Faculty of Applied Sciences
University of Sri Jayewardenepura, Sri Lanka

STA 335 2.0 Generalized Linera Models and Non Linear models

PREDICTING NEAR-FUTURE HEART DISEASE RISK AND IDENTIFYING KEY PREDICTORS USING LOGISTIC REGRESSION

Prepared By

H.S.D FERNANDO
AS2021381

PRESENTED TO:

Dr. Niroshan Withanage
Department of Statistics

Table of Content

1. Introduction

Heart and blood vessels diseases are generally called as heart disease. It is a type of cardiovascular disease. Cardiovascular diseases are the number 1 cause of death globally. 31% of all deaths worldwide are occurred due to cardiovascular diseases. Four out of 5 cardiovascular deaths are due to heart attacks and strokes. Therefore making early detection and prediction crucial for effective treatments. This report analyzes a heart disease classification dataset to predict which patients are most likely to suffer from a heart disease in the near future using the features given.

- age - Displays the age of the individual.
- sex - Displays the gender of the individual using the following format : 1 = male 0 = female
- cp - Chest-pain type: displays the type of chest-pain experienced by the individual using the following format : 0 = typical angina 1 = atypical angina 2 = non — angina pain 3 = asymptotic
- trestbps - Resting Blood Pressure: displays the resting blood pressure value of an individual in mmHg(unit). Anything above 130-140 is typically cause for concern.
- chol - Serum Cholesterol: displays the serum cholesterol in mg/dl (unit)
- fbs - Fasting Blood Sugar: compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then : 1 (true) else : 0 (false) '>126' mg/dL signals diabetes
- restecg - Resting ECG : displays resting electrocardiographic results 0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy
- thalach - Max heart rate achieved : displays the max heart rate achieved by an individual.
- exang - Exercise induced angina : 1 = yes 0 = no
- oldpeak - ST depression induced by exercise relative to rest: displays the value which is an integer or float.
- slope - Slope of the peak exercise ST segment : 0 = upsloping: better heart rate with exercise (uncommon) 1 = flat: minimal change (typical healthy heart) 2 = downsloping: signs of unhealthy heart
- ca - Number of major vessels (0–3) colored by fluoroscopy: displays the value as integer or float.
- target - Displays whether the individual is suffering from heart disease or not 1=yes 0=no

2. Methodology

This is a study related to heart disease classification. There are six qualitative predictors (sex, cp, fbs, restecg, exang, slope) and six quantitative predictors (age, trestbps, chol, thalach,oldpeak,ca). In data preprocessing part missing values and misleading values will checked and the all the qualitative variables will be coded appropriately.

At the beginning of the analysis, the variables will be explored. For the qualitative variables, the composition of the sample will be assessed using pie charts. For the quantitative variables, boxplots will be plotted to identify any potential outliers, and dot plots will be drawn to examine the distribution of the data. Descriptive statistics will also be calculated. If any outliers are detected, they will be treated as missing values for further analysis. After explore all predictors, outcome variable will be analyze using a pie chart also. It will give an idea about the number of individual is suffering from heart disease in the sample.

For the data analysis process, the association between each qualitative variable and the response variable (target) will be assessed using two-way frequency tables and the chi-square test. Before checking the associations between the outcome variable and the qualitative predictors, box plots will be created for each quantitative variable, grouped by the levels of the outcome variable, to detect outliers and visualize the median difference for each category. If any outliers are detected, they will be treated as missing values. Then, the association between quantitative variables and the binary response variable will be assessed using two methods. First, the dataset will be split into subgroups based on the levels of the outcome variable. Then, the normality of the selected quantitative variable will be checked within each subgroup. The association with the response will be assessed using an independent t-test, provided that the normality assumption is satisfied. If the normality assumption is not satisfied, the Mann-Whitney U test will be used instead.

Logistic regression will be used as the primary modeling technique to predict the likelihood of heart disease, incorporating both quantitative and qualitative predictor variables. Prior to modeling, rows with missing values and outliers will be removed. Forward selection, using the Akaike Information Criterion (AIC), will be employed to select predictor variables. In each step, the predictor with the lowest AIC will be added to the model, and this process will continue until the current AIC is lower than the AIC after adding the next predictor. After selecting the main predictors, interaction effects will be added. The same AIC-driven forward selection will be used to identify significant interactions. All descriptive statistical analyses will be performed using MINITAB and Logistic regression will be done by using R language.

3. Data exploration

3.1 Composition of the sample

3.1.1 Composition of the sample with respect to sex

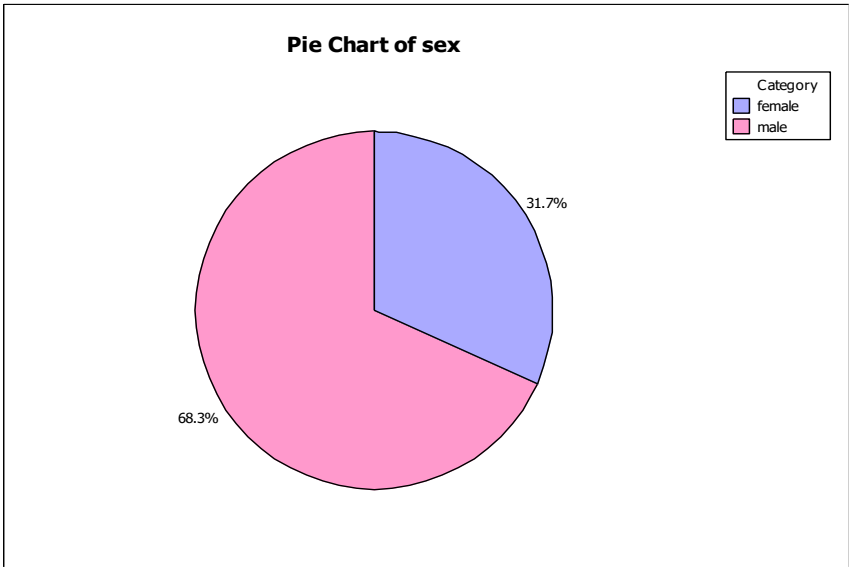

Figure 01 – Pie chart of sex

According to figure 01, male population is more than two times than female population in the sample and it is approximately 68%.

3.1.2 Composition of the sample with respect to Chest-pain type

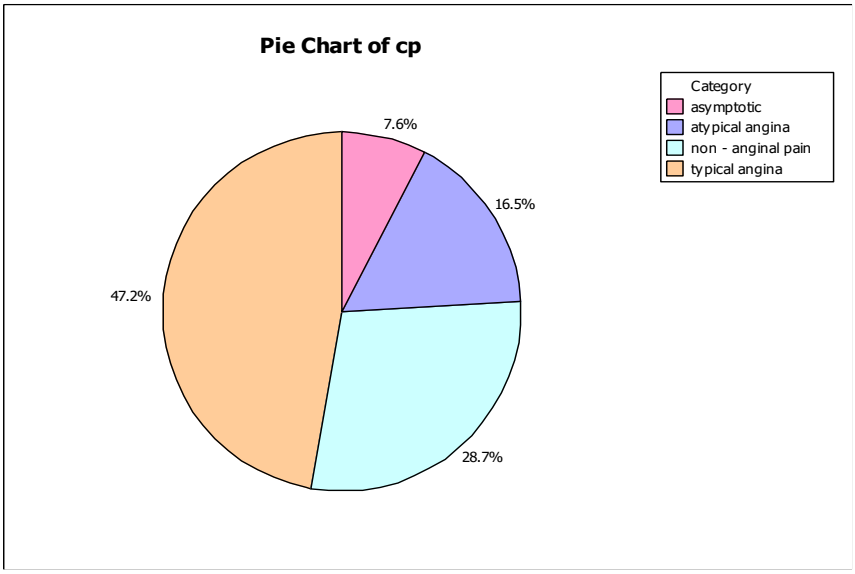

Figure 02 – Pie chart of Chest-pain type

Figure 02 shows, majority of peoples experienced chest pain due to typical angina while that pain can be non-angina pain more than 28%. It implies approximately, 75% percent of chest pain is due to typical angina or non-angina pain.

