MATERNAL HEALTH RISK PREDICTION

Jagdish Patil

2022-11-02

INTRODUCTION

Many pregnant women die from pregnancy issues as a result of a lack of information on maternal health care during and after pregnancy. It is more common in rural regions and among lower-middle-class families in emerging countries. During pregnancy, every minute should be observed to ensure the proper growth of the baby and the safe delivery.

Pregnancy complications may be caused by conditions women have before pregnancy or conditions women develop during pregnancy. The impact of pregnancy complications on maternal and neonatal outcomes is difficult to estimate because pregnancy complications are made up of a broad range of conditions with varying levels of severity. Every year, an estimate of 358,000 maternal mortalities is recorded worldwide, with about 99% cases occurring in poor developing countries. Early screening is expected to reduce maternal mortality rates.

OBJECTIVE

The purpose of this project, is to analyze the risk factors associated with risk level in the dataset and also to fit models with the capacity to predict maternal complications.

IMPORTING DATA SET

```
d=read.csv("C://Users//jagdish//Downloads//Maternal Health Risk Data S
et.csv")
d=as.matrix(d)
head(d)
```

```
##      ï..Age SystolicBP DiastolicBP BS      BodyTemp HeartRate RiskLevel
## [1,] "25"   "130"      " 80"       "15.00" " 98.0"  "86"      "high risk"
## [2,] "35"   "140"      " 90"       "13.00" " 98.0"  "70"      "high risk"
## [3,] "29"   " 90"      " 70"       " 8.00" "100.0"  "80"      "high risk"
## [4,] "30"   "140"      " 85"       " 7.00" " 98.0"  "70"      "high risk"
## [5,] "35"   "120"      " 60"       " 6.10" " 98.0"  "76"      "low risk"
## [6,] "23"   "140"      " 80"       " 7.01" " 98.0"  "70"      "high risk"
```

```
n=nrow(d);n

## [1] 1014

p=ncol(d);p

## [1] 7
```

DATA DESCRIPTION

Data is collected from data site Kaggle. Data has been collected from different hospitals, community clinics, maternal health cares through the loT (Internet of Things) based risk monitoring system.

Age: Age in years when a woman is pregnant.

Systolic Blood Pressure: Upper value of Blood Pressure in mmHg.

Diastolic BP: Lower Value of Blood Pressure in mmHg.

Blood Sugar: Blood glucose levels is in terms of a molar concentration, mmol/L.

Body temperature: Temperature of body.

Heart Rate: A normal resting heart rate in beats per minute.

Risk Level: Predicted Risk Intensity Level during pregnancy considering the previous attribute.

Type of variables:

Categorical variable: Risk levels

Continuous variable: Age, Systolic Blood Pressure, Diastolic Blood Pressure, Blood Sugar, Body Temperature, Heart Rate.

Conversion of risk level in to numerical categories

```
for(i in 1:n)
{
  if(d[i,7]=="high risk"){  d[i,7]=3 }
  if(d[i,7]=="mid risk"){  d[i,7]=2 }
  if(d[i,7]=="low risk"){  d[i,7]=1 }
}
head(d)

##      ï..Age SystolicBP DiastolicBP BS      BodyTemp HeartRate RiskLevel
## [1,] "25"   "130"      " 80"       "15.00" " 98.0"  "86"      "3"
## [2,] "35"   "140"      " 90"       "13.00" " 98.0"  "70"      "3"
## [3,] "29"   " 90"      " 70"       " 8.00" "100.0"  "80"      "3"
## [4,] "30"   "140"      " 85"       " 7.00" " 98.0"  "70"      "3"
## [5,] "35"   "120"      " 60"       " 6.10" " 98.0"  "76"      "1"
## [6,] "23"   "140"      " 80"       " 7.01" " 98.0"  "70"      "3"
```

Low risk is denoted by 1.

Mid risk is denoted by 2.

High risk is denoted by 3

CHECKING MISSING VALUES

```
which(is.na(d)==TRUE)

## integer(0)
```

There is no missing value in dataset.

```
d=as.numeric(d)
d=matrix(d,nrow=1014,ncol=7,byrow=F)
head(d)

##      [,1] [,2] [,3]  [,4] [,5] [,6] [,7]
## [1,]   25  130   80 15.00   98   86    3
## [2,]   35  140   90 13.00   98   70    3
## [3,]   29   90   70  8.00  100   80    3
## [4,]   30  140   85  7.00   98   70    3
## [5,]   35  120   60  6.10   98   76    1
## [6,]   23  140   80  7.01   98   70    3
```

