Coronary Heart Risk Study

Table of Content

1. Introduction

The data set used in this project is provided by the medical practitioner. It consist of 16 variables viz. male, age, education, currentSmoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartRate, glucose, TenYearCHD. A brief summary of all the variables is given below.

Table-1.1: Variable Description

Category of variable	Name of the column	Meaning	Type of variable
Demographic	sex	It represents the sex of the patient whether male or female ("1" means Male, "0" means Female)	Nominal
	age	It represents the age of the patient	Continuous
	education	It represents the education level ("1" means High School, "2" means Bachelors, "3" means Masters and "4" means Professional/PhD)	Continuous
Behavioral	currentSmoker	It represents whether the person is a current smoker or not ("1" means Yes, "0" means No)	Nominal
	cigsPerDay	It represents average number of cigarettes that a person smoke in a day	Continuous
Medical(history)	BPMeds	It indicates whether the patient was under blood pressure medication or not ("1" means Yes, "0" means No)	Nominal
	prevalentStroke	It indicates that whether the patient had stokes or not ("1" means Yes, "2" means No)	Nominal
	prevalentHyp	It indicates that whether the patient was hypertensive or not ("1" means Yes, "2" means No)	Nominal
	diabetes	It indicates whether the patient had diabetes or not ("1" means Yes, "2" means No)	Nominal
Medical(current)	totChol	Represents total cholesterol level of the patient	Continuous
	sysBP	Represents systolic blood pressure of the patient	Continuous
	diaBP	Represents diastolic blood pressure of the patient	Continuous
	BMI	Represents body mass index of the patient	Continuous
	heartRate	Represents the heart rate of the patient	Continuous
	glucose	Represents glucose level in the patient	Continuous
Predict variable (desired target)	TenYearCHD	Represents 10 year coronary heart risk disease ("1" means Yes, "2" means No).	Nominal

1.1 Problem Statement:

- The aim of this project work is to analyze the data set received from the medical practitioner in order to identify whether the person either male or female will have a heart risk disease after ten years given the set of variables.
- The project work also identifies the most important variables which are the main causes of heart attacks and suggests the appropriate solution to reduce the risk of heart attack in the next ten years.

1.2 Problem Objective:

- The objective of this study is to find out the probability of a patient getting a heart attack.
- The objective of this study is to reduce the risk of patient getting a heart attack in the next ten year.
- Give some actionable recommendations to the Doctors, pharmaceutical companies, arranging awareness program/campaigns to spread the information regarding important factors for getting a heart attack.

2. Exploratory Data Analysis

2.1 Univariate Analysis

The univariate analysis was done only on the continuous variables such as – age, education, totChol, sysBp, diaBP, BMI, heartRate and glucose. Below is shown the histogram plot of the continuous variables (Fig.2.1 and Fig 2.2).

