Statistical Data Mining Fall 2019

**Life Insurance Risk Assessment**

horizontal line

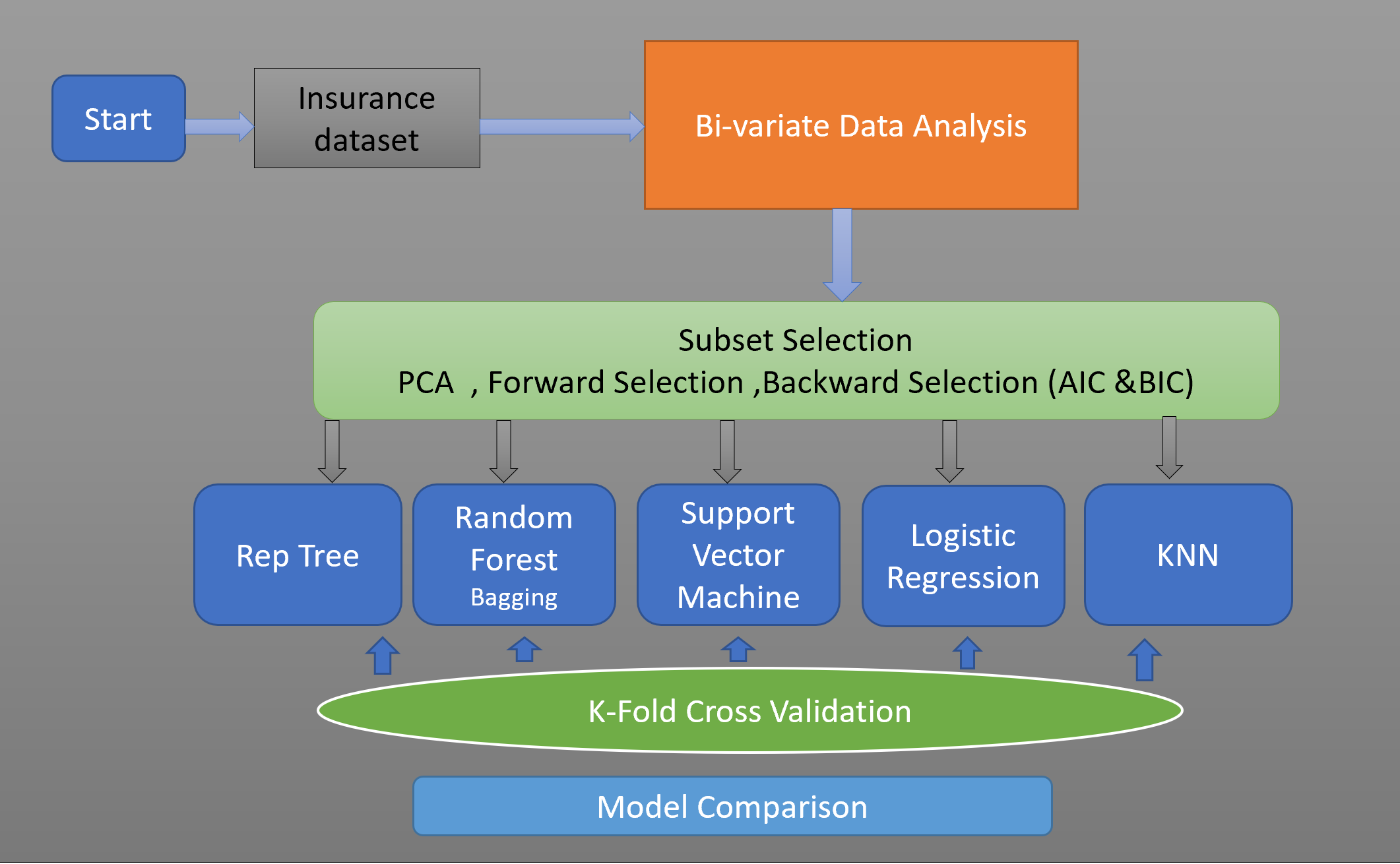
# Introduction

This project targets on building a Life insurance predictive analytics on underwriting process which involves potential customer risk assessment analysis. This involves risk classification of a new customer based on their personal details as well as medical and financial history. This improve decision-making and make the underwriting process faster and more economical. Therefore , helps the life insurance business to enhance customer acquisition and customer retention.

