Report on Diabetes dataset

Introduction:

The dataset containing data on Diabetes. Our objective is to predict whether a person is diabetic or not from other variables.First ,the target variable is identified which gives the information whether the person is diabetic or non-diabetic or create if it is not available.

Method used and Assumptions

Variable "glyhb" is taken as target variable. As per standard glyhb(Glycosylated hemoglobin) greater than 7 consider as diabetic and below it non-diabetic.Therefore it is assumed that "glyhb>7 is diabetic and"glyhb"=<7 non-diabetic. The dataset is examined using descriptive statistics(mean,median, mode,min maximum and percentile), correlation table to show how continuous variables are correlated with target variable, high magnitude means highly correlated, different plots are used to visualize the dataset like histogram,scatter-plots (for continuous vs target variable),barplot(categorical vs target variable) and frequency table to show count and percentage of each level of categorical variable in each level of taregt variable(i,e diabetic and non-diabetic).Then Perform the necessary data pre-processing like missing value imputation ,outlier detection for preparing the data for the predictive modeling. Finally Random Forest method is used to build the model for prediction and it is validated with test dataset using confusion matrix and ROC plot.

Results with R codes

First, set the working directory and then import the dataset in R studio.

```
# set working directory to this R-script location
setwd("~/Downloads")


# read CSV-file
data<-read.csv("diabetes_v2 (2).csv",na.strings = c("","NA"))
```

head of dataset

```
head(data)

##     id chol stab.glu hdl ratio glyhb   location age gender height
weight
## 1 1000  203       82  56   3.6  4.31 Buckingham  46 female     62
```

```
121
## 2 1001  165       97  24   6.9  4.44 Buckingham  29 female      64
218
## 3 1002  228       92  37   6.2  4.64 Buckingham  58 female      61
256
## 4 1003   78       93  12   6.5  4.63 Buckingham  67   male      67
119
## 5 1005  249       90  28   8.9  7.72 Buckingham  64   male      68
183
## 6 1008  248       94  69   3.6  4.81 Buckingham  34   male      71
190
##    frame bp.1s bp.1d bp.2s bp.2d waist hip time.ppn
## 1 medium   118    59    NA    NA    29  38      720
## 2  large   112    68    NA    NA    46  48      360
## 3  large   190    92   185    92    49  57      180
## 4  large   110    50    NA    NA    33  38      480
## 5 medium   138    80    NA    NA    44  41      300
## 6  large   132    86    NA    NA    36  42      195
```

Descriptive Statistics

Data Structure

```
# get the data structure
str(data)
```

```
## 'data.frame':    403 obs. of  19 variables:
##  $ id      : int  1000 1001 1002 1003 1005 1008 1011 1015 1016 1022
...
##  $ chol    : int  203 165 228 78 249 248 195 227 177 263 ...
##  $ stab.glu: int  82 97 92 93 90 94 92 75 87 89 ...
##  $ hdl     : int  56 24 37 12 28 69 41 44 49 40 ...
##  $ ratio   : num  3.6 6.9 6.2 6.5 8.9 ...
##  $ glyhb   : num  4.31 4.44 4.64 4.63 7.72 ...
##  $ location: Factor w/ 2 levels "Buckingham","Louisa": 1 1 1 1 1 1 1
1 1 1 ...
##  $ age     : int  46 29 58 67 64 34 30 37 45 55 ...
##  $ gender  : Factor w/ 2 levels "female","male": 1 1 1 2 2 2 2 2 2 1
...
##  $ height  : int  62 64 61 67 68 71 69 59 69 63 ...
##  $ weight  : int  121 218 256 119 183 190 191 170 166 202 ...
##  $ frame   : Factor w/ 3 levels "large","medium",..: 2 1 1 1 2 1 2 2
1 3 ...
##  $ bp.1s   : int  118 112 190 110 138 132 161 NA 160 108 ...
##  $ bp.1d   : int  59 68 92 50 80 86 112 NA 80 72 ...
##  $ bp.2s   : int  NA NA 185 NA NA NA 161 NA 128 NA ...
##  $ bp.2d   : int  NA NA 92 NA NA NA 112 NA 86 NA ...
##  $ waist   : int  29 46 49 33 44 36 46 34 34 45 ...
##  $ hip     : int  38 48 57 38 41 42 49 39 40 50 ...
##  $ time.ppn: int  720 360 180 480 300 195 720 1020 300 240 ...
```

The ‘data.frame’ has 403 observations of 19 variables.Three variables “frame”,“gender” and ‘location are categorical variables and all other are continuous(numeric/integers).Variable ’glyhb’ means glycosylated hemoglobin and this gives information about whether is person would be diabetic or non diabetic. Generally glyhb >7 consider as diabetic and below it non diabetic.