SUMMARY OF DATA

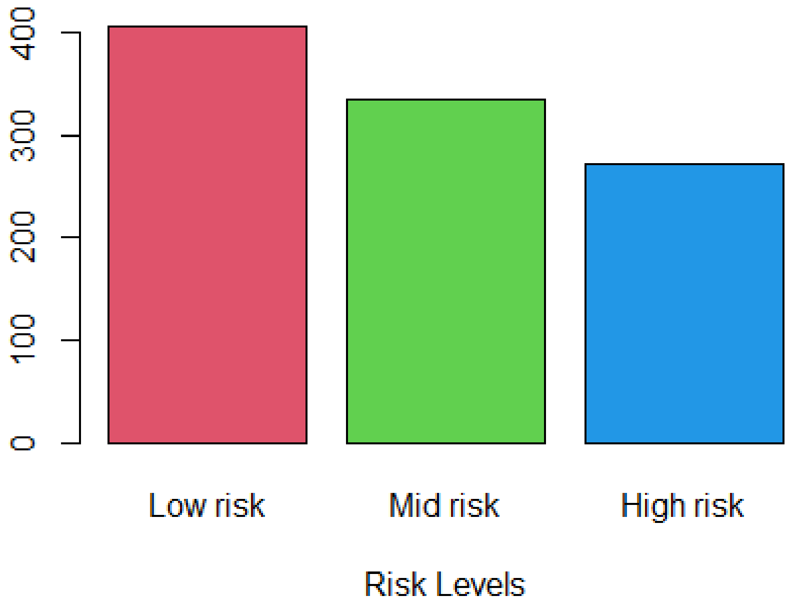
```
summary(d)

##        V1              V2              V3                V4
##  Min.   :10.00   Min.   : 70.0   Min.   : 49.00   Min.   : 6.000
##  1st Qu.:19.00   1st Qu.:100.0   1st Qu.: 65.00   1st Qu.: 6.900
##  Median :26.00   Median :120.0   Median : 80.00   Median : 7.500
##  Mean   :29.87   Mean   :113.2   Mean   : 76.46   Mean   : 8.726
##  3rd Qu.:39.00   3rd Qu.:120.0   3rd Qu.: 90.00   3rd Qu.: 8.000
##  Max.   :70.00   Max.   :160.0   Max.   :100.00   Max.   :19.000
##        V5               V6              V7
##  Min.   : 98.00   Min.   : 7.0   Min.   :1.000
##  1st Qu.: 98.00   1st Qu.:70.0   1st Qu.:1.000
##  Median : 98.00   Median :76.0   Median :2.000
##  Mean   : 98.67   Mean   :74.3   Mean   :1.868
##  3rd Qu.: 98.00   3rd Qu.:80.0   3rd Qu.:3.000  
##  Max.   :103.00   Max.   :90.0   Max.   :3.000
```

EXPLORATORY DATA ANALYSIS

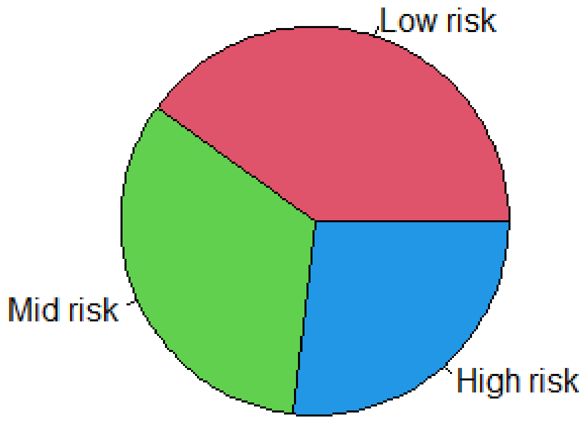
RISK LEVELS

For categorical variables, we'll just checking the frequency distribution of the data using bar plot. Another way to show the relationships between classes or categories of a variable is in a pie or circle chart. In a pie chart, each "slice" represents the proportion of the total phenomenon that is due to each of the classes or groups.

```
Y=d[,7]
summary(Y)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   1.000   2.000   1.868   3.000   3.000

a=c(length(which(Y==1)),length(which(Y==2)),length(which(Y==3)))
a

## [1] 406 336 272



percentage=(a*100)/sum(a)
Risklevel=c("Low risk","Mid risk","High risk")
data.frame(Risklevel,percentage)



##   Risklevel percentage
## 1  Low risk   40.03945
## 2  Mid risk   33.13609
## 3 High risk   26.82446



barplot(a,col=c(2,3,4),names.arg =Risklevel,border=T,xlab ="Risk Level
s")
```



```
pie(percentage,Risklevel,radius=1,col=c(2,3,4))
```



Looks like most pregnant women in this dataset mostly has low health risk. Out of 1014 observations, 406 (40%) of pregnant women has low risk, 336 (33.1%) has mid risk, and 272 (26.8%) has high risk.We will explore the data to get more insights and see why pregnant women has different health risk. We will try to check each variable that can affect it.

HISTOGRAM

It is approximate representation of distribution of data. Categorical variables only has a few values that represent different classes/categories, numerical variables has a continuous value. Therefore to understand the distribution of data on each variable, we'll use histogram instead of bar chart.

BOX PLOT

A boxplot is a standardized way of displaying the dataset based on the five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles .

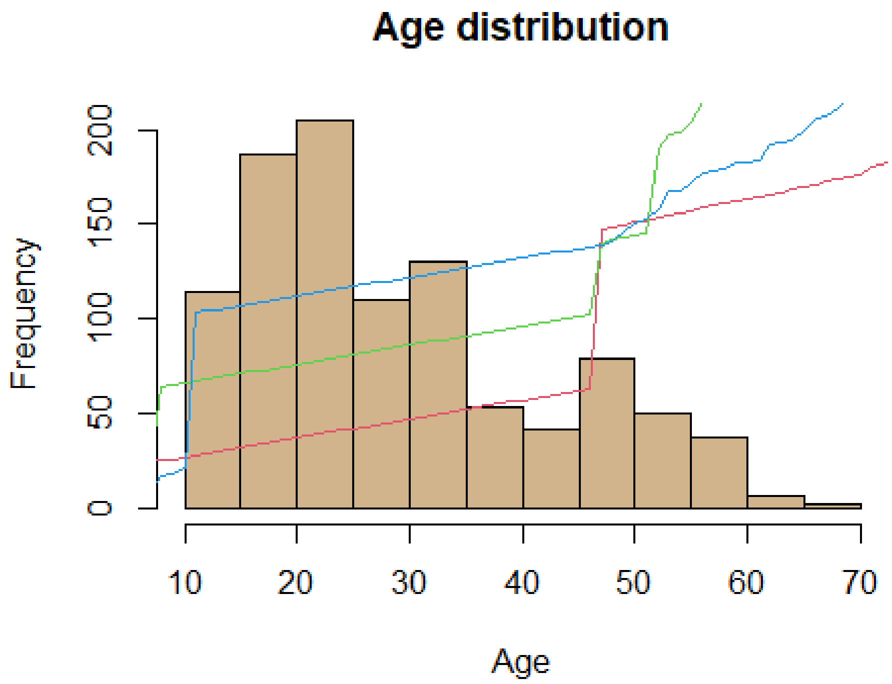
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. If the outliers are non-randomly distributed, they can decrease normality (making the graph skewed). It increases the error variance and reduces the power of statistical tests. They can cause bias and/or influence estimates. We will use box plot to visualize continous data to find out whether there are outliers.

