Cairo University

Faculty of Economics and Political Science

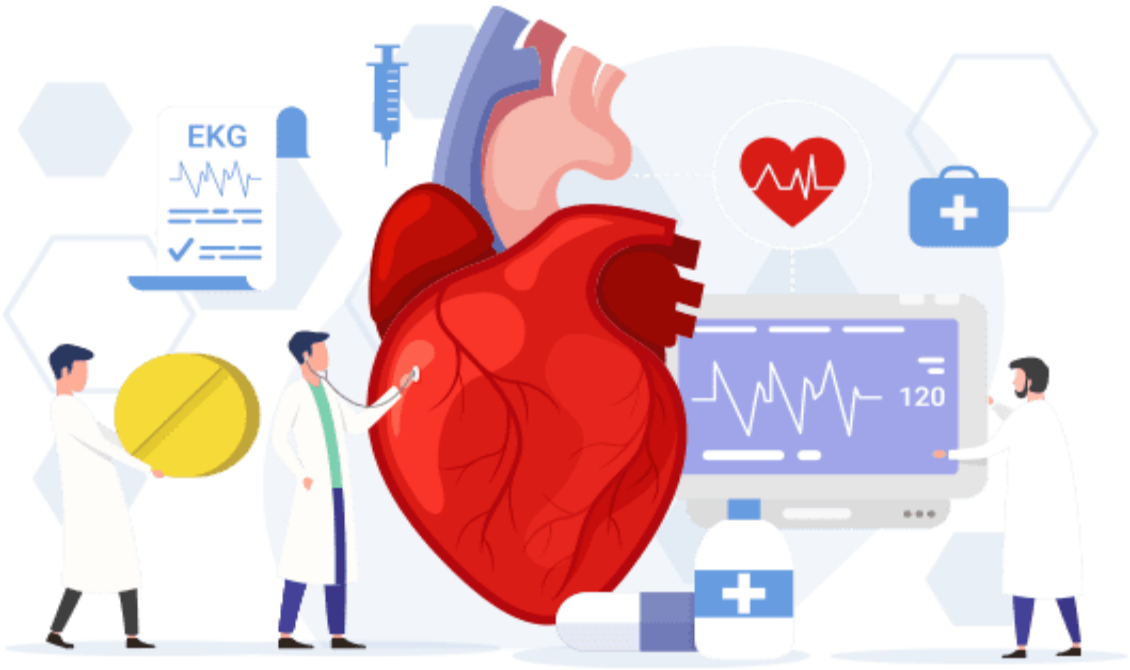
English Section

Statistics Department

4th Year

DATA SCIENCE PROJECT

"Analyzing Factors Contributing to Heart Disease Diagnosis"



Submitted by:

Saif Essam Abdelmaboud (5200356)

Youssef Hatem Mohamed (5200836)

Mareez Maher Maawad Damian (5200524)

Under The Supervision of:

Dr. Sara Osama

TA: Rawan Ibrahem

Table of Contents

LIST OF FIGURES

LIST OF TABLES

I. Introduction:

Cardiovascular diseases remain a significant global health concern, contributing to a substantial burden on healthcare systems and affecting the quality of life for individuals. Accurate prediction and early detection of heart disease play a pivotal role in preventive healthcare strategies. This academic report delves into the application of various machine learning techniques to analyze a dataset related to heart disease, aiming to develop predictive models for identifying individuals at risk.

The dataset encompasses diverse quantitative and qualitative variables, such as age, blood pressure, cholesterol levels, and exercise-induced angina, among others. Three distinct prediction techniques have been employed and evaluated for their effectiveness in predicting heart disease: Binary Logistic Regression, K-Nearest Neighbors (KNN) Algorithm, and Decision Tree.

Each method is meticulously explored, including preprocessing steps, model development, and evaluation metrics. The report provides a comprehensive analysis of the predictive performance of these models, shedding light on their strengths and limitations. By comparing the results, we aim to guide the selection of an optimal predictive model for heart disease based on the dataset's characteristics and the desired balance between accuracy, interpretability, and other relevant metrics.

The ultimate goal of this report is to contribute to the growing body of knowledge in cardiovascular health prediction, offering insights into the potential applications and effectiveness of different machine learning approaches in this critical domain.

II. Dataset Description:

A. Data Source and Population:

1. Data Source:

The dataset was obtained from Kaggle, a platform for data science and machine learning competitions. Specifically, the dataset is available at Heart Disease.

2. Population:

The population under consideration in this dataset consists of individuals who have undergone diagnosis for heart disease. The dataset includes a diverse range of patients, providing a representative sample for the analysis of factors contributing to heart disease diagnosis.

B. Data Collection Methodology:

The data collection methodology for this dataset is commonly in medical datasets to gather information during routine clinical examinations, which may involve physical examinations, medical history interviews, and diagnostic tests. The dataset likely includes information collected from individuals who presented with symptoms or risk factors associated with heart disease.

C. Variables and Their Explanations:

The dataset consists of 918 entries and 11 features. Key features include patient demographics, vital signs, symptoms, and potential risk factors associated with heart conditions. The dataset includes variables that contribute on the heart disease diagnosis. The features encompass a variety of patient demographics, vital signs, symptoms, and potential risk factors associated with heart disease diagnosis. The following is a detailed explanation of each variable:

1. Age (Continuous):
    - Description: The age of the patient in years.
    - Type: Continuous numerical variable.
2. Sex (Binary):
    - Description: Gender of the patient.
    - Type: Categorical variable with two levels - "0" representing female and "1" representing male.

3. Chest Pain Type (Categorical):
    - Description: The type of chest pain experienced by the patient.
    - Type: Categorical variable with four levels, indicating different types of chest pain.
4. Resting Blood Pressure (Continuous):
    - Description: Resting blood pressure measured in mm Hg.
    - Type: Continuous numerical variable.
5. Serum Cholesterol (Continuous):
    - Description: Serum cholesterol levels in mg/dL.
    - Type: Continuous numerical variable.
6. Fasting Blood Sugar (Binary):
    - Description: Fasting blood sugar level greater than 120 mg/dL or not.
    - Type: Categorical variable with two levels - "0" representing false (not greater than 120 mg/dL) and "1" representing true (greater than 120 mg/dL).
7. Resting Electrocardiographic Results (Categorical):
    - Description: Resting electrocardiographic results indicating normal, abnormality related to ST-T wave, or left ventricular hypertrophy.
    - Type: Categorical variable with three levels.
8. Maximum Heart Rate Achieved (Continuous):
    - Description: Maximum heart rate achieved during exercise.
    - Type: Continuous numerical variable.
9. Exercise-Induced Angina (Binary):
    - Description: Presence or absence of exercise-induced angina.
    - Type: Categorical variable with two levels - "0" representing no exercise-induced angina and "1" representing exercise-induced angina.
10. ST Depression Induced by Exercise Relative to Rest (Continuous):
    - Description: ST depression induced by exercise relative to rest.
    - Type: Continuous numerical variable.
11. Slope of the Peak Exercise ST Segment (Categorical):
    - Description: Slope of the peak exercise ST segment indicating upsloping, flat, or down sloping.

- Type: Categorical variable with three levels.

12. Heart Disease (Binary):
    - Description: Presence or absence of heart disease.
    - Type: Binary variable with "0" representing no heart disease and "1" representing the presence of heart disease.

The dataset's features provide a comprehensive set of information that can be used to explore relationships and patterns associated with heart disease diagnosis. The combination of continuous and categorical variables enables a diverse range of statistical analyses and modeling techniques.

III. Data Preparation:

Firstly, the type of variables has been checked ensuring the correct definition of variable types. In R, categorical variables are often defined as factors to enhance their representation and facilitate statistical analysis. In this dataset, the variables “Chest Pain Type”, “Heart Disease”, “Fasting BS”, “Resting ECG”, and “ST Slope” were initially identified as categorical. However, to ensure proper handling in R, the type of these categorical variables was confirmed, and they were converted from character to factors. Now all variables are truly defined.

Secondly, a thorough check was performed to determine if any entries were missing. Fortunately, there were no missing values in the dataset, simplifying subsequent analyses and ensuring the integrity of the results.