Data Summary

```
# get statistical summary
summary(data)
```

```
##        id             chol          stab.glu          hdl
##  Min.   : 1000   Min.   : 78.0   Min.   : 48.0   Min.   : 12.00
##  1st Qu.: 4792   1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00
##  Median :15766   Median :204.0   Median : 89.0   Median : 46.00
##  Mean   :15978   Mean   :207.8   Mean   :106.7   Mean   : 50.45
##  3rd Qu.:20336   3rd Qu.:230.0   3rd Qu.:106.0   3rd Qu.: 59.00
##  Max.   :41756   Max.   :443.0   Max.   :385.0   Max.   :120.00
##                  NA's   :1                       NA's   :1
##      ratio            glyhb             location        age
##  Min.   : 1.500   Min.   : 2.68   Buckingham:200   Min.   : 19.00
##  1st Qu.: 3.200   1st Qu.: 4.38   Louisa    :203   1st Qu.: 34.00
##  Median : 4.200   Median : 4.84                    Median : 45.00
##  Mean   : 4.522   Mean   : 5.59                    Mean   : 47.74
##  3rd Qu.: 5.400   3rd Qu.: 5.60                    3rd Qu.: 60.00
##  Max.   :19.300   Max.   :16.11                    Max.   :400.00
##  NA's   :1        NA's   :13
##     gender        height          weight           frame
##  female:234   Min.   :52.00   Min.   :   99.0   large :103
##  male  :169   1st Qu.:63.00   1st Qu.:  151.0   medium:184
##               Median :66.00   Median :  173.0   small :104
##               Mean   :66.02   Mean   :  202.2   NA's  : 12
##               3rd Qu.:69.00   3rd Qu.:  200.0
##               Max.   :76.00   Max.   :10000.0
##               NA's   :5       NA's   :1
##      bp.1s           bp.1d             bp.2s           bp.2d
##  Min.   : 90.0   Min.   : 48.00   Min.   :110.0   Min.   : 60.00
##  1st Qu.:121.2   1st Qu.: 75.00   1st Qu.:138.0   1st Qu.: 84.00
##  Median :136.0   Median : 82.00   Median :149.0   Median : 92.00
##  Mean   :136.9   Mean   : 83.32   Mean   :152.4   Mean   : 92.52
##  3rd Qu.:146.8   3rd Qu.: 90.00   3rd Qu.:161.0   3rd Qu.:100.00
##  Max.   :250.0   Max.   :124.00   Max.   :238.0   Max.   :124.00
##  NA's   :5       NA's   :5        NA's   :262     NA's   :262
##      waist           hip           time.ppn
##  Min.   :26.0   Min.   :30.00   Min.   :   5.0
##  1st Qu.:33.0   1st Qu.:39.00   1st Qu.:  90.0
##  Median :37.0   Median :42.00   Median : 240.0
##  Mean   :37.9   Mean   :43.04   Mean   : 341.2
##  3rd Qu.:41.0   3rd Qu.:46.00   3rd Qu.: 517.5
##  Max.   :56.0   Max.   :64.00   Max.   :1560.0
##  NA's   :2      NA's   :2       NA's   :3
```

These results show that the most of variables has missing values.NA represents number missing value in that variable.The target variable glyhb also has some missing values. But the target variable should not have any missing values.Therefore missing rows should be removed.Summary statistic (mean, median ,1st qu, 3rd qu,minimum amd maximum value) of continuous variables and counts of levels of categorical variables are calculated. Variable 'id' is a person ID and not giving any information therefore should be removed,

Remove some unimportant variables As id variable is just an unique id not giving information it should be removed.Apart from this the variables more than 50 percent missing values should be removed for better prediction.As variable bp.2s and bp.2d has high missing values(262).

```
# remove id variable
data<-data[,-1]
#remove variables more than 50% missing values
data<-data[, -which(colMeans(is.na(data)) > 0.5)]
```

Remove NA from glyhb

```
data<-data[!is.na(data$glyhb), ]
```

Columnwise missing values see columnwise missing values in the dataset

```
colSums(is.na(data))

##     chol stab.glu      hdl    ratio    glyhb location      age 
gender 
##        1        0        1        1        0        0        0 
0 
##   height   weight    frame    bp.1s    bp.1d    waist      hip 
time.ppn 
##        5        1       11        5        5        2        2 
3
```

Missing value imputation

KNN is most advanced and reliable method for missing values imputaion

```
library(VIM)

data1<-kNN(data,variable = colnames(data))

data <- subset(data1, select = -c(17:32))
```

Now all missing values in every variables has been imputed As it this package creates duplicate variables of every variables,the second code is used to remove it.

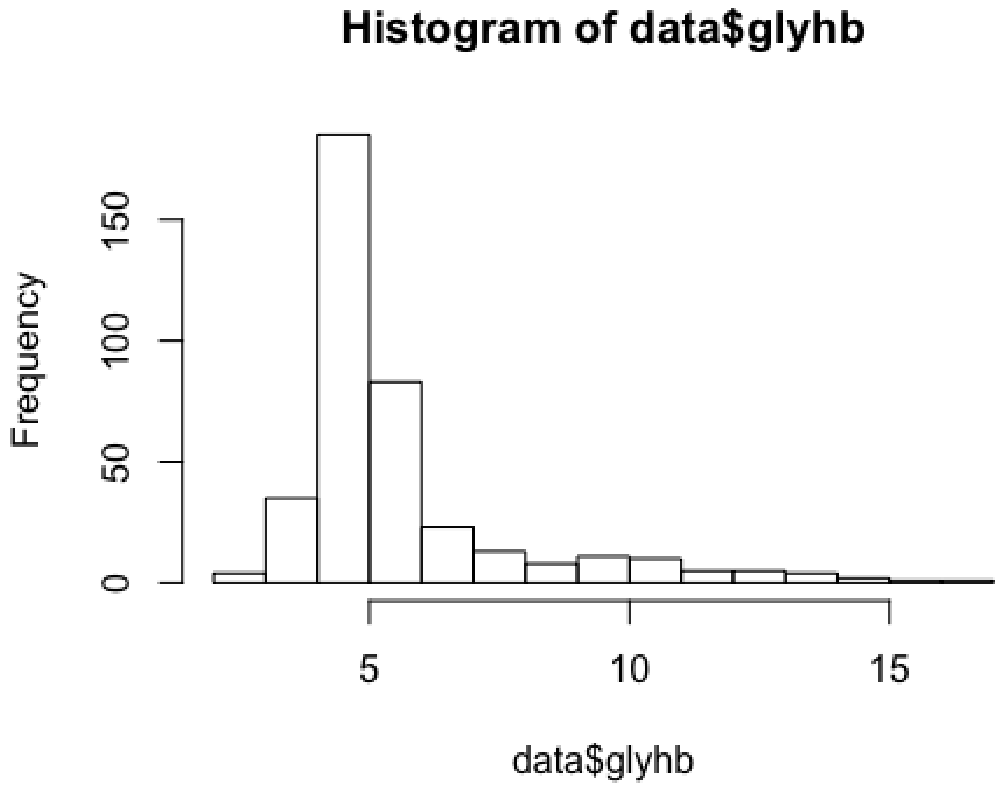
Data visulisation(plots) and outlier detection

```
library(funModeling)#require library




library(tidyverse)

#correlation table for how the other variables are related to target
variable.
correlation_table(data,"glyhb")

##    Variable glyhb
## 1     glyhb  1.00
## 2  stab.glu  0.75
## 3     ratio  0.33
## 4      chol  0.25
## 5       age  0.24
## 6     waist  0.22
## 7     bp.1s  0.20
## 8       hip  0.14
## 9    height  0.06
## 10    bp.1d  0.03
## 11 time.ppn  0.03
## 12   weight -0.02
## 13      hdl -0.15
```

correlation table shows that the variable stab.glu is highly correlated with the target variable(.75).