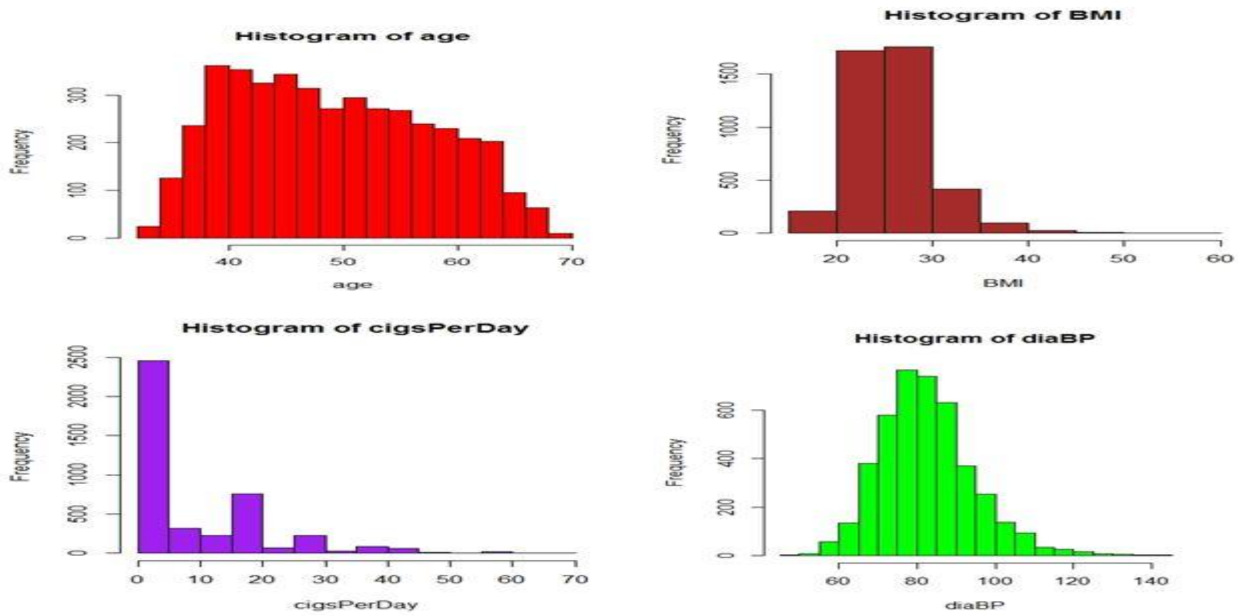


Fig.2.1: Histogram

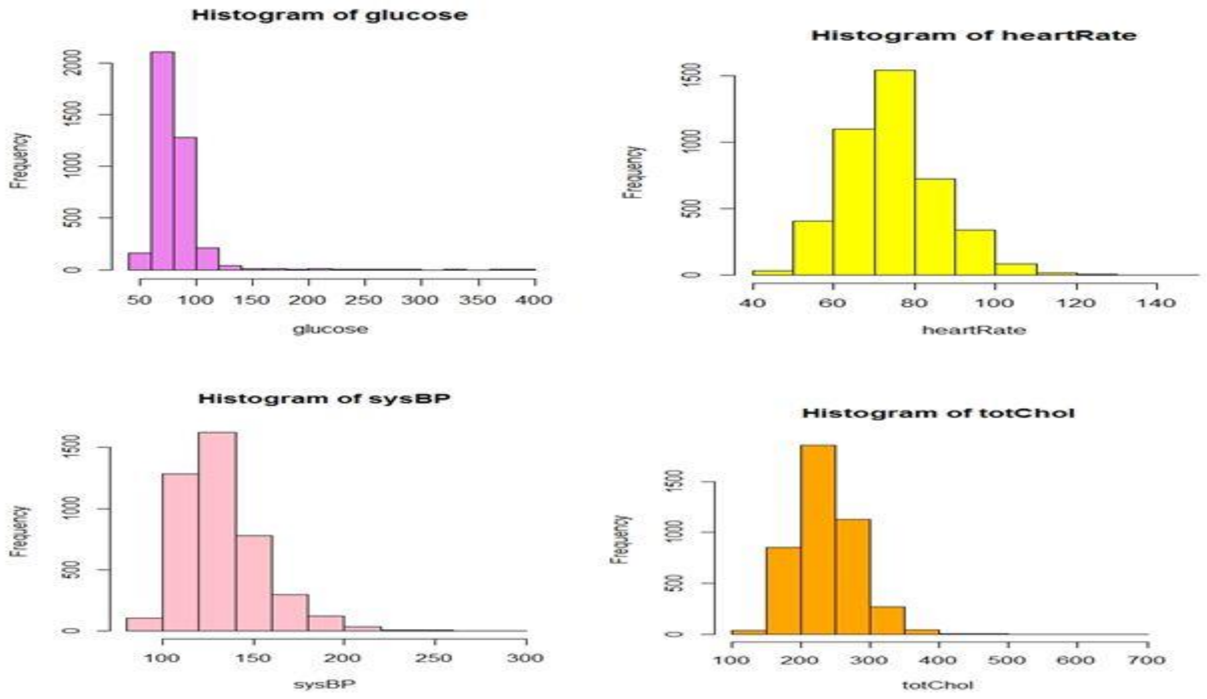


Fig.2.2: Histogram

- Variable age has a right skewed distribution (Fig.2.1).
- Variable BMI also has right skewness (Fig.2.1).
- Variable cigsPerDay has highly right skewed distribution (Fig.2.1).
- Variable diaBP has normal distribution (Fig.2.1).
- Variable glucose is also highly right skewed (Fig.2.2).
- Variable heartRate also has right skewed distribution (Fig.2.2).
- Variable sysBP has right skewness (Fig.2.2).
- Variable totChol is also right skewed (Fig.2.2).

2.2 Bivariate Analysis:

- Variable “age” is a continuous variable. It is not normally distributed over the entire range of distribution (Fig.2.3).
- Education v/s age plot shows that the number of patients who have high school education are less than the patients who have Bachelors, Masters and Professioal/PhD (Fig.2.3).
- The variable “totChol” has a right skewed distribution and it can also be seen from Fig.2.3 that the patients who are under blood pressure medication or diabetic have relatively higher level of total cholesterol.

- “sysBP” is a right skewed distribution. It can be seen (Fig.4) that the patients who are smokers and hypertensive have high systolic blood pressure.
- “diaBP” is a right skewed distribution. It can be seen (Fig.2.3) that the patients who are smokers and hypertensive have high diasystolic blood pressure.
- The variable BMI is a right skewed distribution. The scatter plot between BMI and diaBP shows that the patients with high diastolic blood pressure have high body mass index (Fig.2.3).
- The plot between BMI and sysBP shows that patients with high BMI have high value of systolic blood pressure (Fig.2.3).
- The plot between “sysBP and “diaBp” indicates that the patients who are having high diastolic blood pressure also have high stolic blood pressure (Fig.2.3).