**Abstract**

Comparative Analysis of Ensemble and Non-Ensemble methods of prediction on a Prudential life insurance potential customer dataset is the prime focus of this project. The data provides us an insight of the potential customer whose case is to be assessed . The data is analysed to remove the outliers , bi-variate analysis is performed to find out highly correlated variables. Data is reduced to best subset using various mechanisms and then it is fit into models which is validated using k fold cross validation. The models are therefore compared in terms of mean squared test error, sensitivity , time taken to build the model and classification accuracy. The risk class is ranging from 1 to 8 , so the ensemble techniques gives good model interpretation and for this particular dataset , we see that the model performance is decent as compared to ensemble methods , but not satisfactory.

**Project Overview**



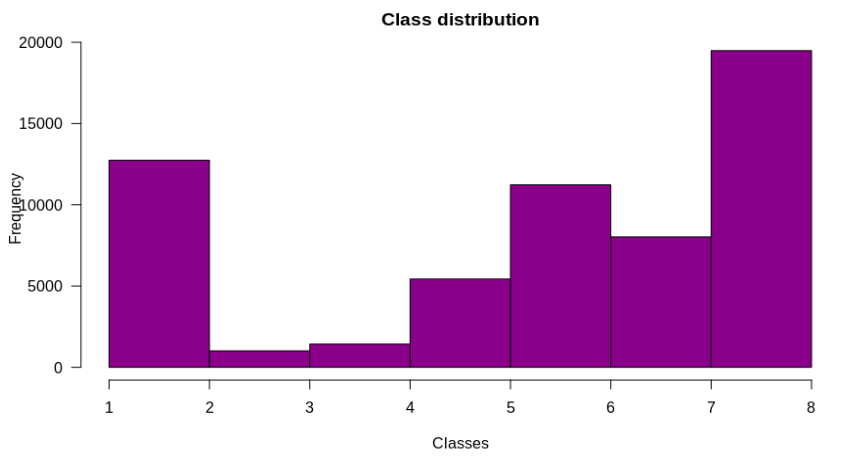
**Exploratory Data Analysis and Data Cleaning**

We performed data cleaning as a part of Exploratory Data Analysis on the given dataset in the following manner:

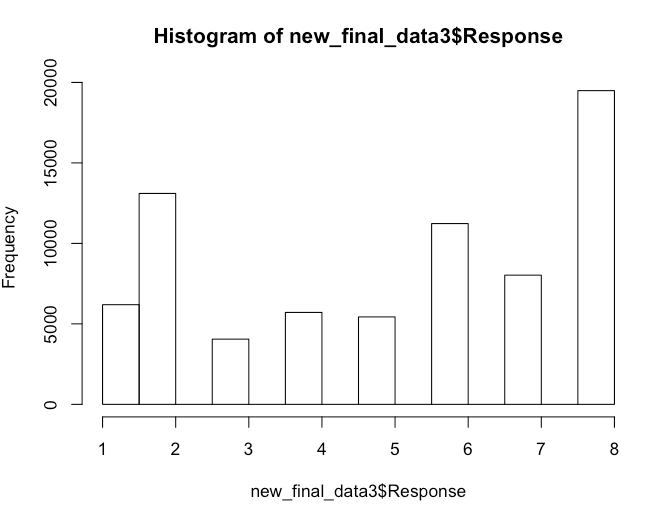
* First traverse through all the columns to find the NAs and the missing data
* Findings say that 12 columns have NAs in their respective data
* In 11 columns the count is more than 10% of the complete data
* The threshold is set at 10% and columns with more than 10% of NAs have been removed
* The rows with NA values are dropped as well
* The data now is devoid of any NA values