Visualisation or univariate analysis of target variable.

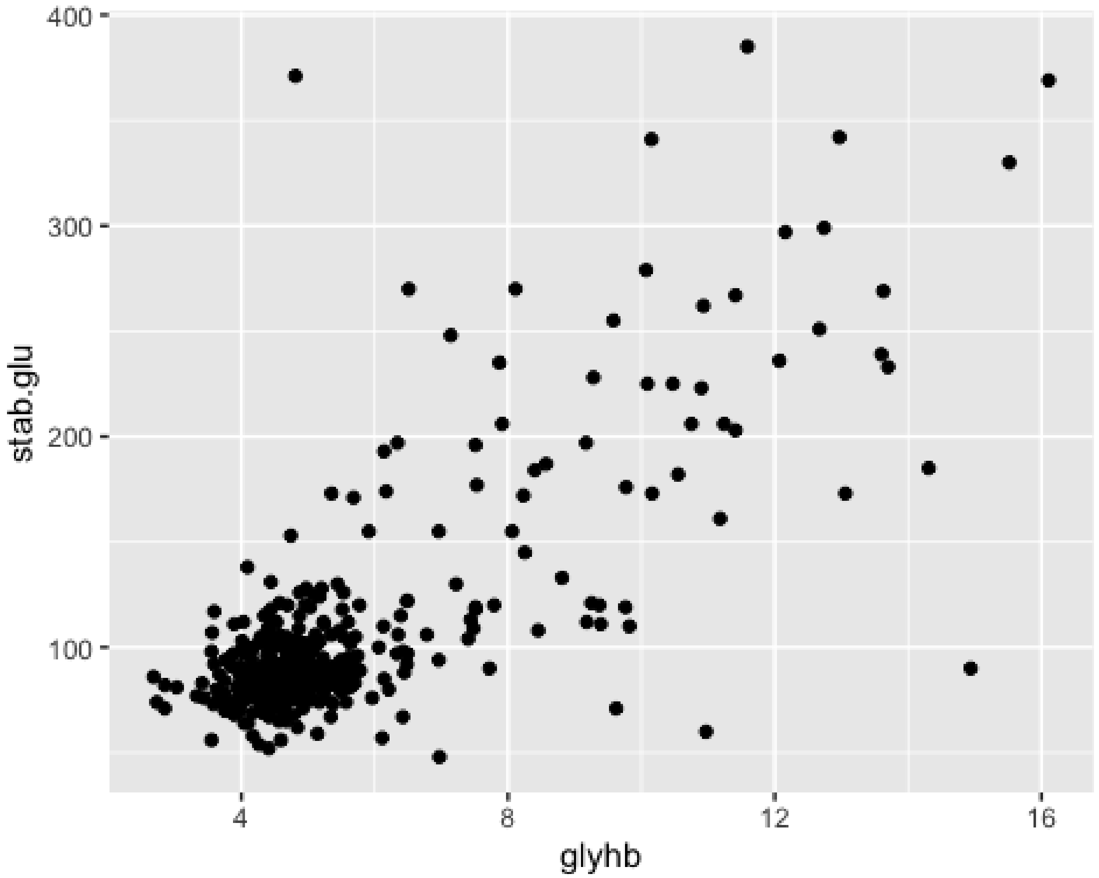
```
hist(data$glyhb)
```



The histogram shows that the target variable is not normally distributed ,it has right long tail means positivly skewed.

scatterplot of glyhb vs stab.glu

```
p <-ggplot(data, aes(glyhb,stab.glu))
p +geom_point()
```

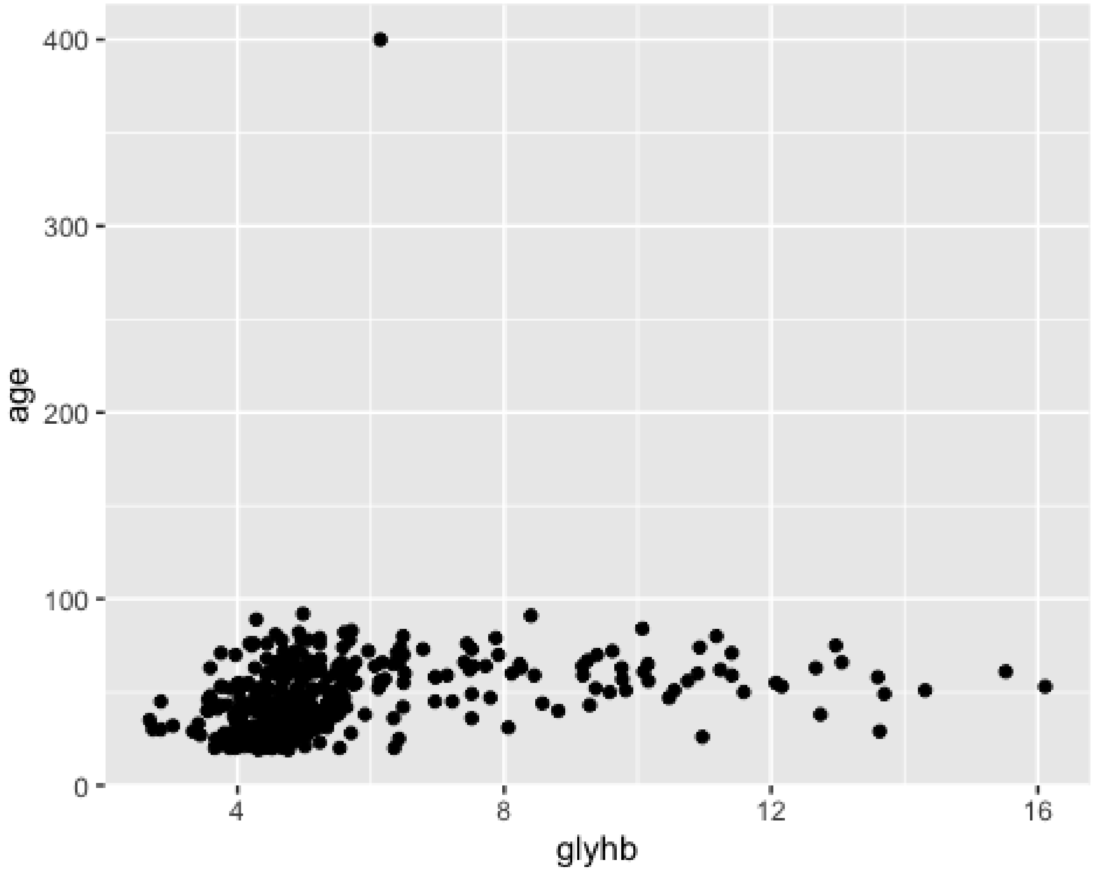


```
#delete two point as outliers
data<-data[!(data$stab.glu>300 & data$glyhb<8),]
data<-data[!(data$stab.glu<100 & data$glyhb>12),]
```

