

**Submitted by:**

33296085 Vajratiya Vajrobol

33553377 Sai Kumar Vadde

33669079 Rajesh Jyothi

*Machine Learning and Neural Network Techniques Using Normalization, Feature Selection and Dimensionality Reduction (Using PCA analysis)*

# Abstract

*The Diagnosis of Malignant and Benign is a challenging task for Doctors if the number of test reports is increasing in the count. The data analysis of this manual process may take longer time and efficiency of the results cannot be as promised. Using the advanced AI, Machine Learning and Deep Learning techniques, the diagnosis of the best cancer is possible in less more with more efficient results. In this report, we performed Breast Cancer Diagnosis on Wisconsin Breast Cancer Diagnosis dataset (WBCD) from the UCI repository. The dataset is collected based in the Fine Needle Aspirate (FNA) tests, by penetrating a fine needle in the beast to collect the samples. Over 569 patients were tested based on the initial results. The images of the tumour captured are computerised for easy data analysis. We performed Machine Learning algorithms such as Decision Tree, Random Forest, and Neural Networks models on this dataset to detect the diagnosis of cancer. Later we normalised the data, implemented feature selection techniques using Correlation Matric and Random Forest. PCA is also used for Dimensionality reduction. The results showed that the Random forest model performed well all the four cases with an average of 95.5% accuracy, unaffected with the overfitting of training data. The Neural Networks performed well using Principal Components.*

***Keywords****: Breast Cancer, Diagnosis, Breast Cancer Diagnosis, Machine Learning, Neural Networks, AI techniques, PCA analysis, Feature Selection, Correlation Matrix, Random Forest, Decision Tree.*

# Introduction

Cancer is the second main cause of deaths in the world which is responsible for 9.6 million deaths in 2018 (World Health Organisation, 2018). Comparatively, Breast Cancer is the most affected cancer in women alone and causing deaths in millions. As per the WHO report 2018, 2.09 million deaths occurred due to Breast Cancer. The scientific studies state that if the breast cancer was detected early stage, The chances of cure increases. This can also help in providing a personalized treatment based on the level of disease.

With advanced computational hardware and software technologies, Machine Learning, Deep Learning and Artificial Intelligence techniques, helps us in understanding the hidden features and patterns in the data. These patterns help in better diagnosis of the disease and level of the tumour in the breasts. So that the personalised treatment can be arranged based on the severity of the Cancer.

## About Wisconsin Breast Cancer Diagnosis Dataset.

The dataset is collected based in the Fine Needle Aspirate (FNA) tests, by penetrating a fine needle in the beast to collect the samples. About 569 one such samples are collected who expected to have a tumour in the breast (Street et al., 1993). The data is collected by computerizing the image features such as radius, texture, perimeter, area, smoothness, Compactness, concavity, Concave points, symmetry, and fractal dimension, of the breast. The Mean, Standard Error and Worst are calculated for each of these variables and tabulated with diagnosis status (Street et al., 1993). The two groups of diagnosis, Benign and Malignant are the tumour developed in the breasts. If this tumour grows larger, that leads to Breast Cancer, also called Malignant. On the other hand, the tumour that stops growth after a certain point of time is called Benign tumour. It is harmless and can be removed with surgical treatments. The whole dataset has 569 observations with 32 columns including patient ID and Diagnosis status. A detailed description of all the variables is provided in the Appendix section.

This paper deals with the Breast Cancer diagnosis using Machine Learning and Neural Network models on Wisconsin Breast Cancer Diagnosis dataset (WBCD). This Data analysis aims to find a better model that can help in finding the diagnosis of the patients based on the features provided. The Machine Learning models such as Decision Tree, Random Forest, and Neural Networks. Initially, prototype models are designed, and the performance of the models are noted. In the later stages, we aimed to improve the performance of the model using AI techniques including, Normalization of the data, Feature selection using the Random Forest algorithm & correlation matrix, Dimensionality Reduction using Principal Component Analysis (PCA). These heuristics techniques can help in reducing the number of parameters in a dataset by selecting the significant ones. this solves the multicollinearity problem to the maximum extend and improves the performance. Normalization is the Data Preparation stage, performed on the numerical data to bring the attributes on a single scale (0 to 1) this helps in understanding the distribution of the data. The PCA analysis helps in dimensionality reduction by reducing the dimensions of the principal components, without losing the variance of each component. Initially, the attributes are converted into the Principal components, the top few components which capture the highest variance are selected for model design.

The results showed that The random forest model performed well in all the cases with an average of 95.5% accuracy, whereas Decision Tree, overfitting the training dataset and gave good results with competing or raw data, and the results are not steady during Feature selection and Dimensionality reduction. Neural networks showed the highest accuracy in the PCA analysis with 95.07% accuracy.

# Background

The Criticality increase for a doctor, in manual diagnosis, when the number of tests reports or the data increases. Solving this problem, many studies were conducted on Breast Cancer Diagnosis and found the Machine Learning and AI techniques that can solve the problem of diagnosis. Here are a few of the Research Analysis reports that contributed to our analysis.