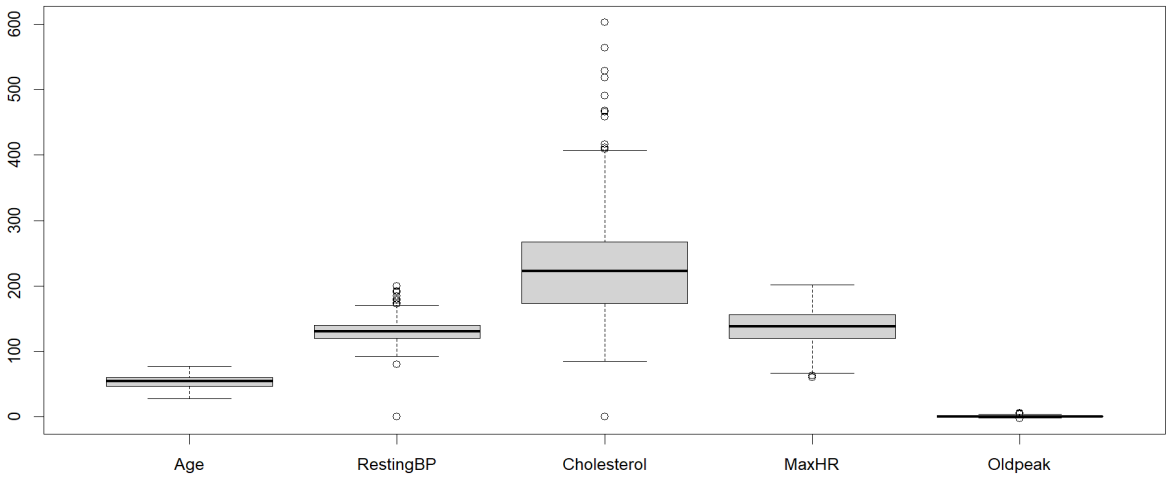


Figure 1: Box Plots for all Quantitative Variables

Finally, from Figure (1), a comprehensive analysis was conducted by creating boxplots for all quantitative variables. The variables "Resting BP" and "Serum Cholesterol" were found to contain outliers. However, the percentage of outliers was calculated for each variable and it was observed that the percentage of outliers in these variables was approximately 5%, representing a relatively small proportion of the dataset.

Therefore, given the modest percentage of outliers and by checking those points considering their potential impact on the overall dataset, a decision was made to remove these outliers from the dataset so that the model will not be overfitting or misleading. This process ensures that the data used for analysis is more robust and less influenced by extreme values, which could skew statistical measures and model outcomes.

IV. Descriptive Measures:

1. Plots for qualitative variables:

1. Gender:

Levels	Frequency	percentage
Male	694	79.2%
Female	182	20.8%

Table 1: Frequency and Relative Freq. Table for Gender

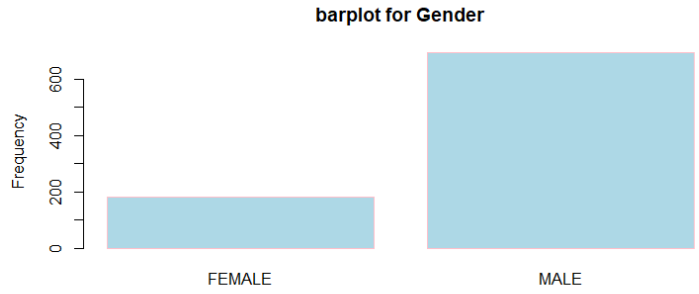


Figure 2: Bar Chart for Gender

- From Table (1) and Figure (2): we can notice that the majority of data are males with 79% and the females are only 21%.

2. **Chest Pain Type:**

Levels	Frequency	Percentage
ASY	471	53.8%
NAP	196	22.4%
ATA	167	19.1%
TA	42	4.8%

Table 2: Frequency and Relative Freq. Table for Chest Pain Type

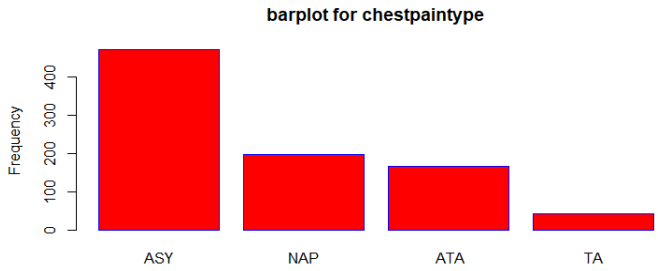


Figure 3: Bar Chart for Chest Pain Type

- From Table (2) and Figure (3): we can notice that ASY has the highest percentage in data which is 54%, followed by NAP. Which is 22%, followed by ATA which is 19%, the most uncommon one is TA which is only 5%.

3. **Fasting BS:**

Levels	frequency	percentage
No Blood Pressure	675	77.1%
Have Blood Pressure	201	22.9%

Table 3: Frequency and Relative Frequency Table for Fasting Blood Sugare

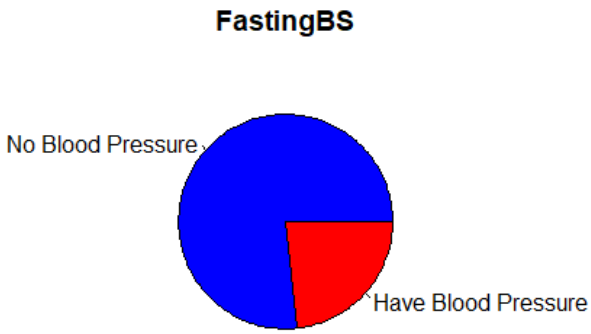


Figure 4: Pie Chart for Fasting Blood Sugare

- From Table (3) and Figure (4): we can notice that the majority have don't have fasting blood pressure which are 77% of the data and there are only 23% who have fasting blood pressure.

4. **Resting ECG:**

Levels	Frequency	Percentage
Normal	534	61%
LVH	177	20.2%
ST	165	18.8%

Table 4: Frequency and Relative Freq Table for Resting ECG

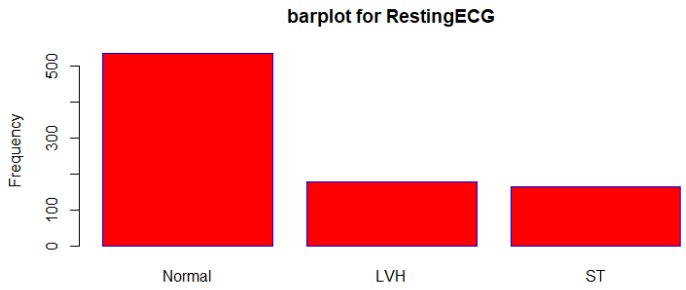


Figure 5: Bar Chart for Resting ECG

From Table (4) and Figure (5): we can notice that the majority of data have Normal Resting electrocardiography which is 61%, followed by 20.2% have LVH Resting electrocardiography and only 18.8% have ST abnormalities.

5. **Exercise Angina:**

Levels	Frequency	Percentage
Normal	534	61%
LVH	177	20.2%
ST	165	18.8%

Table 5: Frequency and Relative Frequency Table for Exercise-Included Angina

Figure 6: Pie Chart for Exercise-Included Angina

- From Table (5) and Figure (6): we can notice that the majority of data is Normal exercise-included angina which is 61% of data, 20% is LVH exercise angina and 18.8% are ST abnormalities.

6. **ST-Slope:**

Levels	Frequency	Percentage
Flat	435	49.7%
Up	385	43.9%
Down	56	6.4%

Table 6: Frequency and Relative Freq. Table for ST Slope

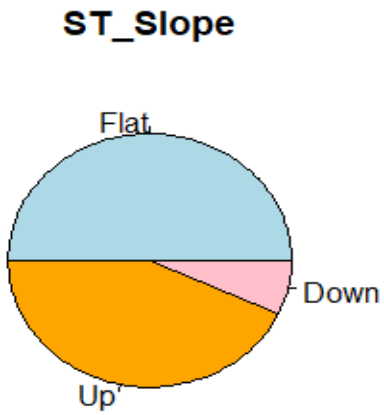


Figure 7: Pie Chart for ST Slope

- From Table (6) and Figure (7): we can notice that the majority of data has flat ST-slope which is 50% of data,44% has Up ST-slope and 6% has down ST-slope.