The scatter plot between glyhb and stab.glu show that there is linear relationship between both.BUT few points are coming as outlier and it should be removed.

scatterplot of glyhb vs age

```
p <-ggplot(data, aes(glyhb,age))
p +geom_point()
```

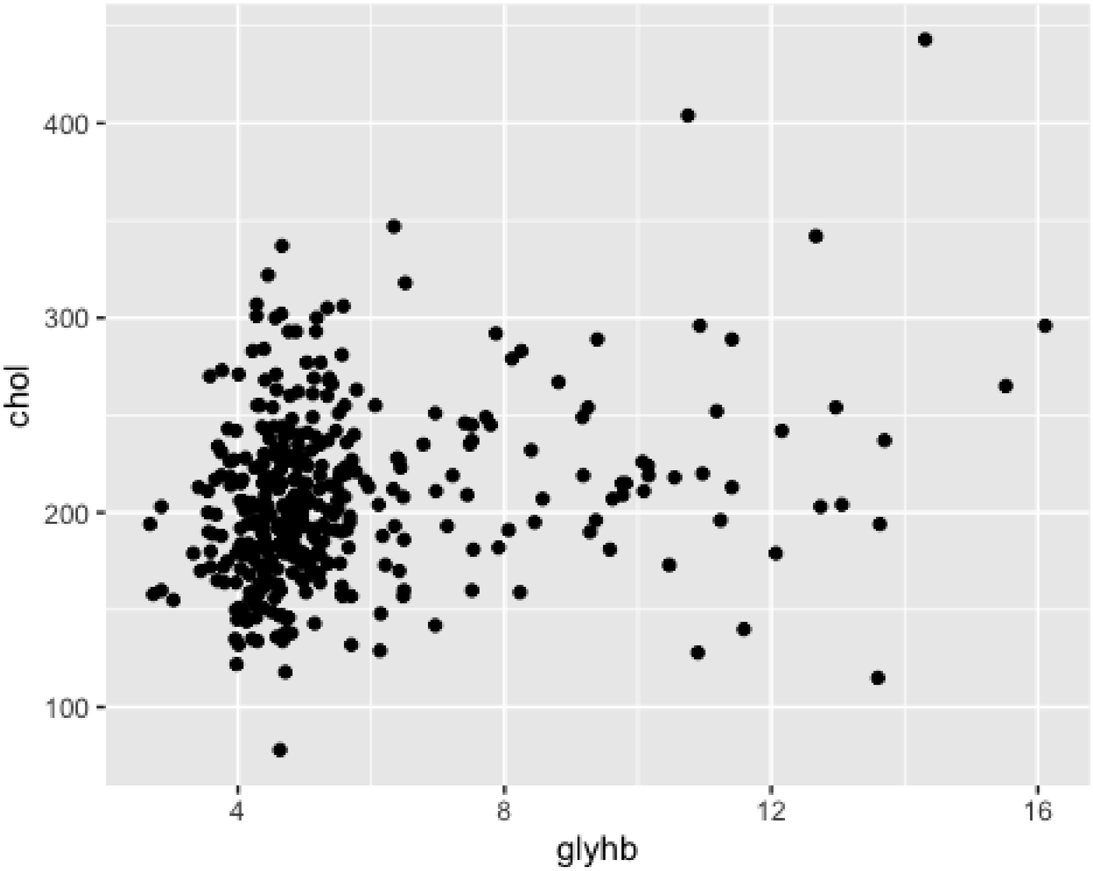


```
# remove outlier(age=400)
data<-data[data$age<100,]
```

The scatter plot shows that one person has age 400 which is impossible and need to be removed