3.1.3 Composition of the sample with respect to Resting ECG

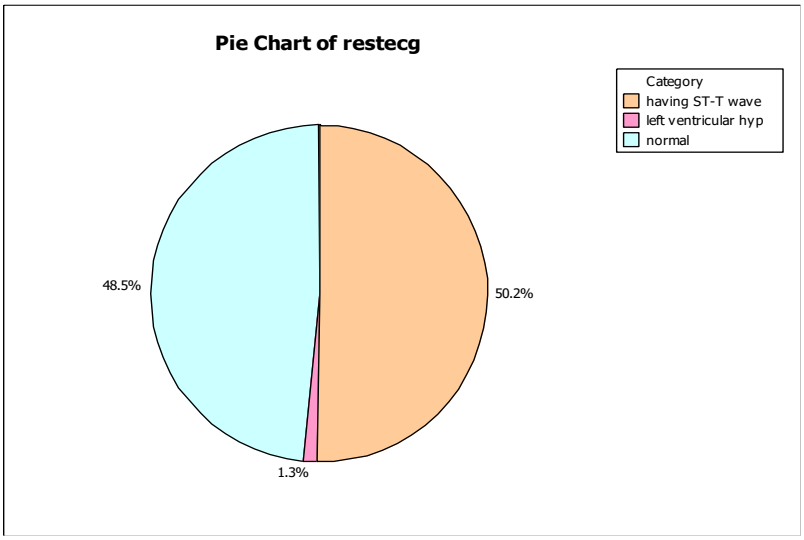

Figure 03 – Pie chart of Resting ECG

Figure 03 represents that resting ECG shows left ventricular hypertrophy very rarely than normal and ST-T wave abnormality. It is only around 1% of all test results.

3.1.4 Composition of the sample with respect to Fasting Blood Sugar

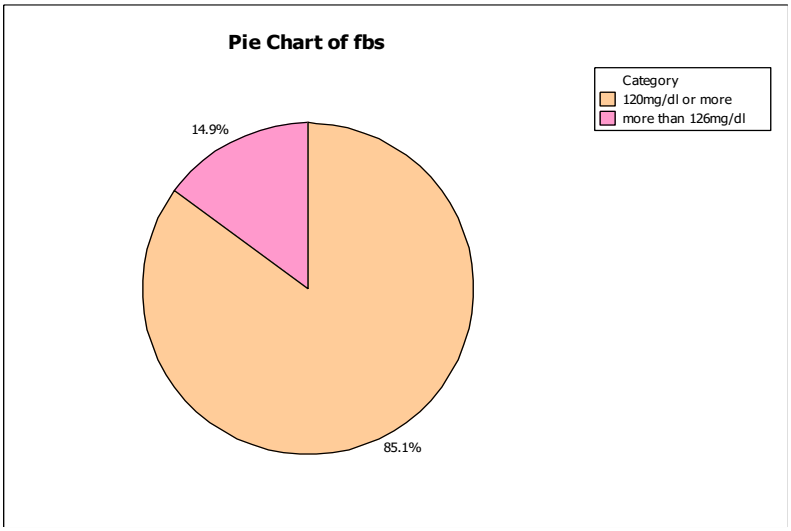

Figure 04 – Pie chart of Fasting Blood Sugar

According to figure 04, 15% of individuals have diabetes signals although all of individual have more than 120mg/dl FBS level.

3.1.5 Composition of the sample with respect to Exercise induced angina

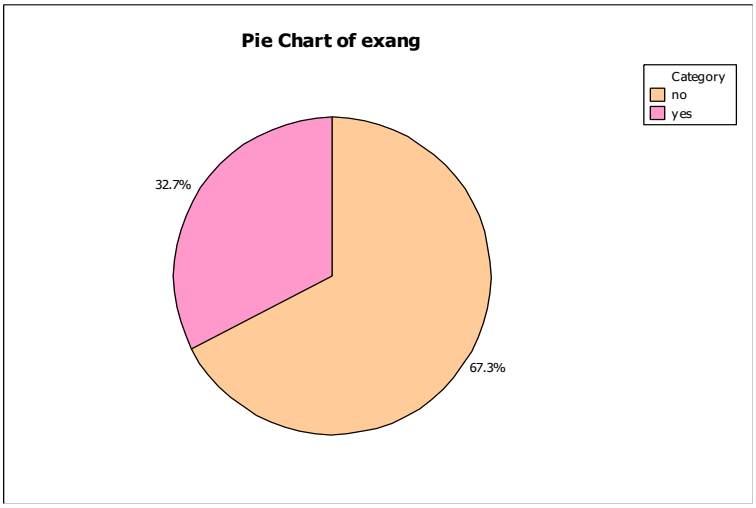

Figure 05 – Pie chart of Exercise induced angina

According to Figure 05, the number of individuals who did not experience exercise-induced angina is more than twice the number who did not.

3.1.6 Composition of the sample with respect to Slope of the peak exercise ST segment

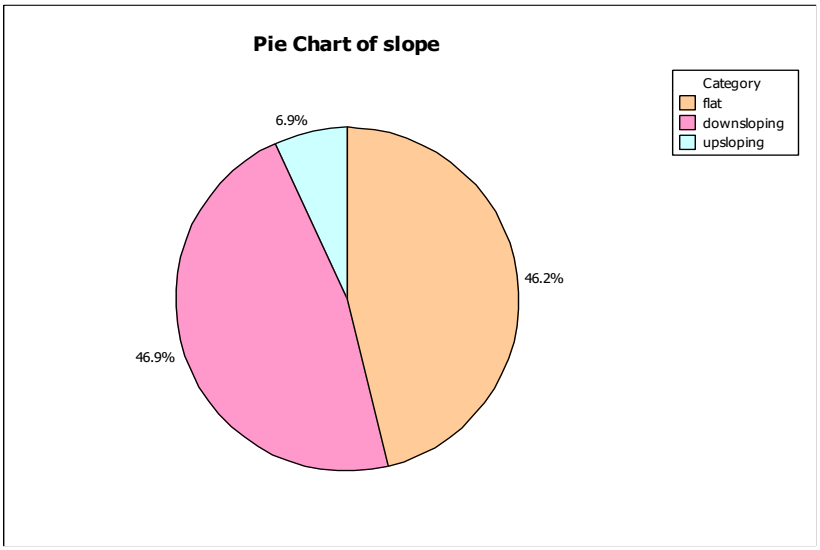

Figure 06 – Pie chart of Slope of the peak exercise ST segment

Figure 06 shows that more than 53% of individuals have healthy heart but only 7% individuals have better heart rate among them.

3.2 Distribution of quantitative independent variables

3.2.1 Distribution of Resting Blood Pressure

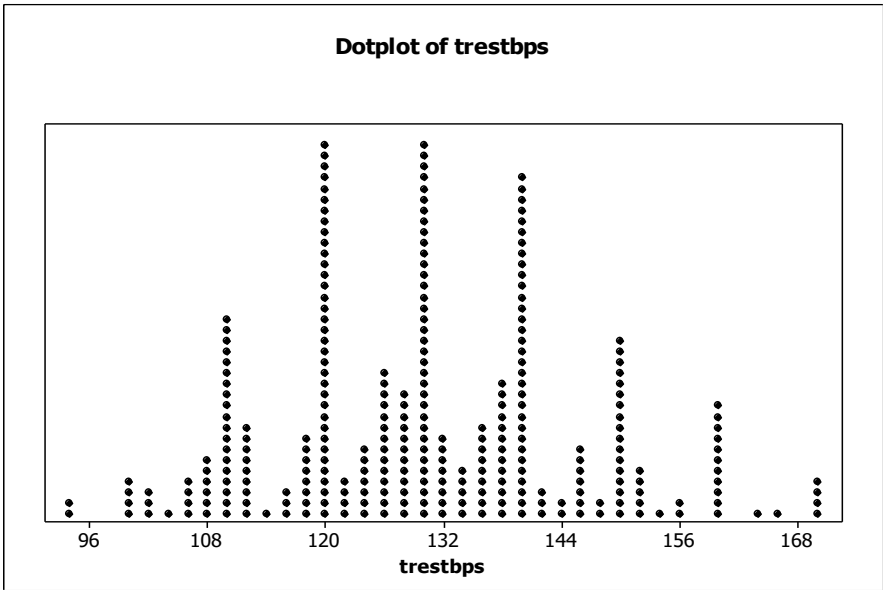

Figure 07 – Dot plot of Resting Blood Pressure

Table 01 - Descriptive Statistics of Resting Blood Pressure

Variable	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
trestbps	130.17	0.907	15.44	94.00	120.00	130.00	140.00	170.00

According to figure 07 and table 01 Resting Blood Pressure is symmetrically distributed around mean 130.17 after treating for all outliers.