Salama and team conducted a research analysis in 2012 on three different Breast Cancer Datasets for diagnosing Malignant or Benign. They implemented different classifier models such as Decision Tree, SVM, Multi-Layer perceptron, Naïve Bayes classifier and KNN. They also implemented a PCA analysis of the models to get better results. The results show that the SVM-RBF kernel gave better results with 96.84 % accuracy on dataset 1, 99.0% on dataset 2 and 96.99% on the dataset 3 (Salama et al., 2012).

Nguyen in 2013, reported promising model results with 100% accuracy and 99.8% accuracy on an average. The results are obtained by implementing the Random Forest algorithm with feature selection techniques on the Wisconsin Dataset (Nguyen et al, 2013).

Zheng and team, in 2014, implemented a K-means algorithm to find the patterns in the Wisconsin Breast Cancer dataset and a powerful Machine Learning algorithm called Support Vector Machine (SVM). The optimization techniques, Feature selection, feature extraction and Dimensionality Reduction. The analysis ended up with an accuracy of 97.38% suing Kernel SVM (Zheng et al., 2014).

The motivation for our data analysis is from the above-mentioned Research analysis conducted. So, in our data analysis, we used feature selection methods, Dimensionality reduction using PCA along with algorithm.

# AI Technique

In this analysis, we have implemented Machine Learning algorithms such as Decision Tree, Random Forest, and Neural Networks with heretic AI techniques like Normalization, Feature Selection and Principal Component Analysis for Dimensionality Reduction

Decision Tree algorithm is a tree-based structure, that analyses the data based on given input conditions. It is a supervised learning algorithm that can be used for classification (output is categorical) and Regression (output is continuous) problems (DI et al., 2009). In this analysis, the parameters are for Decision Tree asset default. For tuning the Decision tree, **CP value** is calculated, and **pruning** is then implemented.

Random Forest is the collection of tree structures, or simply a collection of Decision Trees. The algorithm internally works on all the possible combinations of the decision trees and best tree result is chosen. It is a bagging algorithm and implements Ensemble Learning techniques. The overfitting problem in Decision Tree is reduced in the Random Forest algorithm (Chen et al, 2012). For our analysis Test and error, curves are plotted to understand the error rate concerning the tree.

Neural network the basic Deep Learning technique, which has an input, hidden and output layer. The input layer has the independent attributes and the output layer gives the target variables. Each input layer is connected to the multiple hidden layers for processing. It can also be used for regression and classification problems (Shekhawat & Sheetal, 2011). In this analysis, we used the neural networks in classification mode by setting **linout = FALSE,** this implies that NN now operates in classification mode. If set as **TRUE**, it operates in a regression model.

*Normalization* is one of the basic Data transformation technique, used mainly on numerical data. As, all the variables in this dataset are not on same scale, we scale them between 0 to 1 using a user-defined normalization formulae:  *(x - min(x)) / (max(x) - min(x)),* This data is used for further analysis.

The *feature selection* is implemented by

1. *Filter method – Correlation Matrix and*
2. *Embedded Method – Random Forest.*

* Using the correlation matrix, if coefficients are close to +1 or -1, defines a high correlation between the variables. And 0 represent there is no correlation, and the variables are independent of each other. The direction of correlation is defined by the positive and negative signs. The feature selection is done, by removing one of the highly correlated variables among two, so that the other variable is independent of each other variable. This reduces the complexity of the model and makes it simpler, and on the other hand, it reduces the *multiple collinearity* problem. In this process, we selected 19 variables for the further modelling process. The selected variables are shown in the appendix section.
* The Random Forest model has the predefined feature selection methods, called *VarImp(),* and *importance().* Based on the input data and the tree structure, the important variables are suggested. Modelling is done based on these selected features. The selected variables in this method are shown in the appendix.

## The motivation for PCA:

The above two methods reduce the complexity of the models by limiting the input variables. The problem here is, though we are reducing the complexity of the model, we are losing the data withheld in other variables, So, we implemented PCA analysis, where each component captures variance of all the variables. So, that when limited to the top 10 components, the variance data held with other variables is not lost.

In this analysis. The principal components are limited to 12, to capture more variance. The screen plot and Biplot help in understanding the variance captured in each component and their direction tells the percentage of variance captures in each dimension.

# Evaluation Method and Results

Figure 1 shows the complete architectural view and flow of the data analysis.

System configurations used

Table 1: The below mentioned are the basic requirements used

|  |  |
| --- | --- |
| **Sys Config** | **Requirements** |
| OC | Windows 10 |
| RAM | 8 GB |
| Program | R |
| GUI used | Rstudio |
| Version of GUI | v1.2.5019 |

For the Breast cancer dataset, we used the windows 10 operating system with 8 GB of RAM. the programming is done In R using Rstudio platform. The all the updated packages are installed, and respective libraries are imported.