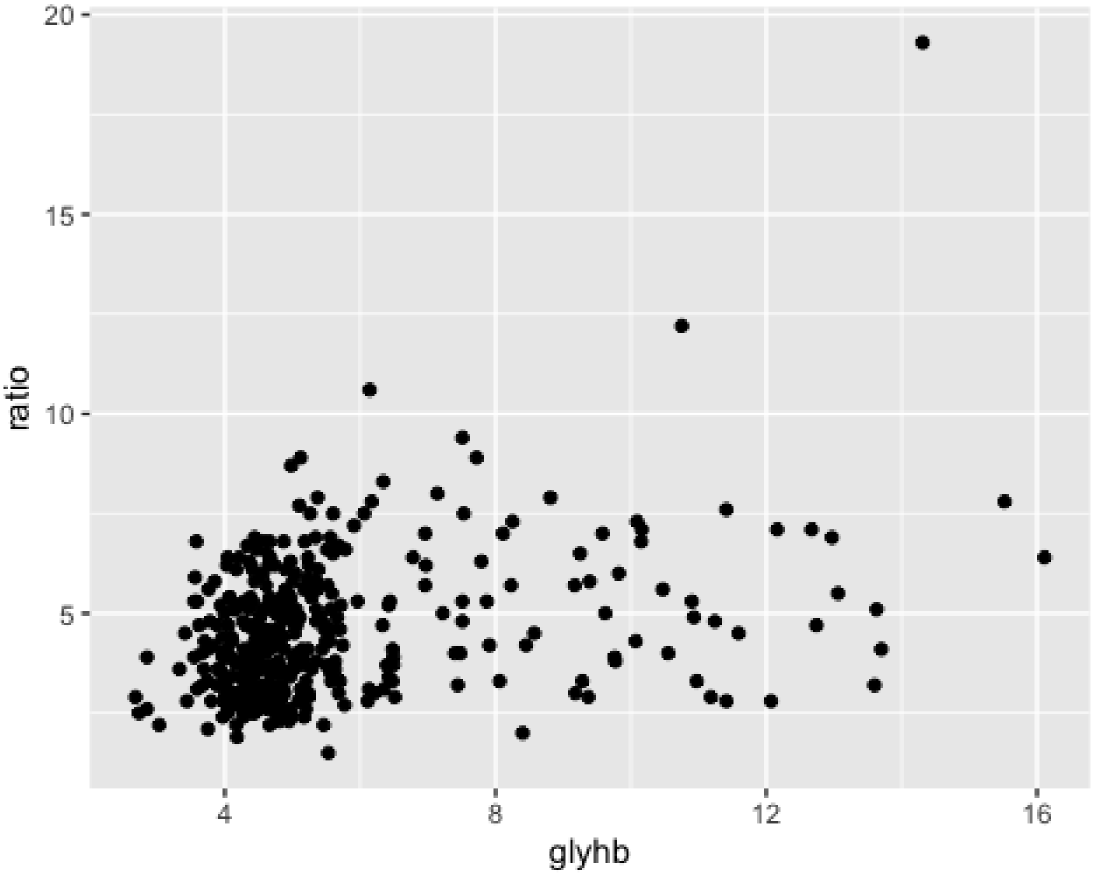
scatterplot of glyhb vs chol

```
p <-ggplot(data, aes(glyhb,chol))
p +geom_point()
```



visulisation and outlier detection in variable “ratio”

```
p <-ggplot(data, aes(glyhb,ratio))
p +geom_point()
```



```
#delete the outlier which is greater than 15 ratio and greater than 12
glyhb
data<-data[!(data$ratio>15 & data$glyhb>12),]
```

Create new variable “diabetic”..whether person is diabetic or not.

```
data$diabetic<-ifelse(data$glyhb>7,"diabetic","non-diabetic")
data$diabetic<-as.factor(data$diabetic)
data<-data%>%select(-glyhb)
```

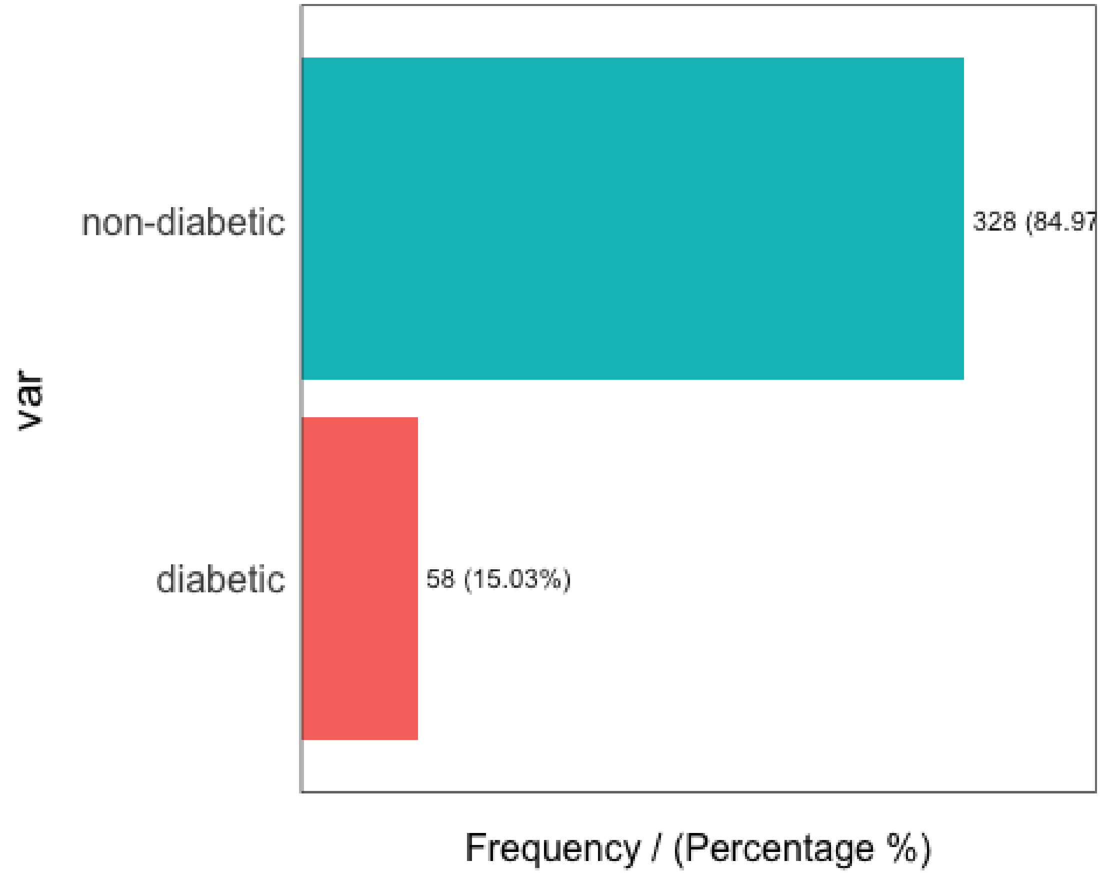
As per standard a person having glyhb >7 consider as diabetic and glyhb<=7 as non diabetic. Removed variable ‘glyhb’ now.