3.2.2 Distribution of age

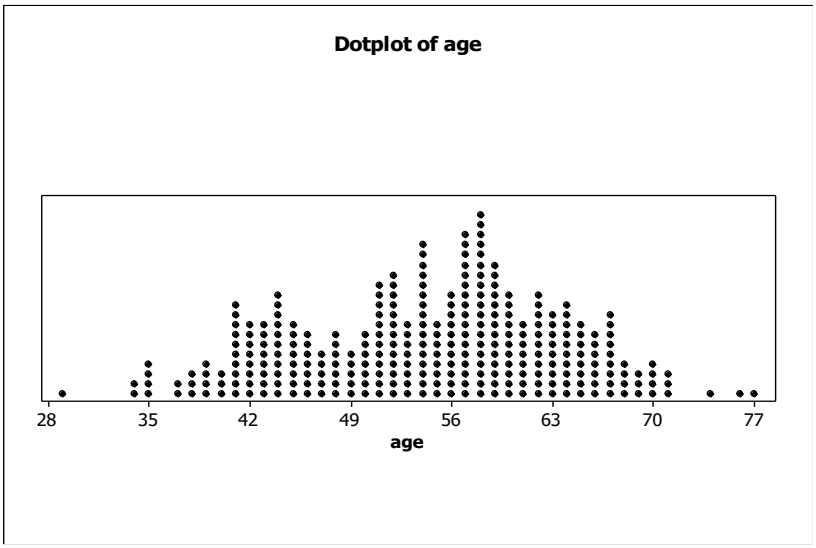

Figure 08 – Dot plot of age

Table 02 - Descriptive Statistics of age

Variable	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
age	54.366	0.522	9.082	29.000	47.000	55.000	61.000	77.000

Age is distributed nearly symmetrically with mean 54.336 and it is spread from 29 to 77 range according to figure 08 and table 02.

3.2.3 Distribution of Serum Cholesterol

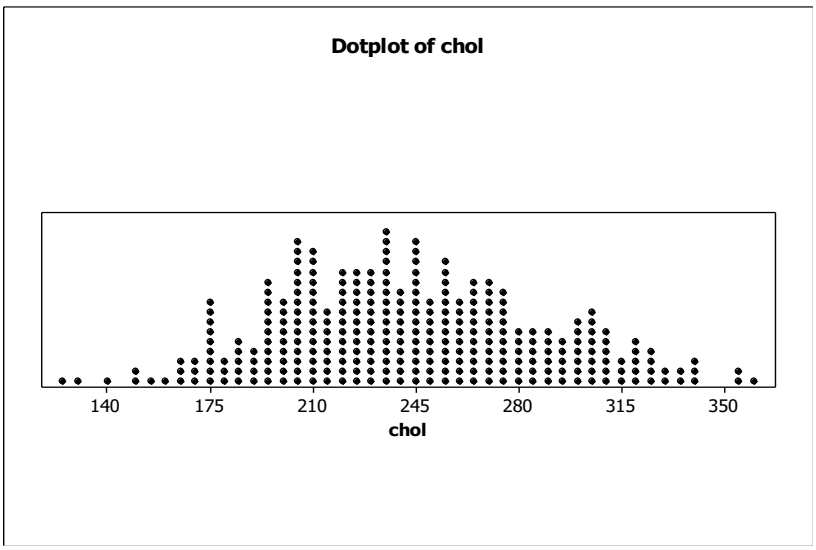

Figure 09 – Dot plot of Serum Cholesterol

Table 03 - Descriptive Statistics of Serum Cholesterol

Variable	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
chol	243.09	2.62	45.16	126.00	210.50	240.00	273.50	360.00

Figure 09 and table 03 shows, Serum Cholesterol level is approximately symmetrically distributed around mean 243.09 after treating for all outliers.

3.2.4 Distribution of Max heart rate achieved

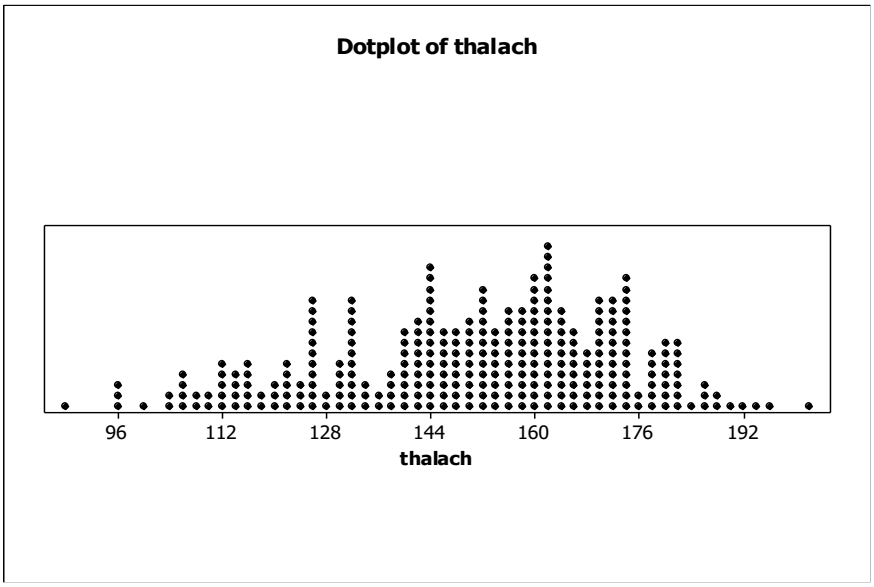

Figure 10 – Dot plot of Max heart rate achieved

Table 04 - Descriptive Statistics of Max heart rate achieved

Variable	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
thalach	150.13	1.28	22.13	88.00	135.00	153.00	166.50	202.00

Figure 10 and table 04 represent that max heart rate achieved has nearly negative skewed distribution with median 153 and the maximum of Max heart rate achieved is 202.

3.2.5 Distribution of ST depression induced by exercise relative to rest

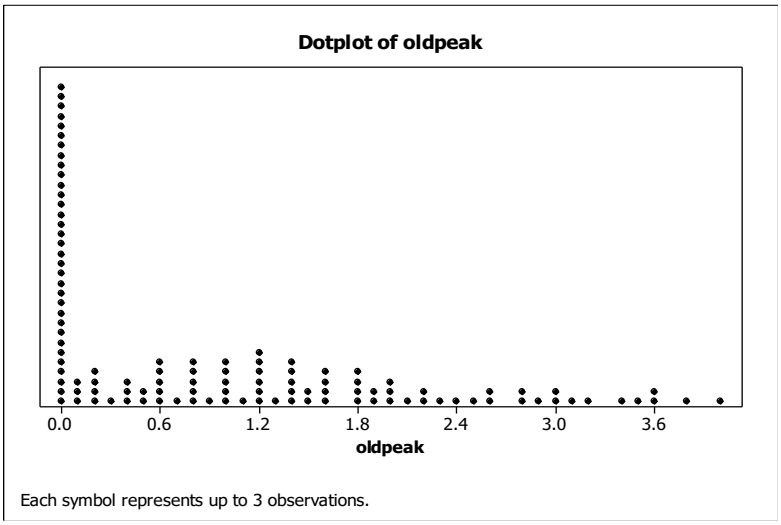

Figure 11 – Dot plot of ST depression induced by exercise relative to rest

Table 05 - Descriptive Statistics of ST depression induced by exercise relative to rest

Variable	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
oldpeak	0.9745	0.0608	1.0496	0.0000	0.0000	0.6500	1.6000	4.0000

Based on Figure 11 and Table 04, the ST depression induced by exercise relative to rest is asymmetrically distributed, with a median of 0.65. Additionally, zero appears more frequently than other values.

3.2.6 Distribution of Number of major vessels colored by fluoroscopy

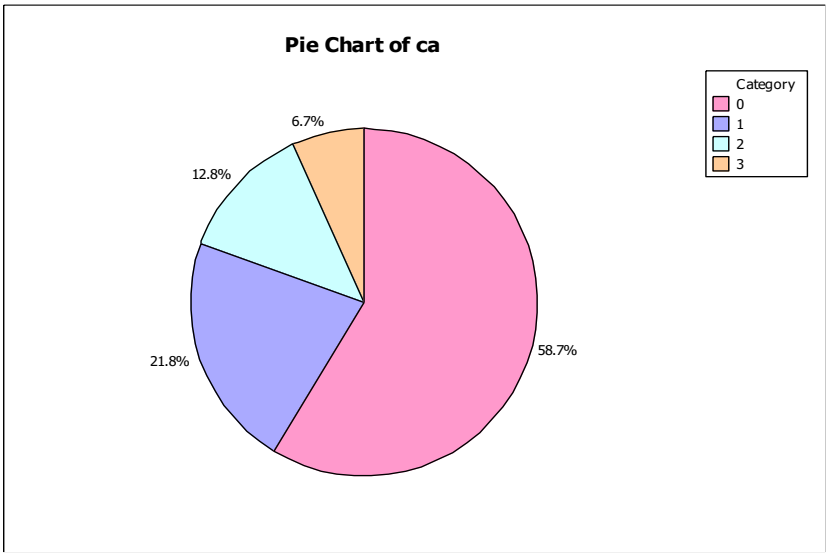

Figure 12 – Pie charts of Number of major vessels colored by fluoroscopy

Figure 12 shows that most of the times no vessels colored by fluoroscopy but more than 40% tests color at least one vessel.