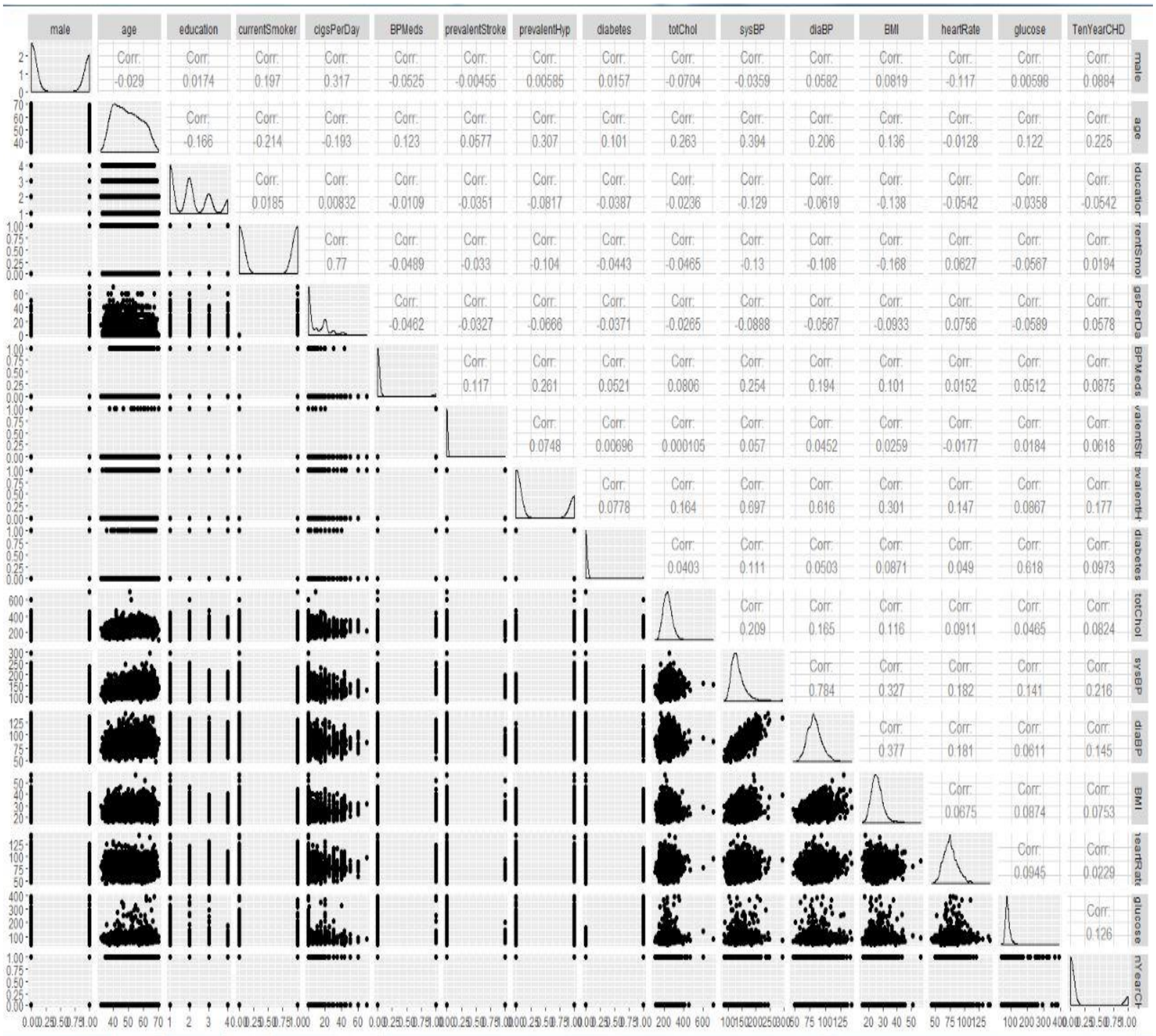


Fig.2.3: Bivariate Analysis graphs

2.3 Correlation:

- For correlation plot, we are considering only the continuous variables viz. totChol, sysBP, diaBP, BMI, heartRate, glucose, age, education and cigsPerDay.
- The correlation plot is shown in the Fig.2.4.
- From Fig.5 it can be seen that
  - Variables sysBP and diaBP has 78% correlation.
  - Variables sysBP and age has 39% correlation.
  - Variables and age has 26% correlation

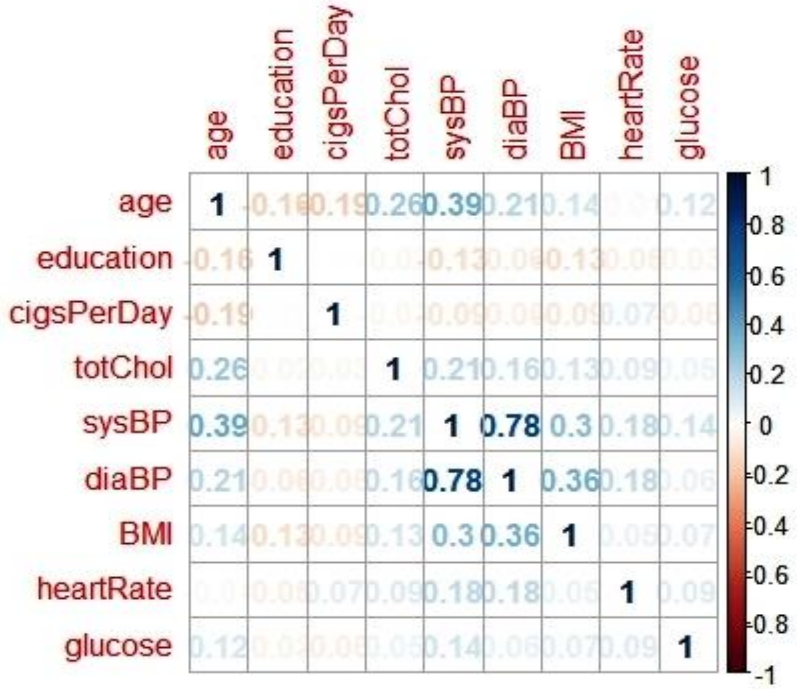
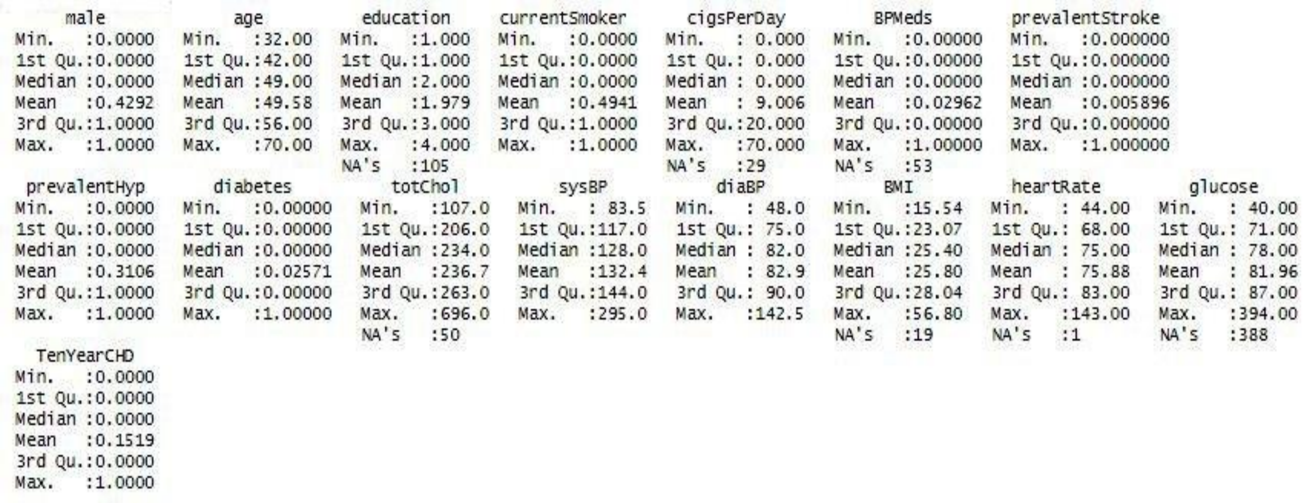


Fig.2.4: Correlation plot

2.4 Summary of EDA:

- Age is a continuous variable with minimum value of 32 and maximum value of 70. It has mean of 49.58.
- The data set contains the “NA” values for the columns “educations”, “cigsPerDay”, “BPMeds”, “totChol”, “BMI”, “heartRate”, “glucose”.
- totChol has a minimum value of 107 and maximum value of 696. It also has a mean value of 236.7.
- glucose has a minimum value of 40 and a maximum value of 394. The mean value of glucose is 81.96.
- heartRate is also found to be a very important variable as it has large variation between minimum and maximum value. The minimum value is 44 and maximum value is 143. The mean value of heart rate is 75.88.