Visualisation of categorical variables with respect to target variable “diabetic”.

check frequency table and plot for variable “diabetic”

```
table(data$diabetic)

##
##     diabetic non-diabetic
##           58          328
```

```
freq(data$diabetic)
```

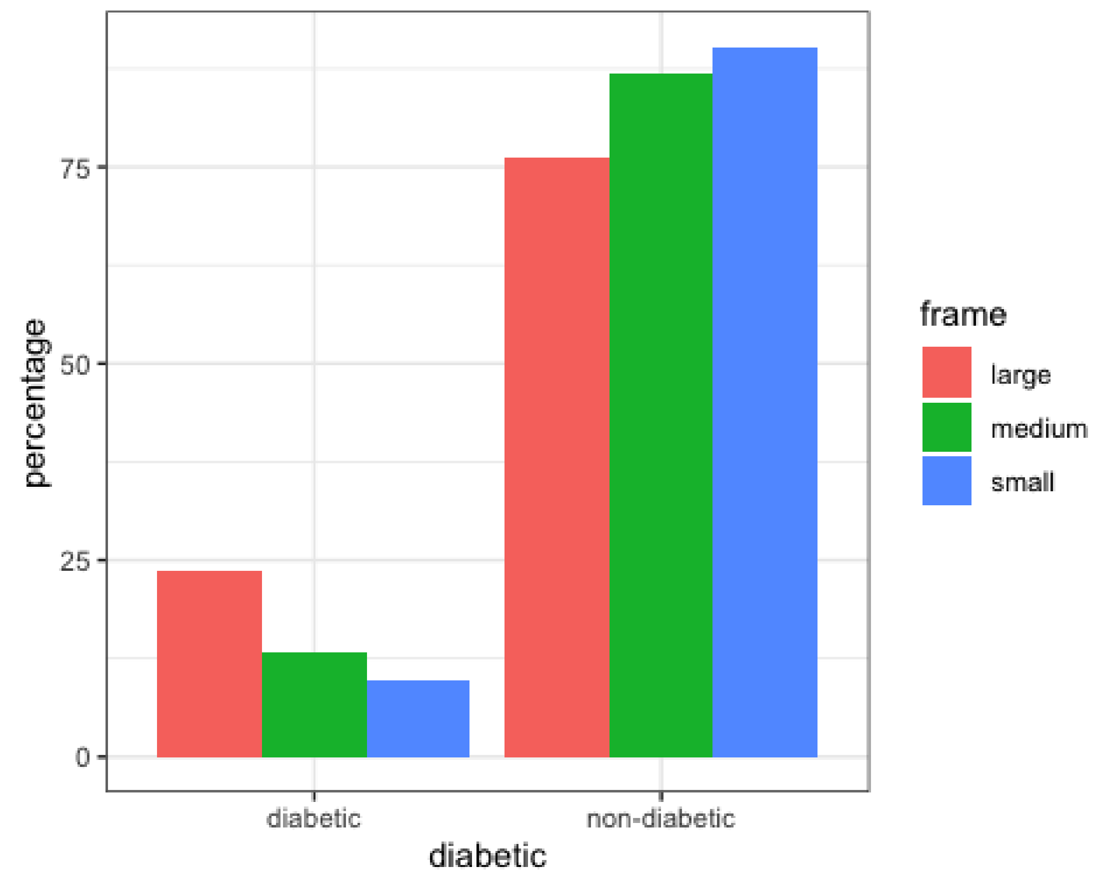


```
##            var frequency percentage cumulative_perc
## 1 non-diabetic       328      84.97           84.97
## 2     diabetic        58      15.03          100.00
```

table shows that 328 out of 386(84.97%) are non diabetic only 58(15.03%)are diabetic in dataset.