7. **Heart Disease:**

Levels	Frequency	Percentage
No	397	45.3%
Yes	479	54.7%

Table 7: Frequency and Relative Frequency Table for Heart Disease

Figure 8: Pie Chart for Heart Disease

- From Table (7) and Figure (8): we can notice that the majority of data has heart disease which is 55% and 45% doesn't have heart disease.

2. For cross sectional tables:

1. Between Sex and HeartDisease:

```
====================================
             HeartDisease
Sex               0       1   Total
------------------------------------
F               139      43     182
row prop.     0.764   0.236   0.208
col prop.      0.35    0.09
table prop.   0.159   0.049
------------------------------------
M               258     436     694
row prop.     0.372   0.628   0.792
col prop.      0.65    0.91
table prop.   0.295   0.498
------------------------------------
Total           397     479     876
              0.453   0.547
====================================
```

```
$measure
   odds ratio with 95% C.I.
Sex estimate    lower    upper
  F 1.000000       NA       NA
  M 5.462773 3.754266 7.948793

$p.value
   two-sided
Sex midp.exact fisher.exact   chi.square
  F         NA           NA           NA
  M          0 1.432106e-21 3.226856e-21
```

- Odds Ratio: 5.462773 with a 95% Confidence Interval (3.754266, 7.948793).
- P-value < 0.05, indicating a significant relationship between sex and heart disease.
- This means that Females are less likely to have heart disease compared to males.

2. Between ExerciseAngina and HeartDisease:

```
=======================================
                HeartDisease
ExerciseAngina       0       1   Total
---------------------------------------
N                  346     184     530
row prop.        0.653   0.347   0.605
col prop.        0.872   0.384
table prop.      0.395   0.210
---------------------------------------
Y                   51     295     346
row prop.        0.147   0.853   0.395
col prop.        0.128   0.616
table prop.      0.058   0.337
---------------------------------------
Total              397     479     876
                 0.453   0.547
=======================================
```

```
$measure
             odds ratio with 95% C.I.
ExerciseAngina estimate    lower    upper
             N  1.00000       NA       NA
             Y 10.87702 7.688874 15.38712

$p.value
             two-sided
ExerciseAngina midp.exact fisher.exact   chi.square
             N         NA           NA           NA
             Y          0 3.085261e-52 7.451256e-49
```

- Odds Ratio: 10.87702 with a 95% Confidence Interval (7.688874, 15.38712)
- P-value < 0.05, indicating a significant relationship between exercise-induced angina and heart disease.
- This means that individuals with exercise-induced angina are more likely to have heart disease compared to those without.

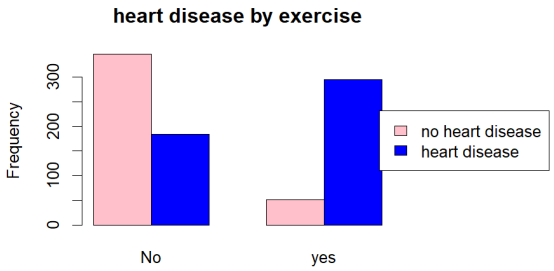


Figure 9: Bar Chart between Heart Disease and Exercise-Included Angina

3. Between Age and HeartDisease:

```
=======================================
                HeartDisease
categorize_age       0       1   Total
---------------------------------------
(15,30]              4       0       4
row prop.        1.000   0.000   0.005
col prop.        0.010   0.000
table prop.      0.005   0.000
---------------------------------------
(30,45]            128      61     189
row prop.        0.677   0.323   0.216
col prop.        0.322   0.127
table prop.      0.146   0.070
---------------------------------------
(45,60]            207     267     474
row prop.        0.437   0.563   0.541
col prop.        0.521   0.557
table prop.      0.236   0.305
---------------------------------------
(60,77]             58     151     209
row prop.        0.278   0.722   0.239
col prop.        0.146   0.315
table prop.      0.066   0.172
---------------------------------------
Total              397     479     876
                 0.453   0.547
---------------------------------------
```

```
        Goodman-Kruskal's gamma for ordinal categorical data

data:  c
Z = 8.6632, p-value < 2.2e-16
95 percent confidence interval:
 0.3681831 0.5618757
sample estimates:
Goodman-Kruskal's gamma
              0.4650294
```

- We might relate having heart disease for the old people.
- From p-value we conclude that there is a significant monotone relationship between age and heart disease.
- There is a moderate positive relationship between age and heart disease. it is more likely to have heart disease for older people (which is our assumption).

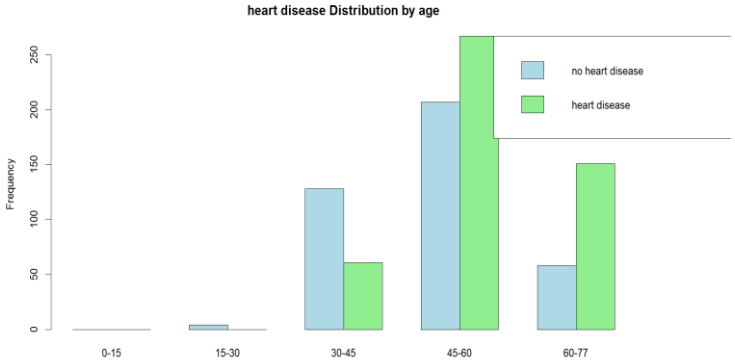


Figure 10: Bar Chart between Heart Disease and Age

4. Between Sex and Chest Pain Type:

Sex	ChestPainType ASY	ATA	NAP	TA	Total
F	62	58	53	9	182
row prop.	0.341	0.319	0.291	0.049	0.208
col prop.	0.132	0.347	0.270	0.214	
table prop.	0.071	0.066	0.061	0.010	
M	409	109	143	33	694
row prop.	0.589	0.157	0.206	0.048	0.792
col prop.	0.868	0.653	0.730	0.786	
table prop.	0.467	0.124	0.163	0.038	
Total	471	167	196	42	876
	0.538	0.191	0.224	0.048	

```
> assocstats(d)
                    X^2 df   P(> X^2)
Likelihood Ratio 40.213  3 9.6013e-09
Pearson          41.024  3 6.4620e-09

Phi-Coefficient   : NA
Contingency Coeff.: 0.212
Cramer's V        : 0.216
```

- Sometimes we relate chest pain type as females are more likely to have lung cancer. We can test it. From p-value of likelihood ratio test and pearson chi square test we can conclude that there is a significant relationship between gender and chest pain type.
- From cramer's V we can conclude that there is a weak-moderate relationship between chestpaintype and Gender. Chest pain type is affected by Gender whether male or female.

3. For quantitative variables:

1. Age:

Mean	Mean after removing first 10% and last 10%	Median	Range	MAD	IQR	St.dev	skew	kurtosis
53.34	53.50142	54.00	49.00	10.38	13.00	9.47	-0.16	-0.44

Table 8: Descriptive Statistics for Age

- From Table (8): we can see that the average age of individuals in the dataset is 53.34, with a relatively normal distribution.
- The modified mean after removing the first and last 10% of data is 53.50142, indicating a stable central tendency.
- The median age is 54.00, and the range of ages is 49.00, suggesting a moderate spread.
- The skewness is slightly negative, indicating a slight leftward skew, and the kurtosis is within the normal range.

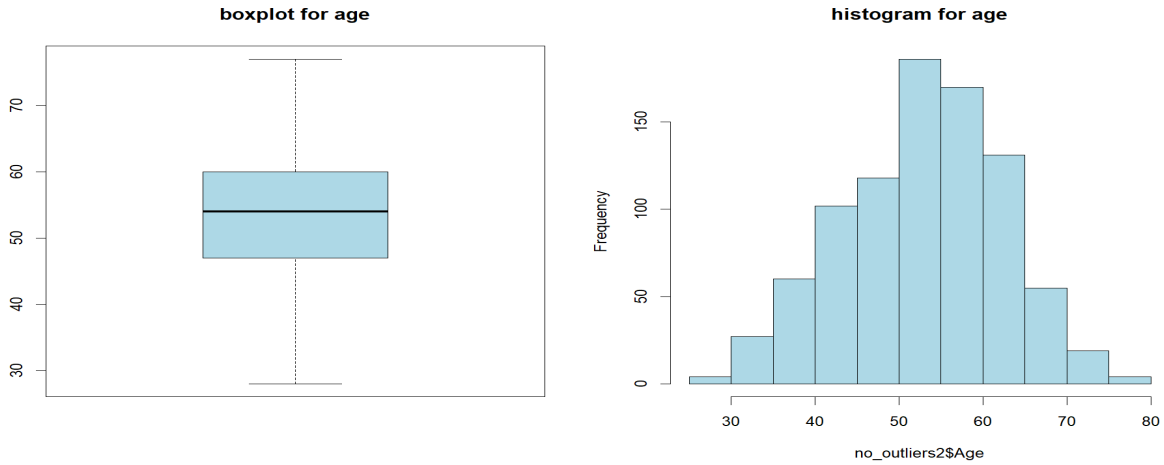


Figure 11: Box Plot and Histogram for Age

- From Figure (11): the same conclusion we can get from graphs. From boxplot we can notice there are no outliers. From histogram we can notice that the data for age are approximately normally distributed.