3.3 Analyze the target variable

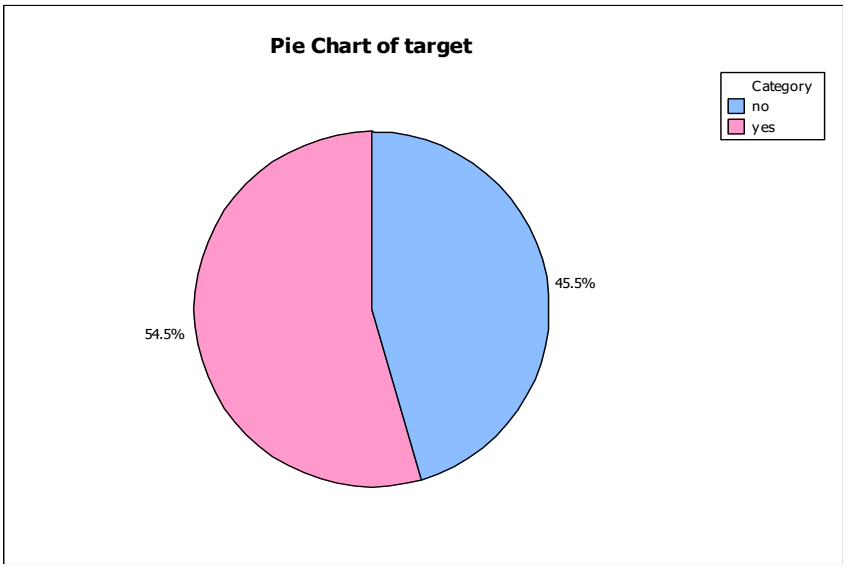

Figure 13 – Pie charts of target variable

According to figure 13,more than 50% of individuals have heart disease but there is no considerable difference between individuals who has heart disease and not.

4. Data analysis and Interpretation

4.1 Association between qualitative predictors and target variable

4.1.1 Association between sex and target

Table 06 - Tabulated statistics between target and sex

```
Rows: target   Columns: sex

       female    male     All

no         24     114     138
        17.39   82.61  100.00
        25.00   55.07   45.54

yes        72      93     165
        43.64   56.36  100.00
        75.00   44.93   54.46

All        96     207     303
        31.68   68.32  100.00
       100.00  100.00  100.00

Cell Contents:      Count
                    % of Row
                    % of Column


Pearson Chi-Square = 23.914, DF = 1, P-Value = 0.000
Likelihood Ratio Chi-Square = 24.841, DF = 1, P-Value = 0.000
```

Table 06 shows that 75% of females have heart diseases among all females but it is around 45% for males.

Hypothesis to be tested,

H0: There is no relationship between gender and target
H1: There is a relationship between gender and target

According to table 06, p-value is less than 0.0001. Therefore, we can reject Ho at 5% level of significance. It conclude that there is a relationship between gender and target.

4.1.2 Association between Chest-pain type and target

Table 07 - Tabulated statistics between target and Chest-pain type

```
Rows: target   Columns: cp

                            non -
                atypical  anginal  typical
      asymptotic   angina     pain   angina     All
```

```
no              7         9        18       104       138
             5.07      6.52     13.04     75.36    100.00
            30.43     18.00     20.69     72.73     45.54

yes            16        41        69        39       165
             9.70     24.85     41.82     23.64    100.00
            69.57     82.00     79.31     27.27     54.46

All            23        50        87       143       303
             7.59     16.50     28.71     47.19    100.00
           100.00    100.00    100.00    100.00    100.00

Cell Contents:      Count
                    % of Row
                    % of Column


Pearson Chi-Square = 81.686, DF = 3, P-Value = 0.000
Likelihood Ratio Chi-Square = 85.941, DF = 3, P-Value = 0.000
```

Hypothesis to be tested,

H0: There is no relationship between target and Chest-pain type
H1: There is a relationship between target and Chest-pain type

Table 07 shows that p-value is less than 0.0001. Therefore, we can reject Ho at 5% level of significance. It conclude that there is a relationship between target and Chest-pain type.

4.1.3 Association between Fasting Blood Sugar and target

Table 08 - Tabulated statistics between target and Fasting Blood Sugar

```
Rows: target   Columns: fbs

       120mg/dl    more than
        or more    126mg/dl    All

no          116          22     138
          84.06       15.94  100.00
          44.96       48.89   45.54

yes         142          23     165
          86.06       13.94  100.00
          55.04       51.11   54.46

All         258          45     303
          85.15       14.85  100.00
         100.00      100.00  100.00

Cell Contents:      Count
                    % of Row
                    % of Column
```

```
Pearson Chi-Square = 0.238, DF = 1, P-Value = 0.625
Likelihood Ratio Chi-Square = 0.238, DF = 1, P-Value = 0.626
```

Hypothesis to be tested,

H0: There is no relationship between target and Fasting blood sugar
H1: There is a relationship between target and Fasting blood sugar

Table 08 shows that p-value is 0.625 since it is greater than 0.05 we do not have enough evidence to reject Ho at 5% level of significance. It conclude that there is a no relationship between target and Chest-pain type.

4.1.4 Association between Resting ECG and target

Table 09 - Tabulated statistics between target and Resting ECG

```
Rows: target   Columns: restecg

                          left
          having  ventricular
       ST-T wave          hyp  normal     All

no            56            3      79     138
           40.58         2.17   57.25  100.00
           36.84        75.00   53.74   45.54

yes           96            1      68     165
           58.18         0.61   41.21  100.00
           63.16        25.00   46.26   54.46

All          152            4     147     303
           50.17         1.32   48.51  100.00
          100.00       100.00  100.00  100.00

Cell Contents:      Count
                    % of Row
                    % of Column


Pearson Chi-Square = 10.023, DF = 2, P-Value = 0.007
Likelihood Ratio Chi-Square = 10.113, DF = 2, P-Value = 0.006

* NOTE * 2 cells with expected counts less than 5
```

Hypothesis to be tested,

H0: There is no relationship between target and Resting ECG
H1: There is a relationship between target and Resting ECG

According to table 09, p-value is 0.007. Therefore, we can reject Ho at 5% level of significance. It conclude that there is a relationship target and resting ECG.

4.1.5 Association between target and Exercise induced angina

Table 10 - Tabulated statistics between target and Exercise induced angina

```
Rows: target   Columns: exang

           Yes     no     All

no          76     62     138
         55.07  44.93  100.00
         76.77  30.39   45.54

yes         23    142     165
         13.94  86.06  100.00
         23.23  69.61   54.46

All         99    204     303
         32.67  67.33  100.00
        100.00 100.00  100.00

Cell Contents:      Count
                    % of Row
                    % of Column


Pearson Chi-Square = 57.799, DF = 1, P-Value = 0.000
Likelihood Ratio Chi-Square = 59.735, DF = 1, P-Value = 0.000
```

Hypothesis to be tested,

H0: There is no relationship between target and Exercise induced angina
H1: There is a relationship between target and Exercise induced angina

Table 10 shows that p-value is less than 0.0001. Therefore, we can reject Ho at 5% level of significance. It conclude that there is a relationship target and Exercise induced angina.

4.1.6 Association between target and Slope of the peak exercise ST segment

Table 11 - Tabulated statistics between target and Slope of the peak exercise ST segment

```
Rows: target   Columns: slope

          flat  downsloping  upsloping     All

no          91           35         12     138
         65.94        25.36       8.70  100.00
         65.00        24.65      57.14   45.54
```

```
yes         49          107           9     165
         29.70        64.85        5.45  100.00
         35.00        75.35       42.86   54.46

All        140          142          21     303
         46.20        46.86        6.93  100.00
        100.00       100.00      100.00  100.00

Cell Contents:       Count
                     % of Row
                     % of Column


Pearson Chi-Square = 47.507, DF = 2, P-Value = 0.000
Likelihood Ratio Chi-Square = 49.076, DF = 2, P-Value = 0.000
```

Hypothesis to be tested,

H0: There is no relationship between target and Slope of the peak exercise ST segment
H1: There is a relationship between target and Slope of the peak exercise ST segment

Table 11 shows that p-value is less than 0.0001. Therefore, we can reject Ho at 5% level of significance. It conclude that there is a relationship target and Slope of the peak exercise ST segment.

4.1.7 Association between target and Number of major vessels colored by fluoroscopy

Table 12 - Tabulated statistics between target and Number of major vessels colored by fluoroscopy

```
Rows: target   Columns: ca

             0       1       2       3       All

no          45      44      31      17       137
         32.85   32.12   22.63   12.41    100.00
         25.71   67.69   81.58   85.00     45.97

yes        130      21       7       3       161
         80.75   13.04    4.35    1.86    100.00
         74.29   32.31   18.42   15.00     54.03

All        175      65      38      20       298
         58.72   21.81   12.75    6.71    100.00
        100.00  100.00  100.00  100.00    100.00

Cell Contents:       Count
                     % of Row
                     % of Column


Pearson Chi-Square = 72.922, DF = 3, P-Value = 0.000
Likelihood Ratio Chi-Square = 76.658, DF = 3, P-Value = 0.000
```

Hypothesis to be tested,

H0: There is no relationship between target and Number of major vessels colored by fluoroscopy
H1: There is a relationship between target and Number of major vessels colored by fluoroscopy

Table 12 shows that p-value is less than 0.0001. Therefore, we can reject Ho at 5% level of significance. It conclude that there is a relationship target and Number of major vessels colored by fluoroscopy.