visualisation(barplot) of frame vs diabetic

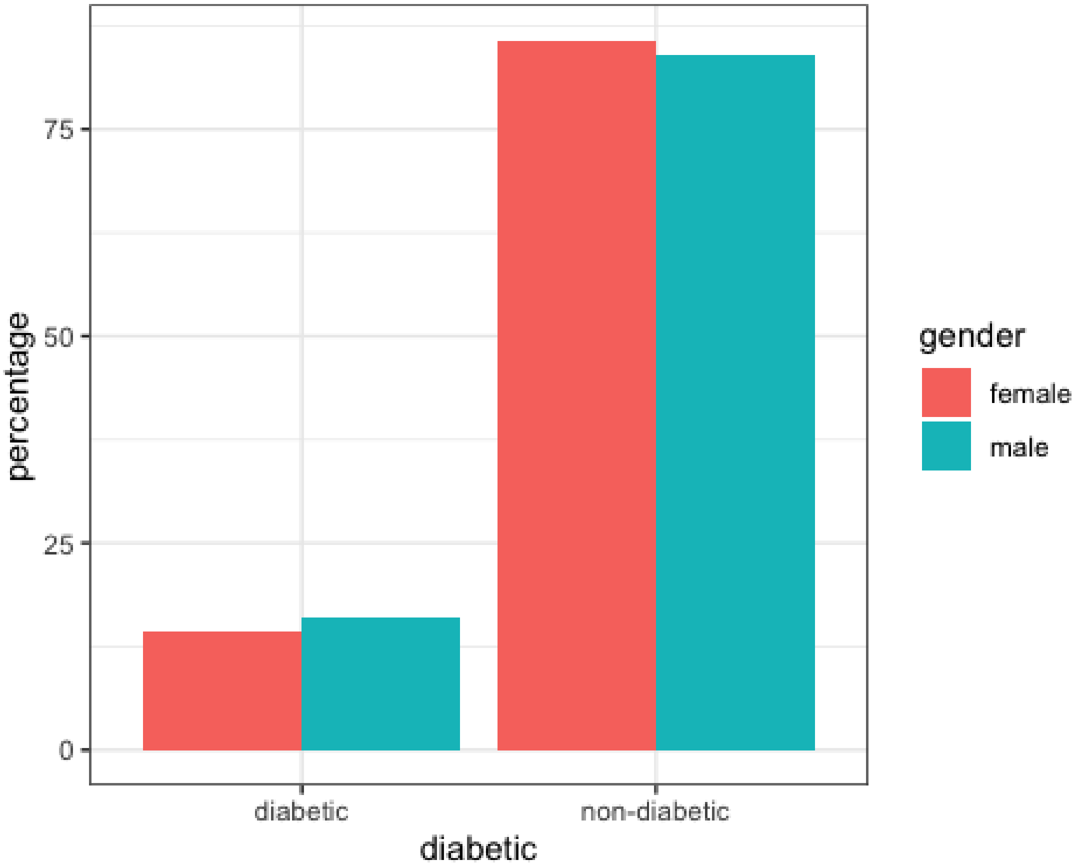
```
df2 <- data %>%
  group_by(frame, diabetic) %>%
  tally() %>%
  complete(diabetic, fill = list(n = 0)) %>%
  mutate(percentage = n / sum(n) * 100)
ggplot(df2, aes(diabetic, percentage, fill = frame)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  theme_bw()
```



The barplot clearly shows that percentage of each levels(large,medium,high) in “frame” are exactly opposite to person diabetic or non diabetic. Means diabetic person has highest percent of level“large” and lowest percentage of level “small”.Similarly non diabetic person has highest percent of level “small”amd lowest percentage f level“large”. This can be most significant variable for prediction.

barplot between “gender” and “diabetic”

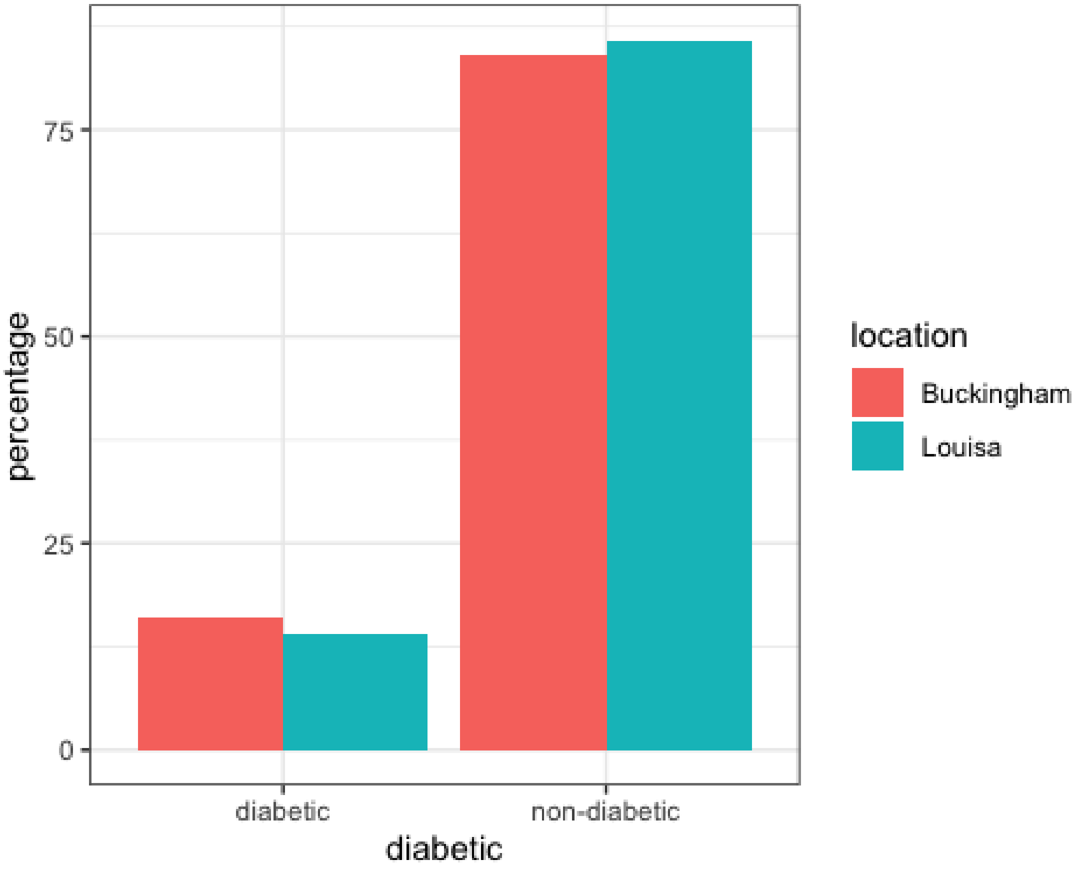
```
table(data$gender)

##
## female   male
##    225    161

df2 <- data %>%
  group_by(gender, diabetic) %>%
  tally() %>%
  complete(diabetic, fill = list(n = 0)) %>%
  mutate(percentage = n / sum(n) * 100)
ggplot(df2, aes(diabetic, percentage, fill = gender)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  theme_bw()
```



Barplot shows that the females are less percentage of diabetes as compare to males

barplot between “location” and “diabetic”

```
df2 <- data %>%
  group_by(location, diabetic) %>%
  tally() %>%
  complete(diabetic, fill = list(n = 0)) %>%
  mutate(percentage = n / sum(n) * 100)

ggplot(df2, aes(diabetic, percentage, fill = location)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  theme_bw()
```



train-test split

```
set.seed(200)
index<-sample(nrow(data),0.70*nrow(data),replace = F)
train<-data[index,]
test<-data[-index,]
```

The dataset has been split into train and test.Train data is used to build model and test to validate or test the model.

