Prediction of Breast Cancer Wisconsin (Diagnosis)

Contents

Abstract

Breast cancer is most frequently found in women and found worldwide. The purpose of the project is to detect tumor is cancer or not. Early detection of cancer helps to mitigate the extreme state and provides prevention rather than cure. In this study, we deployed linear and non-linear classification models to predict whether cancer is benign or malignant. The linear models like logistic regression, LDA, PLS-DA, glmnet, and SparseLDA, non-linear models like MDA, RDA, neural networks, SVM, FDA, KNN, Naive Bayes, and random forest are applied for train sets, selected two best models from each linear and non-linear classification techniques. The performance of three models on the test set had a kappa statistic of 86% and the neural network model provided the best performance with a kappa statistic of 88.5% approximately.

1. Introduction

The study of the Breast Cancer Wisconsin (Diagnosis) has the characteristics of cell nuclei to understand its structure and its change (Figure 1.1). The objective of the project is to predict whether the tumor is benign (not cancer) or malign (cancer) which implies that it is a two-class classification problem.

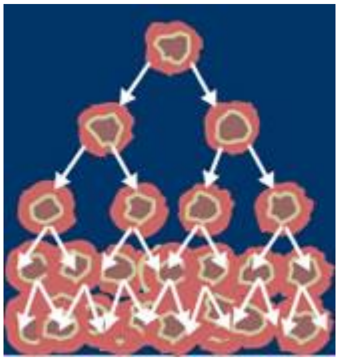


Figure 1.1: Structure of cell nuclei (Image source [1])

1.1 Dataset Description

The variables in the dataset describe the properties of cell nuclei. The properties of each cell are derived from an image of a breast mass through a fine needle aspirate (FNA). The 30 predictors are computed from the real ten-valued features of each image mean, standard error, and worst or largest (average of the three largest values). For instance, value 3 is Mean Radius, value 13 is Radius SE and value 23 is Worst Radius [1].

The description of each predictor is as follows:

- **Radius:** Average of distances from the center to points on the perimeter
- **Texture:** Standard deviation of gray-scale values
- **Perimeter:** Size of the core tumor
- **Area:** Region of tumor
- **Smoothness:** Local variation in radius lengths
- **Compactness:** Perimeter^2 / area - 1.0
- **Concavity:** Severity of concave portions of the contour
- **Concave points:** Number of concave portions of the contour
- **Symmetry:** Axis around the tumor
- **Fractal dimension:** coastline approximation - 1

The 30 predictors are derived from the above 10 features. There is a total of 33 predictors, 32 are continuous and the response variable is categorical.

2. Data Pre-Processing

The first step in developing a model is to understand the data and format the data from unstructured to structured form. Data pre-processing helps to understand the data and format it into a structured form where the models are easy to learn the characteristics of data. We followed the steps sequentially and started by handling missing values, dummy variables, degenerate variables, skewness, correlation, and outliers. Our dataset has only continuous predictors and no need for dummy variables and degenerate variables.

2.1 Missing values

Handling missing values is the crucial step in data preprocessing. The dataset did not have any missing information (Figure 2.1), and there was no loss of information at this step. However, a variable called 'X' has its entire data missing. Hence, we removed it.

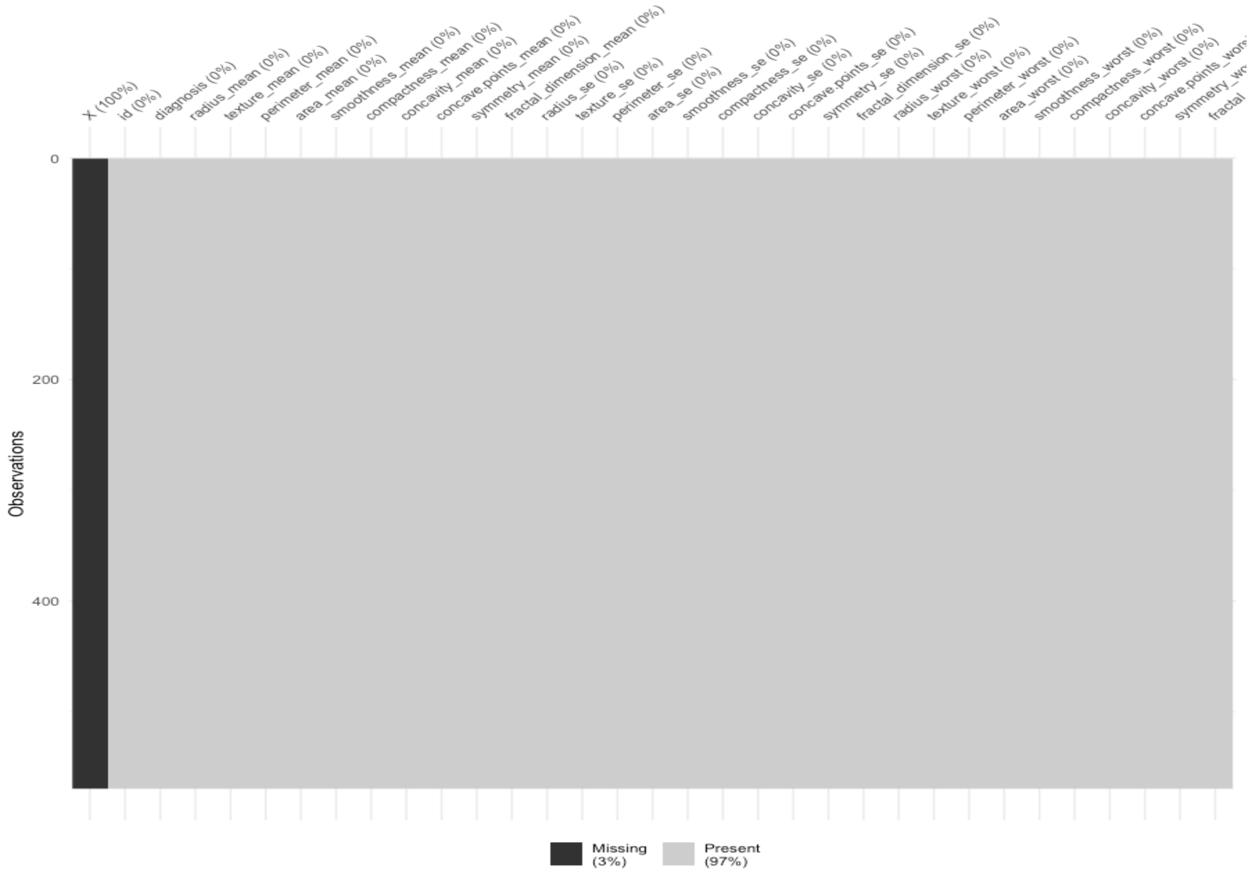


Figure 2.1: Plot visualize missing values

2.2 Skewness

Histograms were used to understand the distribution of our predictor variables. Most of our predictors are heavily and moderately right skewed. Below histograms (Figure 2.2) show the distribution of our data.