- diaBP is also a variable of consideration. It has a minimum value of 48 and maximum value of 142.5. The mean value is 82.9.
- The sysBP also has a very high mean value of 132.4 and maximum value of 295.
- The variable cigsPerDay shows a high skewness (Fig.2.1).
- Variable totChol shows skewness (Fig.2.2).



```
      male              age          education       currentSmoker      cigsPerDay          BPMeds        prevalentStroke   
 Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.00000   Min.   :0.000000  
 1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000  
 Median :0.0000   Median :49.00   Median :2.000   Median :0.0000   Median : 0.000   Median :0.00000   Median :0.000000  
 Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941   Mean   : 9.006   Mean   :0.02962   Mean   :0.005896  
 3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000  
 Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000   Max.   :70.000   Max.   :1.00000   Max.   :1.000000  
                                  NA's   :105                      NA's   :29       NA's   :53                          
  prevalentHyp       diabetes          totChol          sysBP           diaBP            BMI          heartRate         glucose      
 Min.   :0.0000   Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.0   Min.   :15.54   Min.   : 44.00   Min.   : 40.00  
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0   1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00  
 Median :0.0000   Median :0.00000   Median :234.0   Median :128.0   Median : 82.0   Median :25.40   Median : 75.00   Median : 78.00  
 Mean   :0.3106   Mean   :0.02571   Mean   :236.7   Mean   :132.4   Mean   : 82.9   Mean   :25.80   Mean   : 75.88   Mean   : 81.96  
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.0   3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00  
 Max.   :1.0000   Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.5   Max.   :56.80   Max.   :143.00   Max.   :394.00  
                                    NA's   :50                                      NA's   :19      NA's   :1        NA's   :388     
   TenYearCHD    
 Min.   :0.0000  
 1st Qu.:0.0000  
 Median :0.0000  
 Mean   :0.1519  
 3rd Qu.:0.0000  
 Max.   :1.0000  
```