We then calculated the collinearity of the predictors using the correlation matrix and set the intra-correlation threshold at 0.7 and treat the predictors highly correlated if their correlation is above 0.7.This is summarised by the following plot using GGPAIRS

We can therefore conclude that out of the following predictors, we can drop one of them and proceed with the rest to carry out feature selection which is described in Feature Selection process section.The class distribution of the dataset is given by the following histogram::



It can be seen that there aren't as many samples for class 3(1013) and class 4(1428), as there are for other classes. This shows that the classes 2 and 3 are imbalanced as compared to other classes. The following histogram displays the class distribution after balancing.



This histogram shows the classes after data balancing, as we can see here that now the classes 2 and 3 are pretty balanced as compared to other classes.

**Performance measures**

Various scales are used to gauge the performance of the classifiers.

* Classification Accuracy .Any classifier could have an error rate and it may fail to categorize correctly. Classification accuracy is calculated as Correctly classified instances divided by total number of instances multiplied by 100.
* Mean Absolute Error Mean absolute error is the average of the variance between predicted and actual value in all test cases. It is a good measure to gauge the performance.
* Root Mean Square Error Root mean squared error is used to scale dissimilarities between values actually perceived and the values predicted by the model. It is determined by taking the square root of the mean square error.
* Confusion Matrix A confusion matrix encompasses information about actual and predicted groupings done by a classification system.
* The performance is tested out using the Training set as well as using different Cross Validation methods. The class is arrived by considering all 35 attributes of the dataset having total 116 predictors.

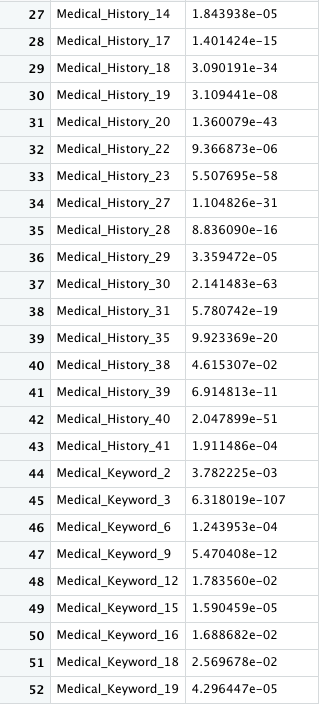
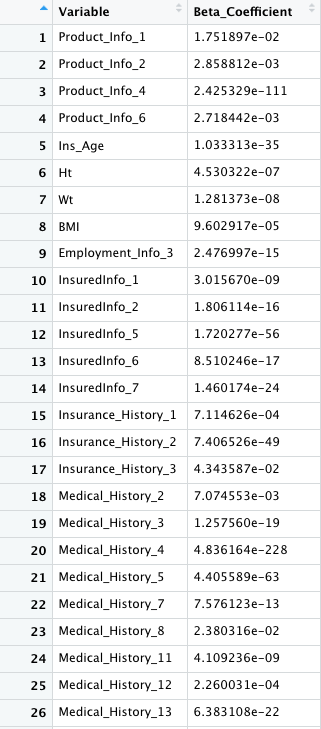
**Methodologies and Algorithms Implemented**

* KNN classification
* Multinomial Logistic Regression
* Polynomial Regression
* Principal Component regression
* Support Vector Machines
* Decision Tree
* Random Forests
* Adaboost