AGE

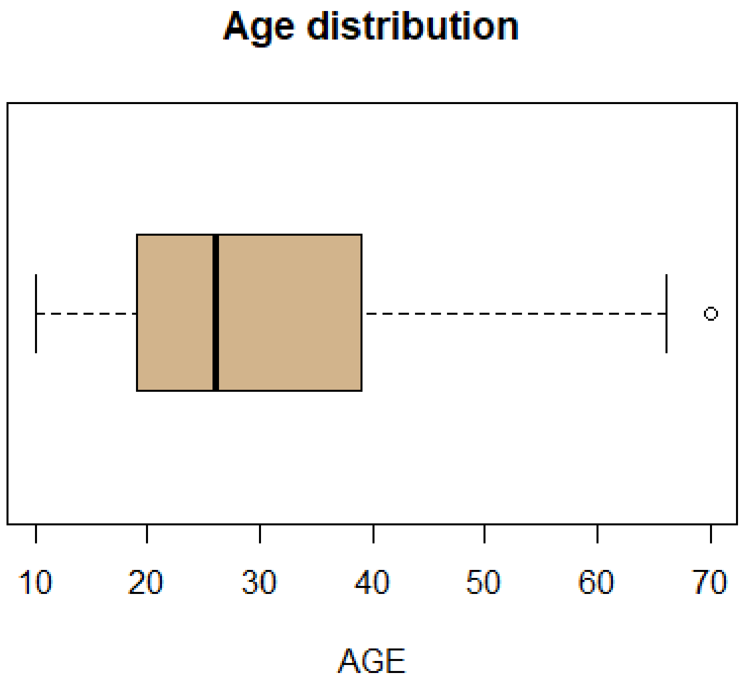
```
Age=d[,1]
summary(Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   19.00   26.00   29.87   39.00   70.00
```

```
hist(Age,col="tan",border=T,main="Age distribution")
lines(which(Y==1),type="l",col="red")
lines(which(Y==2),type="l",col="green")
lines(which(Y==3),type="l",col="blue")
```



```
boxplot(Age,horizontal = T,xlab="AGE",col="tan",main="Age distribution
")
```



The mean of the column age is 30 year.

The median of the column Age is 26 years.

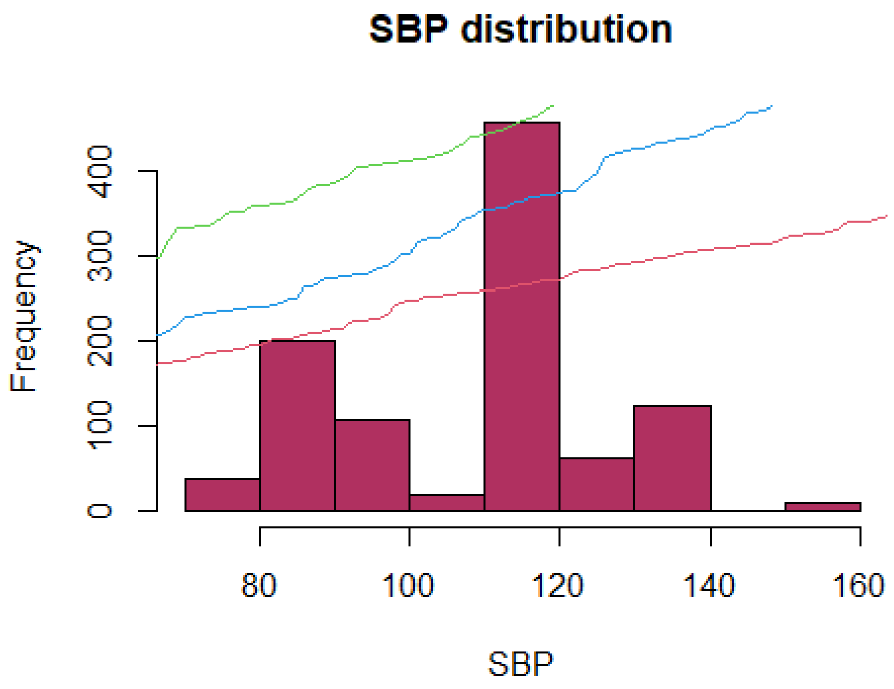
Minimum age of pregnant woman is 10 years and Maximum is 70 years.

There is one outlier present which is 70.