4.2 Association between quantitative predictors and target variable

4.2.1 Association between target and age

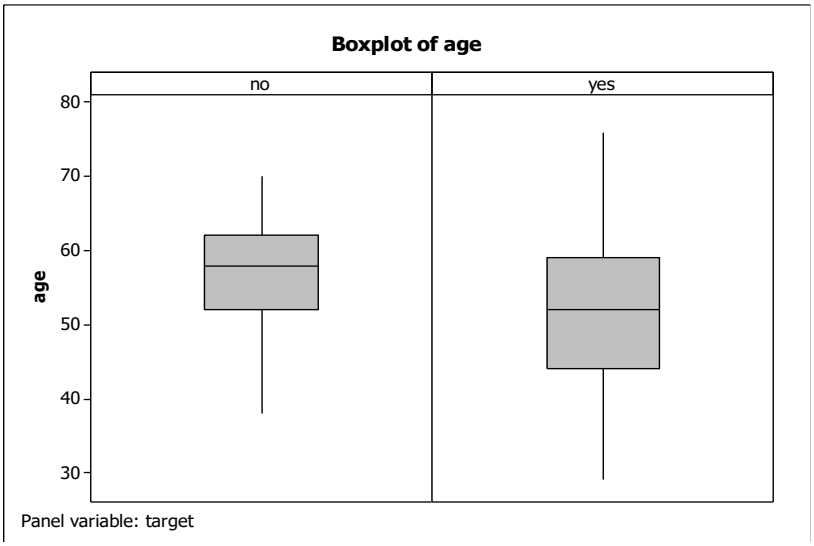

Figure 14- box plot of age by target

According to figure 14, median ages are different between individuals who has heart disease and not.

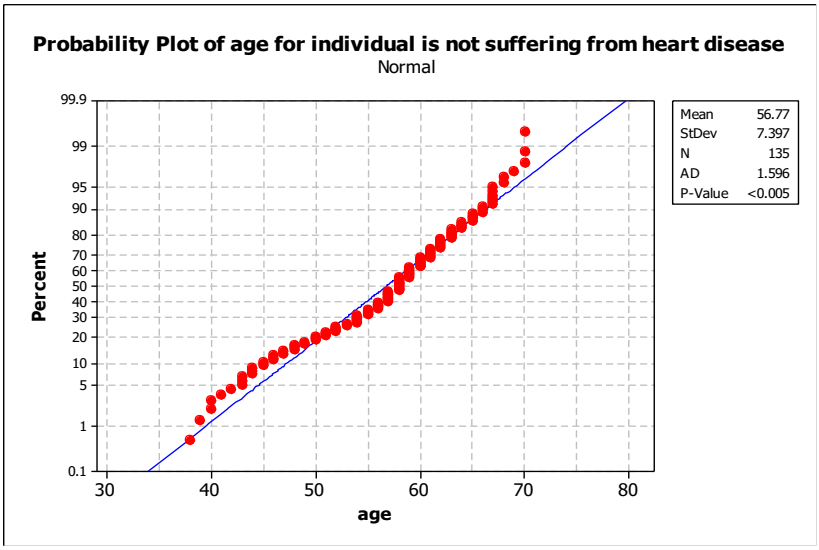

Figure 15- Normality plot of age for individuals who has not heart disease

Hypothesis to be tested,

H0: Age of individuals who has not heart disease is normally distributed.
H1: Age of individuals who has not heart disease is not normally distributed.

According to figure 12, p-value is less than 0.005. Therefore, we can reject Ho at 5% level of significance. It conclude that Age of individuals who has not heart disease is not normally distributed.

Table 13 - Mann-Whitney Test age by target

```
          N  Median
age_N   135  58.000
age_Y   165  52.000


Point estimate for ETA1-ETA2 is 5.000
95.0 Percent CI for ETA1-ETA2 is (2.999,7.001)
W = 23536.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0000
The test is significant at 0.0000 (adjusted for ties)
```

Hypothesis to be tested,

H0: Median age of individuals who has heart disease is equal to Median Age of individuals who has not heart disease
H1: Median age of individuals who has hear disease is not equal to Median Age of individuals who has not heart disease

Table 13 revels that p-value of Mann-Whitney test is less than 0.0001. Therefore, we can reject Ho at 5% level of significance. Since Median age of individuals who has heart disease is not equal to Median Age of individuals who has not heart disease. It conclude that there is an association between target and age.

4.2.2 Association between target and Resting Blood Pressure

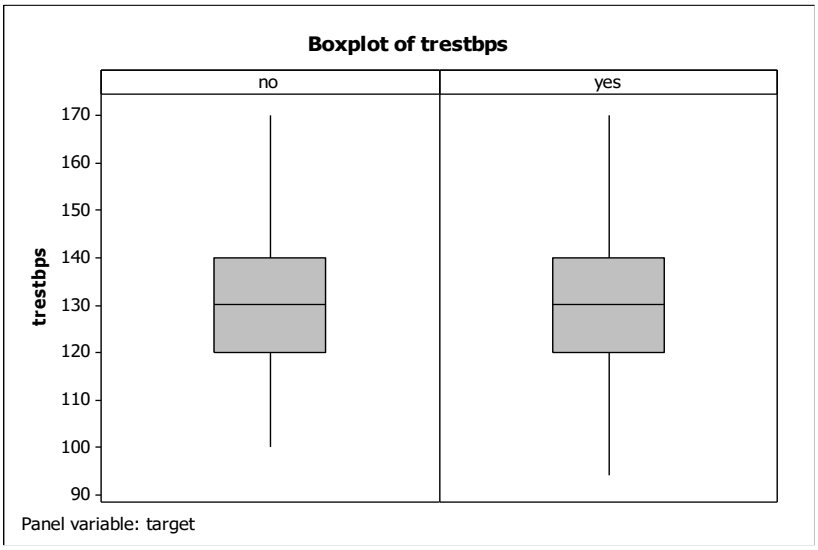

Figure 16- box plot of Resting Blood Pressure by target

According to figure 16, median Resting Blood Pressures are not different between individuals who has heart disease and not.

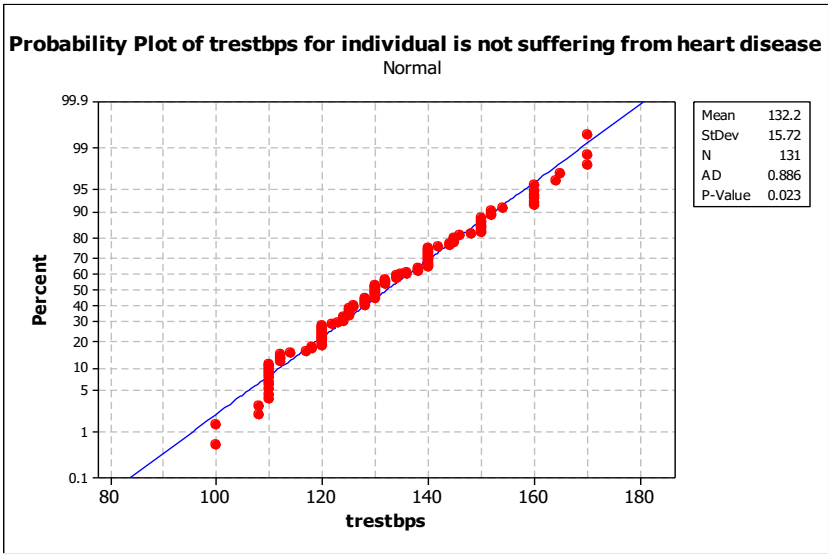

Figure 17- Normality plot of age for individuals who has not heart disease

Hypothesis to be tested,

H0: Resting blood pressure of individuals who has not heart disease is normally distributed.
H1: Resting blood pressure of individuals who has not heart disease is not normally distributed.

Figure 17 shows that p-value is 0.023 and it is less than 0.05. Therefore, we can reject Ho at 5% level of significance. It conclude that resting blood pressure of individuals who has not heart disease is not normally distributed.