The architectural flow of the project is as follows,

* The data is collected from the UCI repository made available for Rstudio.
* For Data understanding, some visualizations are plot and simple commands such as str(), summary, columns(), etc are used
* In Data Cleaning, the unwanted columns “1” and “33” are removed.
* The data partition is done using the *CreateDatapartition ()* method from the caret package, by sub-setting data into 75% for train and 25% for training.
* Followed by Data Modelling, using Decision Tree, Random Forest, and Neural Networks.
* The results are noted for the basic prototype.
* Now Normalization of the done, using raw cleaned data.
* It is followed by feature selection
* And the Dimensionality reduction using PCA.
* Again, after the above two steps, models are designed.

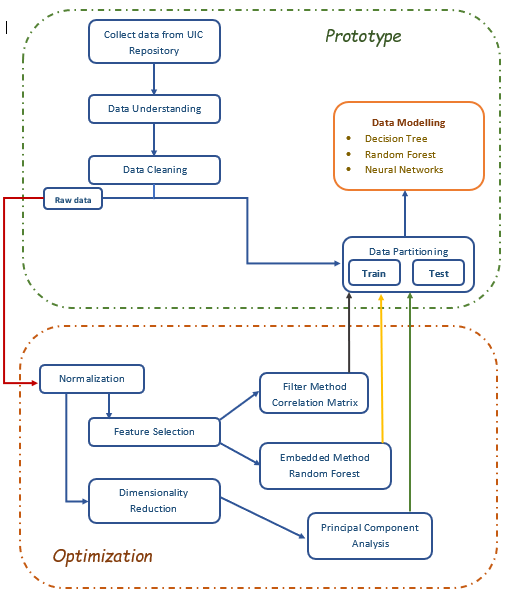


Figure 1: The architectural diagram of the project

## Results:

The prototype results are shown in Table 2, and the performance of the models after two feature selections is shown in Table 3.

|  |  |  |
| --- | --- | --- |
| **Models used** | **Feature Selection 1 Correlation Metrix** | **Feature Selection 2 Random Forest** |
| Random Forest | 95.77 | 95.07 |
| Decision Tree | 93.66 | 94.37 |
| Neural Networks | 94.37 | 93.66 |

Table 2: Results for the prototype model. Table 3: Results comparison between two Feature selection methods

|  |  |
| --- | --- |
| **Models used** | **Base Model** |
| Random Forest | 95.77 |
| Decision Tree | 95.07 |
| Neural Networks | 93.66 |

In the prototype analysis, the complete dataset is used without performing normalisation, we can see that all the three models performed well with 90% above accuracy. Neural Networks performed slightly less compared to the other two models.

When the feature selection techniques, correlation matrix and Random Forest are implemented, the Random forest performed slightly greater above in the first case with 95.77% accuracy, Decision tree performed well in the second case with 94.37% and neural networks performed well in the first case, correlation matrix feature selection.

Table 4: Comparison of all the methods and techniques performed on the models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models used** | **Base Model** | **Feature Selection 1 Correlation Metrix** | **Feature Selection 2 Random Forest** | **Dimensionality Reduction using PCA** |
| Random Forest | 95.77 | 95.77 | 95.07 | 94.37 |
| Decision Tree | 95.07 | 93.66 | 94.37 | 92.96 |
| Neural Networks | 93.66 | 94.37 | 93.66 | 95.07 |

The above Table 4 shows the comparison of all the model performance with different cases. In the PCA analysis, only 12 PCA components are considered for the analysis, which captures above 96% of the variation. And the results are, as shown above, the Neural Networks performed good with an accuracy of 95.07%, whereas Decision tree performed the least, with 92.96% among the three models.

# Discussion

The Data analysis using the machine learning and Neural Networks models are performed, and results show that the Neural networks performed well with the PCA analysis compared with the rest of the three cases. Random Forest, on the other hand, gave pretty good results in all the cases with an average of 95.5% accuracy. Decision tree showed some fluctuations in the results shows that there is some overfitting in the model when the complete dataset is used.

We learnt that, when the raw data is used with all the features, the models such as Decision Tree, overfits the training dataset, and cannot perform well with the new dataset. The Neural Networks and Random Forest gave steady results, shows that they are nor influenced much with overfitting.

To improve the performance of the models, we could use the K Fold Cross-validation, so the models are not biased and give better accuracy after tuning. Additionally, parameter tuning can be implemented for each model. Overall, the models performed as expected, but still, the misclassification rate of the models should be reduced so that, the models give accurate results in Malignant and Benign classification. One misclassification will cost the life of a person.

# References

Chen, Xi, and Hemant Ishwaran. "Random forests for genomic data analysis." *Genomics* 99, no. 6 (2012): 323-329.

DI, Xiang-hong, Zhou-long WANG, Ming-ming ZHANG, and Shu-ping ZHANG. "Extraction of Land Covering Information Based on Decision Tree in Penglai City [J]." *Shandong Land and Resources* 11 (2009).

Nguyen, Cuong, Yong Wang, and Ha Nam Nguyen. "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic." (2013).

Salama, Gouda I., M. Abdelhalim, and Magdy Abd-elghany Zeid. "Breast cancer diagnosis on three different datasets using multi-classifiers." *Breast Cancer (WDBC)* 32, no. 569 (2012): 2.

Shekhawat, Prachitee B., and Sheetal S. Dhande. "A classification technique using associative classification." *International Journal of Computer Applications* 20, no. 5 (2011): 20-28.