2. Resting BP:

Mean	Modified Mean	Median	Range	IQR	St.dev	MAD	skewness	Kurtosis
130.49	130.3063	130.00	73.00	20.00	15.18	14.83	0.06	-0.46

Table 9: Descriptive Statistics for Resting BP

- Table (9): shows that the average resting blood pressure is 130.49, with a stable central tendency even after removing extreme values.
- The modified mean is 130.3063, indicating a slight decrease when removing the first and last 10% of data.
- The median resting blood pressure is 130.00, and the range is 73.00, suggesting a moderate spread.
- The skewness is close to zero, indicating a relatively symmetrical distribution. The kurtosis is within the normal range.

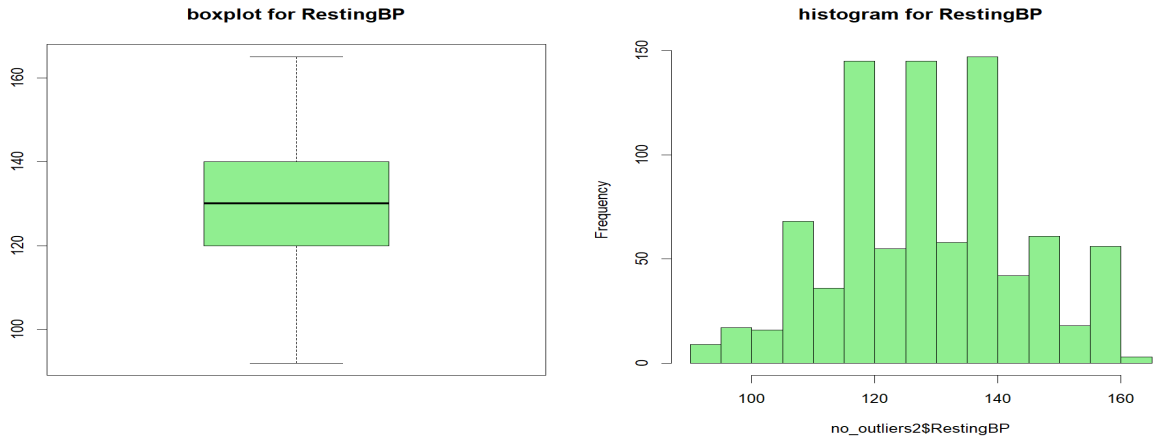


Figure 12: Box Plot and Histogram for Resting BP

- From Figure (12): the same conclusion we can get from graphs. From boxplot we can notice there aren't any outliers and resting BP ranges between (120,140). From histogram we can notice that the data for Resting blood pressure are normally distributed and we can notice that there are more than 150 observations with Resting blood pressure from (135,140) and from (115,120) and from (125,130) units which are the highest frequencies.

3. Cholesterol:

Mean	Modified mean	Median	IQR	Range	St.dev	skewness	kurtosis	MAD
198.64	204.3575	222.00	93.00	603.00	108.72	-0.60	0.16	66.72

Table 10: Descriptive Statistics for Cholesterol

- From Table (9): we can see that the average cholesterol level is 198.64, with an increase in the modified mean after removing extreme values.
- The median cholesterol level is 222.00, and the range is substantial at 603.00, indicating a wide variability.
- The skewness is negative, suggesting a leftward skew, and the positive kurtosis indicates a heavy tail.

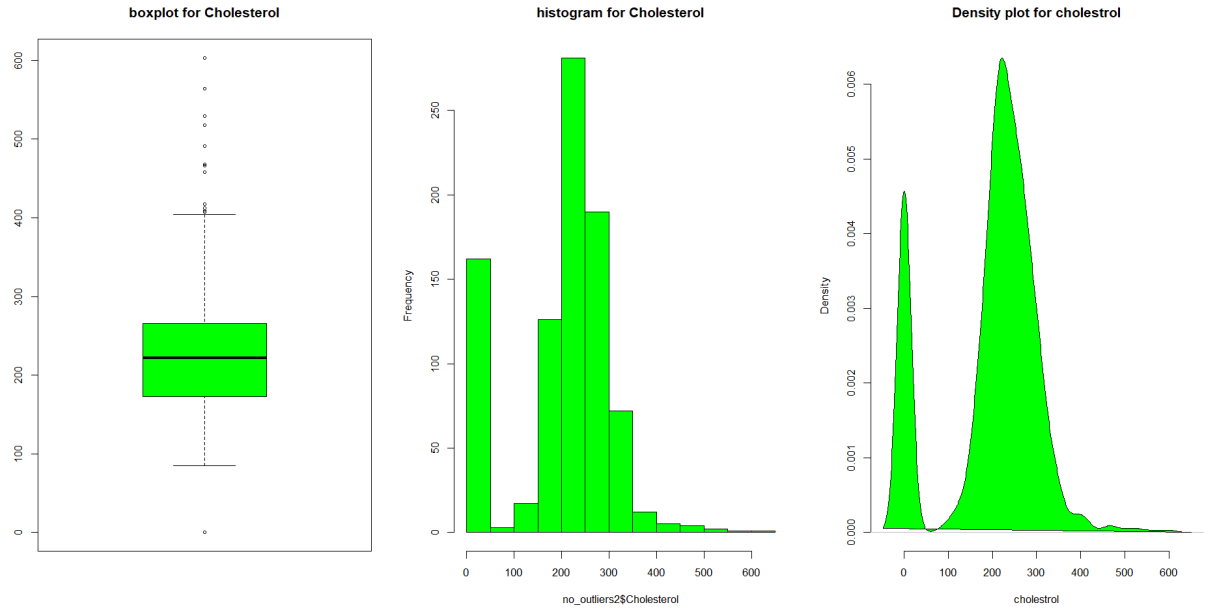


Figure 13: Box Plot, Histogram, and Density Plots for Cholesterol

- From Figure (13): from the boxplot we can notice there are outliers and cholesterol ranges between (180,250). From histogram we can notice that the data for cholesterol are normally distributed and we can notice that there are more than 250 observations with cholesterol from (200,250) units which is the highest frequency. From density plot we can also notice that cholesterol is normally distributed.

4. **Old peak:**

Mean	Mod.mean	Median	Range	IQR	St.dev	MAD	skewness	kurtosis
0.87	0.7220798	0.50	8.80	1.5	1.06	0.74	1.03	1.31

Table 11: Descriptive Statistics for Old Peak

- From Table (11): the average old peak is 0.87, with a decrease in the modified mean after removing extreme values.
- The median old peak is 0.50, and the range is 8.80, indicating variability in the data.
- The skewness is positive, suggesting a rightward skew, and the positive kurtosis indicates a heavy tail.