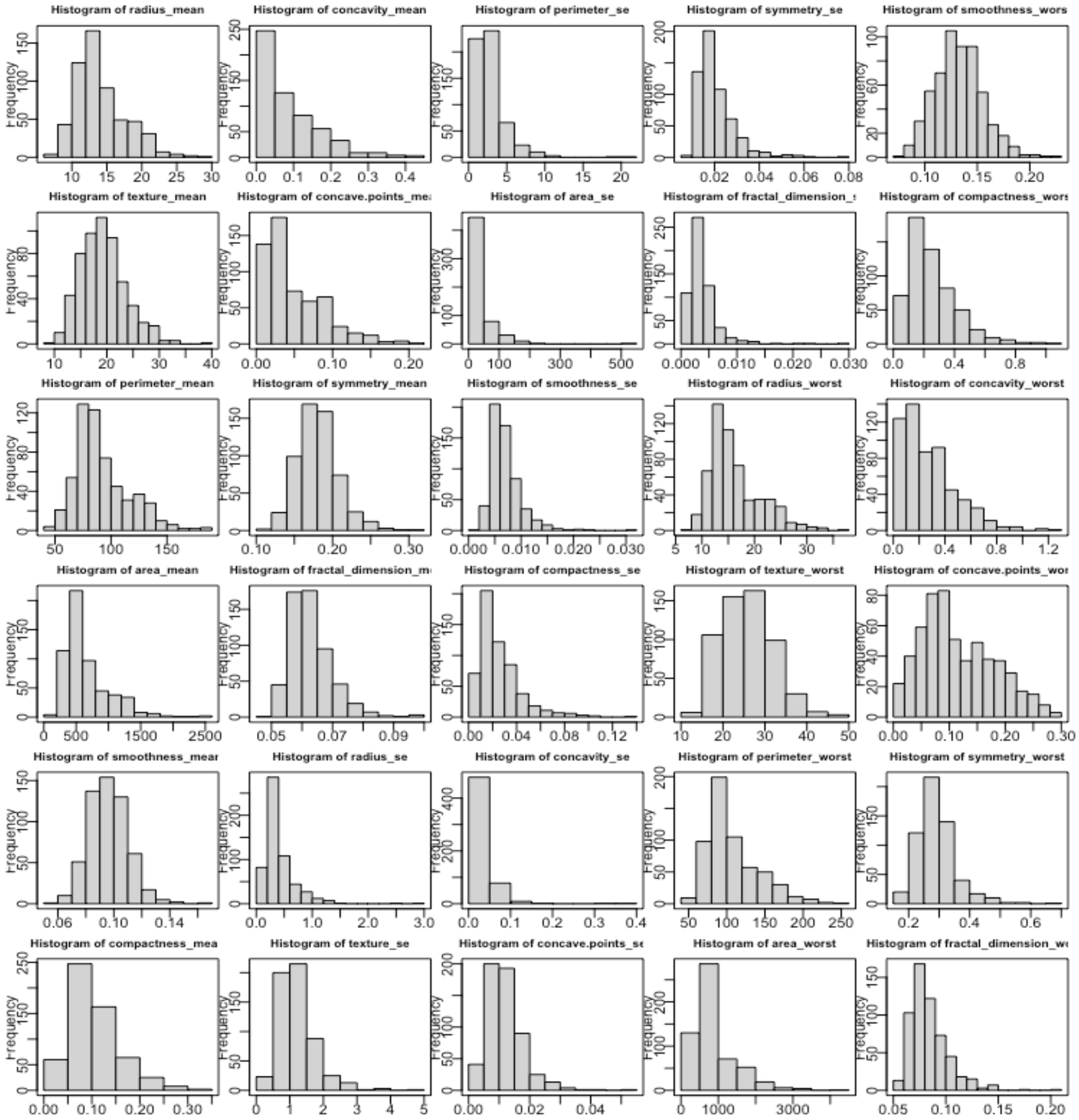


Figure 2.1: Histograms to understand the data distribution

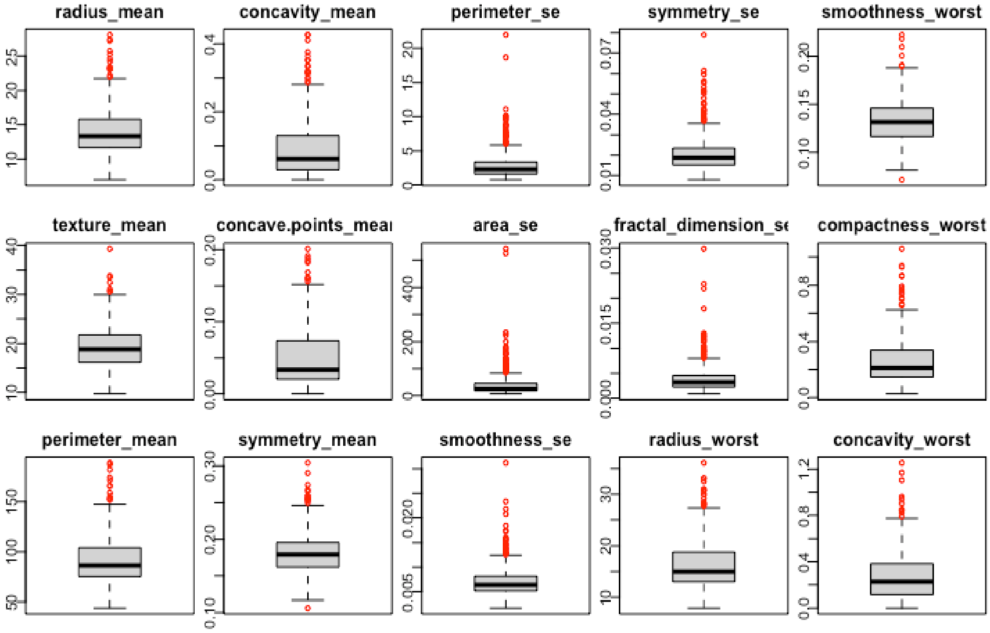
Also, the table 2.1 below has the values of skewness. Strongly skewed values are those greater than 1 or less than -1. There is a moderate skewness in the values between -1 and -1/2 and between 1 and 1/2.