Table 14 - Mann-Whitney Test resting blood pressure by target

```
              N  Median
trestbps_N  131  130.00
trestbps_Y  159  130.00


Point estimate for ETA1-ETA2 is 2.00
95.0 Percent CI for ETA1-ETA2 is (-0.00,8.00)
W = 20288.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0841
The test is significant at 0.0833 (adjusted for ties)
```

Hypothesis to be tested,

H0: Median resting blood pressure of individuals who has heart disease is equal to median resting blood pressure of individuals who has not heart disease
H1: Median resting blood pressure of individuals who has heart disease is not equal to median resting blood pressure of individuals who has not heart disease

Table 13 represents that p-value of Mann-Whitney test is 0.0841 and it is greater than 0.05. Therefore, we have not enough evidence to reject Ho at 5% level of significance. Since Median resting blood pressure of individuals who has heart disease is equal to median resting blood pressure of individuals who has not heart disease. It conclude that there is no significant association between target and resting blood pressure.

4.2.3 Association between target and Serum Cholesterol

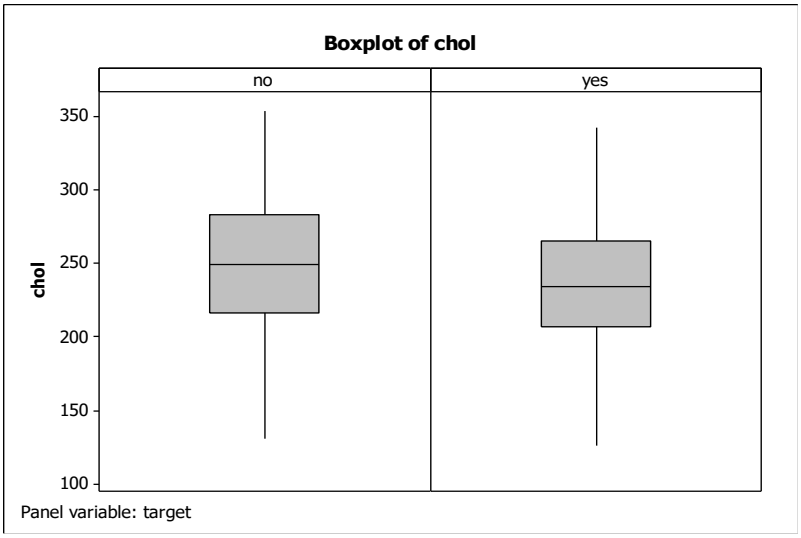

Figure 18- box plot of Resting Blood Pressure by target

According to figure 18, median serum cholesterol levels are slightly different between individuals who has heart disease and not.

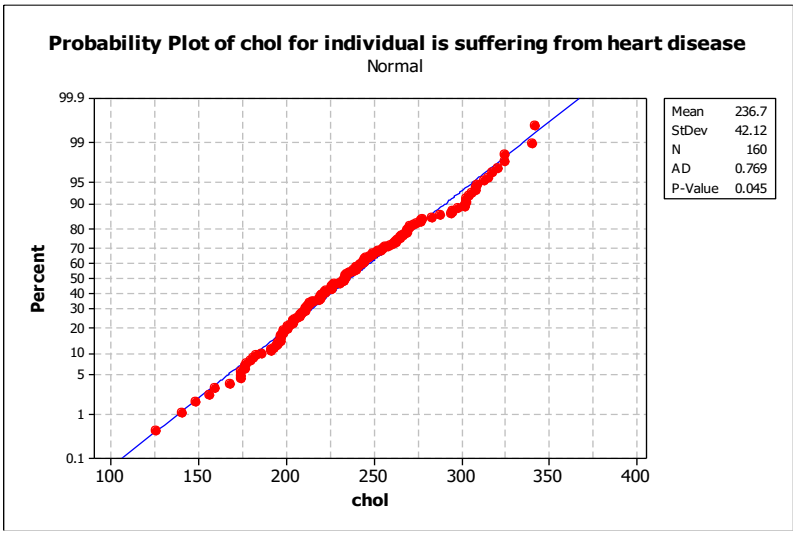

Figure 19- Normality plot of serum cholesterol for individuals who has not heart disease

Hypothesis to be tested,

H0: Serum cholesterol of individuals who has heart disease is normally distributed.
H1: Serum cholesterol of individuals who has heart disease is not normally distributed.

Figure 19 shows that p-value is 0.045 and it is less than 0.05. Therefore, we can reject Ho at 5% level of significance. It conclude that serum cholesterol of individuals who has heart disease is not normally distributed.

Table 15 - Mann-Whitney Test of serum cholesterol by target

```
           N  Median
chol_N1  135  249.00
chol_Y2  160  234.00


Point estimate for ETA1-ETA2 is 13.00
95.0 Percent CI for ETA1-ETA2 is (3.00,24.00)
W = 21760.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0147
The test is significant at 0.0147 (adjusted for ties)
```

Hypothesis to be tested,

H0: Median serum cholesterol of individuals who has heart disease is equal to median serum cholesterol of individuals who has not heart disease
H1: Median serum cholesterol of individuals who has heart disease is not equal to median serum cholesterol of individuals who has not heart disease

In table 15, p-value of Mann-Whitney test is 0.0147 and it is less than 0.05. Therefore, we can reject Ho at 5% level of significance. Since Median serum cholesterol of individuals who has heart disease is not equal to median serum cholesterol of individuals who has not heart disease. It conclude that there is an association between target and serum cholesterol.

4.2.4 Association between target and Max heart rate achieved

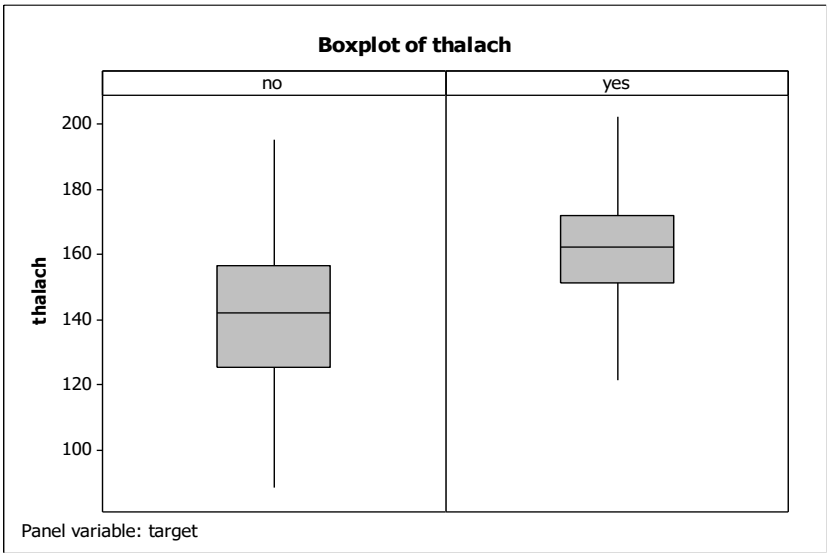

Figure 20- box plot of Max heart rate achieved by target

According to figure 20, individuals who has heart disease have higher median max heart rate than individuals who has not heart disease.

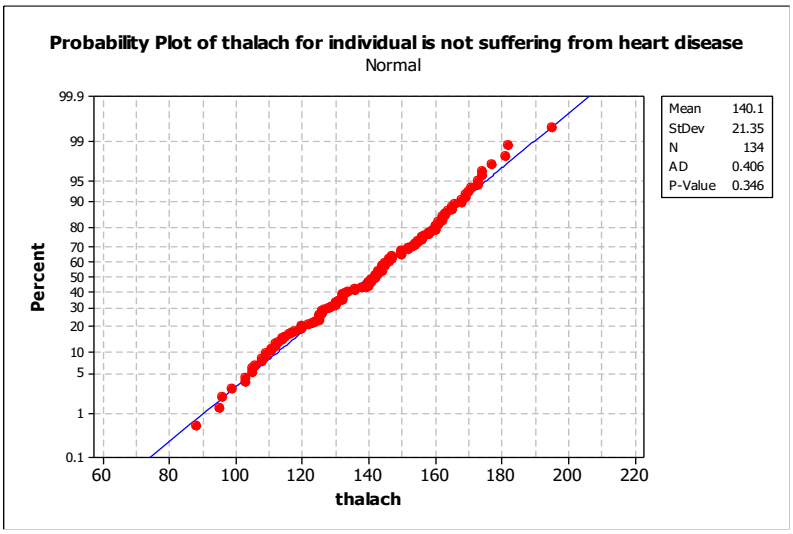

Figure 21- Normality plot of max heart rate achieved for individuals who has not heart disease

Hypothesis to be tested,

H0: Max heart rate achieved of individuals who has not heart disease is normally distributed.
H1: Max heart rate achieved of individuals who has not heart disease is not normally distributed.

Figure 21 shows that p-value is 0.346 and it is greater than 0.05. Therefore, we have not enough evidence to reject Ho at 5% level of significance. It conclude that max heart rate achieved of individuals who has not heart disease is normally distributed.