Fig.2.5: Summary of the data set

3. Data Cleaning and Pre Processing

3.1 Outlier Analysis:

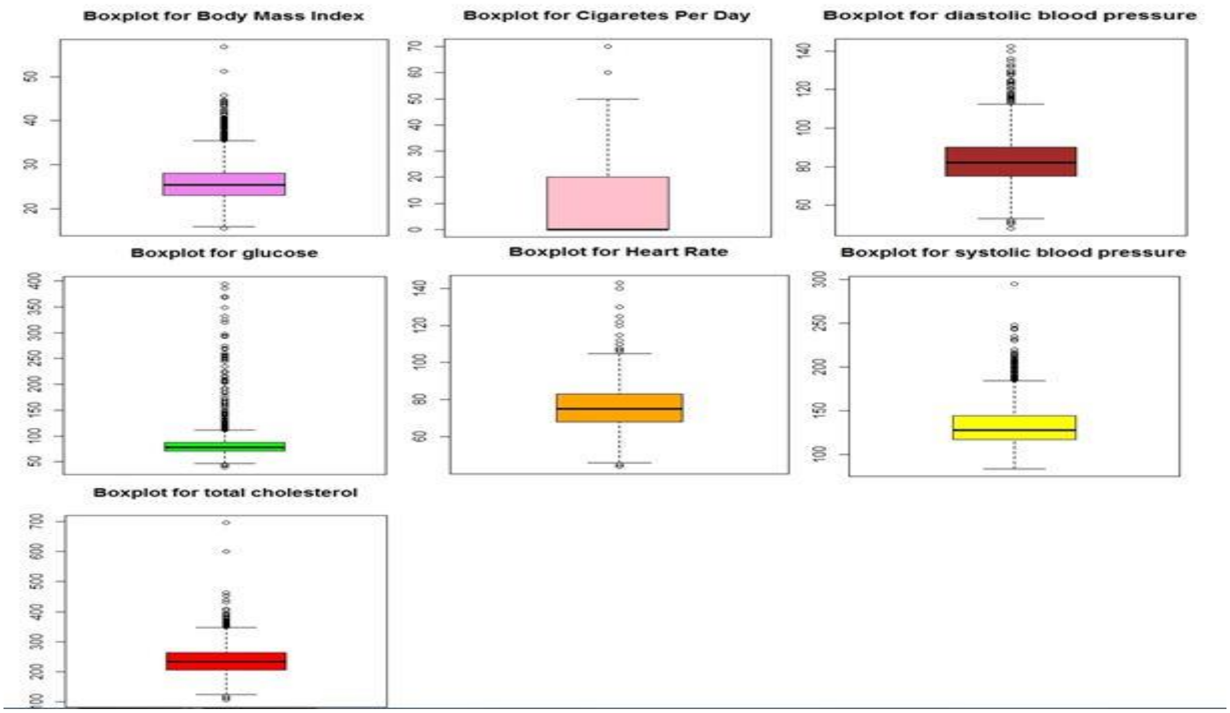


Fig.3.1: Outlier Variables

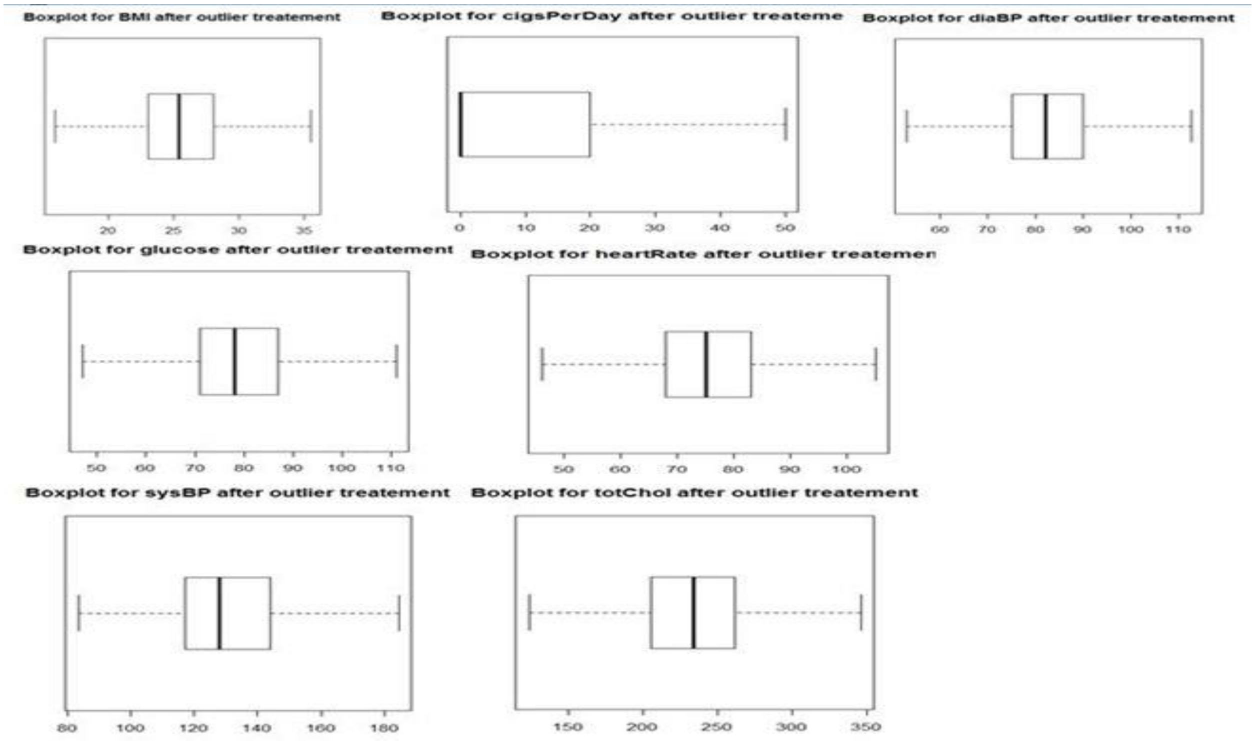


Fig.3.2: Outlier Variable after treatmen

- The boxplot shows that the variables BMI, cigsPerDay, diaBP, glucose, heartRate, sysBP and totChol contains the outlier (Fig.3.1).
- The treatment of outliers involves replacing the outliers by capped values meaning replacing those values which lie outside the lower limit by 5th percentile and those which lie above the upper limit by the value of 95th percentile.
- After outlier treatment, boxplot of all the variables BMI, cigsPerDay, diaBP, glucose, heartRate, sysBP and totChol is plotted again and it shows that the outliers have been removed (Fig.3.2)

3.2 Missing Value Treatment:

- Plot of missing value is shown in Fig.6
- Variable heartRate has 0.02% of missing values.
- Variable BMI has 0.45% of missing values.
- Variable cigsPerDay contains 0.68% of missing value.
- Variable totChol contains 1.18% of missing values.
- Variable BPMeds contains 1.25% of missing values.
- Variable education contains 2.48% of missing values.
- Variable glucose contains 9.15% of missing values.
- In order to have a good model building, it is required to treat these missing values.
- For treatment of missing values, imputation method is deployed.

- Since the variable education does not contain any outlier, therefore for its missing value treatment, imputation by mean is done.
- Rest all the other variables has outliers therefore to treat them imputation by median is done.
- After imputing the missing values, again the plot of missing values is done to verify whether or not any missing value exist in the data set.
- The Fig.3.4 shows that there is no missing value in the data set after imputation.

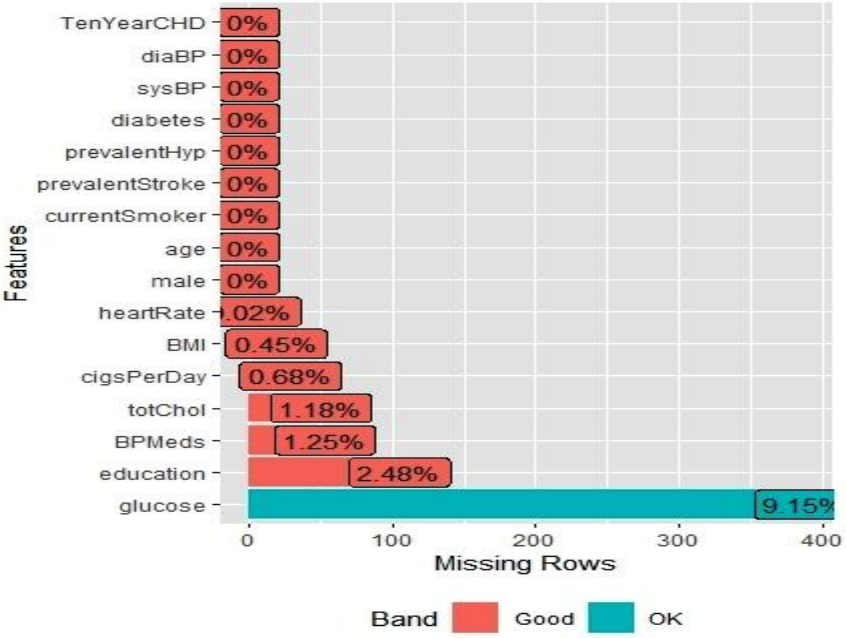


Fig.3.3: Missing Value Plot before imputation

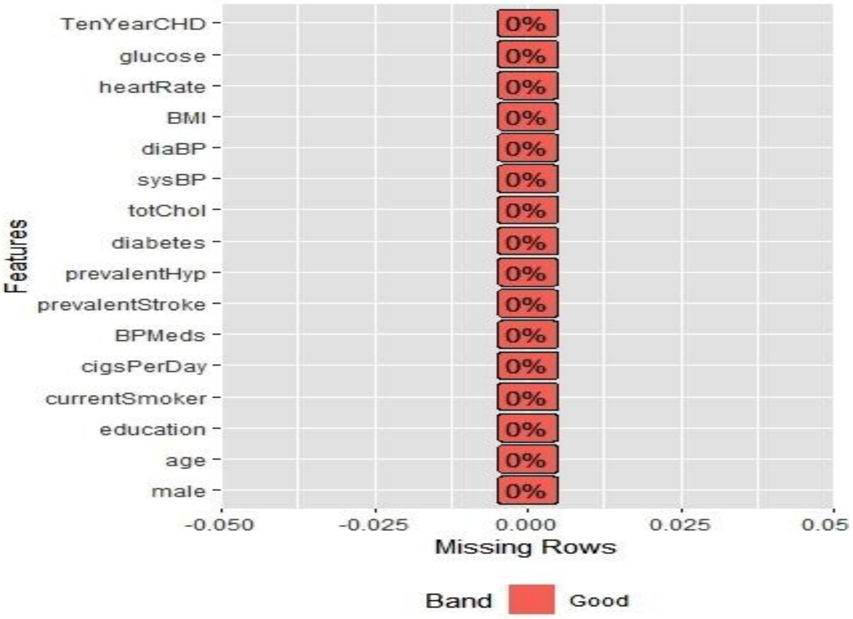


Fig.3.4: Missing value plot after imputation

3.3 Data Imbalancing:

- Exploring the target variable i.e TenYearCHD in the original data set received from medical practitioner. It is seen that the data is distributed in 84:20 ratio meaning that the data is imbalanced (Table-3.1). The data set has a total of 4240 rows.