Variable	Skewness	Variable	skewness	Variable	skewness
radius_mean	0.9374168	symmetry_mean	0.7217877	concavity_se	5.0835502
texture_mean	0.6470241	fractal_dimension_mean	1.2976191	concave.points_se	1.4370701
perimeter_mean	0.9854334	radius_se	3.0723468	symmetry_se	2.1835728
area_mean	1.6370654	texture_se	1.6377733	fractal_dimension_se	3.9033041
smoothness_mean	0.4539207	perimeter_se	3.4254803	radius_worst	1.0973059
compactness_mean	1.1838556	area_se	5.4185001	texture_worst	0.4956970
concavity_mean	1.3938008	compactness_se	1.8922032	perimeter_worst	1.1222227
concave.points_mean	1.1650124	smoothness_se	2.3022616	area_worst	1.8495814
smoothness_worst	0.4132383	compactness_worst	1.4657948	concavity_worst	1.1441794
concave.points_worst	0.4900213	symmetry_worst	1.4263764	fractal_dimension_worst	1.6538237

Table 2.1: Skewness values of continuous predictors

2.3 Handling Outliers

Outliers are detected using boxplots. The red circles in the below plots (Figure 2.2) show the outliers for each predictor variable. The data indicates that it has been tampered with and later needs to be handled. Spatial sign transformation helps to remove these outliers.



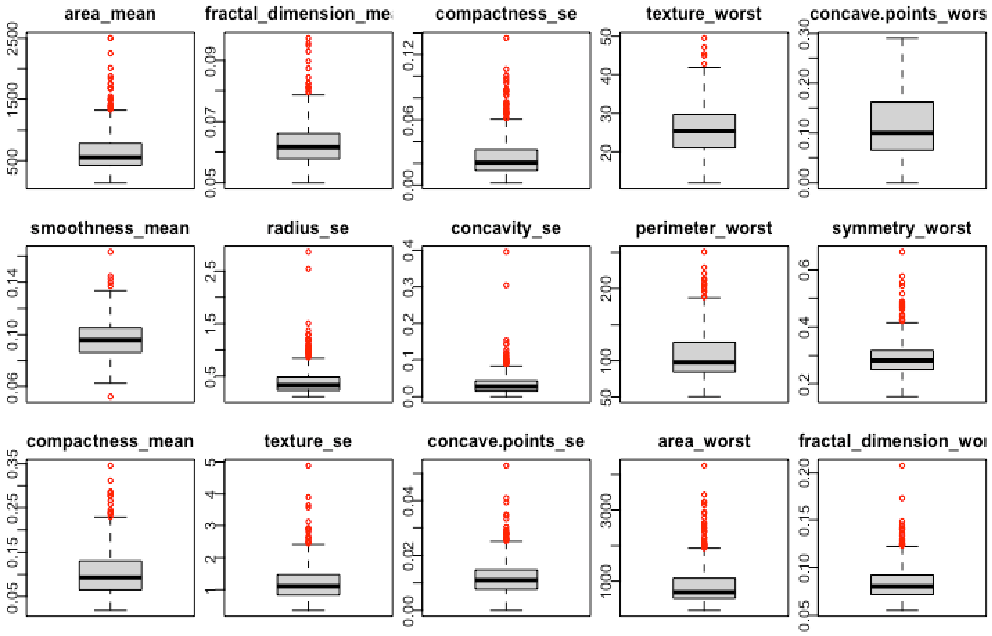


Figure 2.2: Box Plots to visualize outliers

2.4 Data Transformation

Following the transformation, we used box-cox and spatial sign transformations (Figure 2.3) to remove skewness and outliers, respectively. The below table 2.2 contains the values of skewness after applying transformation methods. Here, most of the predictors are approximately symmetric and range between 0 and 0.5 or -0.5. Also, the outliers have vanished.

Variables	Skewness	Variables	Skewness	Variables	Skewness
radius_mean	0.03873987	radius_se	-0.1288336	radius_worst	0.1202914
texture_mean	-0.0820146	texture_se	-0.2780549	texture_worst	-0.0666335
perimeter_mean	0.02635615	perimeter_se	-0.1869524	perimeter_worst	0.1117393
area_mean	0.2105584	area_se	0.03640982	area_worst	0.1505764
smoothness_mean	-0.0180633	smoothness_se	-0.1716217	smoothness_worst	-0.0009101
compactness_mean	-0.1167777	compactness_se	-0.0326025	compactness_worst	-0.0635906
concavity_mean	0.6120693	concavity_se	1.206639	concavity_worst	0.5543184
concave.points_mean	0.57593	concave.points_se	0.5650721	concave.points_worst	0.3180021
symmetry_mean	-0.1675667	symmetry_se	-0.1650464	symmetry_worst	0.00235428
fractal_dimension_mean	-0.0573499	fractal_dimension_se	-0.2021556	fractal_dimension_worst	-0.0170629