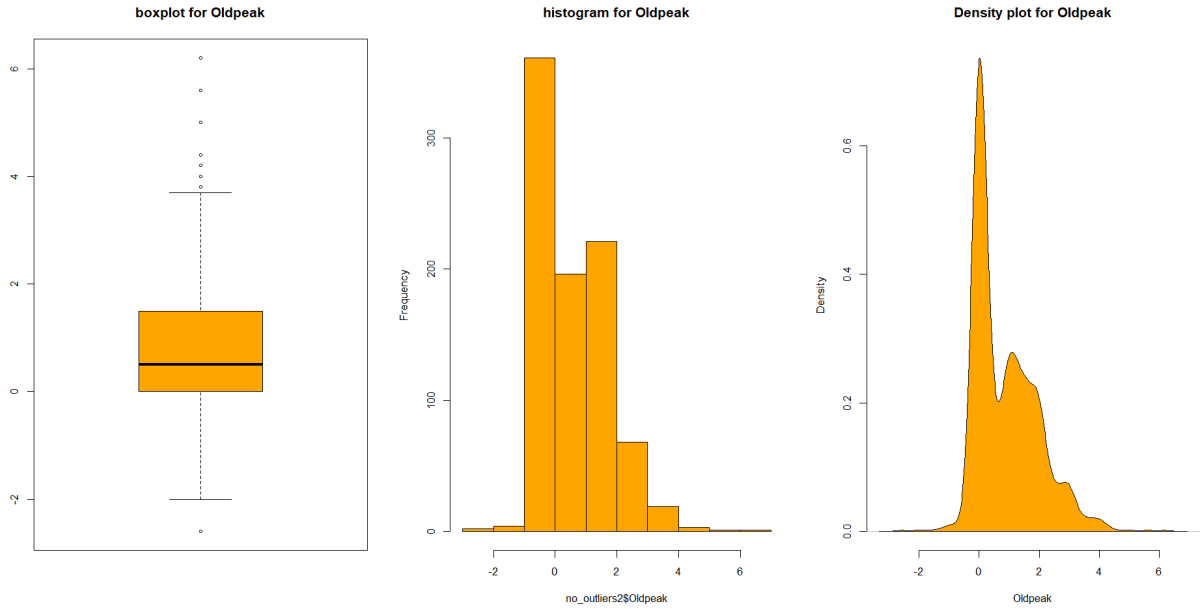


Figure 14: Box Plot, Histogram, and Density Plot for the Old Peak

- From Figure (14): from boxplot we can notice there are outliers and old peak ranges between (0,2). we can notice that there are more than 300 observations with old peak from (-1,0) units which is the highest frequency. From density plot we can also notice that old peak.

5. Max HR:

Mean	Mod.mean	median	IQR	Range	MAD	St.dev	skewness	kurtosis
137.07	137.4929	138.00	36.00	142.00	26.69	25.37	-0.14	-0.49

Table 12: Descriptive Statistics for Max HR

- From Table (12): the average maximum heart rate is 137.07, with a stable central tendency.
- The modified mean after removing extreme values is 137.4929, indicating a slight increase.
- The median maximum heart rate is 138.00, and the range is 142.00, suggesting variability.
- The skewness is slightly negative, indicating a slight leftward skew, and the kurtosis is within the normal range.

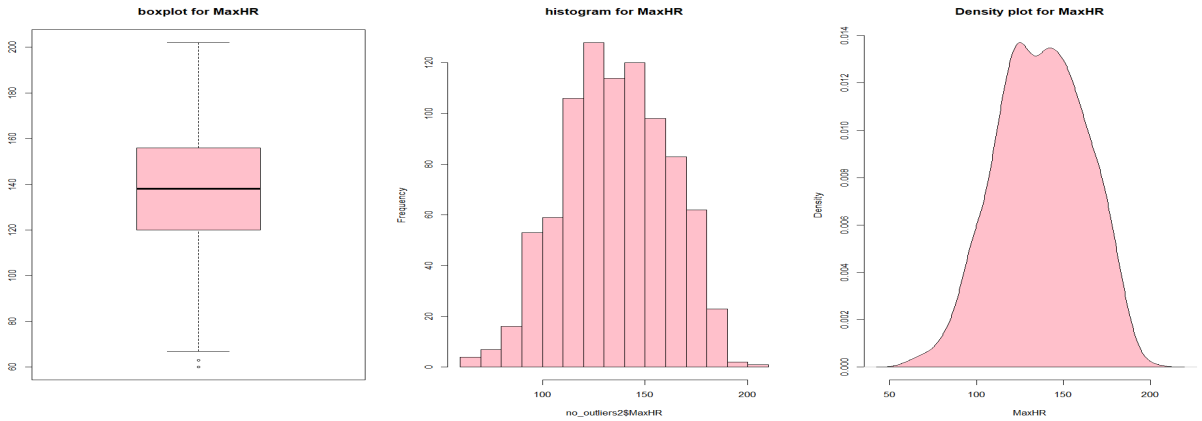


Figure 15: Box Plot, Histogram, and Density Plot for Max HR

- From Figure (15): from boxplot we can notice that there are outliers in Max HR, we can notice from density plot that data for Max HR are normally distributed, we can notice from histogram that there are more than 120 observations ranges between (120-130) units.

4. Correlation between variables:

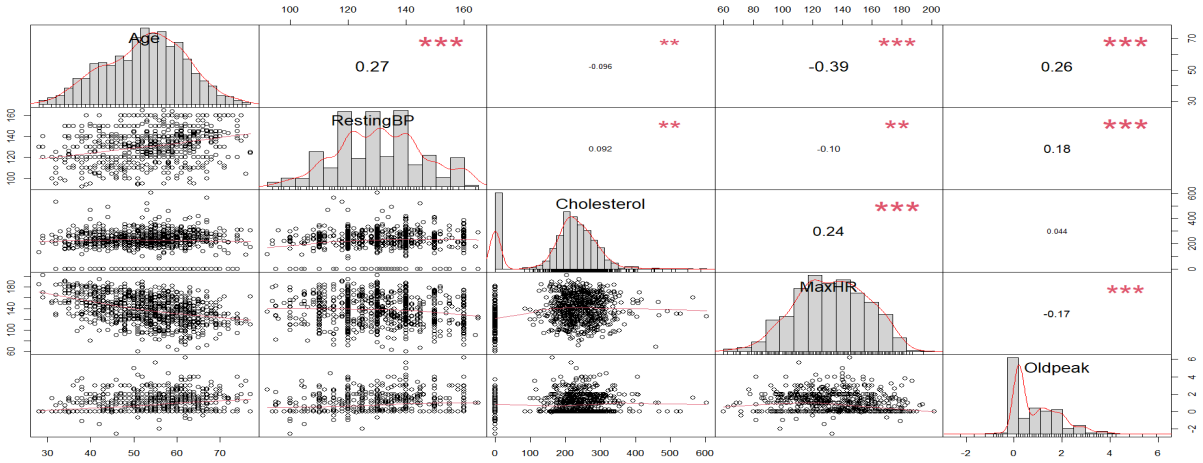


Figure 16: Correlation Matrix Chart between Quantitative Variables

From Figure (16) we can see that:

- There is significant weak positive linear relationship between age and resting BP.
- There is significant negative weak (very weak) linear relationship between age and cholesterol.
- There is significant moderate positive linear relationship between Max HR and age.
- There is significant positive weak linear relationship between age and old peak.
- There is significant weak positive linear relationship between cholesterol and Resting BP.
- There is significant negative weak linear relationship between Max HR and Resting BP.
- There is significant positive weak linear relationship between Resting BP and old peak.
- There is significant positive weak linear relationship between cholesterol and Max HR.
- There is significant negative weak linear relationship between old peak and Max HR.

5. For Quantitative and Qualitative Variables:

1. Between Cholesterol and Exercise Angina:

We can notice that lower levels of cholesterol for people who do exercise than people who not do exercise.

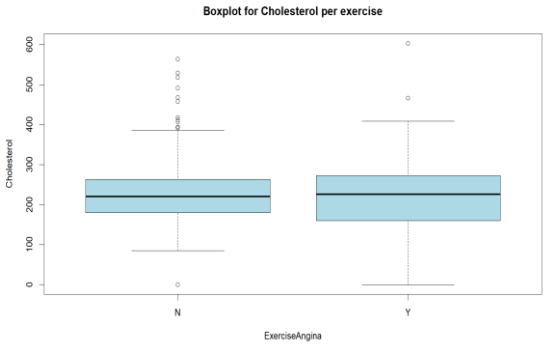


Figure 17: Box Plot for Cholesterol Per Exercise

2. Between Max HR and Exercise Angina:

- We can notice that exercise decrease the Max HR, people who do exercise has heart rate less than people who not do exercise

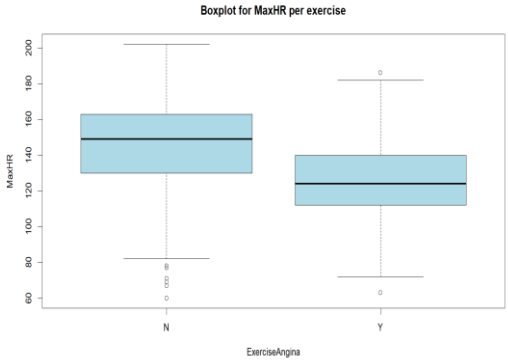


Figure 18: Box Plot for Max HR per Exercise

3. Between Cholesterol and Sex:

- We can notice that females are more likely to have high cholesterol level.