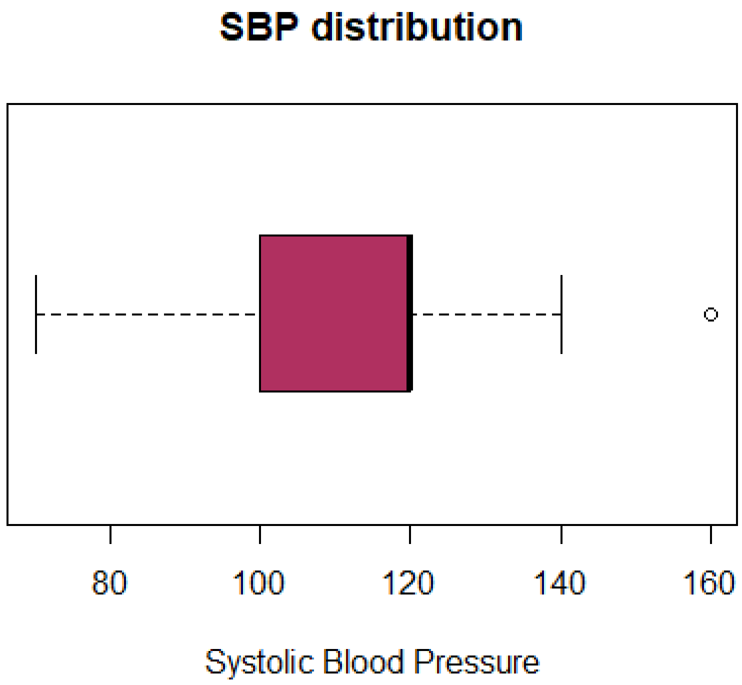
Systolic blood pressure

The systolic blood pressure measures the force of blood against the artery walls while the ventricles - the lower two chambers of your heart- squeeze, pushing blood out to the rest of the body.

```
SBP=d[,2]
summary(SBP)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    70.0   100.0   120.0   113.2   120.0   160.0

hist(SBP,col="maroon",border=T,main="SBP distribution")
lines(which(Y==1),type="l",col="red")
lines(which(Y==2),type="l",col="green")
lines(which(Y==3),type="l",col="blue")
```



```
boxplot(SBP,horizontal = T,xlab="Systolic Blood Pressure",col="maroon"
,main="SBP distribution")
```

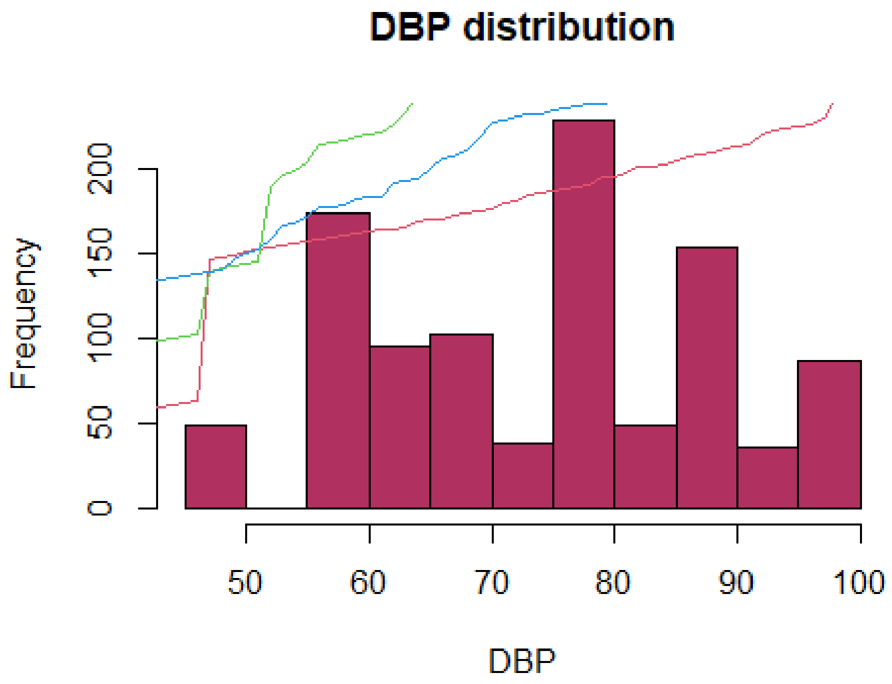


The normal range of systolic blood pressure is less than 120 mmHg. Range of SBP in data is 70 mmHg to 160 mmHg. 75% of pregnant women have normal systolic blood pressure in data.

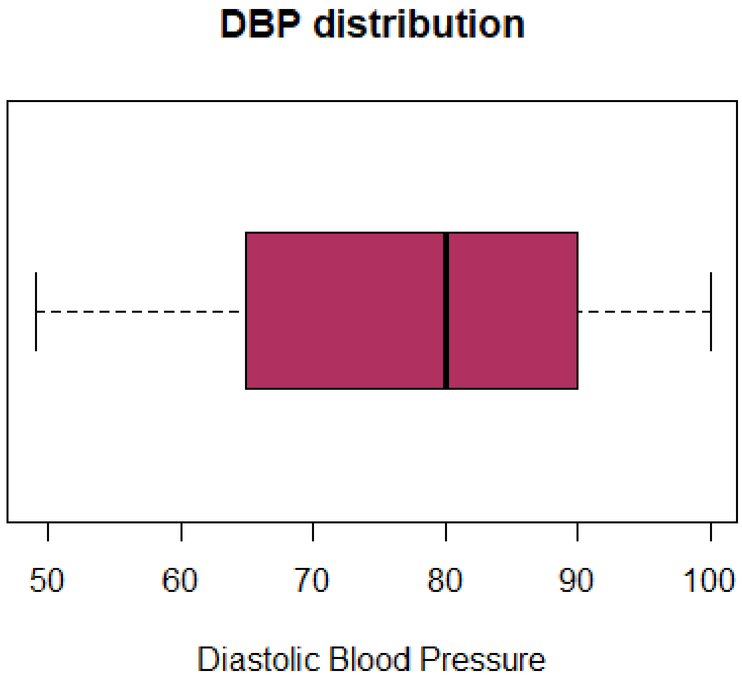
Diastolic blood pressure

The diastolic blood pressure measures the force of blood against the artery walls as the heart relaxes and the ventricles are allowed to refill with blood

```
DBP=d[,3]
summary(DBP)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.00   65.00   80.00   76.46   90.00  100.00

hist(DBP,col="maroon",border=T,main="DBP distribution")
lines(which(Y==1),type="l",col="red")
lines(which(Y==2),type="l",col="green")
lines(which(Y==3),type="l",col="blue")
```



```
boxplot(DBP,horizontal = T,xlab="Diastolic Blood Pressure",col="maroon",main="DBP distribution")
```