Table 2.2: Skewness values post box-cox transformation

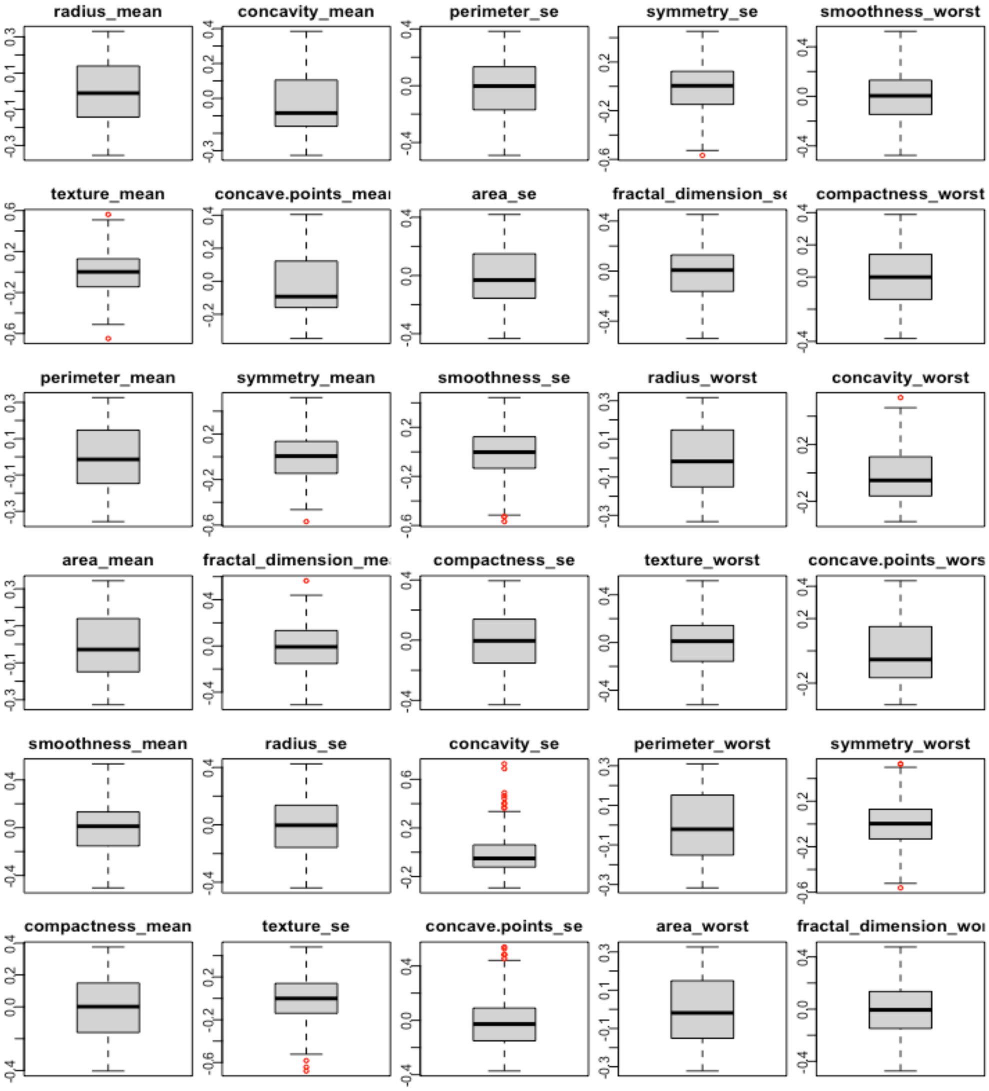


Figure 2.3: Box Plots post spatial sign transformation

2.5 Correlation

Identifying the relationship between the predictors is easier with the below plot (Figure 2.4). With a 90% cutoff value, we identify ten highly correlated predictors. We checked with 75% and 85% cutoff values as well, but we have a huge data loss with these two cutoff values. Principle component analysis helped determine if any significant predictors had been excluded and found that none of the ten variables were vital. The first ten components explain 95% of variance and it can visualize using scree plot (Figure 2.5)