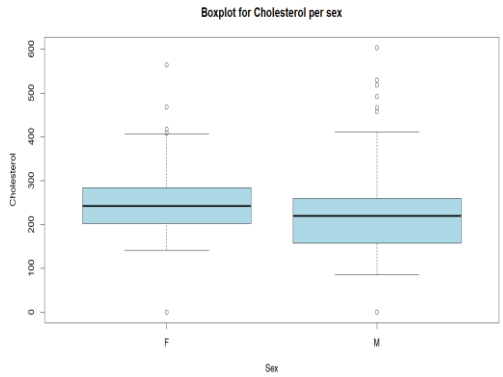


Figure 19: Box Plot for Cholesterol and Sex

V. Binary Logistic Regression Model:

We have applied binary logistic regression model to response variable heart disease using all the explanatory variable. Then we applied stepwise and compared between them as they are nested models with likelihood ratio test:

$$H_0: \beta_i = 0 \qquad \forall\ i=1,6,9,10$$

$$H_1: \beta_i \neq 0 \qquad \forall\ i=1,6,9,10$$

```
Likelihood ratio test

Model 1: train$HeartDisease ~ Age + Sex + ChestPainType + RestingBP +
    Cholesterol + FastingBS + RestingECG + MaxHR + ExerciseAngina +
    Oldpeak + ST_Slope
Model 2: train$HeartDisease ~ Sex + ChestPainType + Cholesterol + FastingBS +
    MaxHR + ExerciseAngina + Oldpeak + ST_Slope
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  16 -197.67
2  12 -198.71 -4 2.0688     0.7231
```

From the Log likelihood ratio, the decision is: **Don't** reject Ho at level of significance 0.05, indicating that extra parameters don't add extra information and doesn't have significant effect on response heart disease.

1. Fitted Model:

$$\mathbf{Ln\left(\frac{\Pi(xJ)}{1-\Pi(xJ)}\right) = \alpha + \beta 1D1 + \beta 2D2 + \beta 3D3 + \beta 4D4 + \beta X1 + \beta 6D5 + \beta 7X2 + \beta 8D6 + \beta 9X3 + \beta 10D7 + \beta 11D8}$$

2. The Estimated Model:

$$\mathbf{Ln\left(\frac{\Pi(xJ)}{1-\Pi(xJ)}\right) = 2.5438 + 1.26D1 - 1.35D2 - 1.911D3 - 1.15D4 - 0.003X1 + 1.07D5 - 0.011X2 + 1.03D6 + 0.333X3 - 2.365D7 - 1.14D8}$$

3. Checking significance and interpreting parameters:

$$H_0: \beta_J = 0 \qquad \forall\ J= 1, 2, \dots, 11$$

$$H_1: \beta_J \neq 0 \qquad \forall\ J= 1,2, \dots, 11$$

```
Call:
glm(formula = train$HeartDisease ~ Sex + ChestPainType + Cholesterol +
    FastingBS + MaxHR + ExerciseAngina + Oldpeak + ST_Slope,
    family = binomial, data = train)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.543809   0.899873   2.827 0.004701 **
Sex1             1.260399   0.339242   3.715 0.000203 ***
ChestPainType2  -1.357111   0.317701  -4.272 1.94e-05 ***
ChestPainType3  -1.911907   0.420175  -4.550 5.36e-06 ***
ChestPainType4  -1.150442   0.514833  -2.235 0.025444 *
Cholesterol     -0.003897   0.001308  -2.980 0.002879 **
FastingBS1       1.070094   0.345321   3.099 0.001943 **
MaxHR           -0.011835   0.005886  -2.011 0.044349 *
ExerciseAngina1  1.034764   0.298113   3.471 0.000518 ***
Oldpeak          0.333821   0.136878   2.439 0.014735 *
ST_Slope2       -2.365204   0.293477  -8.059 7.68e-16 ***
ST_Slope3       -1.149071   0.576062  -1.995 0.046076 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 840.85  on 609  degrees of freedom
Residual deviance: 397.41  on 598  degrees of freedom
AIC: 421.41

Number of Fisher Scoring iterations: 5
```

- It takes five iterations weighted least squares to obtain maximum likelihood estimates of the parameters.
- Reject H0 for every **(β_j v j=1, …, 11)** at level of significance 0.05 as p value for each parameter less than 0.05 so these parameters have significance effect on the heart disease.
- According to p-value of chi-square that less than 0.05 that model is better than the intercept-only model and by comparing between the predicted values from the fitted model to those from the null model we found pseudo-R squared equal 0.5273675.

4. Interpreting the significant parameters:

✓ $e^{\beta 1=1.26039}$ = **3.526796:** the estimated odds of having heart disease among males is 3.526796 times from females holding other variables constant.

✓ $e^{\beta 2=-1.35711}$ = **0.2574:** the estimated odds of having heart disease among having chest pain type NAP is 0.2574 times from having chest pain type ASY holding other variables constant.

✓ $e^{\beta 3=-1.911907}$ = **0.147798:** the estimated odds of having heart disease among having chest pain type ATA is 0.147798 times from having chest pain type ASY holding other variables constant.

- ✓ $e^{\beta 4=-1.150442}$ = **0.31649:** the estimated odds of having heart disease among having chest pain type TA is 0.31649 times from having chest pain type ASY holding other variables constant.
- ✓ $e^{\beta 5=-0.003897}$ = **0.99611:** the estimated odds of having heart disease decreases with a fraction of 0.388% with each one-unit increase in cholesterol holding other variables constant.
- ✓ $e^{\beta 6=1.07}$ = **0.2574:** the estimated odds of having heart disease among people having blood sugar is 0.2574 times from people having no blood sugar holding other variables constant.
- ✓ $e^{\beta 7=-0.011835}$ = **0.98823:** the estimated odds of having heart disease decreases with a fraction of 1.176% with each one-beat increase in max heart rate holding other variables constant.
- ✓ $e^{\beta 8=1.03}$ = **2.8144:** the estimated odds of having heart disease among people having exercise angina type Y is 2.8144 times from people having exercise angina type N other variables constant.
- ✓ $e^{\beta 9=0.33821}$ = **1.402:** the estimated odds of having heart disease increases with a fraction of 40.24% with each one-unit increase in old peak holding other variables constant.
- ✓ $e^{\beta 10=-2.36}$ = **0.0944:** the estimated odds of having heart disease among people having st slope type up 0.0944 times from people having st slope type flat holding other variables constant.
- ✓ $e^{\beta 11=-1.14}$ = **0.3198:** the estimated odds of having heart disease among people having st slope type down 0.3198 times from people having st slope type flat holding other variables constant.

5. Goodness of fit and multicollinearity:

By checking multicollinearity of the model, we found that:

```
> car::vif(selection_2)
                    GVIF Df GVIF^(1/(2*Df))
Sex             1.098683  1        1.048181
ChestPainType   1.155218  3        1.024340
Cholesterol     1.181068  1        1.086769
FastingBS       1.087927  1        1.043037
MaxHR           1.142167  1        1.068722
ExerciseAngina  1.187123  1        1.089552
Oldpeak         1.271087  1        1.127425
ST_Slope        1.408764  2        1.089456
```

- All variables have VIF less than 10 so there is no multicollinearity in the model.

6. Summarizing the predictive power of the model:

We will construct classification table to determine:

1- Sensitivity which is the probability of correctly predicting a success
2- Specificity which is the probability of correctly predicting a failure
3- Determine misclassification error
4- Determine the accuracy of correct classification

But first of all, we need to determine optimal cutoff point as classification table very sensitive to the cutoff point.

7. Determining the optimal cutoff point:

- We plot the curve of the accuracy of each cutoff point against all possible cutoff points and from the plot.
- From Figure (20): we found that at point 0.4154449 have the maximum accuracy.