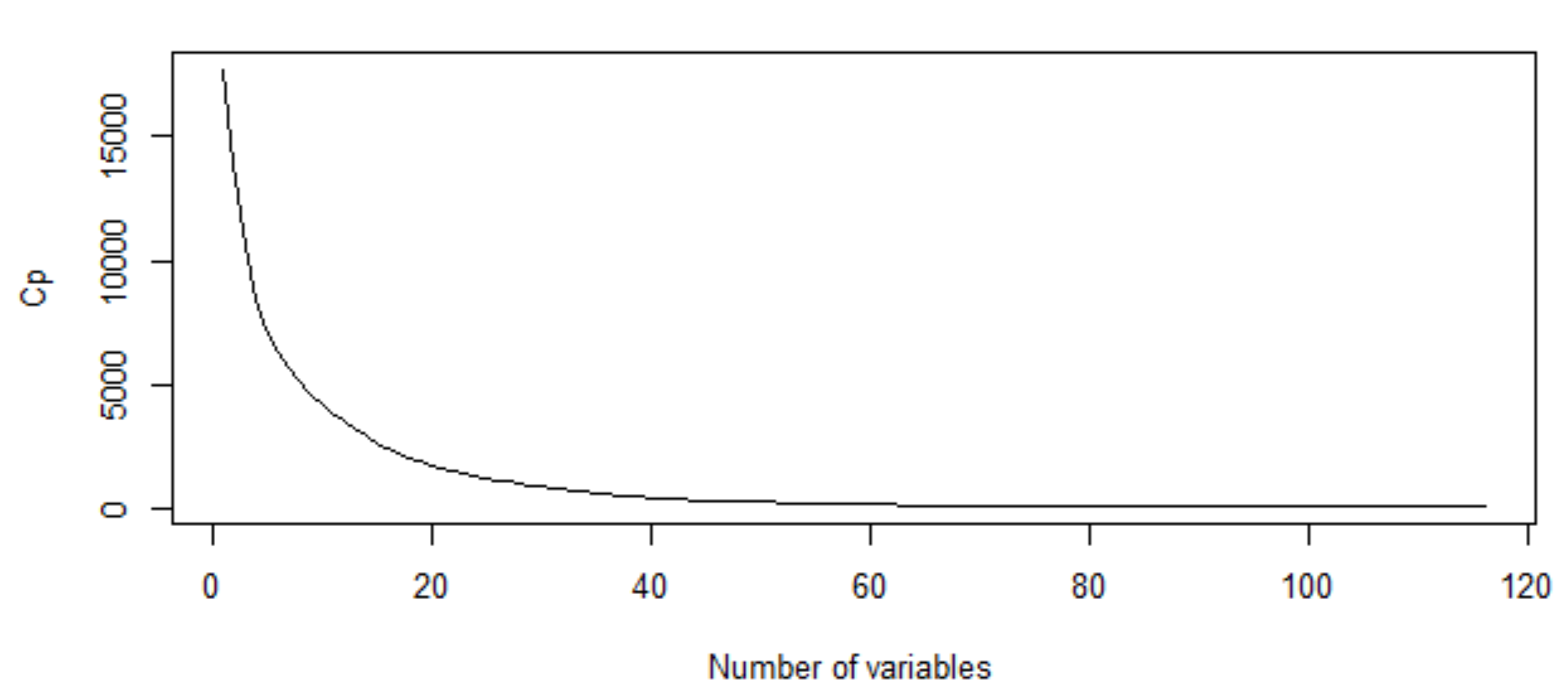
