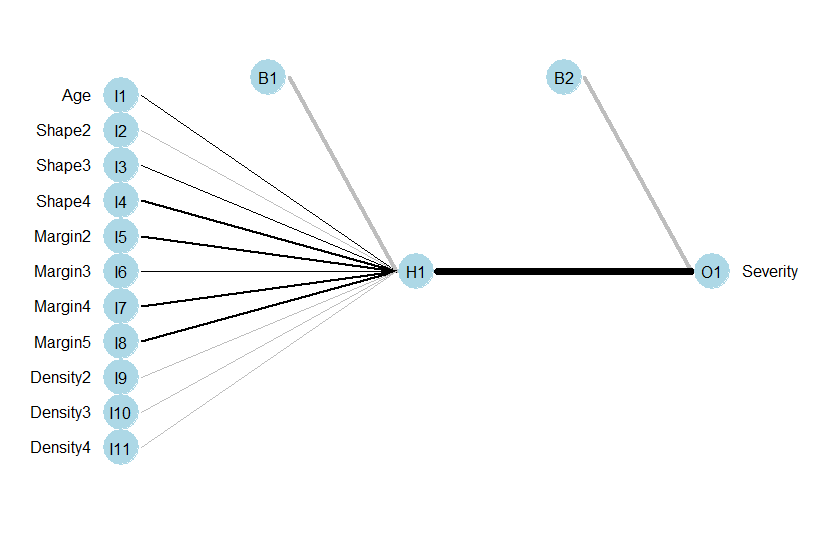


Figure 13: Neural Networks

# Interpretation of the Results:

Random Sampling allocates the observations randomly to the training and testing subsets. So, for every iteration the samples will be shuffled, as a result, the accuracy is not going to be same, every time we design the model. In this case, Considering the average/mean accuracy of the model would be one of the options.

After training the models on the dataset, 10 iterations were run on the models to check the performance of the model. As we have done random sampling, we achieved different accuracies for different iterations for every model. The results are as shown in table 1 below

Table 3: Accuracy of the model for 10 iterations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models Accuracy in %** | | | | |
| **No. of Iterations** | **Decision Tree** | | **Random Forest** | **Neural Networks** |
| **Before**  **Prune** | **After Prune** |
| 1 | 84.68 | 84.68 | **86.69** | 85.48 |
| 2 | 83.06 | 83.47 | 83.87 | 83.06 |
| 3 | 82.26 | 82.66 | 83.06 | 83.06 |
| 4 | 84.68 | **86.70** | 84.27 | **87.10** |
| 5 | 84.27 | 82.66 | 85.48 | 85.48 |
| 6 | **77.02** | **79.43** | **79.84** | 81.05 |
| 7 | **85.89** | 85.88 | 85.89 | 85.89 |
| 8 | **85.89** | 85.08 | 85.48 | 85.08 |
| 9 | 83.87 | 82.25 | 84.27 | 83.06 |
| 10 | 80.24 | 81.45 | 80.65 | **79.03** |
| **Mean** | **83.19** | **83.43** | **83.95** | **83.82** |

From the above table,

## Decision Tree before Prune

The minimum accuracy of the Decision Tree model is observed to be 77.02%, whereas the Maximum is 85.89% observed in 7th and 8th iterations.

## Decision Tree After Prune

After prune, the minimum model accuracy observed is 79.43% and the maximum is 86.70% in 4th iteration.

## Random Forest

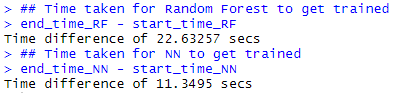
The model performed well in the iterations, except in 6th and 10th iterations it has shown 79.84 and 80.65% accuracy. And maximum is found in the first iteration with 86.69%.

## Neural Networks

All the models reported the least accuracy in the 6th Iteration, whereas Neural Networks performed well with 81.05% accuracy. The maximum accuracy of all the models is observed to be 87.10% in the 4th iteration.

we can observe that there is no fixed accuracy for any of the model. So, the mean/average accuracy of the models are calculated. It is observed that all the 3 model results are almost the same. Among these, Random Forest and Neural Networks performed well with 83.95 and 83.82 % accuracy respectively. The Decision Tree performed slightly less. Compared to the rest of the models with 83.19 % before pruning and 83.43% after pruning.

***Time is taken to run the models.***



When it comes to time take for the model to get trained, Random Forest has taken 22.63 secs to train the model on the other hand Neural networks have taken 11.34 secs for the model to get trained.

## Raised Questions:

The questions for the analysis are answered through the exploratory data analysis. In this, we have found that:

1. The tissue type is found to be irregular, this shows that for a Breast Cancer patient most commonly the tissue was irregular.
2. The most common density type in Density in malignant patients is **Low.** This is more dominating than any other Density type.

# Conclusion

The computer technology used in the medical field has many advantages, mainly in the diagnosis of a disease. A breast cancer diagnosis of mammographic data is nothing old, but using high-performance machine learning algorithms, it increases the chances for better treatment. In this experimental research on the mammographic dataset, we found that Random Forest and Neural networks performed well compared to the Decision tree. But, when it comes to time taken to train the model. The decision tree model is faster enough to get trained, compared to Random Forest and Neural Networks. Random Forest took 22.63 seconds were as Neural Networks took 11.34 secs.

If the model performance is considered, then Random Forest would be the best option. If the simplicity of the model is the concern, then the Decision Tree can be the best option. Though Decision Tree is easy to interpret the results, it has high chances of Overfitting. So, additionally, better performance tuning should be incorporated with the Decision Tree.

As a part of my future studies, I would like to work on the performance of the models which is simple and gives high performance. To achieve this, better Machine Learning and Deep Learning techniques such as SVM, Dimensionality Reduction methods like PCA & LDA, and Deep neural networks like Multi-Layer Perceptions will be implemented.

# Reference:

Wang, Defeng, Lin Shi, and Pheng Ann Heng. "Automatic detection of breast cancers in mammograms using structured support vector machines." *Neurocomputing* 72, no. 13-15 (2009): 3296-3302.