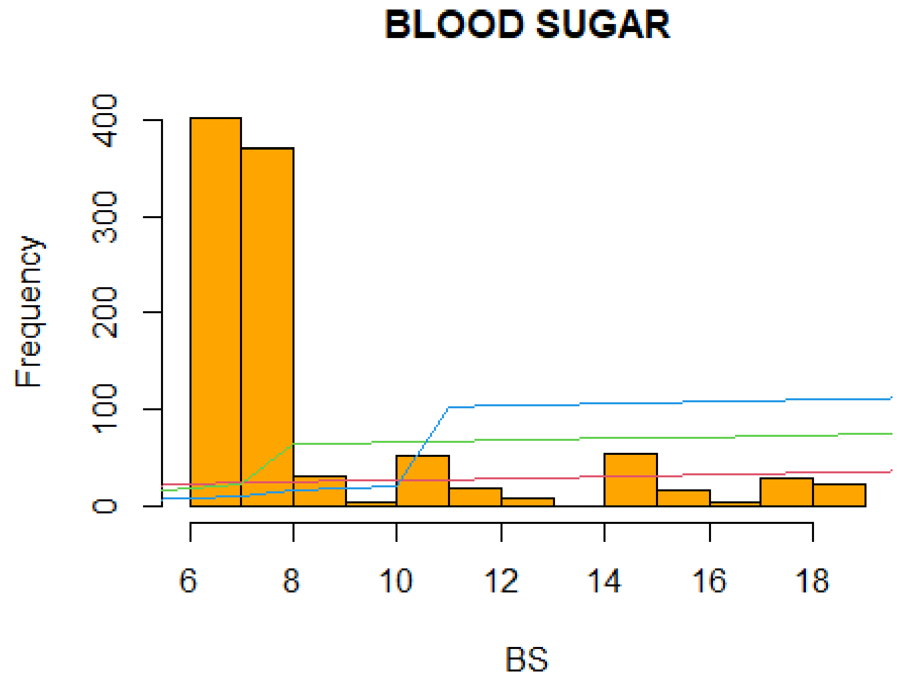


The normal range of diastolic blood pressure is <90 mmHg. Range of DBP in data is 49 to 100 we can see that third quartile of DBP is 90 ,therefore 75% of population have normal DBP .

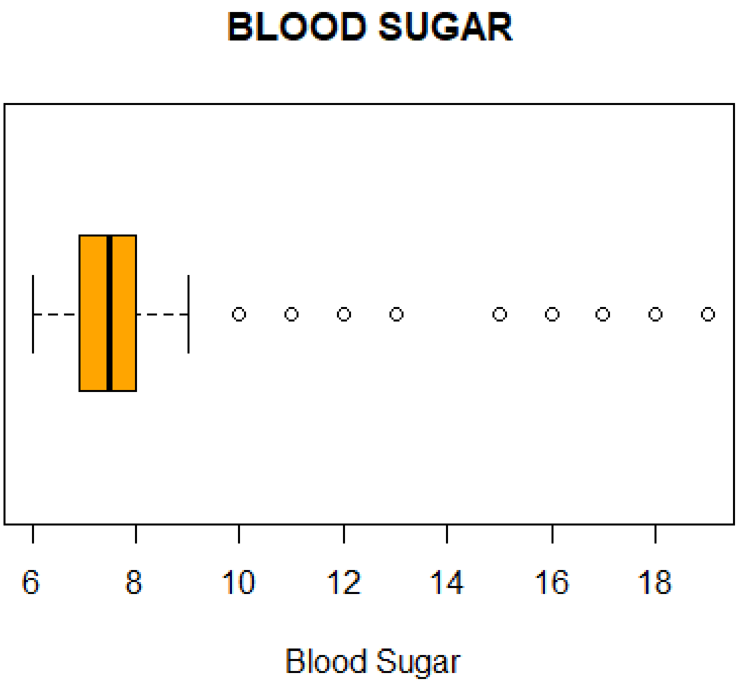
BLOOD SUGAR

The blood sugar level is the amount of glucose in the blood. Glucose is a sugar that comes from the foods we eat, and it’s also formed and stored inside the body. It’s the main source of energy for the cells of our body, and it’s carried to each cell through the bloodstream.

```
BS=d[,4]
summary(BS)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.000   6.900   7.500   8.726   8.000  19.000

hist(BS,col="orange",border=T,main="BLOOD SUGAR")
lines(which(Y==1),type="l",col="red")
lines(which(Y==2),type="l",col="green")
lines(which(Y==3),type="l",col="blue")
```



```
boxplot(BS,horizontal = T,xlab=" Blood Sugar",col="orange",main="BLOOD
SUGAR")
```