Model building(Random Forest)

Random forest classifier is used to build the model

```
#install.packages("randomForest")
library(randomForest)

rf<-randomForest(diabetic~.,data = train)
rf

##
## Call:
##  randomForest(formula = diabetic ~ ., data = train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
## 
##         OOB estimate of  error rate: 8.15%
## Confusion matrix:
##              diabetic non-diabetic class.error
## diabetic           29           16  0.35555556
## non-diabetic        6          219  0.02666667
```

rf is our random forest model.This model is tested with test data in next step.

validation

```
#validation with test data

#install.packages("irr")
library(irr)



#install.packages("e1071")
library(e1071)
```

The model will be validated by confusion matric and ROc plot (area under curve) with new dataset (test dataset)

prediction from test data(new data)

```
prediction_rf<-predict(rf,test)
```

prediction_rf is predition of "diabetic" from test dataset. Now it will be checked and compare with its original value.

validation by confusion matrix

```
confusionMatrix(prediction_rf,test$diabetic)

## Confusion Matrix and Statistics
## 
##               Reference
## Prediction     diabetic non-diabetic
##   diabetic            7            2
##   non-diabetic        6          101
##                                           
##                Accuracy : 0.931
##                  95% CI : (0.8686, 0.9698)
##     No Information Rate : 0.8879
##     P-Value [Acc > NIR] : 0.08664
##
##                   Kappa : 0.5997
##
##  Mcnemar's Test P-Value : 0.28884
##
##             Sensitivity : 0.53846
##             Specificity : 0.98058
##          Pos Pred Value : 0.77778
##          Neg Pred Value : 0.94393
##              Prevalence : 0.11207
##          Detection Rate : 0.06034
##    Detection Prevalence : 0.07759
##       Balanced Accuracy : 0.75952
##
##        'Positive' Class : diabetic
##
```

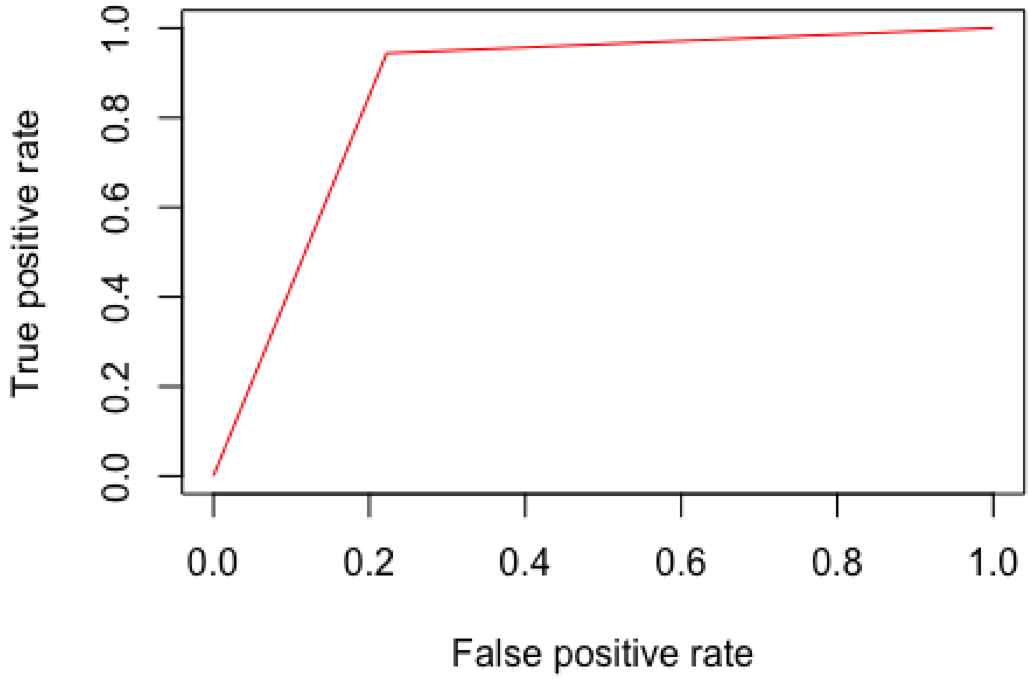
kappa=aprox .6 which is good value for a decent model and accuracy is .931

From confusion matrix our model is predicting 7 as dibetic which are actually diabetic and 2 as diabetic which are actually non diabetic. silimarly predicting 6 as non-diabetic which are actaully diabetic and 101 as non-diabetic which are actually non diabetic.

```
#install.packages("ROCR")
library(ROCR)

#install.packages("pROC")
library(pROC)
```

validation by ROC plot (area under curve)

```
predict_rf<-
prediction(as.numeric(test$diabetic),as.numeric(prediction_rf))
perf<-performance(predict_rf,"tpr","fpr")
plot(perf,col="red")
```



```
#Area under curve
auc(test$diabetic,as.numeric(prediction_rf ))


## Area under the curve: 0.7595
```

The area under curve is 0.7595 which is a good sign .Our model is working well.

Conclusion

To predict diabetes in a person “stab.glu” is most significant(positively significant) variable amongs continuous variables(numeric/integer) and “frame” is most significant variable amongs categorical variables.Females are less diabetic as compare to males.Location “lousia” showing less diabetic percentage as compare to “buckingham”. Old age people are more diabetic as “glyhb” is increasing with age.

Our model(random forest) predicts, whether a person is diabetic or not. The entire dataset is split into train(70%)and test(30%).The train data is used to build the model for prediction and test is used as new datset for validation of model. Two metric “Confusion matrix” and “ROC plot” are used to validate the result.From confusion matrix Accuracy of model is .931 and Kappa is .6 and from roc curve area under the curve is coming as 0.7595.All are giving good sign.