Table-3.1: Original distribution of data

TenYearCHD	
0	1
3596	644

- In order to balance the data, SMOTE analysis is deployed. In SMOTE analysis, over sampling by 3000 and under sampling by 250 was utilized. It breaks the data set in the ratio 70:30 (Table-3.2). SMOTE contains a total data set of 68264 rows.

Table-3.2: Distribution of data after SMOTE

TenYearCHD	
0	1
48300	19964

3.4 Variable Transformation:

- Variable scaling is need to be done because the variables like age, cigsPerDay, totChol, sysBP, diaBP, heartRate, glucose, BMI are continuous variables whose minimum and maximum value ranges upto 100 or 200.
- On the other hand, rest other variables are factor variables having output “1” or “0”.
- Therefore it is necessary to bring all the variables at the same scale to avoid any kind of biasing in the model building.
- The variables are scaled with mean = 0 and standard deviation = 1.

4. Model Building

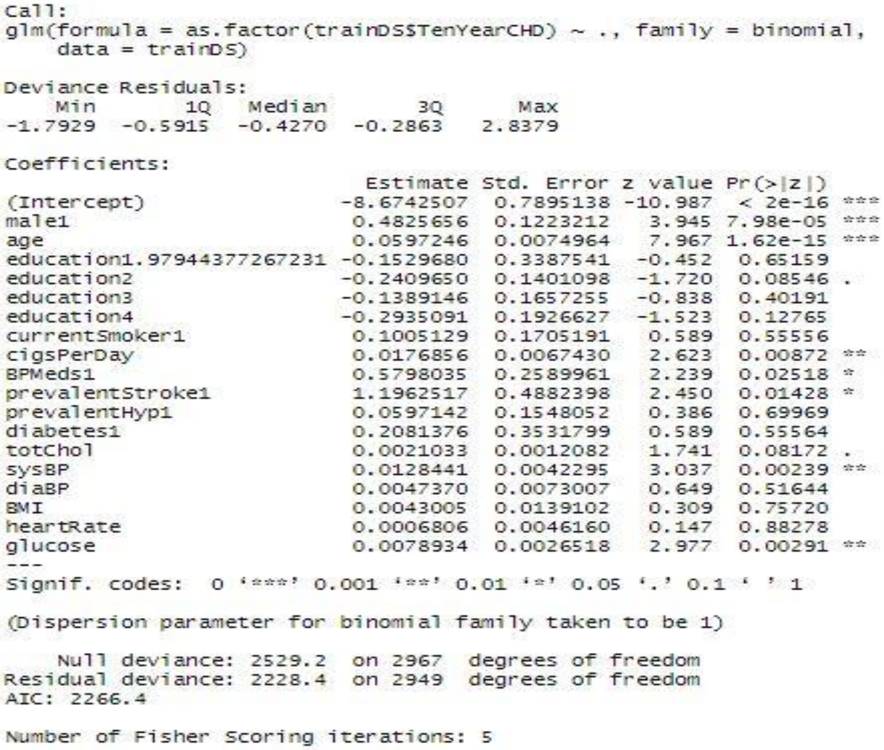
4.1 Data Split:

- The given data set contains 4240 observations.
- The data set was split into train and test data set. The split used was 70:30 ratio.
- Train data set contains 2968 observations.
- Test data set contains 1272 observations.

- Since the given data set contains majority and minority class in the ratio of 85:15 therefore SMOTE method was applied to balance the data set. After SMOTE, the data set was divided in the ratio 70:30.
- For SMOTE analysis, the over sampling by a factor of 3000 was utilized and under sampling by a factor 250 is used.
- This SMOTE data is again divided into smote train and smote test data set.
- The smote train data set contains 47784 observations.
- The smote test data set contains 20479 observations.

4.2 Logistic Regression Model:

Logistic regression model is used to solve the classification problem to identify the probability of occurrence of any class or event. It contains a set of predictor variable and an output variable which is binary classification.



```
Call:
glm(formula = as.factor(trainDS$TenYearCHD) ~ ., family = binomial, 
    data = trainDS)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7929  -0.5915  -0.4270  -0.2863   2.8379  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -8.6742507  0.7895138 -10.987  < 2e-16 ***
male1                     0.4825656  0.1223212   3.945 7.98e-05 ***
age                       0.0597246  0.0074964   7.967 1.62e-15 ***
education1.97944377267231 -0.1529680  0.3387541  -0.452  0.65159    
education2               -0.2409650  0.1401098  -1.720  0.08546 .  
education3               -0.1389146  0.1657255  -0.838  0.40191    
education4               -0.2935091  0.1926627  -1.523  0.12765    
currentSmoker1            0.1005129  0.1705191   0.589  0.55556    
cigsPerDay                0.0176856  0.0067430   2.623  0.00872 ** 
BPMeds1                   0.5798035  0.2589961   2.239  0.02518 *  
prevalentStroke1          1.1962517  0.4882398   2.450  0.01428 *  
prevalentHyp1             0.0597142  0.1548052   0.386  0.69969    
diabetes1                 0.2081376  0.3531799   0.589  0.55564    
totChol                   0.0021033  0.0012082   1.741  0.08172 .  
sysBP                     0.0128441  0.0042295   3.037  0.00239 ** 
diaBP                     0.0047370  0.0073007   0.649  0.51644    
BMI                       0.0043005  0.0139102   0.309  0.75720    
heartRate                 0.0006806  0.0046160   0.147  0.88278    
glucose                   0.0078934  0.0026518   2.977  0.00291 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2529.2  on 2967  degrees of freedom
Residual deviance: 2228.4  on 2949  degrees of freedom
AIC: 2266.4

Number of Fisher Scoring iterations: 5
```

Fig.4.1: p-value after simple logistic regression

- A logistic regression model considering all the variables is built and the p-values were observed to identify the significant variables.
- From Fig.4.1, it can be seen that the p-value of variables male, age, cigsPerDay, BPMeds, prevalentStroke, sysBP and glucose is less than 0.05 and hence these are significant variables in building the logistic regression.
- Second logistic regression model was again built considering only the significant variables and the results obtained were shown in Table-4.1.