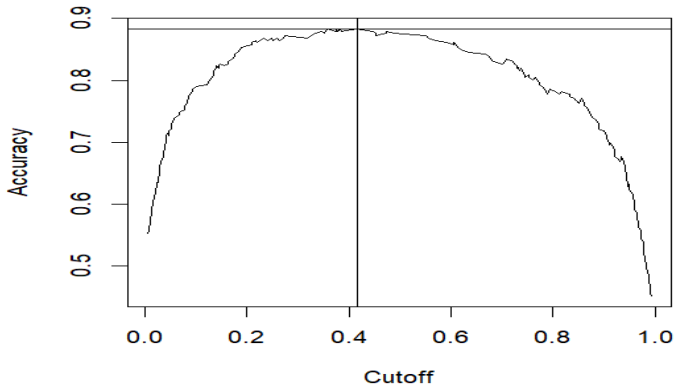


Figure 20: Accuracy Curve

8. The classification table:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 101  18
         1  14 133

               Accuracy : 0.8797
                 95% CI : (0.8344, 0.9162)
    No Information Rate : 0.5677
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7559

 Mcnemar's Test P-Value : 0.5959

            Sensitivity : 0.8783
            Specificity : 0.8808
         Pos Pred Value : 0.8487
         Neg Pred Value : 0.9048
             Prevalence : 0.4323
         Detection Rate : 0.3797
   Detection Prevalence : 0.4474
      Balanced Accuracy : 0.8795

       'Positive' Class : 0
```

From the classification table we got this:

1- The model classifies 87.97% of the observation correctly which greater than 60% which is good.
2- The probability of correctly predicting a having heart disease is 0.8783 which is greater than 0.6.
3- The probability of correctly predicting no having heart disease is 0.8808 which is greater than 0.6.
4- The model classifies 12.03% of the observation wrong.

9. Roc curve and area under the curve:

- We will plot Roc curve as it more powerful than classification table as it plots sensitivity as a function in (1-specificity) for the all-possible cutoff points. Thus, it is more informative than a classification table and determine area under the curve which called concordance index.

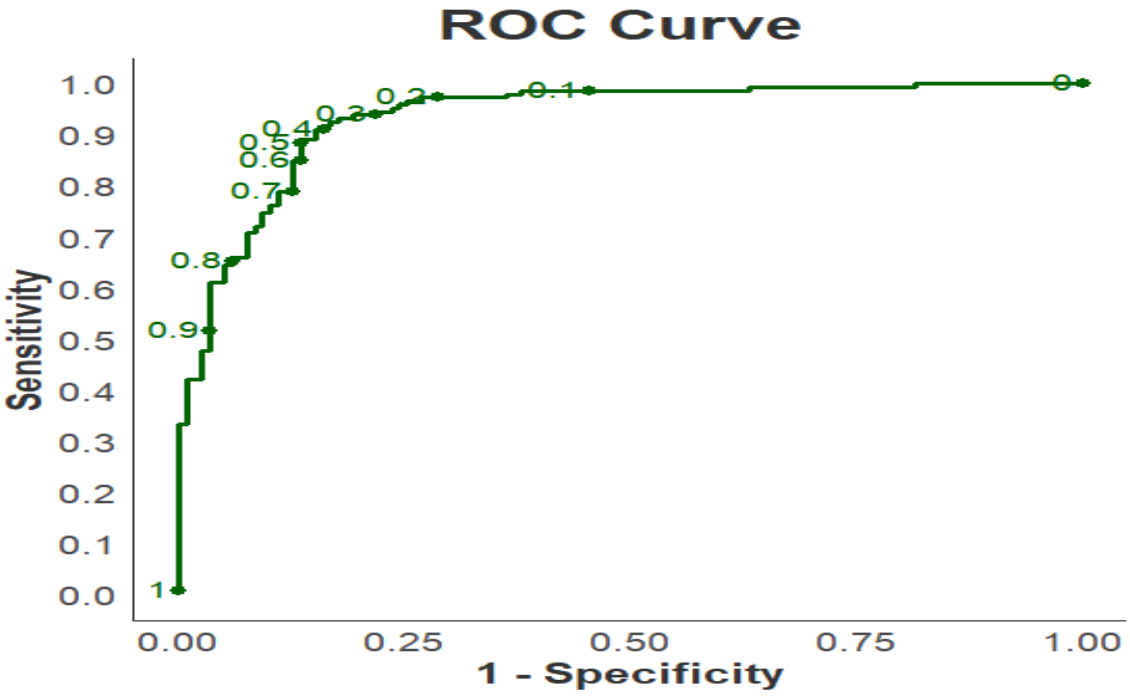


Figure 21: Roc Curve

- From Figure (21): we found that it above 45-degree line which the model is better from model with intercept only and the curve close too much from point (0,1) that model has good predictive power.
- Moreover, the area under the curve (concordance index) is 0.935 which is higher than 0.5 which the model is better from model with intercept only as 0.935 is the probability that the predictions and outcomes are concordant.

VI. Machine Learning Techniques:

1. K-Nearest Neighbors (KNN) Algorithm:

✓ Optimal K Determination:

- To implement the KNN algorithm for predicting heart disease diagnosis, the optimal value of K was determined through a thorough analysis. After evaluating different values of K, it was found that the optimal K for this dataset is 9.
- It is also clear from the Figure (22) that K=9 have the lowest misclassification error.

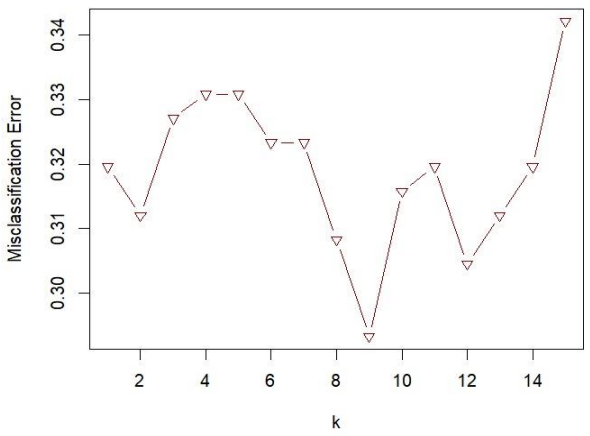


Figure 22: Plot for Misclassification Error Vs K

✓ Model Performance:

- After implementing the KNN algorithm with K=9, the model's performance was assessed using various metrics:

- Misclassification Error: 0.2932331
- Accuracy: 0.706766917293233