Street, W. Nick, William H. Wolberg, and Olvi L. Mangasarian. "Nuclear feature extraction for breast tumour diagnosis." In *Biomedical image processing and biomedical visualization*, vol. 1905, pp. 861-870. International Society for Optics and Photonics, 1993.

World Health Organisation (n.d.). *Breast Cancer*. Retrieved from WHO: <https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>

Zheng, Bichen, Sang Won Yoon, and Sarah S. Lam. "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms." *Expert Systems with Applications* 41, no. 4 (2014): 1476-1482.

# Appendix

1. The below shown is the histogram for Mean, Standard Error and worst data.

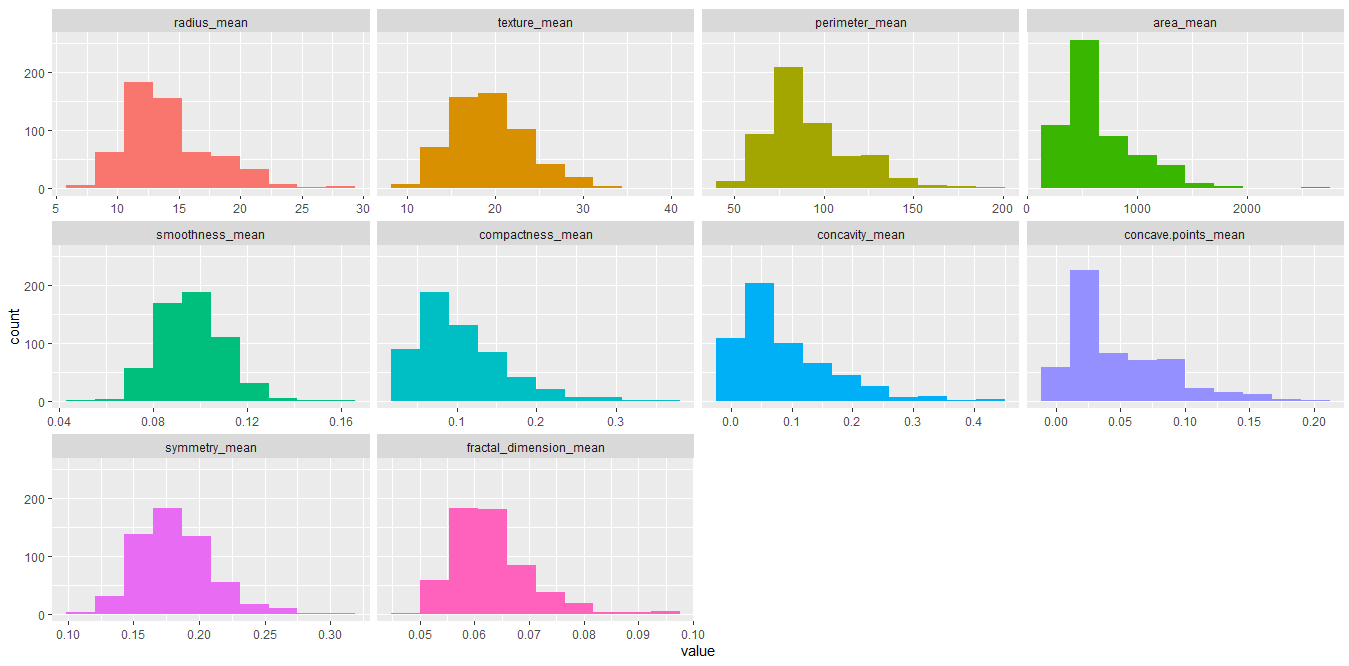


Figure 2: Histogram For Mean data

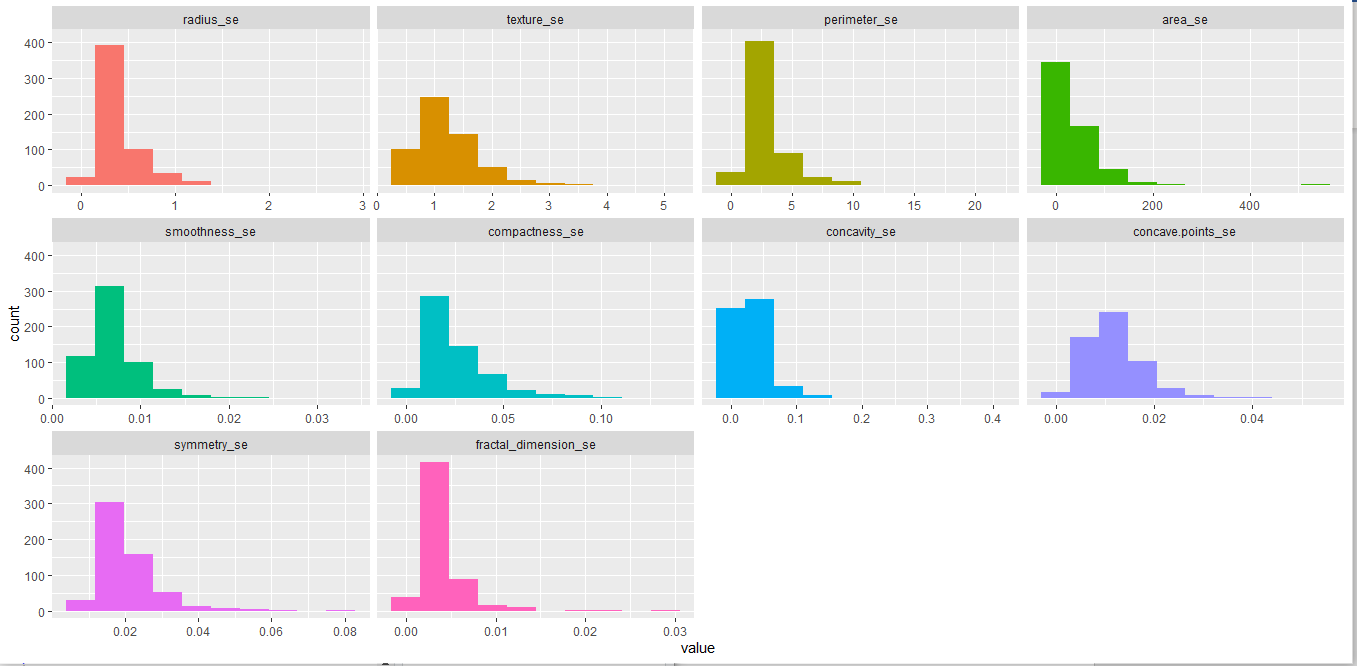


Figure 3: histogram For Standard Error data

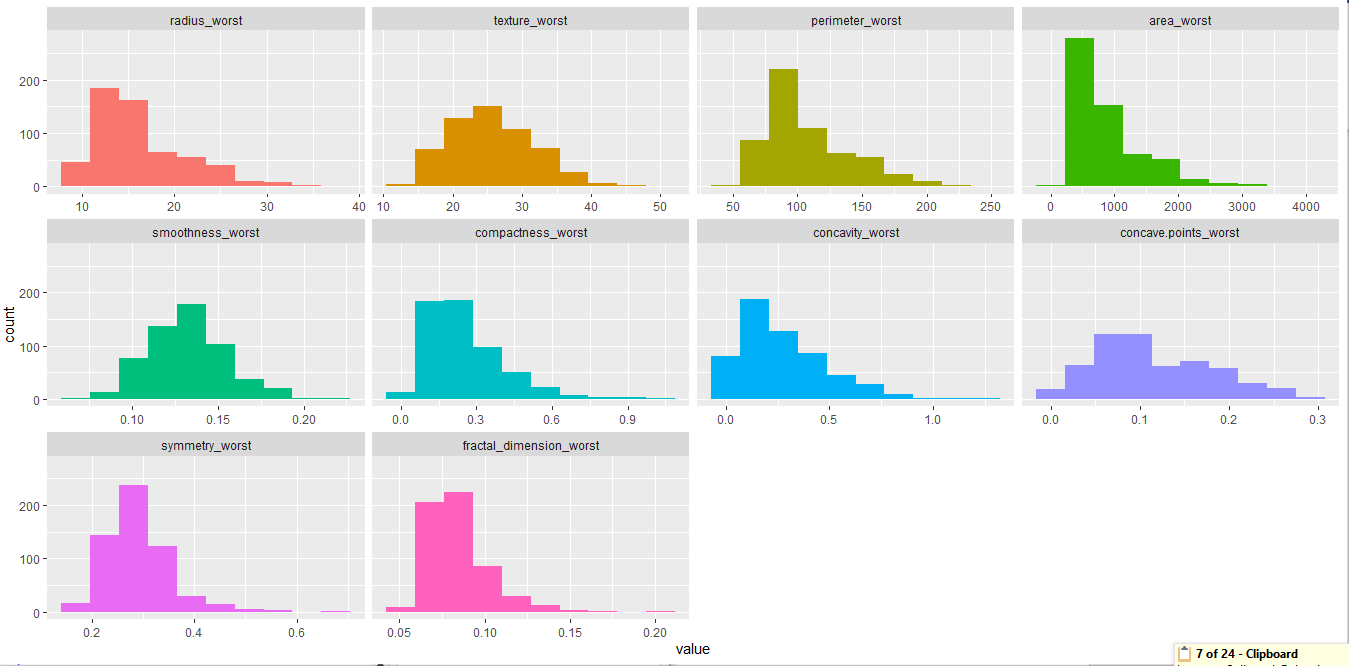


Figure 4:histogram For Worst data

1. The Below shown are the correlation matrix for the Mean, standard Error and Worst data, and the complete raw data.