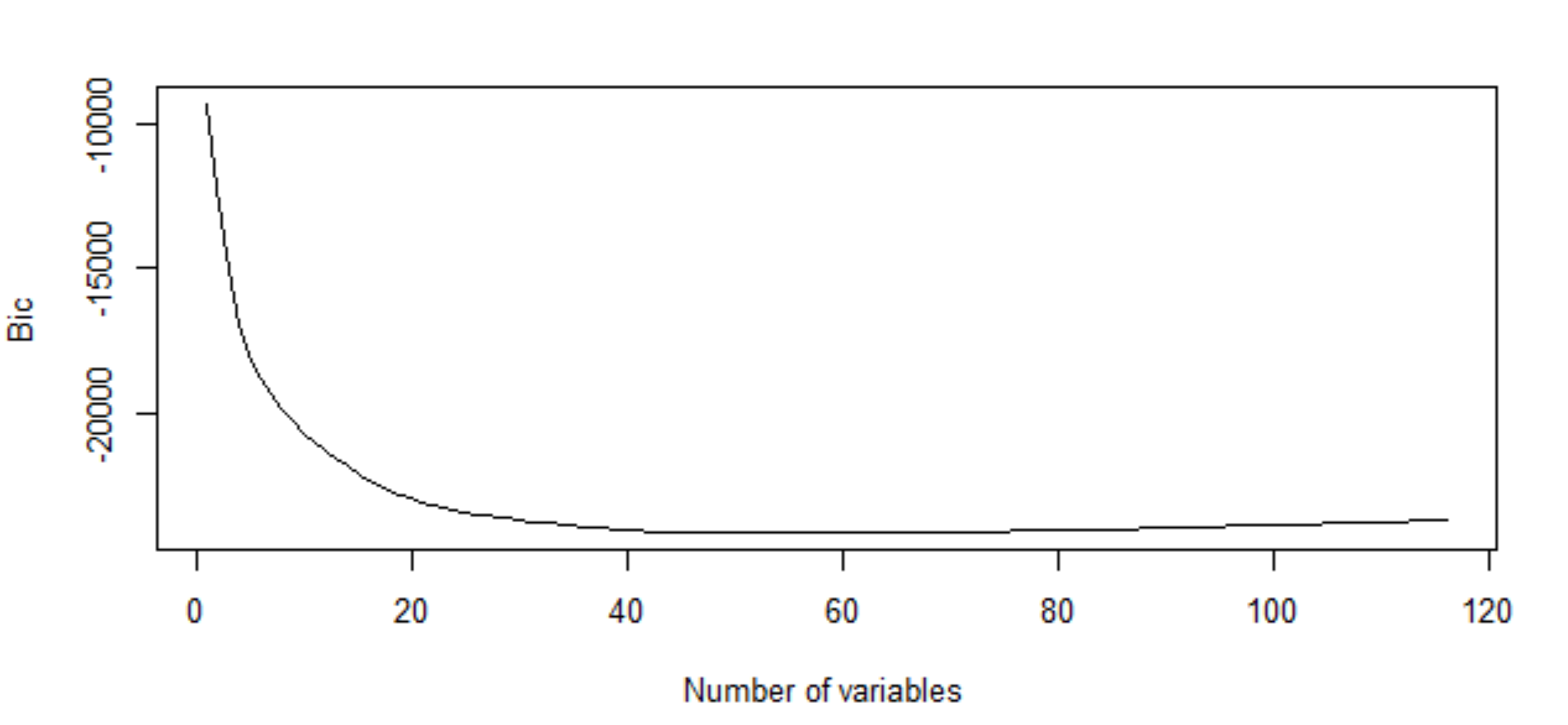
**Feature selection Process**