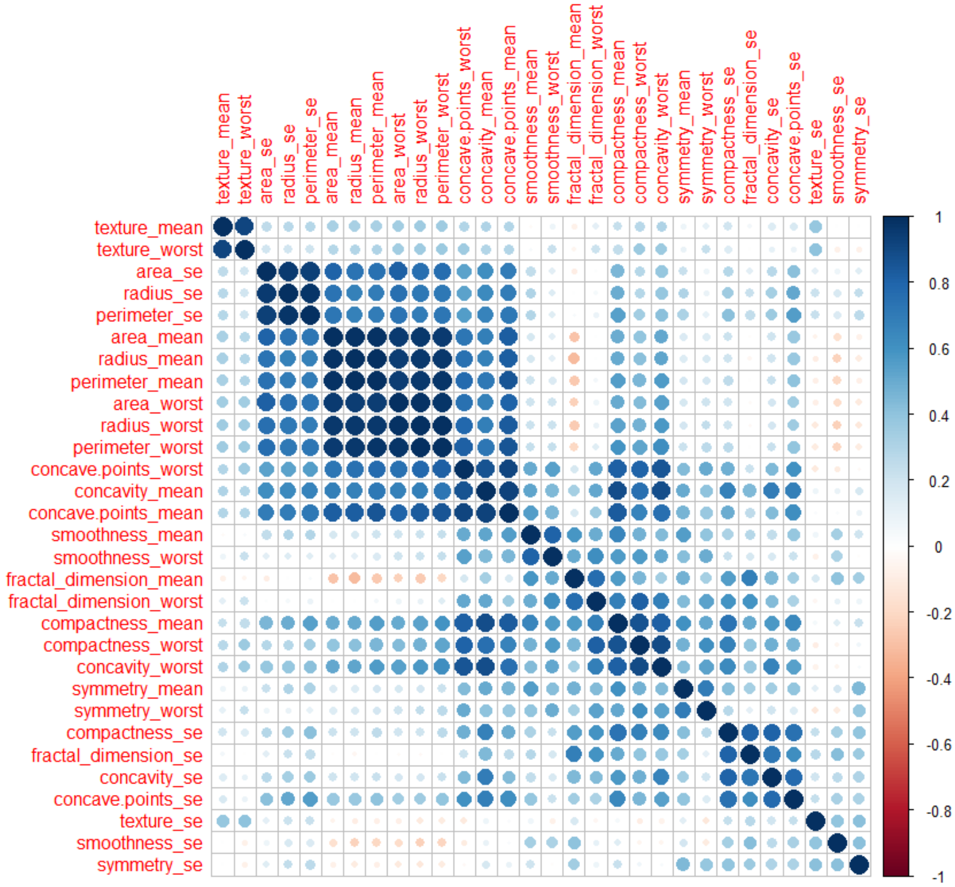


Figure 2.4: Correlation plot to detect multicollinearity

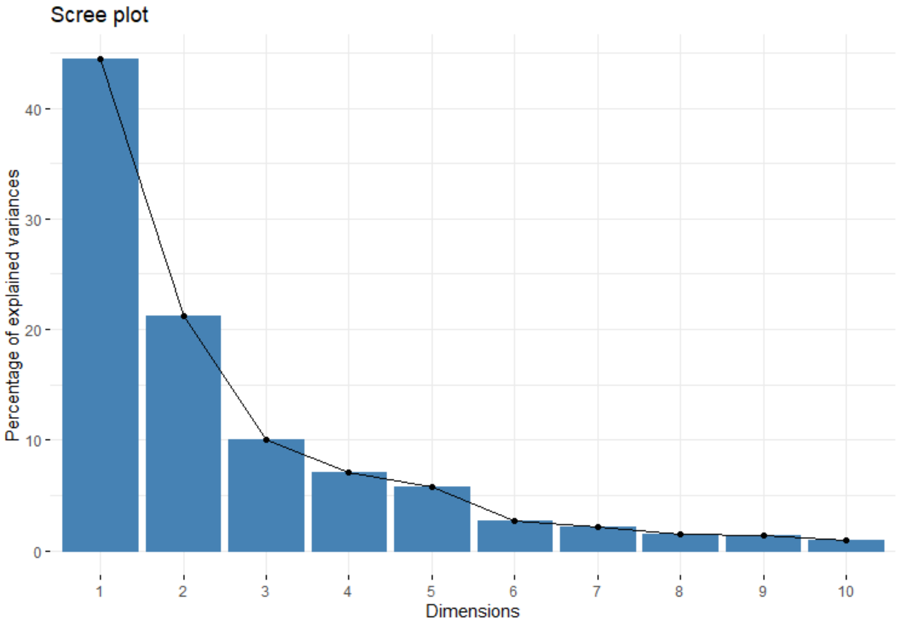


Figure 2.5: Scree Plot to visualize top 10 PC's

2.6 Data Splitting

The response variable diagnosis has two labels Benign (no cancer) and Malign (cancer). There is a huge difference in frequency of each class label, 62.8% for benign and 37.2% for malign which indicates that the response variable is pretty imbalanced. The distribution of each class label is shown in the frequency plot, Figure 2.6. Therefore, stratified random sampling technique has been used while splitting the data into training and testing sets. The ratio of splitting is considered the standard format of 80:20. The training set has 456 observations, and the test set has 113 observations.

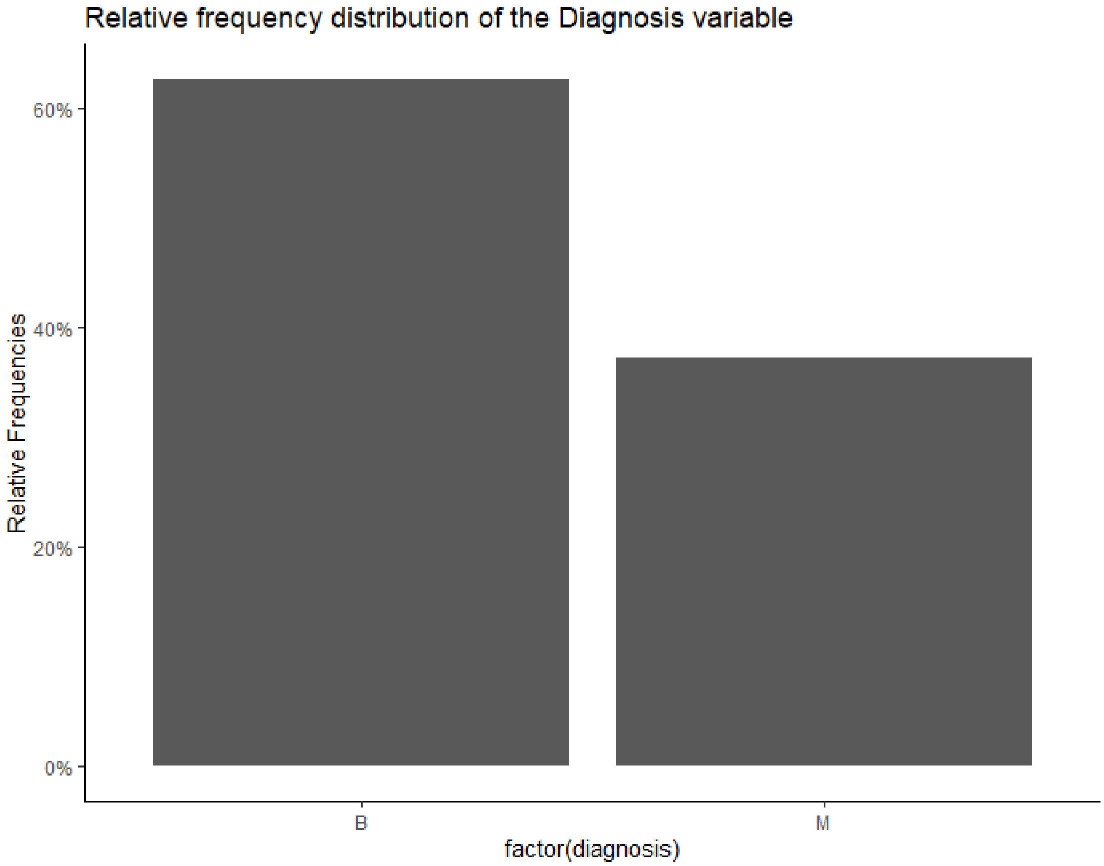


Figure 2.6: Distribution of response "diagnosis" variable

3. Model Development

To predict whether the tumor is malign or benign, we considered linear and non-linear classification models. As the response variable is imbalanced, accuracy as a test statistic may mislead the analysis. Therefore, we considered Kappa as a statistic measure to assess the performance of chosen classification models.

3.1 Linear Classification Models

The models which classify the data into respective labels through-line or plane or hyperplane using the linear combination of input features are known as linear classification models. In this project, we developed five linear classification models on the training set and selected the two best models for the test set, considering kappa as a statistic measure. While training the models, the 10-fold cross-validation technique was considered for resampling. The tuning parameters of each model are provided in Appendix 1 for reference. The summary statistics of training data are shown in the Table 3.1.

MODEL	TUNING PARAMETER	ACCURACY	KAPPA
LOGISTIC REGRESSION	No tuning parameter	0.9603	0.9157
LINEAR DISCRIMINANT ANALYSIS	No tuning parameter	0.9538	0.8989
PLSDA	ncomp = 4	0.9560	0.9040
GLMNET	α = 0.1, λ = 0.01	0.9604	0.9141
SPARSE LDA	NumVars = 4, λ = 0.01 (held constant)	0.9517	0.8941

Table 3.1: Summary statistics of Linear Classification Models (training set)

There are no tuning parameters for logistic regression and linear discriminant analysis models. The lambda is held constant at 0.01 for the Sparse LDA model. The performance assessment metric Kappa is higher for Logistic Regression and GLMNET models.