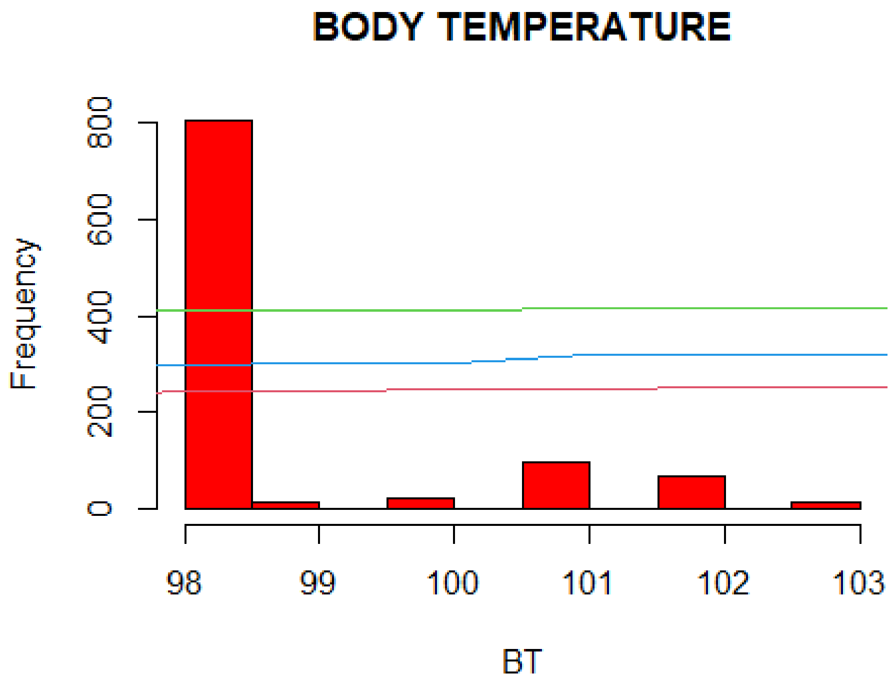
A blood sugar level less than 140 mg/dl. (7.8 mmol/L) is normal. A reading of more than 200 mg/dL (11.1 mmol/L) after two hours indicates diabetes. A reading between 140 and 199 mg/dL (7.8 mmol/L and 11.0 mmol/L) indicates prediabetes.

From histogram we can see that most of pregnant women have normal blood sugar level.

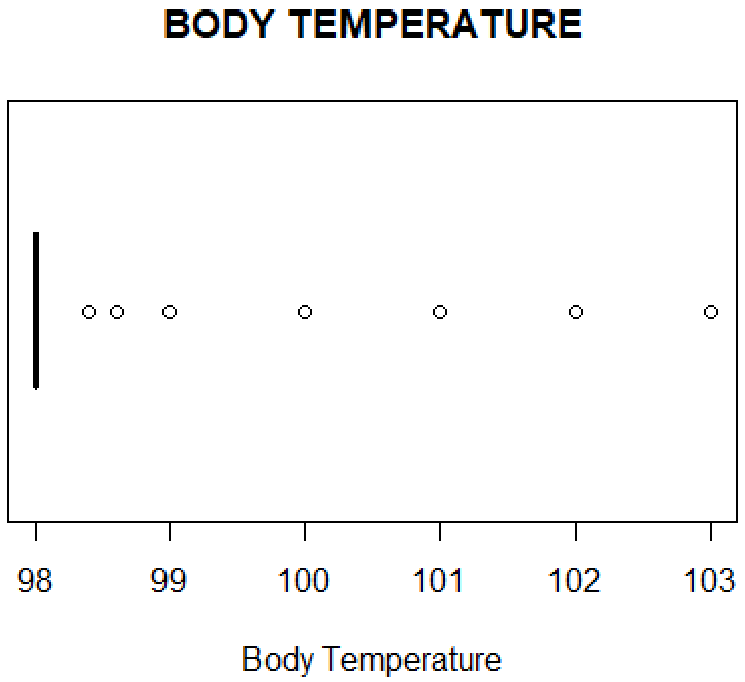
From boxplot we can see there are some outliers are present in data.

BODY TEMPERATURE

```
BT=d[,5]
summary(BT)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   98.00   98.00   98.00   98.67   98.00  103.00

hist(BT,col="red",border=T,main="BODY TEMPERATURE")
lines(which(Y==1),type="l",col="red")
lines(which(Y==2),type="l",col="green")
lines(which(Y==3),type="l",col="blue")
```



```
boxplot(BT,horizontal = T,xlab="Body Temperature",col="red",main="BODY
TEMPERATURE")
```

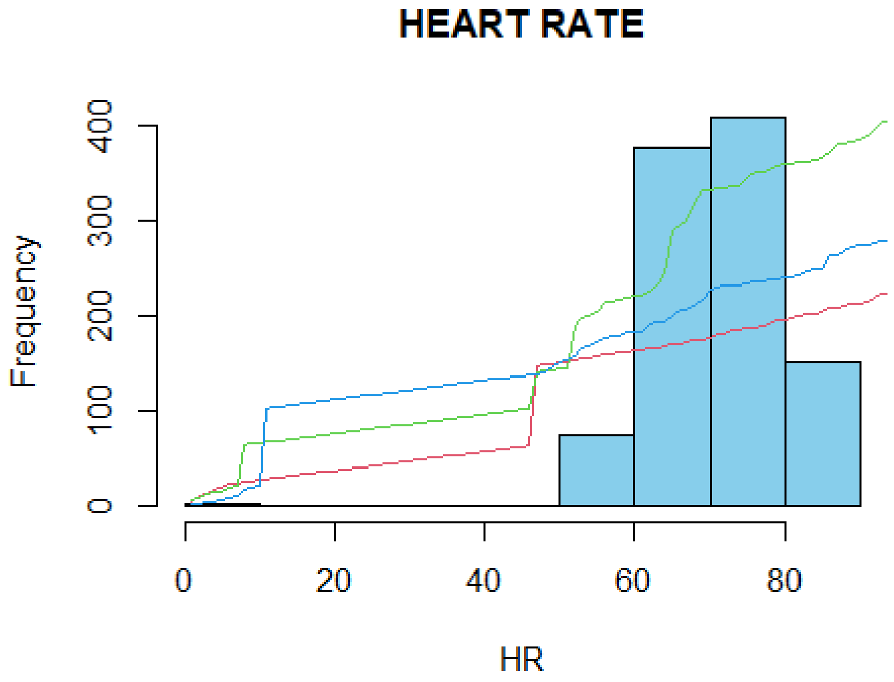


The normal body temperature is around 98 fahrenheit. Almost all preganant women have normal body temperature.

