
# AMERICAN INTERNATIONAL UNIVERSITY

# BANGLADESH

## Final Term Project

### Introduction To Data Science

**Name: Shakib Sadat Shanto**

**ID: 20-43074-1**

**Section: D**

## Project Overview:

In this project, basic data analysis and preprocessing was conducted on selected dataset and K Nearest Neighbors algorithm was applied on the dataset using R programming language. Confusion matrix of the model have also been discussed in the Confusion Matrix section of the report.

## Dataset Overview:

The dataset used for the project was collected from Kaggle. The name of the dataset is **Heart Failure Prediction Dataset**. Here is the URL of the dataset:

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

The dataset contains 918 observations. It has 11 independent variables.

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

1 dependent or class variable.

1. HeartDisease: output class [1: heart disease, 0: Normal]

# Data Analysis and Preprocess:

1. Importing Dataset:

Code:

```
dataset = read.csv('heart.csv')
```

Output:

| | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 40 | M | ATA | 140 | 289 | 0 | Normal | 172 | N | 0.0 | Up | 0 |
| 2 | 49 | F | NAP | 160 | 180 | 0 | Normal | 156 | N | 1.0 | Flat | 1 |
| 3 | 37 | M | ATA | 130 | 283 | 0 | ST | 98 | N | 0.0 | Up | 0 |
| 4 | 48 | F | ASY | 138 | 214 | 0 | Normal | 108 | Y | 1.5 | Flat | 1 |
| 5 | 54 | M | NAP | 150 | 195 | 0 | Normal | 122 | N | 0.0 | Up | 0 |
| 6 | 39 | M | NAP | 120 | 339 | 0 | Normal | 170 | N | 0.0 | Up | 0 |
| 7 | 45 | F | ATA | 130 | 237 | 0 | Normal | 170 | N | 0.0 | Up | 0 |
| 8 | 54 | M | ATA | 110 | 208 | 0 | Normal | 142 | N | 0.0 | Up | 0 |
| 9 | 37 | M | ASY | 140 | 207 | 0 | Normal | 130 | Y | 1.5 | Flat | 1 |
| 10 | 48 | F | ATA | 120 | 284 | 0 | Normal | 120 | N | 0.0 | Up | 0 |
| 11 | 37 | F | NAP | 130 | 211 | 0 | Normal | 142 | N | 0.0 | Up | 0 |
| 12 | 58 | M | ATA | 136 | 164 | 0 | ST | 99 | Y | 2.0 | Flat | 1 |
| 13 | 39 | M | ATA | 120 | 204 | 0 | Normal | 145 | N | 0.0 | Up | 0 |
| 14 | 49 | M | ASY | 140 | 234 | 0 | Normal | 140 | Y | 1.0 | Flat | 1 |
| 15 | 42 | F | NAP | 115 | 211 | 0 | ST | 137 | N | 0.0 | Up | 0 |
| 16 | 54 | F | ATA | 120 | 273 | 0 | Normal | 150 | N | 1.5 | Flat | 0 |
| 17 | 38 | M | ASY | 110 | 196 | 0 | Normal | 166 | N | 0.0 | Flat | 1 |
| 18 | 43 | F | ATA | 120 | 201 | 0 | Normal | 165 | N | 0.0 | Up | 0 |
| 19 | 60 | M | ASY | 100 | 248 | 0 | Normal | 125 | N | 1.0 | Flat | 1 |
| 20 | 36 | M | ATA | 120 | 267 | 0 | Normal | 160 | N | 3.0 | Flat | 1 |
| 21 | 43 | F | TA | 100 | 223 | 0 | Normal | 142 | N | 0.0 | Up | 0 |
| 22 | 44 | M | ATA | 120 | 184 | 0 | Normal | 142 | N | 1.0 | Flat | 0 |

2. Structure of the Dataset:

Code:

```
str(dataset)
```

Output:

```
> str(dataset)
'data.frame':   918 obs. of  12 variables:
 $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
 $ Sex           : chr  "M" "F" "M" "F" ...
 $ ChestPainType : chr  "ATA" "NAP" "ATA" "ASY" ...
 $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
 $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
 $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ RestingECG    : chr  "Normal" "Normal" "ST" "Normal" ...
 $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
 $ ExerciseAngina: chr  "N" "N" "N" "Y" ...
 $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
 $ ST_Slope      : chr  "Up" "Flat" "Up" "Flat" ...
 $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

3. Attributes:

Code:

```
ls(dataset)
```

Output:

```
> ls(dataset)
 [1] "Age"           "ChestPainType"  "Cholesterol"    "ExerciseAngina" "FastingBS"      "HeartDisease"   "MaxHR"
 [8] "Oldpeak"       "RestingBP"      "RestingECG"     "Sex"            "ST_Slope"
```

4. Unique Values of Categorical Attributes:

Code:

```
unique(dataset$Sex)
unique(dataset$ChestPainType)
unique(dataset$RestingECG)
unique(dataset$ExerciseAngina)
unique(dataset$ST_Slope)
```

Output:

```
> unique(dataset$Sex)
[1] "M" "F"
> unique(dataset$ChestPainType)
[1] "ATA" "NAP" "ASY" "TA"
> unique(dataset$RestingECG)
[1] "Normal" "ST"     "LVH"
> unique(dataset$ExerciseAngina)
[1] "N" "Y"
> unique(dataset$ST_Slope)
[1] "Up"   "Flat" "Down"
```

5. Check for missing values:

Code:

```
colSums(is.na(dataset))
```

Output:

```
> colSums(is.na(dataset))
          Age           Sex  ChestPainType      RestingBP    Cholesterol      FastingBS     RestingECG 
            0             0             0             0             0             0             0 
        MaxHR ExerciseAngina       Oldpeak      ST_Slope   HeartDisease 
            0             0             0             0             0 
```

6. Encoding Categorical Values into Numeric Values:

Code:

```
dataset$Sex = factor(dataset$Sex,
                     levels = c("M","F"),
                     labels = c(0,1))
dataset$ChestPainType = factor(dataset$ChestPainType,
                      levels = c("ATA","NAP","ASY","TA"),
                      labels = c(1,2,3,4))
dataset$RestingECG = factor(dataset$RestingECG,
                      levels = c("Normal","ST","LVH"),
                      labels = c(1,2,3))
dataset$ExerciseAngina = factor(dataset$ExerciseAngina,
                        levels = c("N","Y"),
                        labels = c(0,1))
dataset$ST_Slope = factor(dataset$ST_Slope,
                       levels = c("Up","Flat","Down"),
                       labels = c(1,2,3))
```

Output:

| | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 40 | 0 | 1 | 140 | 289 | 0 | 1 | 172 | 0 | 0.0 | 1 | 0 |
| 2 | 49 | 1 | 2 | 160 | 180 | 0 | 1 | 156 | 0 | 1.0 | 2 | 1 |
| 3 | 37 | 0 | 1 | 130 | 283 | 0 | 2 | 98 | 0 | 0.0 | 1 | 0 |
| 4 | 48 | 1 | 3 | 138 | 214 | 0 | 1 | 108 | 1 | 1.5 | 2 | 1 |
| 5 | 54 | 0 | 2 | 150 | 195 | 0 | 1 | 122 | 0 | 0.0 | 1 | 0 |
| 6 | 39 | 0 | 2 | 120 | 339 | 0 | 1 | 170 | 0 | 0.0 | 1 | 0 |
| 7 | 45 | 1 | 1 | 130 | 237 | 0 | 1 | 170 | 0 | 0.0 | 1 | 0 |
| 8 | 54 | 0 | 1 | 110 | 208 | 0 | 1 | 142 | 0 | 0.0 | 1 | 0 |
| 9 | 37 | 0 | 3 | 140 | 207 | 0 | 1 | 130 | 1 | 1.5 | 2 | 1 |
| 10 | 48 | 1 | 1 | 120 | 284 | 0 | 1 | 120 | 0 | 0.0 | 1 | 0 |
| 11 | 37 | 1 | 2 | 130 | 211 | 0 | 1 | 142 | 0 | 0.0 | 1 | 0 |
| 12 | 58 | 0 | 1 | 136 | 164 | 0 | 2 | 99 | 1 | 2.0 | 2 | 1 |
| 13 | 39 | 0 | 1 | 120 | 204 | 0 | 1 | 145 | 0 | 0.0 | 1 | 0 |
| 14 | 49 | 0 | 3 | 140 | 234 | 0 | 1 | 140 | 1 | 1.0 | 2 | 1 |
| 15 | 42 | 1 | 2 | 115 | 211 | 0 | 2 | 137 | 0 | 0.0 | 1 | 0 |
| 16 | 54 | 1 | 1 | 120 | 273 | 0 | 1 | 150 | 0 | 1.5 | 2 | 0 |
| 17 | 38 | 0 | 3 | 110 | 196 | 0 | 1 | 166 | 0 | 0.0 | 2 | 1 |
| 18 | 43 | 1 | 1 | 120 | 201 | 0 | 1 | 165 | 0 | 0.0 | 1 | 0 |
| 19 | 60 | 0 | 3 | 100 | 248 | 0 | 1 | 125 | 0 | 1.0 | 2 | 1 |
| 20 | 36 | 0 | 1 | 120 | 267 | 0 | 1 | 160 | 0 | 3.0 | 2 | 1 |
| 21 | 43 | 1 | 4 | 100 | 223 | 0 | 1 | 142 | 0 | 0.0 | 1 | 0 |

7. Feature Scaling (Normalization):

Code:

```
install.packages("caret")
library(caret)
process = preProcess(as.data.frame(dataset), method=c("range"))
dataset = predict(process, as.data.frame(dataset))
```

Output:

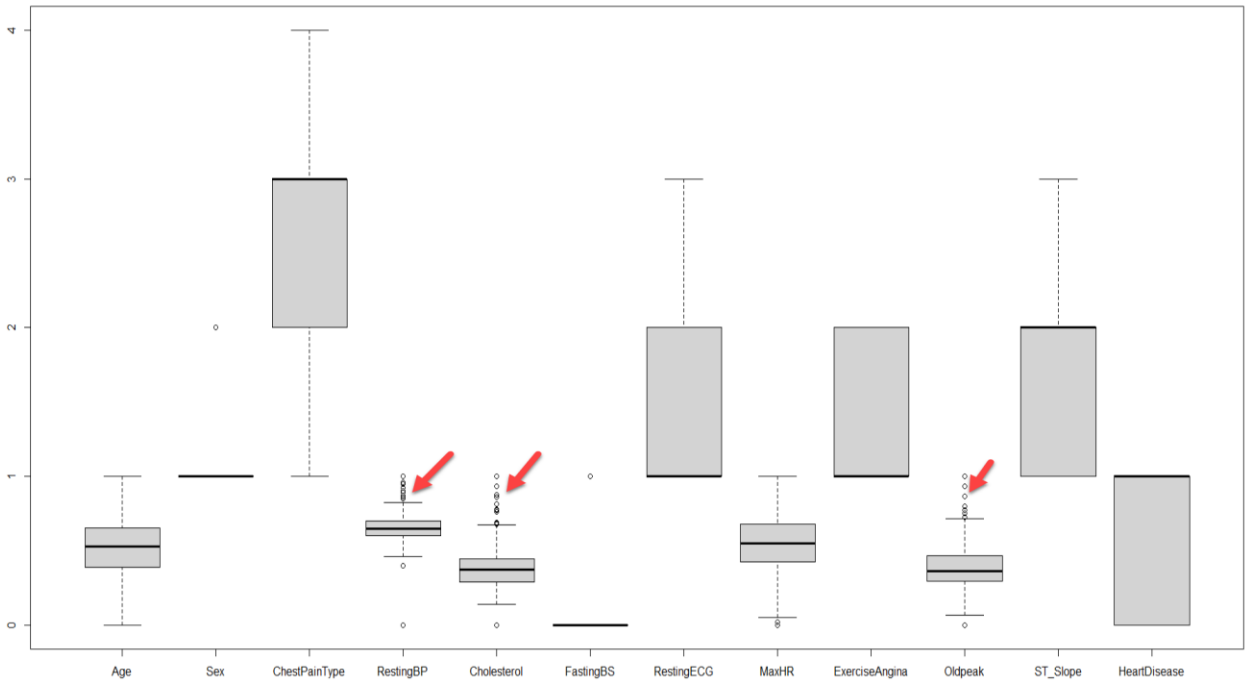
| | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.24489796 | 0 | 1 | 0.700 | 0.4792703 | 0 | 1 | 0.7887324 | 0 | 0.2954545 | 1 | 0 |
| 2 | 0.42857143 | 1 | 2 | 0.800 | 0.2985075 | 0 | 1 | 0.6760563 | 0 | 0.4090909 | 2 | 1 |
| 3 | 0.18367347 | 0 | 1 | 0.650 | 0.4693201 | 0 | 2 | 0.2676056 | 0 | 0.2954545 | 1 | 0 |
| 4 | 0.40816327 | 1 | 3 | 0.690 | 0.3548922 | 0 | 1 | 0.3380282 | 1 | 0.4659091 | 2 | 1 |
| 5 | 0.53061224 | 0 | 2 | 0.750 | 0.3233831 | 0 | 1 | 0.4366197 | 0 | 0.2954545 | 1 | 0 |
| 6 | 0.22448980 | 0 | 2 | 0.600 | 0.5621891 | 0 | 1 | 0.7746479 | 0 | 0.2954545 | 1 | 0 |
| 7 | 0.34693878 | 1 | 1 | 0.650 | 0.3930348 | 0 | 1 | 0.7746479 | 0 | 0.2954545 | 1 | 0 |
| 8 | 0.53061224 | 0 | 1 | 0.550 | 0.3449420 | 0 | 1 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |
| 9 | 0.18367347 | 0 | 3 | 0.700 | 0.3432836 | 0 | 1 | 0.4929577 | 1 | 0.4659091 | 2 | 1 |
| 10 | 0.40816327 | 1 | 1 | 0.600 | 0.4709784 | 0 | 1 | 0.4225352 | 0 | 0.2954545 | 1 | 0 |
| 11 | 0.18367347 | 1 | 2 | 0.650 | 0.3499171 | 0 | 1 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |
| 12 | 0.61224490 | 0 | 1 | 0.680 | 0.2719735 | 0 | 2 | 0.2746479 | 1 | 0.5227273 | 2 | 1 |
| 13 | 0.22448980 | 0 | 1 | 0.600 | 0.3383085 | 0 | 1 | 0.5985915 | 0 | 0.2954545 | 1 | 0 |
| 14 | 0.42857143 | 0 | 3 | 0.700 | 0.3880597 | 0 | 1 | 0.5633803 | 1 | 0.4090909 | 2 | 1 |
| 15 | 0.28571429 | 1 | 2 | 0.575 | 0.3499171 | 0 | 2 | 0.5422535 | 0 | 0.2954545 | 1 | 0 |
| 16 | 0.53061224 | 1 | 1 | 0.600 | 0.4527363 | 0 | 1 | 0.6338028 | 0 | 0.4659091 | 2 | 0 |
| 17 | 0.20408163 | 0 | 3 | 0.550 | 0.3250415 | 0 | 1 | 0.7464789 | 0 | 0.2954545 | 2 | 1 |
| 18 | 0.30612245 | 1 | 1 | 0.600 | 0.3333333 | 0 | 1 | 0.7394366 | 0 | 0.2954545 | 1 | 0 |
| 19 | 0.65306122 | 0 | 3 | 0.500 | 0.4112769 | 0 | 1 | 0.4577465 | 0 | 0.4090909 | 2 | 1 |
| 20 | 0.16326531 | 0 | 1 | 0.600 | 0.4427861 | 0 | 1 | 0.7042254 | 0 | 0.6363636 | 2 | 1 |
| 21 | 0.30612245 | 1 | 4 | 0.500 | 0.3698176 | 0 | 1 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |

## 8. Outlier Detection and Removing:

Code:

boxplot(dataset)

Output:



## Code:

```
summary(dataset$RestingBP)
Iqr_restingbp = .70-.60
upfen_restingbp = .70+1.5*Iqr_restingbp
low_restingbp = .60-1.5*Iqr_restingbp
upfen_restingbp
low_restingbp
summary(dataset$Cholesterol)
Iqr_Cholesterol = 0.4428-0.2873
upfen_Cholesterol = .4428+1.5*Iqr_Cholesterol
low_Cholesterol = 0.2873 -1.5*Iqr_Cholesterol
upfen_Cholesterol
low_Cholesterol
summary(dataset$Oldpeak)
Iqr_Oldpeak = 0.4659-0.2955
upfen_Oldpeak = 0.4659+1.5*Iqr_Oldpeak
low_Oldpeak = 0.2955 -1.5*Iqr_Oldpeak
upfen_Oldpeak
low_Oldpeak
```

Output:

```
> summary(dataset$RestingBP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.600   0.650   0.662   0.700   1.000
> Iqr_restingbp = .70-.60
> upfen_restingbp = .70+1.5*Iqr_restingbp
> low_restingbp = .60-1.5*Iqr_restingbp
> upfen_restingbp
[1] 0.85
> low_restingbp
[1] 0.45
>
> summary(dataset$Cholesterol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2873  0.3698  0.3297  0.4428  1.0000
> Iqr_Cholesterol = 0.4428-0.2873
> upfen_Cholesterol = .4428+1.5*Iqr_Cholesterol
> low_Cholesterol = 0.2873 -1.5*Iqr_Cholesterol
> upfen_Cholesterol
[1] 0.67605
> low_Cholesterol
[1] 0.05405
>
> summary(dataset$Oldpeak)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2955  0.3636  0.3963  0.4659  1.0000
> Iqr_Oldpeak = 0.4659-0.2955
> upfen_Oldpeak = 0.4659+1.5*Iqr_Oldpeak
> low_Oldpeak = 0.2955 -1.5*Iqr_Oldpeak
> upfen_Oldpeak
[1] 0.7215
> low_Oldpeak
[1] 0.0399
```

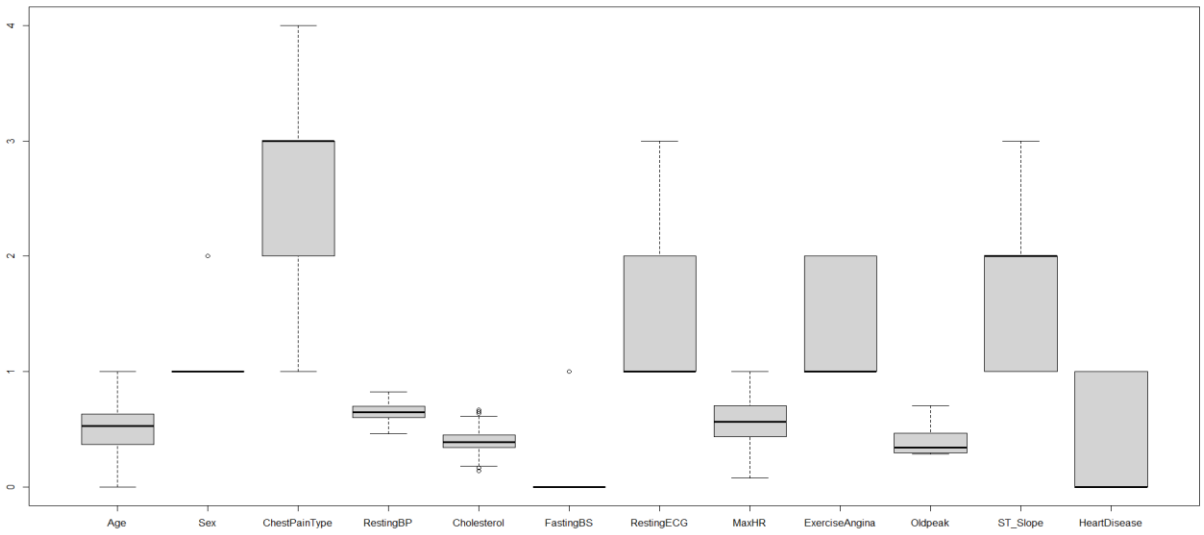
Code:

```
dataset_rm_outlier = subset(dataset, dataset$RestingBP > low_restingbp &
dataset$RestingBP < upfen_restingbp & dataset$Cholesterol > low_Cholesterol &
dataset$Cholesterol < upfen_Cholesterol
                & dataset$Oldpeak > low_Oldpeak & dataset$Oldpeak < upfen_Oldpeak )

boxplot(dataset_rm_outlier)

dataset = dataset_rm_outlier
```

# Output:




| | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.24489796 | 0 | 1 | 0.700 | 0.4792703 | 0 | 1 | 0.7887324 | 0 | 0.2954545 | 1 | 0 |
| 2 | 0.42857143 | 1 | 2 | 0.800 | 0.2985075 | 0 | 1 | 0.6760563 | 0 | 0.4090909 | 2 | 1 |
| 3 | 0.18367347 | 0 | 1 | 0.650 | 0.4693201 | 0 | 2 | 0.2676056 | 0 | 0.2954545 | 1 | 0 |
| 4 | 0.40816327 | 1 | 3 | 0.690 | 0.3548922 | 0 | 1 | 0.3380282 | 1 | 0.4659091 | 2 | 1 |
| 5 | 0.53061224 | 0 | 2 | 0.750 | 0.3233831 | 0 | 1 | 0.4366197 | 0 | 0.2954545 | 1 | 0 |
| 6 | 0.22448980 | 0 | 2 | 0.600 | 0.5621891 | 0 | 1 | 0.7746479 | 0 | 0.2954545 | 1 | 0 |
| 7 | 0.34693878 | 1 | 1 | 0.650 | 0.3930348 | 0 | 1 | 0.7746479 | 0 | 0.2954545 | 1 | 0 |
| 8 | 0.53061224 | 0 | 1 | 0.550 | 0.3449420 | 0 | 1 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |
| 9 | 0.18367347 | 0 | 3 | 0.700 | 0.3432836 | 0 | 1 | 0.4929577 | 1 | 0.4659091 | 2 | 1 |
| 10 | 0.40816327 | 1 | 1 | 0.600 | 0.4709784 | 0 | 1 | 0.4225352 | 0 | 0.2954545 | 1 | 0 |
| 11 | 0.18367347 | 1 | 2 | 0.650 | 0.3499171 | 0 | 1 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |
| 12 | 0.61224490 | 0 | 1 | 0.680 | 0.2719735 | 0 | 2 | 0.2746479 | 1 | 0.5227273 | 2 | 1 |
| 13 | 0.22448980 | 0 | 1 | 0.600 | 0.3383085 | 0 | 1 | 0.5985915 | 0 | 0.2954545 | 1 | 0 |
| 14 | 0.42857143 | 0 | 3 | 0.700 | 0.3880597 | 0 | 1 | 0.5633803 | 1 | 0.4090909 | 2 | 1 |
| 15 | 0.28571429 | 1 | 2 | 0.575 | 0.3499171 | 0 | 2 | 0.5422535 | 0 | 0.2954545 | 1 | 0 |
| 16 | 0.53061224 | 1 | 1 | 0.600 | 0.4527363 | 0 | 1 | 0.6338028 | 0 | 0.4659091 | 2 | 0 |
| 17 | 0.20408163 | 0 | 3 | 0.550 | 0.3250415 | 0 | 1 | 0.7464789 | 0 | 0.2954545 | 2 | 1 |
| 18 | 0.30612245 | 1 | 1 | 0.600 | 0.3333333 | 0 | 1 | 0.7394366 | 0 | 0.2954545 | 1 | 0 |
| 19 | 0.65306122 | 0 | 3 | 0.500 | 0.4112769 | 0 | 1 | 0.4577465 | 0 | 0.4090909 | 2 | 1 |
| 20 | 0.16326531 | 0 | 1 | 0.600 | 0.4427861 | 0 | 1 | 0.7042254 | 0 | 0.6363636 | 2 | 1 |
| 21 | 0.30612245 | 1 | 4 | 0.500 | 0.3698176 | 0 | 1 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |
| 22 | 0.32653061 | 0 | 1 | 0.600 | 0.3051410 | 0 | 1 | 0.5774648 | 0 | 0.4090909 | 2 | 0 |
| 23 | 0.42857143 | 1 | 1 | 0.620 | 0.3333333 | 0 | 1 | 0.7323944 | 0 | 0.2954545 | 1 | 0 |

9. Summary and Histogram:

## Code:

summary(dataset)

## Output:

```
> summary(dataset)
      Age          Sex      ChestPainType   RestingBP       Cholesterol       FastingBS      RestingECG      MaxHR       
 Min.   :0.0000   0:524   1:159         Min.   :0.4600   Min.   :0.1410   Min.   :0.0000   1:419      Min.   :0.07746  
 1st Qu.:0.3673   1:166   2:163         1st Qu.:0.6000   1st Qu.:0.3416   1st Qu.:0.0000   2:110      1st Qu.:0.43838  
 Median :0.5306           3:332         Median :0.6500   Median :0.3897   Median :0.0000   3:161      Median :0.56690  
 Mean   :0.5025           4: 36         Mean   :0.6544   Mean   :0.3973   Mean   :0.1609              Mean   :0.56847  
 3rd Qu.:0.6327                         3rd Qu.:0.7000   3rd Qu.:0.4511   3rd Qu.:0.0000              3rd Qu.:0.70423  
 Max.   :1.0000                         Max.   :0.8250   Max.   :0.6700   Max.   :1.0000              Max.   :1.00000  
 ExerciseAngina    Oldpeak       ST_Slope  HeartDisease    
 0:434          Min.   :0.2841   1:336    Min.   :0.000  
 1:256          1st Qu.:0.2955   2:325    1st Qu.:0.000  
                Median :0.3409   3: 29    Median :0.000  
                Mean   :0.3886            Mean   :0.458  
                3rd Qu.:0.4659            3rd Qu.:1.000  
                Max.   :0.7045            Max.   :1.000  
```

## Code:

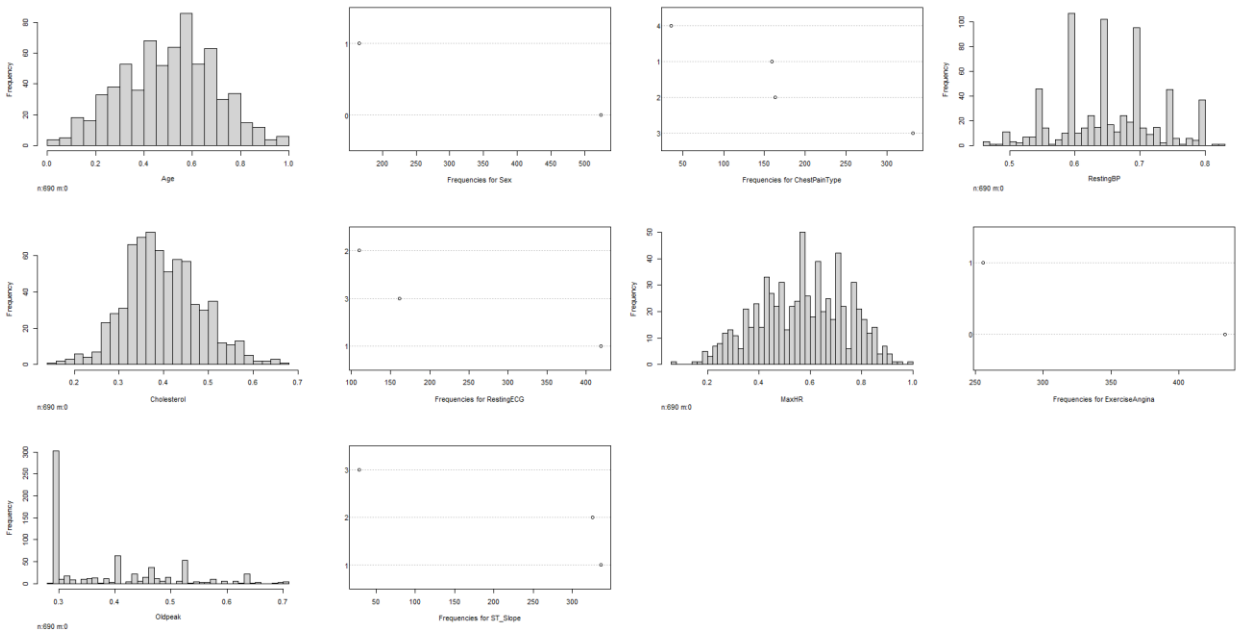
install.packages("Hmisc")

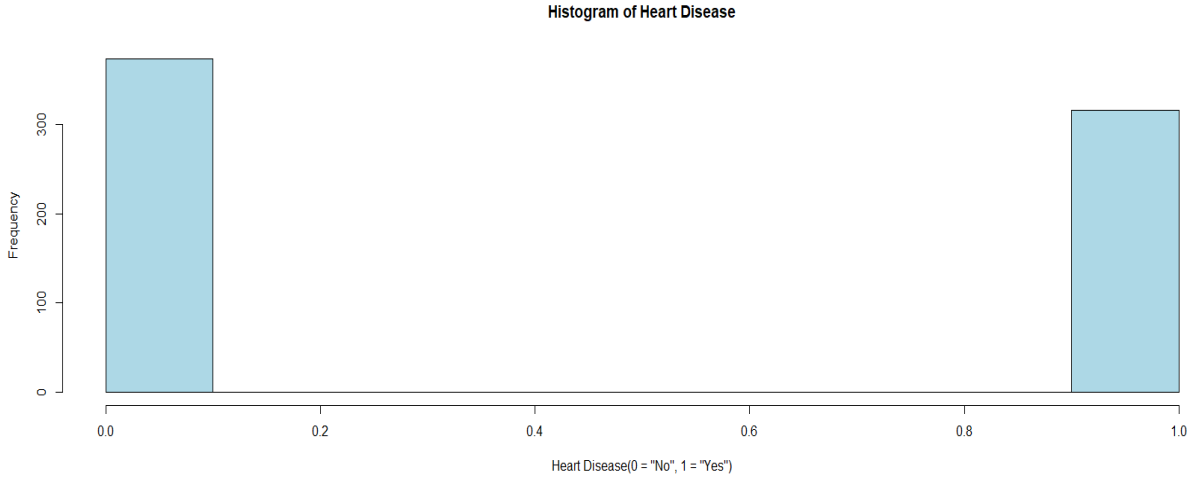
library(Hmisc)

hist.data.frame(dataset)

hist(dataset$HeartDisease,xlab = 'Heart Disease(0 = "No", 1 = "Yes")',main = paste("Histogram of Heart Disease"), col=("lightblue"))

Output:




# Model Building:

1. Splitting Dataset into Training and Test Set:

Code:

```
install.packages("caTools")
library(caTools)
```

```
set.seed(123)
split = sample.split(dataset$HeartDisease, SplitRatio = 0.80)
training_set = subset(dataset, split==TRUE)
test_set = subset(dataset, split==FALSE)
split
```

## Output:

```
> split
  [1]  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
 [22]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [43]  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [64]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
 [85]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE
[106]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
[127]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
[148]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
[169]  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
```

## Training Set:

| | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.24489796 | 0 | 1 | 0.700 | 0.4792703 | 0 | 1 | 0.7887324 | 0 | 0.2954545 | 1 | 0 |
| 2 | 0.42857143 | 1 | 2 | 0.800 | 0.2985075 | 0 | 1 | 0.6760563 | 0 | 0.4090909 | 2 | 1 |
| 5 | 0.53061224 | 0 | 2 | 0.750 | 0.3233831 | 0 | 1 | 0.4366197 | 0 | 0.2954545 | 1 | 0 |
| 8 | 0.53061224 | 0 | 1 | 0.550 | 0.3449420 | 0 | 1 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |
| 9 | 0.18367347 | 0 | 3 | 0.700 | 0.3432836 | 0 | 1 | 0.4929577 | 1 | 0.4659091 | 2 | 1 |
| 10 | 0.40816327 | 1 | 1 | 0.600 | 0.4709784 | 0 | 1 | 0.4225352 | 0 | 0.2954545 | 1 | 0 |
| 12 | 0.61224490 | 0 | 1 | 0.680 | 0.2719735 | 0 | 2 | 0.2746479 | 1 | 0.5227273 | 2 | 1 |
| 13 | 0.22448980 | 0 | 1 | 0.600 | 0.3383085 | 0 | 1 | 0.5985915 | 0 | 0.2954545 | 1 | 0 |
| 14 | 0.42857143 | 0 | 3 | 0.700 | 0.3880597 | 0 | 1 | 0.5633803 | 1 | 0.4090909 | 2 | 1 |
| 15 | 0.28571429 | 1 | 2 | 0.575 | 0.3499171 | 0 | 2 | 0.5422535 | 0 | 0.2954545 | 1 | 0 |
| 18 | 0.30612245 | 1 | 1 | 0.600 | 0.3333333 | 0 | 1 | 0.7394366 | 0 | 0.2954545 | 1 | 0 |
| 19 | 0.65306122 | 0 | 3 | 0.500 | 0.4112769 | 0 | 1 | 0.4577465 | 0 | 0.4090909 | 2 | 1 |
| 20 | 0.16326531 | 0 | 1 | 0.600 | 0.4427861 | 0 | 1 | 0.7042254 | 0 | 0.6363636 | 2 | 1 |
| 21 | 0.30612245 | 1 | 4 | 0.500 | 0.3698176 | 0 | 1 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |
| 22 | 0.32653061 | 0 | 1 | 0.600 | 0.3051410 | 0 | 1 | 0.5774648 | 0 | 0.4090909 | 2 | 0 |
| 23 | 0.42857143 | 1 | 1 | 0.620 | 0.3333333 | 0 | 1 | 0.7323944 | 0 | 0.2954545 | 1 | 0 |
| 24 | 0.32653061 | 0 | 1 | 0.750 | 0.4776119 | 0 | 1 | 0.6338028 | 1 | 0.6363636 | 2 | 1 |
| 26 | 0.16326531 | 0 | 2 | 0.650 | 0.3466003 | 0 | 1 | 0.8309859 | 0 | 0.2954545 | 1 | 0 |
| 27 | 0.51020408 | 0 | 3 | 0.620 | 0.4311774 | 0 | 2 | 0.3661972 | 1 | 0.6363636 | 2 | 0 |
| 28 | 0.48979592 | 0 | 1 | 0.600 | 0.4709784 | 0 | 1 | 0.4084507 | 0 | 0.2954545 | 1 | 0 |
| 33 | 0.53061224 | 0 | 3 | 0.625 | 0.3714760 | 0 | 1 | 0.4366197 | 0 | 0.5227273 | 2 | 1 |
| 34 | 0.26530612 | 0 | 3 | 0.650 | 0.2852405 | 0 | 2 | 0.4929577 | 0 | 0.5227273 | 2 | 1 |
| 35 | 0.30612245 | 1 | 1 | 0.750 | 0.3084577 | 0 | 1 | 0.6619718 | 0 | 0.2954545 | 1 | 0 |

## Test Set:

| | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 0.18367347 | 0 | 1 | 0.650 | 0.4693201 | 0 | 2 | 0.2676056 | 0 | 0.2954545 | 1 | 0 |
| 4 | 0.40816327 | 1 | 3 | 0.690 | 0.3548922 | 0 | 1 | 0.3380282 | 1 | 0.4659091 | 2 | 1 |
| 6 | 0.22448980 | 0 | 2 | 0.600 | 0.5621891 | 0 | 1 | 0.7746479 | 0 | 0.2954545 | 1 | 0 |
| 7 | 0.34693878 | 1 | 1 | 0.650 | 0.3930348 | 0 | 1 | 0.7746479 | 0 | 0.2954545 | 1 | 0 |
| 11 | 0.18367347 | 1 | 2 | 0.650 | 0.3499171 | 0 | 1 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |
| 16 | 0.53061224 | 1 | 1 | 0.600 | 0.4527363 | 0 | 1 | 0.6338028 | 0 | 0.4659091 | 2 | 0 |
| 17 | 0.20408163 | 0 | 3 | 0.550 | 0.3250415 | 0 | 1 | 0.7464789 | 0 | 0.2954545 | 2 | 1 |
| 25 | 0.24489796 | 0 | 2 | 0.650 | 0.3565506 | 0 | 1 | 0.5492958 | 0 | 0.2954545 | 1 | 0 |
| 30 | 0.46938776 | 0 | 1 | 0.625 | 0.3117745 | 0 | 1 | 0.5985915 | 0 | 0.2954545 | 1 | 0 |
| 32 | 0.57142857 | 0 | 2 | 0.650 | 0.2769486 | 0 | 1 | 0.3802817 | 0 | 0.2954545 | 1 | 0 |
| 37 | 0.75510204 | 0 | 3 | 0.700 | 0.5074627 | 1 | 1 | 0.1901408 | 1 | 0.4659091 | 2 | 1 |
| 38 | 0.26530612 | 1 | 1 | 0.550 | 0.4145937 | 0 | 2 | 0.5774648 | 0 | 0.2954545 | 1 | 0 |
| 47 | 0.18367347 | 0 | 3 | 0.600 | 0.3698176 | 0 | 1 | 0.7605634 | 0 | 0.2954545 | 1 | 0 |
| 48 | 0.44897959 | 0 | 1 | 0.700 | 0.3582090 | 0 | 1 | 0.7746479 | 0 | 0.2954545 | 1 | 0 |
| 52 | 0.38775510 | 1 | 3 | 0.600 | 0.3399668 | 0 | 1 | 0.2676056 | 1 | 0.5227273 | 2 | 1 |
| 53 | 0.34693878 | 0 | 1 | 0.700 | 0.3714760 | 1 | 1 | 0.4366197 | 0 | 0.2954545 | 1 | 0 |
| 79 | 0.48979592 | 0 | 1 | 0.700 | 0.1658375 | 0 | 1 | 0.5492958 | 1 | 0.2954545 | 1 | 0 |
| 83 | 0.71428571 | 0 | 3 | 0.750 | 0.3698176 | 0 | 1 | 0.3873239 | 0 | 0.2954545 | 2 | 1 |
| 84 | 0.48979592 | 0 | 1 | 0.800 | 0.3250415 | 0 | 1 | 0.7394366 | 0 | 0.2954545 | 1 | 0 |
| 85 | 0.57142857 | 0 | 3 | 0.750 | 0.3532338 | 1 | 1 | 0.4577465 | 1 | 0.4090909 | 2 | 1 |
| 89 | 0.30612245 | 0 | 4 | 0.600 | 0.4825871 | 0 | 2 | 0.6690141 | 0 | 0.2954545 | 2 | 1 |
| 95 | 0.22448980 | 1 | 2 | 0.550 | 0.3018242 | 0 | 2 | 0.8450704 | 0 | 0.2954545 | 1 | 0 |
| 106 | 0.59183673 | 0 | 1 | 0.700 | 0.4311774 | 1 | 1 | 0.5633803 | 0 | 0.2954545 | 1 | 0 |

## 2. Applying KNN Classifier:

## Code:

```
install.packages("class")
library(class)
y_pred = knn(train = training_set[,-12],
             test = test_set[,-12],
             cl = training_set[,12],
             k = 7)
```

## Confusion Matrix:

## Code:

```
cm = table(test_set[,12], y_pred)
cm
```

Output:

```
> cm = table(test_set[,12], y_pred)
> cm
   y_pred
     0  1
  0 64 11
  1  9 54
```

Accuracy:

Code:

```
accuracy = sum(diag(cm))/nrow(test_set)

accuracy
```

Output:

```
> accuracy = sum(diag(cm))/nrow(test_set)
> accuracy
[1] 0.8550725
```

## Confusion Matrix:

| | |
|---|---|
| Correctly Classified Instances | **118** |
| Incorrectly Classified Instances | **20** |

| Correct Classification | Classified As | |
|---|---|---|
| | **0** | **1** |
| **0** | **64 (True Negatives)** | **11 (False Positives)** |
| **1** | **9 (False Negatives)** | **54 (True Positives)** |

After, applying KNN with value of K = 7, we can observe that it classifies 118 instances correctly from the available 138 instances of test dataset and classifies 20 instances incorrectly. The model was able to classify 64 instances as 0 (Normal) which were actually 0 (Normal) and classify 11 instances as 1 (heart diseases) which were actually 0 (Normal). The model predicts 9 instances as 0 (Normal) which were actually 1 (heart diseases) and predicts 54 instances as 1 (heart diseases) which were actually 1 (heart diseases).

So, finally the accuracy of the model would be, $\frac{Correctly\ Classified\ Instances}{Num\ of\ Instances\ in\ test\ set}$ = 118/138 = .855 * 100% = 85.5%.