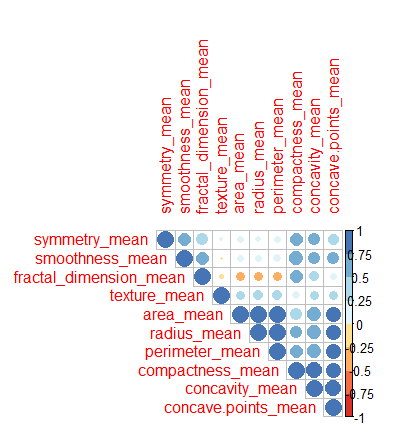


Figure 5: Correlation for mean

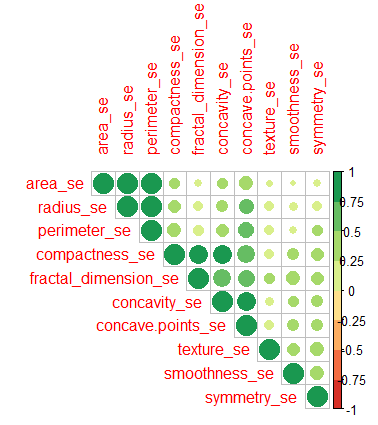


Figure 6: Correlation for standard Error data

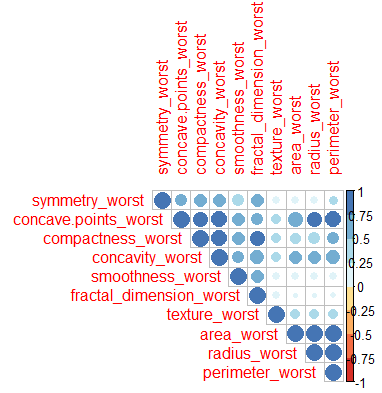


Figure 7: Correlation for worst data

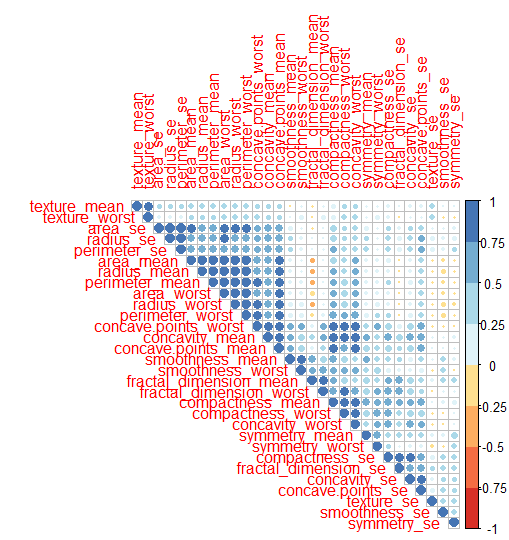
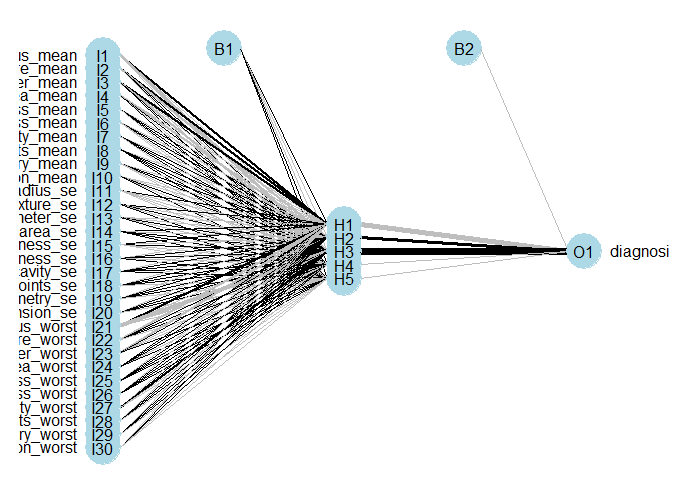
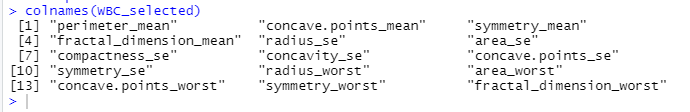


Figure 8: Correlation for raw data

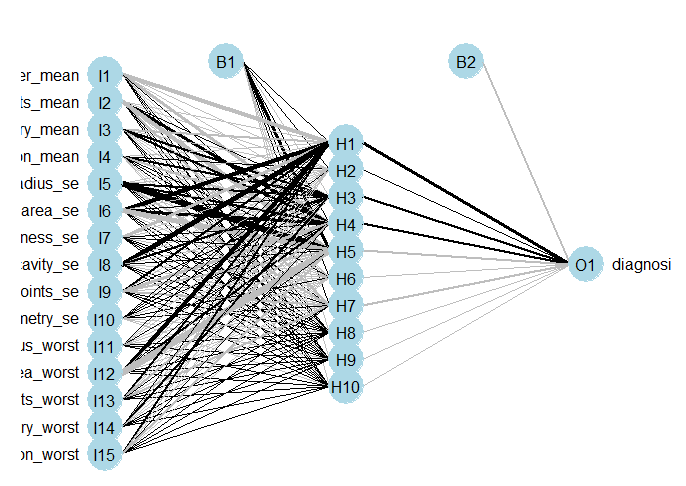
1. Neural networks for Raw data.