* Linear regression-p value significance
* Forward and Backward Subset selection - Based on CP and BIC statistics
* Variable importance plot via Random Forest

Ordinary least square model shows the predictors which are of highest significance. Based on these results , we arrive at the below 52 predictors .



Forward and backward subset selection algorithms are performed on the cleansed dataset and then the AIC and BIC graphs are plotted for upto 116 best subsets. Also the previous results using correlation matrix helped us in identifying which features to drop among all those predictorsBoth the given graphs arrive at similar results which shows that best subset of size 35 can be parsimonious model giving the comparable accuracy as compared to full model.



**Subset selected :** 35 predictors as shown in above graph :

**Model Performance Analysis**

This dataset is is trained using ensemble methods because we are dealing with multiclass classification problem here. Initially , we have trained the model using solitary decision tree and K fold cross validation is performed on it . Also the tree is pruned based on the minimum CP as it enables the removal of unnecessary branches which do not contribute to the overall error variations. This helps to avoid overfitting of the model and better interpretation.

Secondly Random forest and bagging ensemble techniques are used to train the model . It chooses the class based on committee machine which involves N number of trees which are created by bootstrapping the original dataset and the random subsets of predictors are used to choose the node for splitting the data. These newly generated trees are then utilized to classify the test sample and based on majority vote from the trees , the sample is classified. With this approach , we see a little increase in performance . Apart from that , the most important predictor turns out to be BMI and Ins\_Age as per the GINI variable importance graph.

Thirdly we use bagging technique with the mtry value selected as 6 , 7 , 8 as the default value is square root of number of predictors. Mtry helps up to select the number of predictor subsets to be chosen randomly for selection of the node at each step in tree growth.

Below is the results generated for each ensemble method. However , the dataset is huge ,So it takes a long CPU time to train the model with random forest. To make it a little faster , bagging is incorporated with various m values. The misclassification rate is high but it is comparatively better than the non ensemble methods such as logistic regression and LDA.

* **Rep Tree (Grow and prune a decision tree)**

|  |  |  |  |
| --- | --- | --- | --- |
| **K-fold Validation** | **MSE (Mean squared error)** | **Classification accuracy** (%) | **Time Taken build a model** |
| 5 fold | 6.329 | 49.94% | 28 seconds |
| 10 fold | 6.331 | 50.29% | 30 seconds |
| 20 fold | 6.328 | 50.27% | 60 seconds |
| 30 fold | 6.330 | 50.30% | 140 seconds |

* **Random forest**

**Unbalanced Data**

|  |  |  |
| --- | --- | --- |
| **Number of trees** | **MSE (Mean squared error)** | **Classification Accuracy** |
| 50 | 5.955 | 52.31 |
| 1000 | 5.988 | 53.06 |
| 2000 | 5.987 | 53.10 |
| 5000 | 5.983 | 53.24 |
| 10000 | 5.982 | 53.25 |

**Balanced data**

|  |  |  |
| --- | --- | --- |
| **Number of trees in committee machine** | **MSE (Mean squared error)** | **Classification Accuracy** |
| 50 | 3.873 | 65.07 |
| 1000 | 3.847 | 65.72 |
| 2000 | 3.912 | 65.71 |
| 5000 | 3.961 | 65.70 |
| 10000 | 3.994 | 65.71 |

* **Bagging(Number of Trees is set at 1000)**

**Unbalanced Data**

|  |  |  |
| --- | --- | --- |
| **M value selected** | **MSE (Mean squared error)** | **Classification Accuracy** |
| 6 | 5.954 | 53.06 |
| 7 | 5.889 | 53.06 |
| 8 | 5.960 | 53.08 |
| 9 | 5.977 | 52.73 |
| 10 | 5.989 | 52.75 |

**Balanced Data**

|  |  |  |
| --- | --- | --- |
| **M value selected** | **MSE (Mean squared error)** | **Classification Accuracy** |
| 6 | 4.882 | 58.74 |
| 7 | 4.821 | 61.11 |
| 8 | 4.331 | 64.05 |
| 9 | 4.112 | 65.07 |
| 10 | 3.987 | 65.72 |