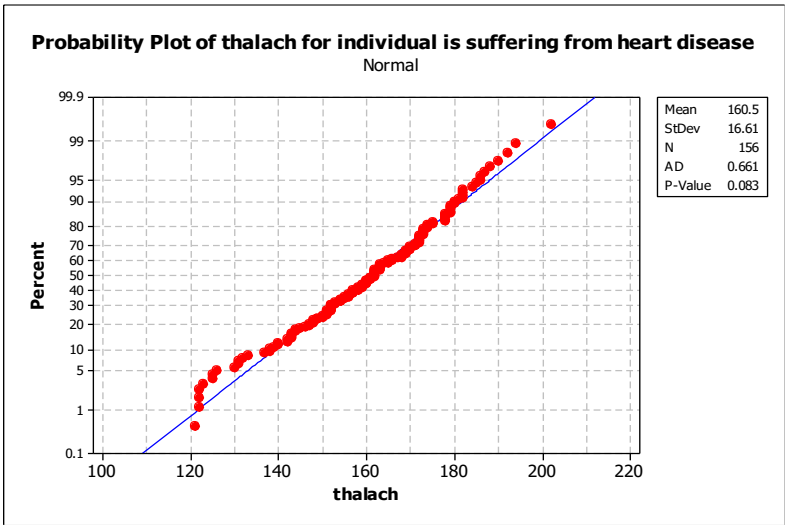

Figure 22- Normality plot of max heart rate achieved for individuals who has heart disease

Hypothesis to be tested,

H0: Max heart rate achieved of individuals who has heart disease is normally distributed.
H1: Max heart rate achieved of individuals who has heart disease is not normally distributed.

Figure 22 represent that p-value is 0.083 and it is greater than 0.05. Therefore, we have not enough evidence to reject Ho at 5% level of significance. It conclude that max heart rate achieved of individuals who has heart disease is normally distributed.

Table 15 - Test and CI for Two Variances between target and max heart rate achieved

```
Method

Null hypothesis         Sigma(no) / Sigma(yes) = 1
Alternative hypothesis  Sigma(no) / Sigma(yes) not = 1
Significance level      Alpha = 0.05


Statistics

target    N   StDev  Variance
no      134  21.353   455.947
yes     156  16.615   276.045

Ratio of standard deviations = 1.285
Ratio of variances = 1.652


95% Confidence Intervals

                                  CI for
Distribution   CI for StDev      Variance
of Data            Ratio           Ratio
Normal        (1.091, 1.517)  (1.191, 2.300)
Continuous    (1.108, 1.575)  (1.228, 2.482)


Tests

                                          Test
Method                          DF1  DF2  Statistic  P-Value
F Test (normal)                 133  155       1.65    0.003
Levene's Test (any continuous)    1  288       9.76    0.002
```

Hypothesis to be tested,

H0 : variances of max heart rate for individuals who has heart disease and variances of max heart rate for individuals who has not heart disease and are equal
H1 : variances of max heart rate for individuals who has heart disease and variances of max heart rate for individuals who has not heart disease and are not equal

According to table 15, p-value of F test is less than 0.05. Therefore, we can reject H0 at 5% level of significance. Hence variances of max heart rate achieved are not equal for individuals who has heart disease and not.

Table 16 - Two-Sample T-Test between target and max heart rate achieved

```
Two-sample T for thalach

target    N   Mean  StDev  SE Mean
no      134  140.1   21.4      1.8
yes     156  160.5   16.6      1.3


Difference = mu (no) - mu (yes)
Estimate for difference:  -20.34
95% CI for difference:  (-24.82, -15.87)
T-Test of difference = 0 (vs not =): T-Value = -8.95  P-Value = 0.000  DF = 249
```

Hypothesis to be tested,

H0: Mean max heart rate achieved of individuals who has heart disease is equal to mean max heart rate achieved of individuals who has not heart disease
H1: Mean max heart rate achieved of individuals who has heart disease is not equal to mean max heart rate achieved of individuals who has not heart disease

According to table 16, p-value of two sample T-test is less than 0.0001. Therefore, we can reject Ho at 5% level of significance. Since Mean max heart rate achieved of individuals who has heart disease is not equal to mean max heart rate achieved of individuals who has not heart disease. It conclude that there is an association between target and max heart rate achieved.

4.2.5 Association between target and ST depression induced by exercise relative to rest

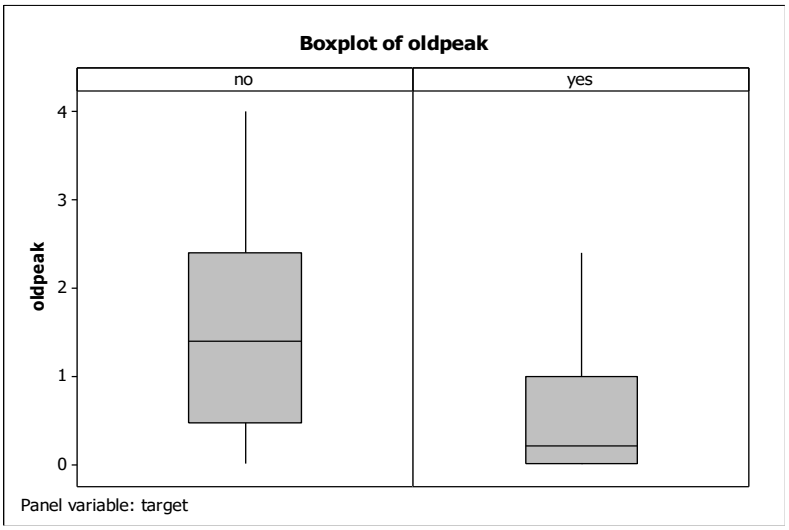

Figure 23- box plot of ST depression induced by exercise relative to rest by target

According to figure 23, individuals who has not heart disease have considerably higher median ST depression induced by exercise relative to rest than individuals who has heart disease.

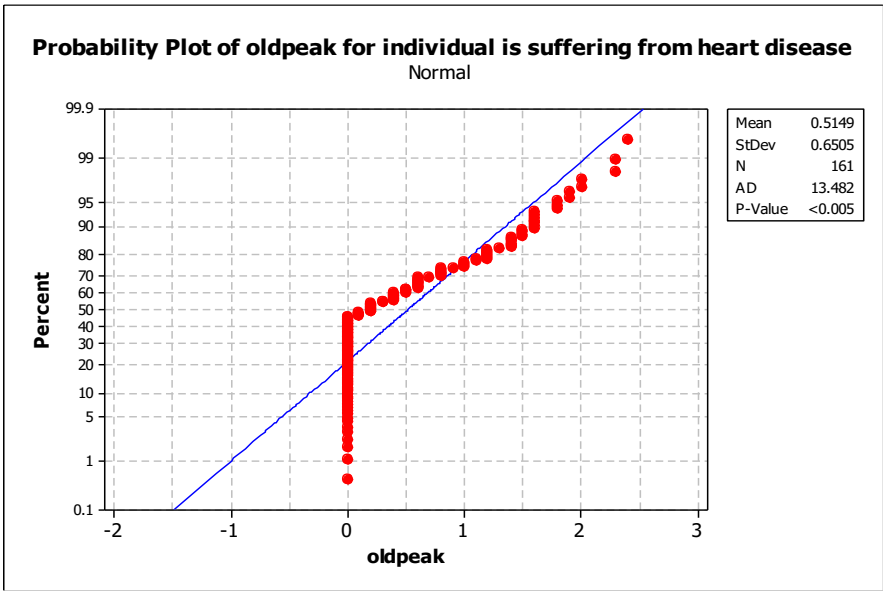

Figure 24- Normality plot of ST depression induced by exercise relative to rest for individuals who has heart disease

Hypothesis to be tested,

H0: ST depression induced by exercise relative to rest of individuals who has heart disease is normally distributed.
H1: ST depression induced by exercise relative to rest of individuals who has heart disease is not normally distributed.

Figure 24 shows that p-value is less than 0.005. Therefore, we can reject Ho at 5% level of significance. It conclude that ST depression induced by exercise relative to rest of individuals who has heart disease is not normally distributed.

Table 17 - Mann-Whitney Test of ST depression induced by exercise relative to rest by target

```
             N  Median
oldpeak_Y  161  0.2000
oldpeak_N  134  1.4000


Point estimate for ETA1-ETA2 is -1.0000
95.0 Percent CI for ETA1-ETA2 is (-1.2000,-0.7000)
W = 18478.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0000
The test is significant at 0.0000 (adjusted for ties)
```

Hypothesis to be tested,

H0: Median ST depression induced by exercise relative to rest of individuals who has heart disease is equal to median ST depression induced by exercise relative to rest of individuals who has not heart disease
H1: Median ST depression induced by exercise relative to rest of individuals who has heart disease is not equal to median ST depression induced by exercise relative to rest of individuals who has not heart disease

In table 17, p-value of Mann-Whitney test is less than 0.0001. Therefore, we can reject Ho at 5% level of significance. Since Median ST depression induced by exercise relative to rest of individuals who has heart disease is not equal to median ST depression induced by exercise relative to rest of individuals who has not heart disease. It conclude that there is an association between target and ST depression induced by exercise relative to rest.