3.2 Non-Linear Classification Models

The models which are not able to classify the data through linearly separable lines are known as non-linear classification models. In this project, we developed eight non-linear classification models on the training set and selected the two best models for the test set, considering kappa as a statistic measure. While training the models, the 10-fold cross-validation technique was considered for resampling. The tuning parameters of each model are provided in Appendix 2 for reference. The summary statistics of training data are shown in the Table 3.2.

MODEL	TUNING PARAMETER	ACCURACY	KAPPA
MIXTURE DISCRIMINANT ANAYSIS	subclasses = 6	0.9692	0.9330
REGULARIZED DISCRIMINANT ANALYSIS	ϒ = 0, λ = 0.0733	0.9582	0.9099
NEURAL NETWORK	size = 6, decay = 0.1	0.9779	0.9525
SUPPORT VECTOR MACHINE	σ = 0.0750 (held constant), C = 4	0.9669	0.9297
FLEXIBLE DISCRIMINANT ANALYSIS	degree = 1, nprune = 5	0.9604	0.9138
K-NEAREST NEIGHBORS	K = 3	0.9582	0.9098
NAÏVE BAYES	fl = 2, usekernel = True, adjust = True (All held constant)	0.9428	0.8779
RANDOM FOREST	mtry = 2	0.9626	0.9194

Table 3.2: Summary statistics of Non-Linear Classification Models (training set)

The Naïve Bayes model doesn’t require pre-processing steps and all tuning parameters were held constant. The performance assessment metric Kappa is higher for Neural Network and MDA models.

4. Model Testing

We considered the two best models each from linear and non-linear classification models. The statistic measure Kappa is used as a test statistic for the final model. The test results (confusion matrix) of four models are shown in Table 4.1.

MODEL	SPECIFICITY	SENSITIVITY	PPV	NPV	ACCURACY	KAPPA
LOGISTIC REGRESSION	0.8810	0.9577	0.9315	0.9250	0.9292	0.8469
GLMNET	0.8810	0.9718	0.9324	0.9487	0.9381	0.8654
NEURAL NETWORK	0.9048	0.9718	0.9452	0.9500	0.9469	0.8852
MIXTURE DISCRIMINANT ANALYSIS	0.8571	0.9859	0.9211	0.9730	0.9381	0.8641

Table 4.1: Summary of top 4 classification models (test set)

From the test results, we concluded that the Neural Network model is the best model which has the highest Kappa value. No model is near to that value that the neural network model achieved. The computational complexity and time complexity are higher for the neural network model as compared to the logistic regression model. As we considered the test statistic as Kappa for performance assessment, the neural network model is considered the best model for evaluating whether the tumor is benign or malign. The confusion matrix of neural network model is shown in Table 4.2.

	REFERENCE	
PREDICTION	**B**	**M**
B	69	4
M	2	38

Table 4.2: Confusion matrix of neural network model

Based on variable importance, the top 5 predictors for the neural network model are shown in Figure 4.1.

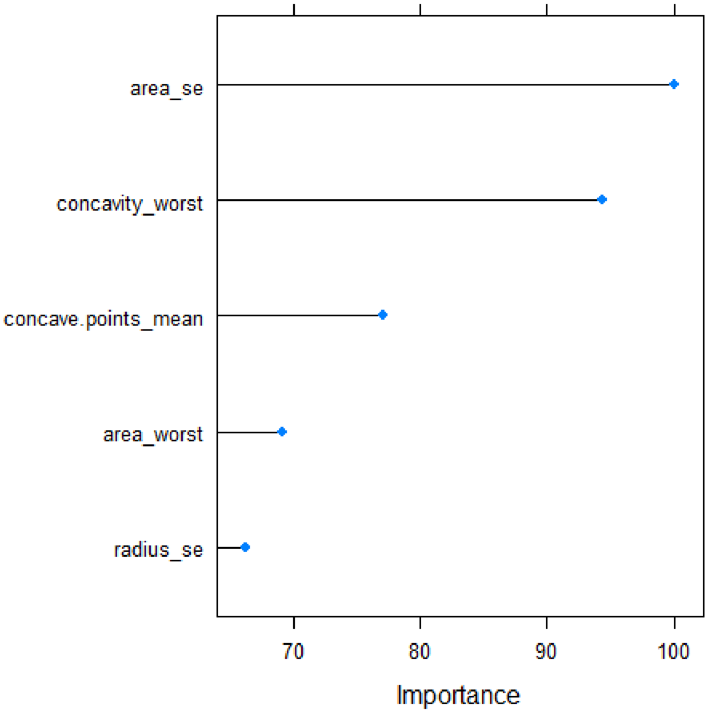


Figure 4.1: Top 5 predictors of cancer dataset

5. Future Work

There is a huge margin difference in training and testing statistic measures. We are trying to identify the reason by verifying each observation and data splitting strategy. The neural network model achieved the kappa of 0.8852. We are looking forward to improving the statistic measure up to 95% approximately.

6. Conclusion

To predict whether the tumor is benign or malign, we have taken a dataset from the UCI repository. We performed data pre-processing steps for skewness, outliers, missing values, and duplicate variables. We applied a box-cox transformation to mitigate the problem of skewness, and spatial sign transformation for outliers and removed highly correlated variables with a cutoff of 0.90. Post-pre-processing, we are left with 21 predictors and 1 response variable. The data split in standard notion 80:20 with applying stratified random sampling technique. The models developed for both linear and non-classification models, applying 10-fold cross-validation as a resampling technique, selected the best two models from each section based on statistic measure kappa. The best four models are tested on test data and finalized the neural network model achieving the highest kappa value of 0.8852 as compared to the other three models.

Appendix

1.1 Linear Classification Models

Partial Least Squares Discriminant Analysis

The final value used for the model was ncomp = 4.