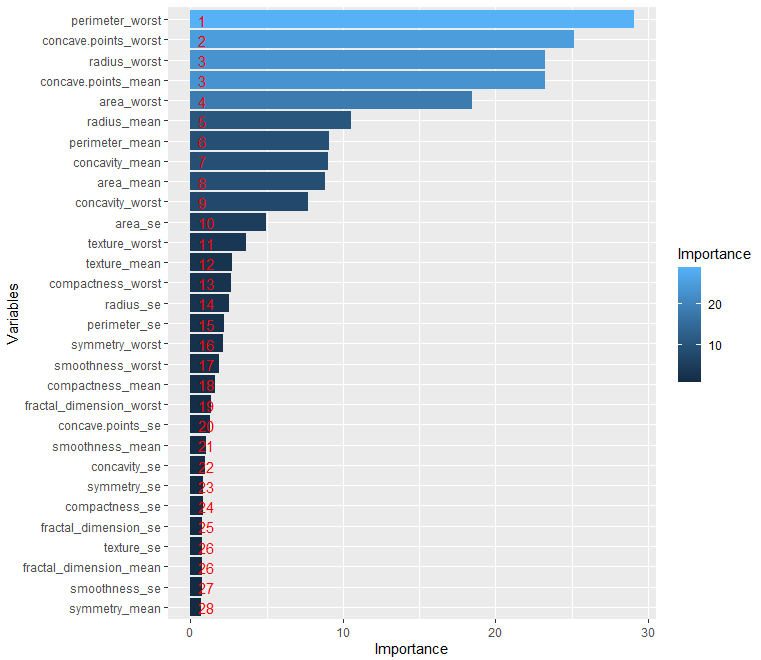
1. Feature selected after correlation matrix



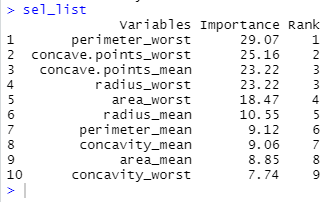
After Feature selection 1: using a Correlation matrix



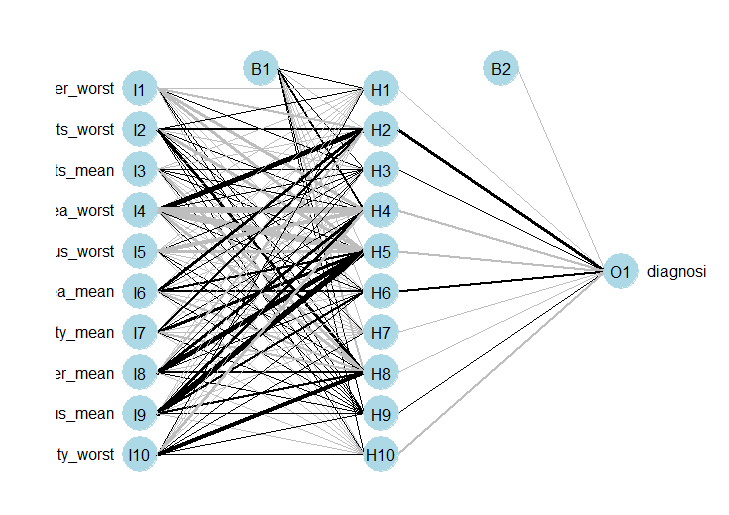
Features selected after feature selection 2