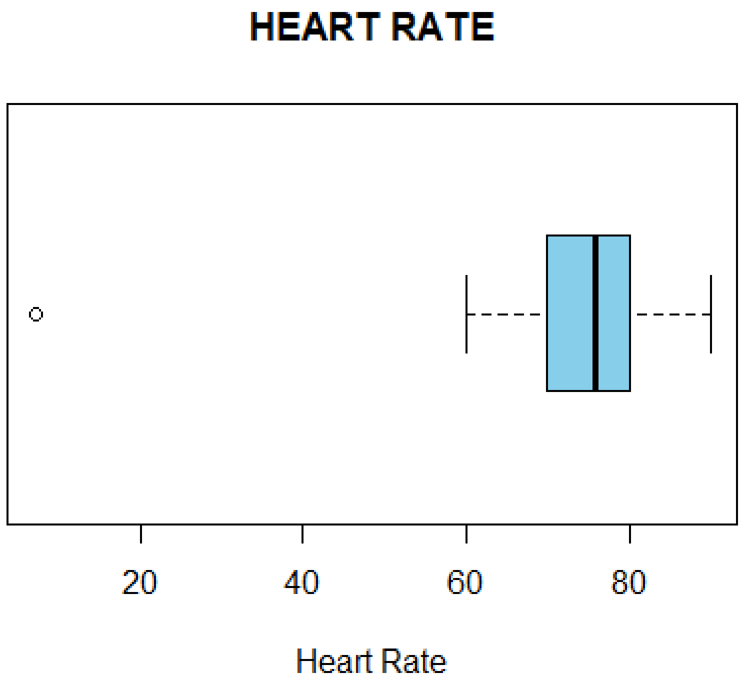
HEART RATE

The number of heartbeats per unit of time, usually per minute. The heart rate is based on the number of contractions of the ventricles the lower chambers of the heart.

```
HR=d[,6]
summary(HR)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.0    70.0    76.0    74.3    80.0    90.0

hist(HR,col="skyblue",border=T,main="HEART RATE")
lines(which(Y==1),type="l",col="red")
lines(which(Y==2),type="l",col="green")
lines(which(Y==3),type="l",col="blue")
```



```
boxplot(HR,horizontal = T,xlab="Heart Rate",col="skyblue",main="HEART
RATE")
```



The ideal heart rate range is 50 to 75 beats per minute. From boxplot we can see that their is outlier with value 7 which is not possible

OUTLIER DETECTION AND HANDLING

Outlier is data point that differ significantly from other observations. As we know from the previous analysis, in this dataset there are several variables that have outlers, but even so most of those values still make sense in real life. The only variable that has an outlier with an unreasonable value is HeartRate. In this vanable, there are two observations that have a heart rate value of 7 bpm (beats per minute). A normal resting heart rate for adults ranges from 60 to 100 beats per minute, and the lowest recorded resting heart rate in human history was 25 bpm. Therefore, we will drop this 2 records that has a heart rate value of 7 because that value doesn't make any sense, and most likely is an input error.

```
l=lm(Y~HR)
yh=l$fitted.values
e=Y-yh
msres=anova(l)$"Mean Sq"[3]
a=rep(1,n)
h=c()
x=as.matrix(cbind(a,HR))
```

```
H=x%*%solve(t(x)%*%x)%*%t(x)
for( i in 1:n)
{
  h[i]=H[i,i]
}

c1=2*p/n
b=which(h>c1)
b

## [1] 500 909
```

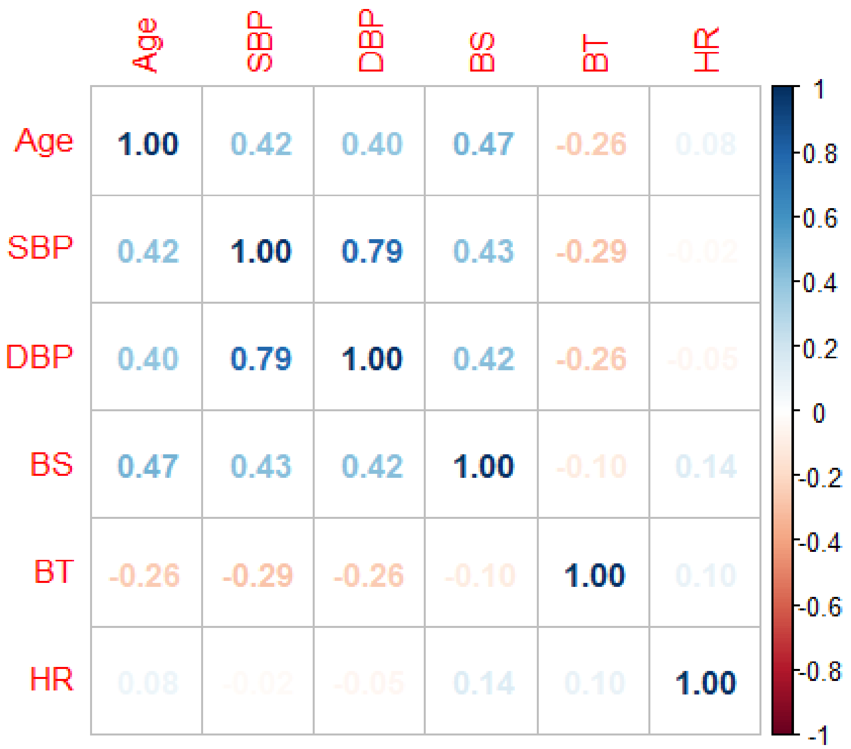
The observation 500 and 909 are outlier so we are goin to remove them from data

```
d=data.frame(Age,SBP,DBP,BS,BT,HR,Y)
d=d[-b,]
```