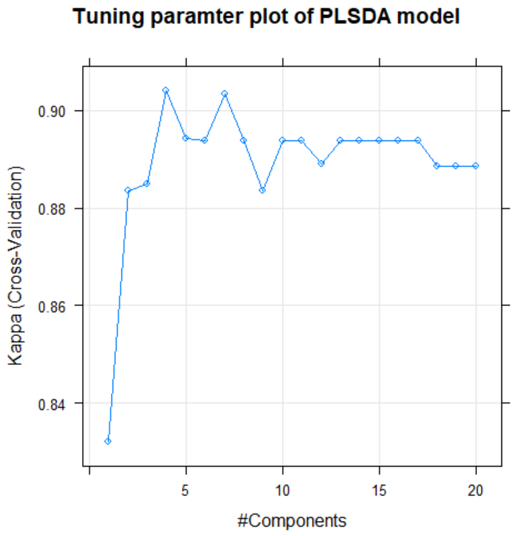


Figure: Plot of tuning parameter of PLSDA model

GLMNET

The final values used for the model were alpha = 0.1 and lambda = 0.01.

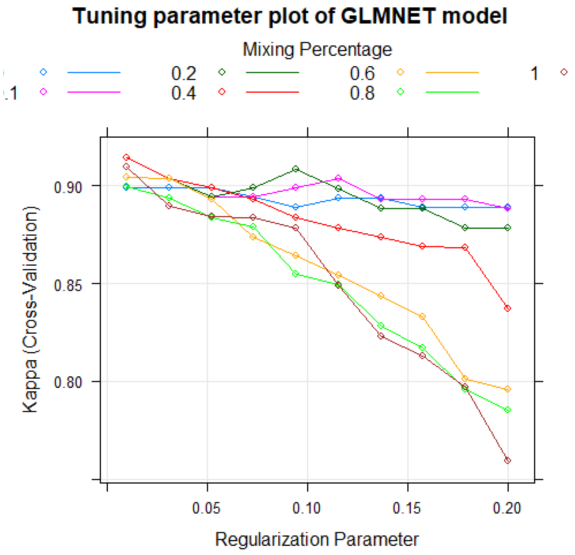


Figure: Plot of tuning parameter of GLMNET model

sparseLDA

Tuning parameter 'lambda' was held constant at a value of 0.01

The final values used for the model were NumVars = 4 and lambda = 0.01.

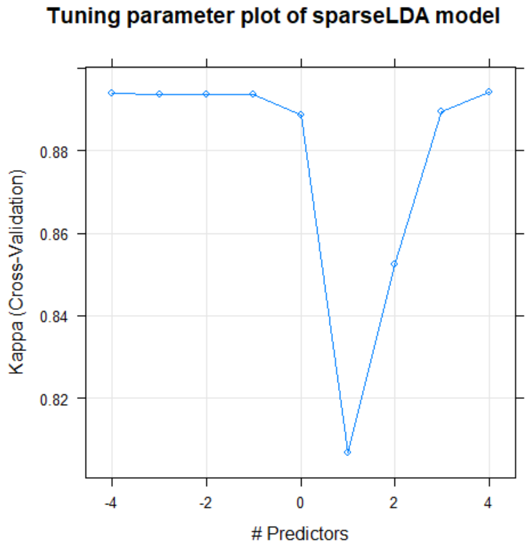


Figure: Plot of tuning parameter of sparseLDA model

1.2 Non-Linear Classification Models

Mixture Discriminant Analysis

The final value used for the model was subclasses = 6.

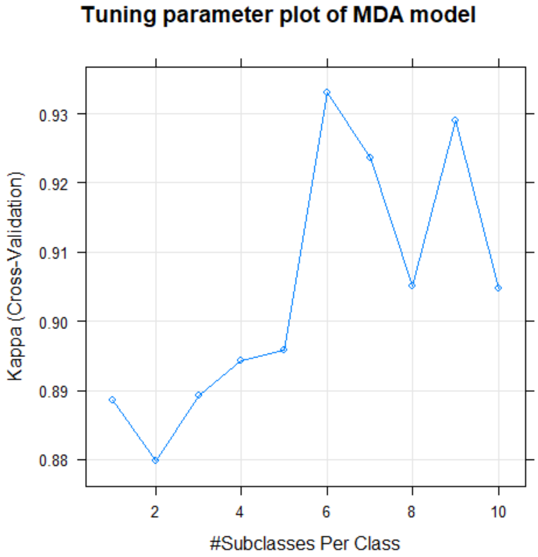


Figure: Plot of tuning parameter of MDA model

Regularized Discriminant Analysis

The final values used for the model were gamma = 0 and lambda = 0.0733.

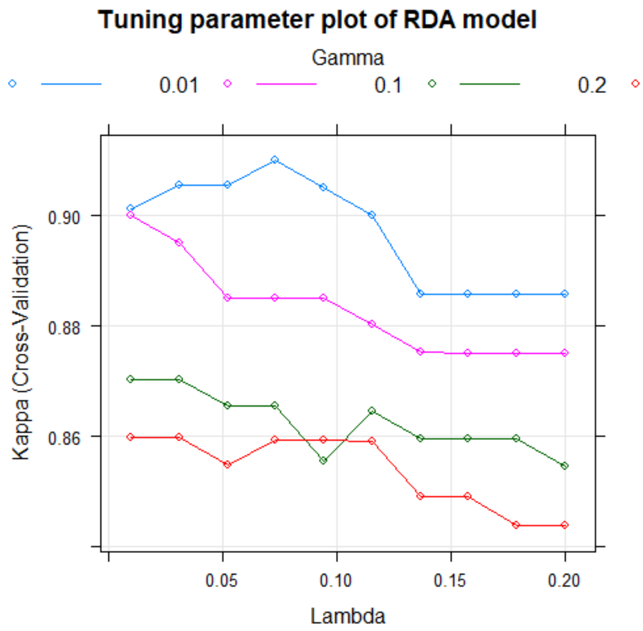


Figure: Plot of tuning parameter of MDA model

Neural Network Model

The final values used for the model were size = 6 and decay = 0.1.