Table-4.1: Confusion Matrix for simple LR

	0	1
0	2496	21
1	422	29

- Third logistic regression model was build on SMOTE data in which all the predictor variables are considered to be significant as all the variables has a p-value less than 0.05.
- The result of the SMOTE logistic regression model is shown in Table-4.2.

Table-4.2: Confusion Matrix LR with SMOTE data

	0	1
0	31947	1863
1	6358	7617

- To test the effectiveness of the performance, the build logistic regression model was tested using K-fold cross validation (assuming K = 10). The result of the cross validation is tabulated in Table-4.3.

Table-4.3: Confusion Matrix LR with SMOTE and cross validation

	0	1
0	13713	2669
1	777	3320

Evaluation:

Table-4.4: Evaluation table for logistic regression model

Model Name	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Logistic Regression with SMOTE	82.79%	83.40%	80.34%	86.55%

- Among all the logistic regression models that were built, it is found that the best model is logistic regression with SMOTE as it can be observed from Table-4.4, the AUC value is 86.55% meaning that the model 86.55% capable of separating true positive values from the false negative values.

4.3 Naïve Bayes Model:

Bayes theorem is another method for calculating the probability of occurrence of any event. This theorem is based on classification problem with an assumption that the predictor variables are independent i.e. occurrence of one variable is not affected by the occurrence of another variables.

- The data set is divided into train and test data set.
- To build the Naïve Bayes model, library "e1071" was utilized.
- To build the Naïve Bayes model, the output variable is converted into factor.
- The model was build considering all the independent variables.
- Confusion matrix is calculated/obtained (Table-4.5).

Table-4.5: Confusion Matrix Naïve Bayes

	0	1
0	986	140
1	93	53

- The second Naïve Bayes Model was build using the SMOTE train data and it was tested on the SMOTE test data set.
- The results of the SMOTE test data set are tabulated in Table-4.6.

Table-4.6: Confusion Matrix Naïve Bayes with SMOTE

	0	1
0	12766	2448
1	1724	3541

- In order to measure the performance of the Naïve Bayes model with SMOTE. The model was test on K-fold cross validation (assuming K = 10).
- The results of the K-fold cross validation is tabulated in Table-4.7.

Table-4.7: Confusion Matrix Naïve Bayes with SMOTE and K-fold cross validation

	0	1
0	13678	2498
1	812	3491

Evaluation:

Table-4.8: Evaluation table for Naïve Bayes model

Model Name	Accuracy	Sensitivity	Specificity	AUC
Naïve Bayes with SMOTE and Cross Validation	83.84%	94.40%	58.29%	87.12%

- The best Naïve Bayes model is found to be Naïve Bayes with SMOTE and cross validation because it has got the AUC value of 87.12% meaning it can separate true positive values from false positive values more effectively compared to other Naïve Bayes model.
- This model also has the sensitivity of 94.40%.

4.4 Random Forest:

Random Forest is an ensemble method for classification, regression and other decision making task. It constructs a number of decision trees on the randomly selected sample data from a large data set. The output of all the decision trees are than looked upon and according to the majority rule, final output is decided.

- Data set is first treated for missing value and outliers.
- Variable conversion is done to transform the variables into correct dimension.
- Data set is then split into train and test data.
- Random forest is built on train data set and its performance is tested on test data set.
- Confusion matrix is shown in Table-4.9.
- Random forest error v/s trees is shown in Fig.4.2.
- Minimum number of trees for tuning is found to be approximately 25 (Fig.4.2).
- Important variables are found to be age, education, cigsPerDay, totChol, sysBP, diaBP, BMI, heartRate, glucose (Fig.4.3).
- Minimum OOB error is found to 0.03% (Fig.4.4) at mtry = 3.

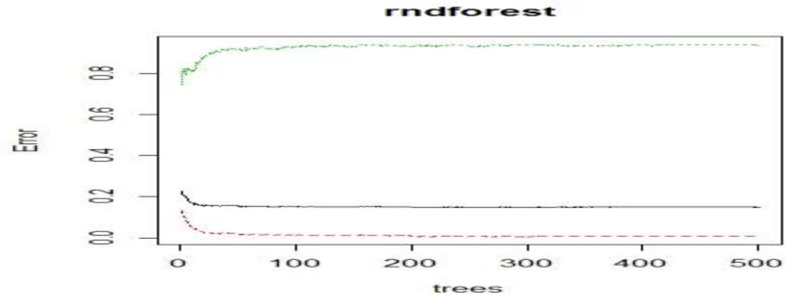
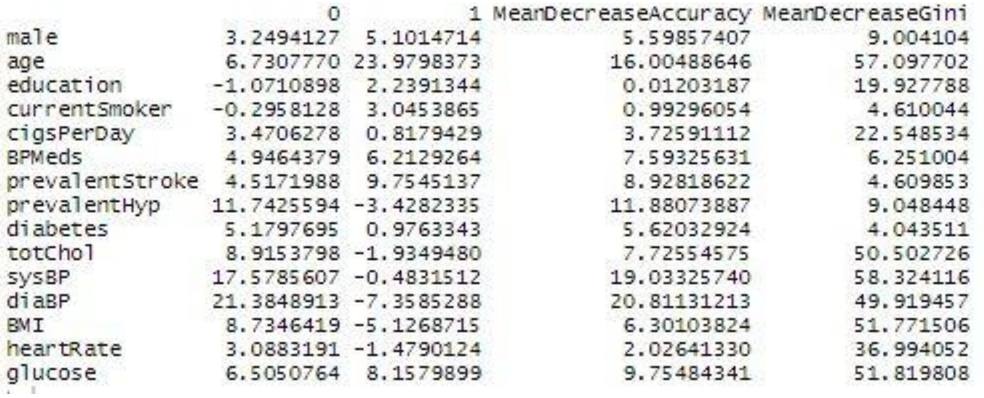