CORRELATION PLOT

Correlation is a statistical measure that expresses the extent to which two variables are linearly related

```
library(corrplot)

## corrplot 0.92 loaded

d=data.frame(Age,SBP,DBP,BS,BT,HR)
cor=cor(d)
corrplot(cor,method="number")
```



Systolic BP and Diastolic BP are highly correlated. As we can see from the graph, they have positive correlation with correlation coefficient value = 0.79. This means that SystolicBP and DiastolicBP variable contains highly similar information and there is very little or no variance in information. This is known as a problem called MultiColinearity, which undermines the statistical significance of an independent variable. We can remove one of them because we don't want a redundant variable while making or training our model. But we will try to dig deeper to decide whether we need to remove these variable, and which variable we should remove.

BUILDING CLASSIFICATION MODELS

1.RANDOM FOREST

Random forest is a versatile machine learning method capable of performing both regression and classification tasks. It is also used for dimensionality reduction, treats missing values, outlier values. It is a type of supervised learning method, where a group of weak models combine to form a powerful model. In Random Forest, we grow multiple trees as opposed to a single tree. To classify a new object based on attributes, each tree gives a classification.The forest chooses the classification having the most votes(Overall the trees in the forest) and in case of regression,it takes the average of outputs by different trees. We choose Random Forest because it's one of the most accurate learning algorithms available, and it's training time is fast. Also, Random Forest can work on dataset that have feature values with different scales, so we don't need to do normalization/feature scaling.

2.SUPPORT VECTOR MACHINE (SVM)

SVM stands for support vector machine, it is a supervised machine learning algorithm which can be used for both Regression and Classification. If you have n features in your training data set, SVM tries to plot it in n-dimensional space with the value of each feature being the value of a particular coordinate. SVM uses hyperplanes to separate out different classes based on the provided kernel function.

3.LOGISTIC REGRESSION

Logistic Regression often referred to as the logit model is a technique to predict the binary outcome from a linear combination of predictor variables.

The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.So to combat overfittting and underfitting we are going to use k-fold cross validation.

K-FOLD CROSS VALIDATION

K-Fold CV is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point.

```
library(randomForest)

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

library(e1071)

library(nnet)

## Warning: package 'nnet' was built under R version 4.1.3

d=data.frame(Age,DBP,BS,BT,HR,Y)
d=d[-b,]
kf=10
n=nrow(d)
p=ncol(d)
```

```
s=sample(1:n,size=n,replace = FALSE)
testrf=testsvm=testml=testnb=c()
for(i in 1:kf)
{
  testd=d[s[((i-1)*(n/kf)+1):(i*(n/kf))],]
  traind=d[-s[((i-1)*(n/kf)+1):(i*(n/kf))],]
  mrf=randomForest(as.factor(Y)~.,traind,importance=T,ntree=1000)
  rfp=predict(mrf,newdata=testd)
  trf=table(rfp,testd[,p])
  testrf[i]=(sum(trf)-sum(diag(trf)))/sum(trf)
   #svm
  msvm=svm(as.factor(Y)~.,data=traind,type='C-classification',kernel='
linear')
  psvm=predict(msvm,newdata=testd)
  tsvm=table(psvm,testd[,p])
  testsvm[i]=(sum(tsvm)-sum(diag(tsvm)))/sum(tsvm)
  #mlr
  ml=multinom(Y~.,data=traind)
  mlp=predict(ml,newdata=testd)
  tml=table(mlp,testd[,p])
  testml[i]=(sum(tml)-sum(diag(tml)))/sum(tml)





}
```

Confusion matrix for random forest model

```
trf

##
## rfp  1  2  3
##   1 31  5  2
##   2  7 31  1
##   3  1  2 21

terf=mean(testrf)*100
terf

## [1] 14.85149

accuracy1=(100-terf)
accuracy1

## [1] 85.14851
```

The accuracy of random forest classification model is 83%

Confusion matrix of SVM model.

```
tsvm

##
## psvm  1  2  3
##    1 36 24  6
##    2  2  9  4
##    3  1  5 14

tesvm=mean(testsvm)*100
tesvm

## [1] 39.70297

accuracy2=(100-tesvm)
accuracy2

## [1] 60.29703
```

The accuracy of support vector machine model is 60.59%

Confusion matrix of logistic regression model

```
tml

##
## mlp  1  2  3
##   1 35 25  3
##   2  2  5  5
##   3  2  8 16

teml=mean(testml)*100
teml

## [1] 39.90099

accuracy3=(100-teml)
accuracy3

## [1] 60.09901
```

The accuracy of logistic regression model is 59.90%

CONCLUSIONS

1.Pregnant women who have high blood glucose level tend to have high health risks.

2. Blood Sugar has positive correlation to Age, Systolic Blood Pressure, and Diastolic Blood Pressure, so pregnant women who have high Age, Systolic Blood Pressure,, and Diastolic Blood Pressure need to be careful.

3. Age is also a fairly important variable, where the health risks of pregnant women seem to start to increase starting from the age of 25 years.

4. For Systolic Blood Pressure, and Diastolic Blood Pressure, these two variables actually have a strong relationship, as evidenced by the correlation coefficient value of 0.79.

5. About Body Temperature, this variable is actually not giving much information because more than 79% of the total value is 98F. But from this variable, we know that pregnant women who have a body temperature above 98.4F tend to have a greater health risk.