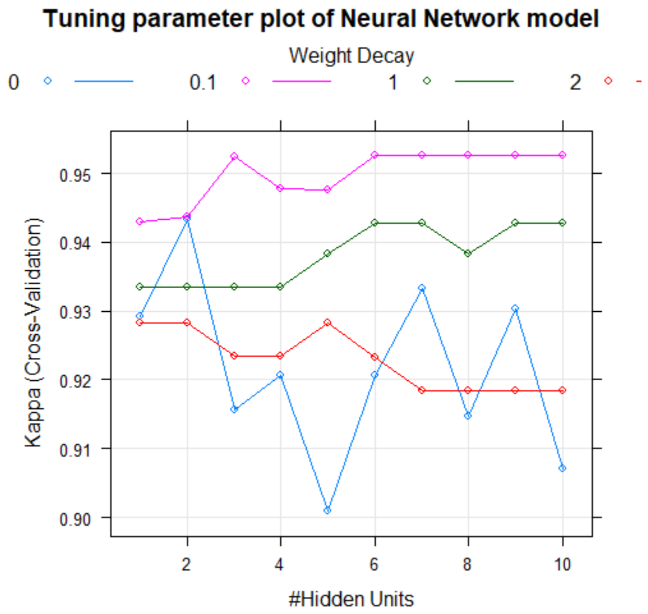


Figure: Plot of tuning parameter of Neural Network model

Support Vector Machine Model

Tuning parameter 'sigma' was held constant at a value of 0.0750.

The final values used for the model were sigma = 0.0750 and C = 4.

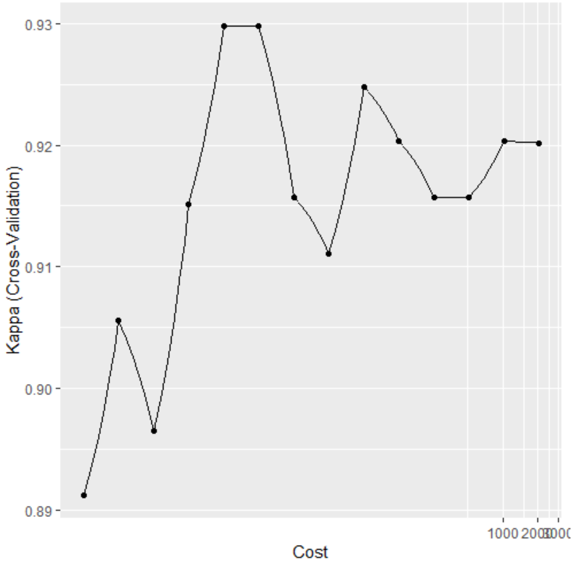


Figure: Plot of tuning parameter of SVM model

Flexible Discriminant Analysis Model

The final values used for the model were degree = 1 and nprune = 5.

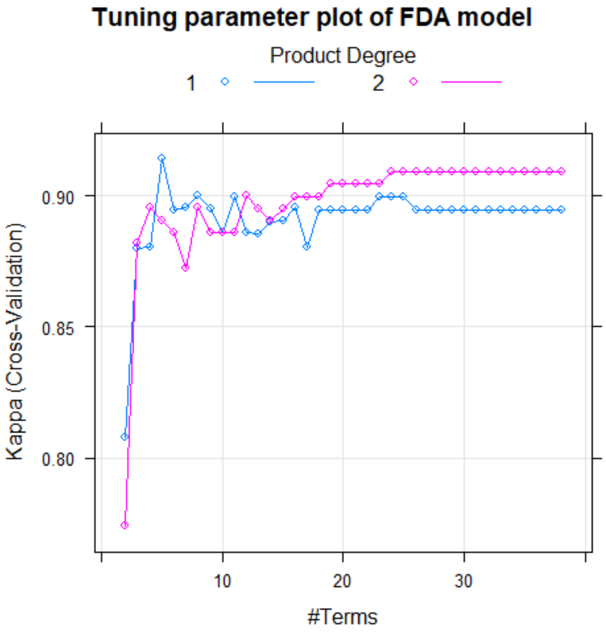


Figure: Plot of tuning parameter of FDA model

K-Nearest Neighbors

The final value used for the model was k = 3.

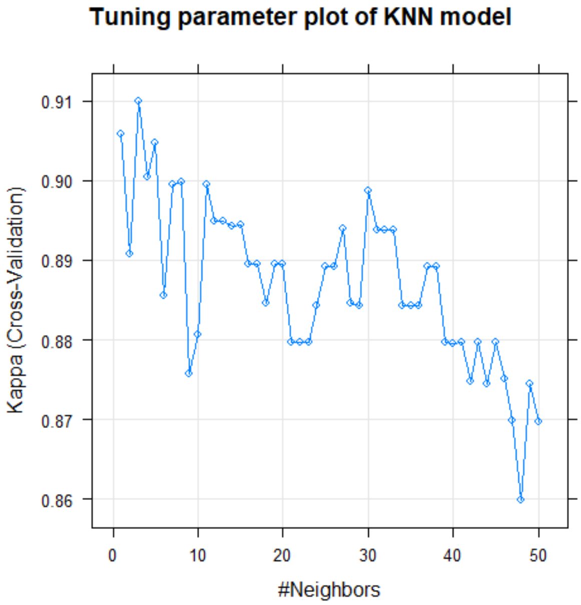


Figure: Plot of tuning parameter of KNN model

Random Forest

The final value used for the model was mtry = 2.

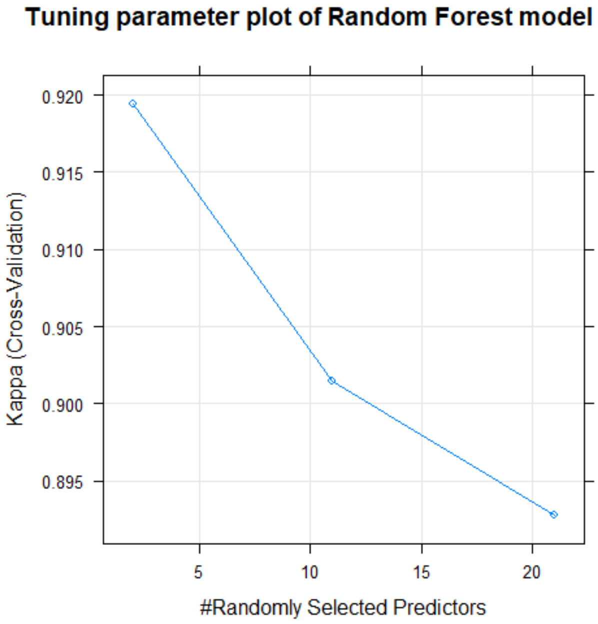


Figure: Plot of tuning parameter of Random Forest model