Fig.4.2: Simple Random Forest error v/s trees

Table-4.9: Confusion Matrix for simple random forest

	0	1
0	1064	15
1	180	13



```
                        0          1 MeanDecreaseAccuracy MeanDecreaseGini
male            3.2494127  5.1014714           5.59857407         9.004104
age             6.7307770 23.9798373          16.00488646        57.097702
education      -1.0710898  2.2391344           0.01203187        19.927788
currentSmoker  -0.2958128  3.0453865           0.99296054         4.610044
cigsPerDay      3.4706278  0.8179429           3.72591112        22.548534
BPMeds          4.9464379  6.2129264           7.59325631         6.251004
prevalentStroke 4.5171988  9.7545137           8.92818622         4.609853
prevalentHyp   11.7425594 -3.4282335          11.88073887         9.048448
diabetes        5.1797695  0.9763343           5.62032924         4.043511
totChol         8.9153798 -1.9349480           7.72554575        50.502726
sysBP          17.5785607 -0.4831512          19.03325740        58.324116
diaBP          21.3848913 -7.3585288          20.81131213        49.919457
BMI             8.7346419 -5.1268715           6.30103824        51.771506
heartRate       3.0883191 -1.4790124           2.02641330        36.994052
glucose         6.5050764  8.1579899           9.75484341        51.819808
```

Fig.4.3: Variable importance

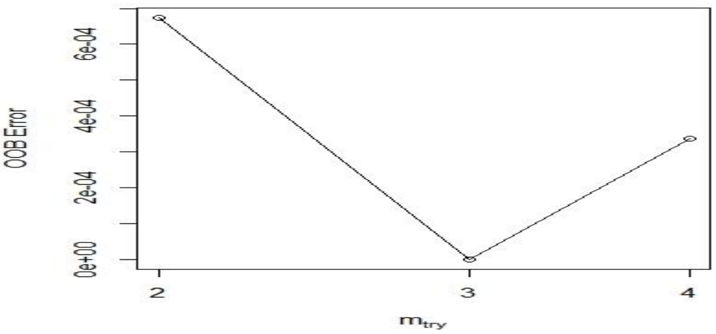


Fig.4.4: OOB error

- The second model that was build is Random Forest with SMOTE.
- The confusion matrix for Random Forest with SMOTE model (Fig.4.10).

Table-4.10: Confusion Matrix random forest with SMOTE

	0	1
0	14490	0
1	340	5649

Evaluation

The Table-4.11 below shows the evaluation table for the best Random Forest model.

Table-4.11: Evaluation table for Random Forest with SMOTE

Model Name	Accuracy	Sensitivity	Specificity	AUC
Random Forest with SMOTE	98.34%	97.71%	100%	99.95%

5. Model Validation

Under the study, different models were build to predict the risk of getting heart rate. In order to validate a model, K fold validation is applied to Logistic Regression model and Naïve Bayes model to check the accuracy, sensitivity and specificity of the model and the results are tabulated (Table-5.1).

In order to validate a Random Forest model, model tuning was applied and the results of model tuning are tabulated (Table-5.1).

In order to decide which model is the best model in predicting the heart attack, area under the curve and sensitivity are considered to be the important variables i.e. the model with highest value of area under curve and sensitivity value is considered as the best model. The comparison of all the different models build is shown in Table-5.1.

Table-5.1: Final Model Comparison

S.No.	Model Name	Accuracy	Sensitivity	Specificity	AUC
1	Simple Logistic Regression	85.07%	85.53%	58%	73.32%
2	Logistic Regression with SMOTE	82.79%	83.40%	80.34%	86.55%
3	Logistic Regression with SMOTE and Cross Validation	83.17%	94.64%	55.43%	86.55%
4	Simple Naïve Bayes	81.68%	91.38%	27.46%	71.68%
5	Naïve Bayes with SMOTE	79.63%	88.10%	59.13%	86.32%
6	Naïve Bayes with SMOTE and Cross Validation	83.84%	94.40%	58.29%	87.12%
7	Simple Random Forest	84.67%	85.53%	46.42%	70.45%
8	Random Forest with SMOTE	98.34%	97.71%	100%	99.95%

- Simple Logistic Regression model has the sensitivity value of 85.53% i.e. the model highly sensitive in detecting the true positive values.

- Logistic Regression model with SMOTE has area under curve of 86.55% i.e. the model can separate true positive values from false positive values with 86.55% probability.
- Logistic Regression model with SMOTE and K fold cross validation also has an area under curve of 86.55%. However, it has a sensitivity value of 94.64%.
- Naïve Bayes model with SMOTE and K fold cross validation has area under curve of 87.12%.
- Random Forest model with SMOTE has an area under curve 99.95% and sensitivity value of 97.71%.

Under the study, it is concluded that the **best model** in predicting the heart attack in a patient after ten year is **Random Forest Model with SMOTE** since it has the highest value of area under curve and sensitivity.

5.1 Insights:

- Data set contain 4240 observations.
- Data set is imbalanced with 85:15 distribution of 0's and 1's.
- Contains missing value and outliers.
- Variables sysBP and diaBP are highly correlated with a correlation of 78%.
- Age and sysBP has 39% correlation.
- If the patient is under medical treatment previously, then they have high probability of getting a heart attack.
- Patients with high BMI value has higher heart rate and glucose level.
- Education level also plays an important role towards the understanding of factors responsible for heart attack.
- Persons above 50 years have high value of sysBP and diaBP.
- Patients who smoke more then 9 cigarettes per day are most likely to get heart attack.

5.2 Recommendations:

- Doctors should work on improving the level of glucose and cholesterol in the patients to reduce heart attacks.
- The patients should be made aware of the fact that smoking is injurious to health, risk related to high cholesterol and glucose level.
- Doctors need to do medication on reducing the hyper tension and blood pressure of the patient as these are the important factor causing heart attack.
- Doctors should also recommend patients to maintain BMI and prepare a proper diet chart for patients with high BMI.
- Pharmaceutical companies should work on manufacturing drugs to boost the immunity level of patients
- Pharma companies should manufacture good quality insulin to control glucose level.
- Regular exercise is also recommended for the patients having significant level of cholesterol, glucose and BMI.