**SVM Accuracy and Mean Squared Error**

**Unbalanced Data**

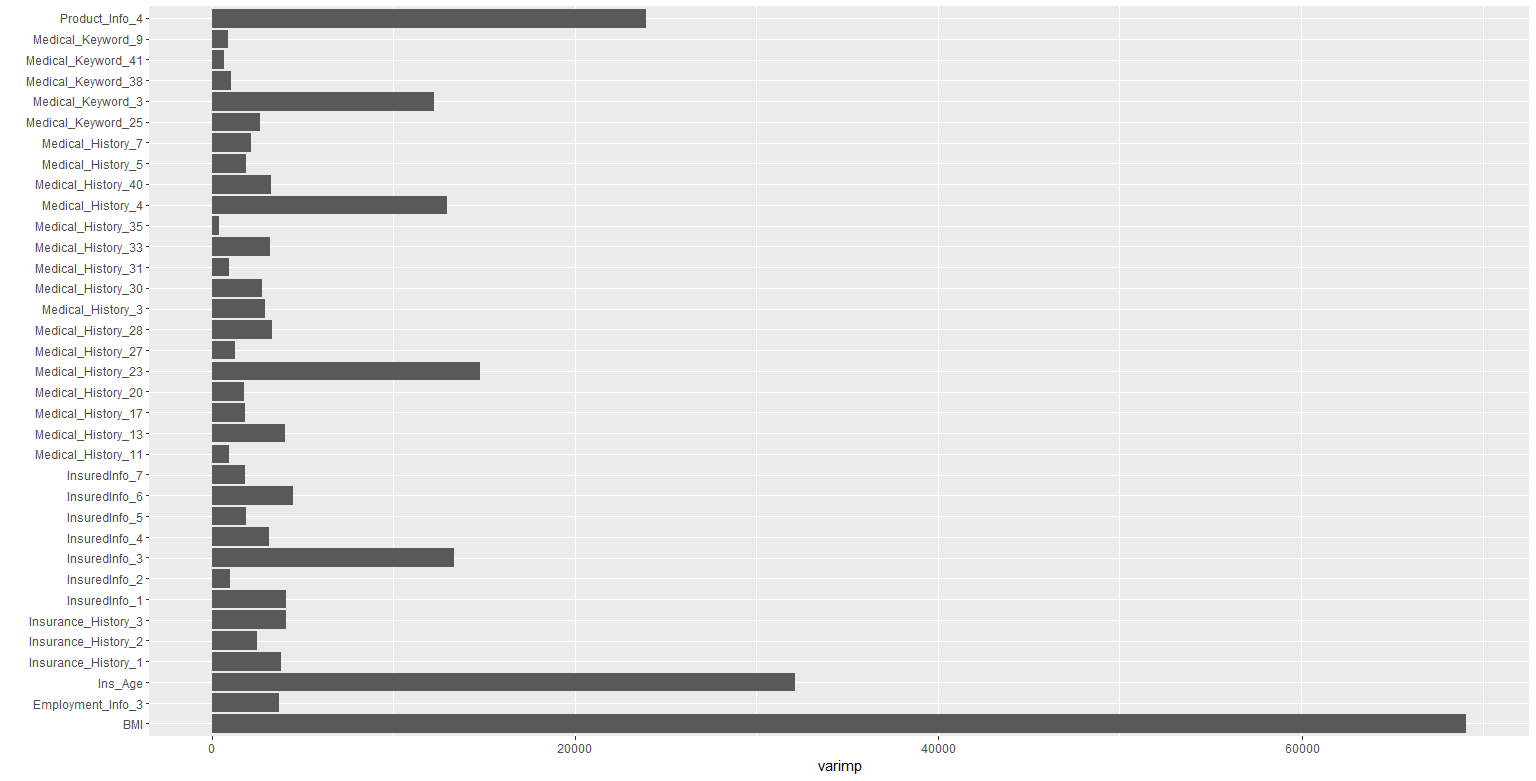
|  |  |  |
| --- | --- | --- |
| **Mode Selected** | **MSE (Mean squared error)** | **Classification Accuracy** |
| Linear | 5.729 | 47.60 |
| Radial | 5.601 | 49.47 |

**Balanced Data**

|  |  |  |
| --- | --- | --- |
| **Mode Selected** | **MSE (Mean squared error)** | **Classification Accuracy** |
| Linear | 6.109 | 44.74 |
| Radial | 5.872 | 48.04 |

**Logistic Regression**

|  |  |  |
| --- | --- | --- |
| **Mode Selected** | **MSE (Mean squared error)** | **Classification Accuracy** |
| Polynomial | 32.7084 | 42.84 |
| Multinomial | 32.7746 | 47.79 |

Variable importance plot using Random Forest

Accuracy Plot for all used methodologies for both balanced and unbalanced dataset