List



NN after feature selection 2



Screen plot in PCA analysis

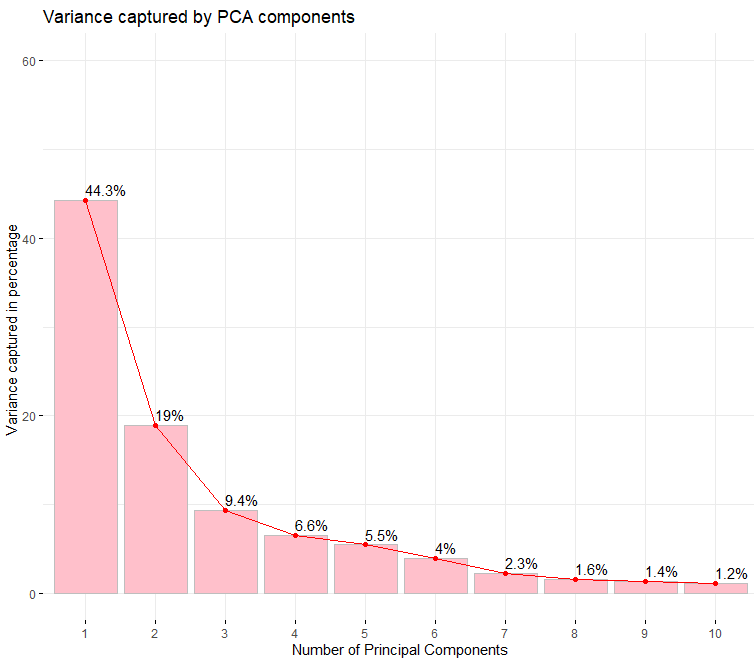
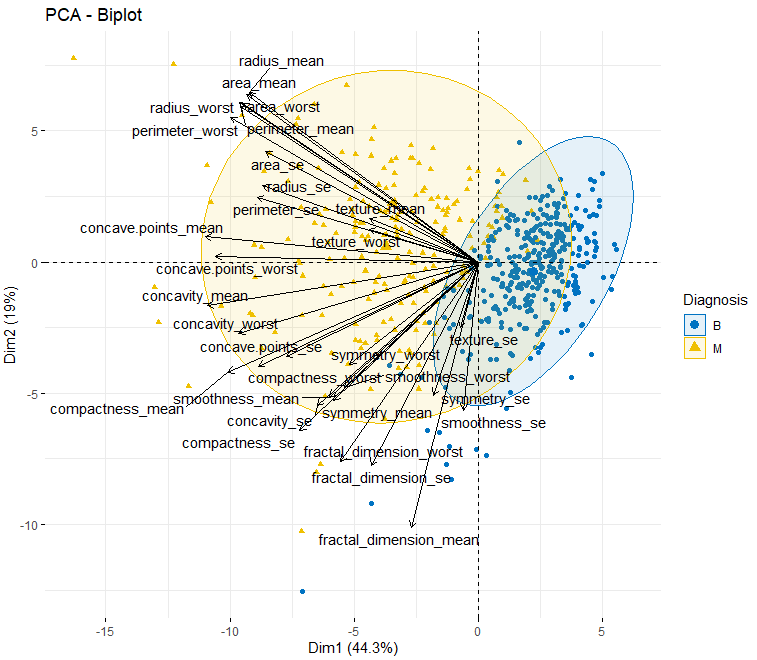
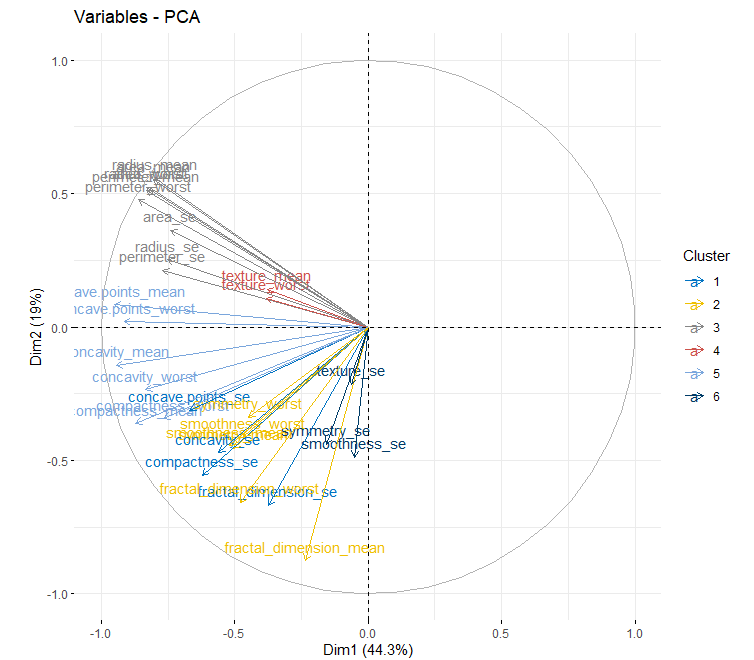


Figure 9: Screen plot for PCA analysis

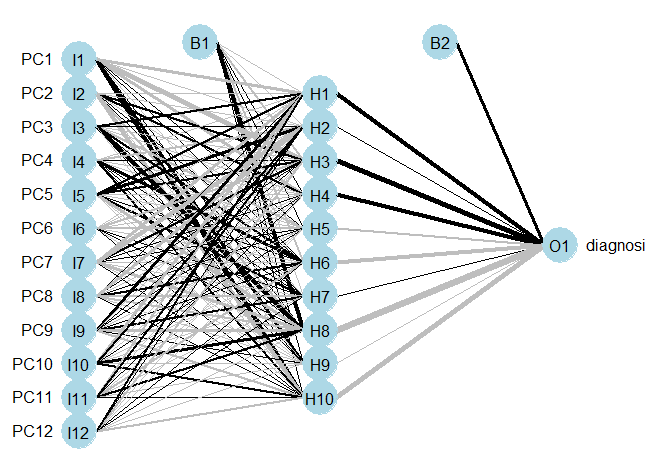
PCA Biplot to capture the Variance in dimensions



Finding clusters in Biplot



NN plot selected PCA components:



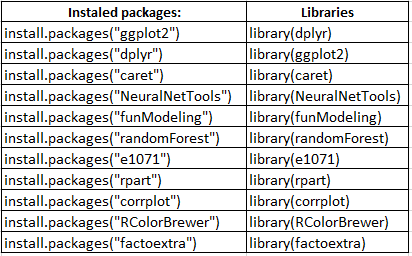
# User Guide:

1. The system configurations used are:

|  |  |
| --- | --- |
| **Sys Config** | **Requirements** |
| OC | Windows 10 |
| RAM | 8 GB |
| Program | R |
| GUI used | Rstudio |
| Version of GUI | v1.2.5019 |

1. The data should be made to the code. Or set the path in the R program to get the data.
2. The packages installed are:

Table5:: Packages and libraries used



1. All the packages used are the latest versions (mostly)
2. The program should be run sequentially line by line