```
 classifier_knn
    0   1
0  82  37
1  41 106
```

✓ Confusion Matrix Statistics:

- The confusion matrix statistics provide a comprehensive overview of the model's performance:

- Accuracy: 0.7068
- 95% Confidence Interval: (0.6481, 0.7608)
- Kappa: 0.4089
- Sensitivity (True Positive Rate): 0.6667
- Specificity (True Negative Rate): 0.7413
- Positive Predictive Value (Precision): 0.6891
- Negative Predictive Value: 0.7211
- Prevalence: 0.4624
- Balanced Accuracy: 0.7040

✓ **Interpretations:**

- The KNN algorithm with an optimal K of 9 achieved an accuracy of approximately 70.68%. This suggests that the model is capable of making accurate predictions, considering the complexity of the dataset. The confusion matrix provides insights into the true positive, true negative, false positive, and false negative predictions.
- The Kappa coefficient, sensitivity, specificity, and other metrics contribute to a comprehensive evaluation of the model's performance. The balanced accuracy, which considers imbalanced class distribution, is also provided.
- In conclusion, the KNN algorithm with K=9 demonstrates promising predictive capabilities for heart disease diagnosis. The model's performance metrics indicate its potential utility in identifying individuals at risk of heart disease.

2. Decision Tree:

A Decision Tree model was constructed with six decision nodes. The most influential predictor for heart disease diagnosis was identified to be the slope of the peak exercise ST segment.

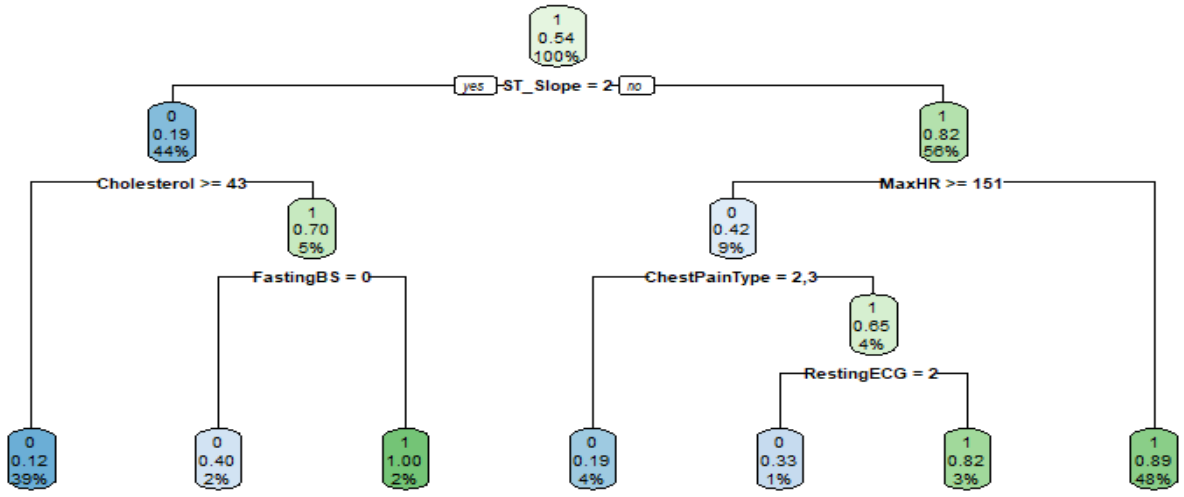


Figure 23: Decision Tree

- From Figure (23): The decision tree revealed the following rules for predicting heart disease:

1. **For Up-sloping ST Segment:**
    - If Serum Cholesterol Level is greater than or equal to 43 mg/dL, the patient is considered to have no heart disease (39% of the sample).
    - While, if Serum Cholesterol Level is less than 43 mg/dL:
        - If Fasting blood sugar level is not greater than 120 mg/dL, the patient is considered to have no heart disease (2% of the sample).
        - If Fasting blood sugar level is greater than 120 mg/dL, the patient is considered to have heart disease (2% of the sample).
2. **For Down-sloping or Flat ST Segment:**
    - If Maximum heart rate is less than 151 beats/min, the patient is considered to have heart disease (48% of the sample).
    - While, if Maximum heart rate is greater than or equal to 151 beats/min:
        - If Chest Pain Type is NAP or ATA, the patient is considered to have no heart disease (4% of the sample).
    - However, if Chest Pain Type is ASY or TA:
        - If the Resting electrocardiographic results are left ventricular hypertrophy (LVH), the patient is considered to have no heart disease (1% of the sample).

- If the Resting electrocardiographic results are normal or ST-T wave, the patient is considered to have heart disease (3% of the sample).

➢ **Confusion Matrix for the Decision Tree Algorithm:**

```
          Reference
Prediction   0   1
         0 101  18
         1  30 117
```

- The confusion matrix statistics provide a comprehensive overview of the model's performance:
- Accuracy: 0.8195
- 95% Confidence Interval: (0.768, 0.8638)
- Kappa: 0.6385
- Sensitivity (True Positive Rate): 0.7710
- Specificity (True Negative Rate): 0.8667
- Positive Predictive Value (Precision): 0.8487
- Negative Predictive Value: 0.7959
- Prevalence: 0.4925
- Balanced Accuracy: 0.8188

➢ **Interpretations:**

The Decision Tree algorithm achieved an accuracy of approximately 81.95%. The decision tree rules provide a clear understanding of the predictors influencing heart disease diagnosis. The model's performance metrics, including sensitivity, specificity, and precision, contribute to a comprehensive evaluation.

The Decision Tree's Kappa coefficient indicates substantial agreement beyond chance, and the balanced accuracy considers the imbalanced class distribution. The high accuracy and balanced metrics suggest that the Decision Tree is effective in predicting heart disease based on the provided features.

VII. Models Comparison:

The binary logistic regression model yielded a significant fit, capturing the relationship between various predictors and the likelihood of heart disease. The model achieved an accuracy of 87.97% and demonstrated good sensitivity and specificity. The area under the ROC curve (AUC) was 0.935, indicating high predictive power.

The KNN algorithm with an optimal K value of 9 resulted in an accuracy of 70.68%. The confusion matrix revealed 82 true negatives, 106 true positives, 37 false negatives, and 41 false positives. The sensitivity and specificity were 66.67% and 74.13%, respectively.

The decision tree model, with six decision nodes, achieved an accuracy of 81.95%. The confusion matrix showed 101 true negatives, 117 true positives, 18 false negatives, and 30 false positives. The model demonstrated a sensitivity of 77.10% and specificity of 86.67%. The AUC was 0.6385, indicating good predictive power.

- **Accuracy:** The binary logistic regression model outperformed both the KNN algorithm and the decision tree in terms of accuracy.
- **Sensitivity and Specificity:** The binary logistic regression and decision tree models exhibited better sensitivity and specificity compared to the KNN algorithm.
- **Area Under the Curve (AUC):** The binary logistic regression model had the highest AUC, followed by the decision tree, while the KNN algorithm had the lowest AUC.

In summary, the binary logistic regression model demonstrated superior performance in predicting heart disease compared to the KNN algorithm and decision tree. However, the choice of the best model depends on the specific goals and requirements of the analysis, considering factors such as interpretability, computational efficiency, and the importance of different evaluation metrics. Moreover, further fine-tuning and validation can enhance the robustness of the models.

VIII. Function Using "For" Loop:

```
patients <- seq(1,876,1)
results <- {}
for (i in 1:length(patients)) {
  results[i] <- ifelse(no_outliers2$Cholesterol[i] >= 200 & no_outliers2$MaxHR[i]> 220- no_outliers2$
                       Age[i]&no_outliers2$RestingBP[i]>120,'GO TO DOCTOR", "need more invistigation")}


k<-as.data.frame(results)
data2<-cbind(k,no_outliers2)
```

- This function provides a decision-making process for those who need to go to doctors instantly and for those who still need further investigation to know their health statement.
- By using this function, we want to investigate the patients using some conditions included in the loop, the loop checks conditions related to Cholesterol, Max Heart Rate, Age, and Resting Blood Pressure, deciding whether a patient should go to the doctor immediately or needs further investigation.
- If those conditions are satisfied therefore, the patient has to go to a doctor. While if not satisfied therefore we need investigate further conditions.

IX. Conclusion:

In conclusion, this academic report has explored and compared three distinct modeling techniques: Binary Logistic Regression, K-Nearest Neighbors (KNN) Algorithm, and Decision Tree, for predicting heart disease based on a comprehensive dataset. Each method has undergone meticulous evaluation, revealing unique insights into their predictive capabilities and interpretability.

The Binary Logistic Regression model demonstrated robustness in parameter estimation, and the significance of the variables was thoroughly examined using likelihood ratio tests. The results highlighted several key predictors of heart disease, such as gender, chest pain type, cholesterol levels, and exercise-induced angina. The model's goodness of fit, assessed through pseudo-R squared and multicollinearity checks, showcased its ability to capture the variance in the data without encountering significant multicollinearity issues.

The KNN Algorithm, with an optimal K value of 9, exhibited a fair predictive accuracy of 70.68%. The confusion matrix and associated metrics illustrated the model's ability to correctly classify instances into heart disease categories. While KNN is a simple and intuitive method, it may face challenges with larger datasets and can be sensitive to irrelevant features.

The Decision Tree model, characterized by six decision nodes, provided a transparent and interpretable decision-making process. By recursively partitioning the data based on features like ST segment slope, cholesterol levels, and maximum heart rate, the tree effectively identified patterns associated with heart disease. The model achieved an accuracy of 81.95%.

Comparatively, the Binary Logistic Regression model emerged as the top-performing model in terms of accuracy. While in terms of interpretability, the Decision Tree Algorithm is better. However, it is essential to consider the trade-offs between model complexity and interpretability when choosing the most suitable approach for a given application.

This report contributes valuable insights to the field of cardiovascular health prediction, emphasizing the significance of model selection based on specific dataset characteristics and predictive goals. Future work may involve refining existing models, exploring ensemble methods, and incorporating additional features to enhance predictive accuracy and robustness. The continued evolution of machine learning techniques holds promise for advancing our ability to identify individuals at risk of heart disease, ultimately contributing to more effective preventive healthcare strategies.

X. References:

- **Our data is from:** https://www.kaggle.com/datasets/amirmahdiabbootalebi/heart-disease
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** Springer.
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, 13(1), 21-27.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984**). Classification and Regression Trees**. CRC press.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). **An Introduction to Statistical Learning.** Springer.