4.3 Selection of the Optimal Logistic Regression Model

4.3.1 Selection of Key Main Effect Predictors

Table 18 – Main effects of Optimal Logistic Regression Model

Step	Model with lowest AIC in each step	Deviance	Df difference	AIC value
1.1	Null	352.64	-	354.64
1.2	cp	280.17	3	288.17
1.3	cp+ oldpeak	237.26	1	247.26
1.4	cp + oldpeak + ca	208.01	1	220.01
1.5	cp + oldpeak + ca + sex	193.93	1	207.93
1.6	cp + oldpeak + ca + sex+ slope	180.38	2	198.38
1.7	cp + oldpeak + ca + sex+ slope +exang	173.01	1	193.01
1.8	cp + oldpeak + ca + sex+ slope +exang +chol	171.00	1	193.00
1.9	cp + oldpeak + ca + sex+ slope +exang +chol + thalach	168.61	1	192.61

4.3.2 Selection of Key Main Effect Predictors with interaction effects

Table 19 – Main and interaction effects of Optimal Logistic Regression Model

Step	Model with lowest AIC in each step	Deviance	Df difference	AIC value
2.1	cp + oldpeak + ca + sex+ slope +exang +chol + thalach	168.61	1	192.61
2.2	cp + oldpeak + ca + sex + slope + exang + chol + thalach + cp:exang	160.89	3	190.89
2.3	cp + oldpeak + ca + sex + slope + exang + chol + thalach + cp:exang + oldpeak:exang	157.19	1	189.19
2.4	cp + oldpeak + ca + sex + slope + exang + chol + thalach + cp:exang + oldpeak:exang + ca:exang	153.68	1	187.68
2.5	cp + oldpeak + ca + sex + slope + exang + chol + thalach + cp:exang + oldpeak:exang + ca:exang + sex:exang	149.34	1	185.34
2.6	cp + oldpeak + ca + sex + slope + exang + chol + thalach + cp:exang + oldpeak:exang + ca:exang + sex:exang + sex:thalach	146.66	1	184.66

Table 20 – summary of Optimal Logistic Regression Model

```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  9.290802   4.621926   2.010   0.0444 *
cpatypical angina           -0.378228   0.806917  -0.469   0.6393
cpnon - anginal pain         0.432795   0.751517   0.576   0.5647
cptypical angina            -1.232145   0.733112  -1.681   0.0928 .
oldpeak                     -0.460842   0.279225  -1.650   0.0989 .
ca                          -1.148244   0.283853  -4.045 5.23e-05 ***
sexmale                    -10.163801   4.694264  -2.165   0.0304 *
slopeflat                   -1.110994   0.514888  -2.158   0.0309 *
slopeupsloping              -0.170922   1.059448  -0.161   0.8718
exangyes                     6.483912   4.578227   1.416   0.1567
chol                        -0.007550   0.005622  -1.343   0.1793
thalach                     -0.018842   0.026300  -0.716   0.4737
cpatypical angina:exangyes   2.003380  12.406356   0.161   0.8717
cpnon - anginal pain:exangyes -5.224849   3.298485  -1.584   0.1132
cptypical angina:exangyes    -8.516258   4.164707  -2.045   0.0409 *
```

```
oldpeak:exangyes              -3.147706   1.594588  -1.974   0.0484 *
ca:exangyes                   -4.806613   2.740162  -1.754   0.0794 .
sexmale:exangyes               4.203283   2.041268   2.059   0.0395 *
sexmale:thalach                0.050069   0.029995   1.669   0.0951 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 352.64  on 255  degrees of freedom
Residual deviance: 146.66  on 237  degrees of freedom
AIC: 184.66
```

Final model:

Logit(p(Target=1)) = log (P(Target2=1) / 1- P(Target2=1)) = $\beta_0+\sum\beta_iX_i$

Logit(p(Target=1)) = 9.2908 – 0.3782 cp $_{\text{atypical angina}}$ **+ 0.4328 cp** $_{\text{non anginal pain}}$

- 1.2321 cp $_{\text{typical angina}}$ **- 0.4608 oldpeak - 1.1482 ca - 10.1638 sex** $_{\text{male}}$

- 1.1110 slope $_{\text{flat}}$ **- 0.1709 slope** $_{\text{unsloping}}$ **+ 6.4839 exang** $_{\text{yes}}$ **- 0.0076 chol**

- 0.0188 thalach + 2.0034 cp $_{\text{atypical angina}}$ **. exang** $_{\text{yes}}$

- 5.2248 cp $_{\text{non anginal pain}}$ **. exang** $_{\text{yes}}$ **- 8.5163 cp** $_{\text{typical angina}}$ **. exang** $_{\text{yes}}$

- 3.1477 oldpeak . exang $_{\text{yes}}$ **- 4.8066 ca . exang** $_{\text{yes}}$

+ 4.2033 sex $_{\text{male}}$ **. exang** $_{\text{yes}}$ **+ 0.0501 sex** $_{\text{male}}$ **. thalach**

Interpretation:

Main Effects

- **cp** $_{\text{atypical angina}}$ **(β = -0.3782):** Compared to the asymptomatic Chest-pain , odds of having heart disease decrease by a factor of 0.685 ($e^{-0.3782}$) for atypical angina when other variables are held constant. It means that individuals with atypical angina have about 31.5% lower odds of having heart disease compared to those with asymptomatic chest pain.
- **cp** $_{\text{non-anginal pain}}$ **(β = 0.4328):** Compared to the asymptomatic Chest-pain, odds of having heart disease increase by a factor of 1.542 ($e^{0.4328}$) for non-angina pain when other variables are held constant. It means that individuals with non-angina pain have about 54.2% higher odds of having heart disease compared to those with asymptomatic chest pain.
- **Cp** $_{\text{typical angina}}$ **(β = -1.2321):** Compared to the asymptomatic Chest-pain, odds of having heart disease decrease by a factor of 0.292 ($e^{-1.2321}$) for typical angina when other

variables are held constant. It means that individuals with typical angina have about 70.8% lower odds of having heart disease compared to those with asymptomatic chest pain.

- **oldpeak (β = -0.4608):** For each one-unit increase in oldpeak (ST depression induced by exercise relative to rest) the odds of having heart disease decreases by 0.631($e^{-0.4608}$) when other variables are held constant.
- **ca (β = -1.1482):** For each one-unit increase in ca (number of major vessels colored by fluoroscopy), the odds of having heart disease decreases by 0.317($e^{-1.1482}$) when other variables are held constant.
- **sex_{male} (β = -10.1638):** Compared to the female individuals, odds of having heart disease decrease by a factor of 0.000038 ($e^{-10.1638}$) for male individuals when other variables are held constant. This very large coefficient might indicate multicollinearity or a very strong effect.
- **$slope_{flat}$ (β = -1.1110):** Compared to the downsloping, odds of having heart disease decrease by a factor of 0.329 ($e^{-1.1110}$) for flat slope when other variables are held constant. It means that individuals with flat slope have about 67.1% lower odds of having heart disease compared to those with a downsloping slope.
- **$slope_{upsloping}$ (β = -0.1709):** Compared to the downsloping, odds of having heart disease decrease by a factor of 0.843 ($e^{-0.1709}$) for upsloping slope when other variables are held constant. It means that individuals with upsloping slope have about 15.7% lower odds of having heart disease compared to those with a downsloping slope.
- **$exang_{yes}$ (β = 6.4839):** Compared to the no exercise-induced angina, odds of having heart disease increase by a factor of 654.7 ($e^{6.4839}$) for exercise-induced angina when other variables are held constant. It means that individuals with exercise-induced angina have very higher odds (about 654 times) of having heart disease compared to those with no exercise-induced angina.
- **chol (β = -0.0076):** For each one-unit increase in chol (serum cholesterol in mg/dl), the odds of having heart disease decreases by 0.992($e^{-0.0076}$) when other variables are held constant.
- **thalach (β = -0.0188):** For each one-unit increase in thalach (maximum heart rate achieved), the odds of having heart disease decreases by 0.981($e^{-0.0188}$) when other variables are held constant.

Interaction Effects

- **$cp_{atypical\ angina}$ * $exang_{yes}$ (β = 2.0034):** Compared to the asymptomatic chest pain without exercise-induced angina, odds of having heart disease increase by a factor of exp(-0.3782+2.0034+6.4839) =3324 for individuals those who are having exercise-induced angina with atypical angina when other variables are held constant.
- **$cp_{non\text{-}anginal\ pain}$ * $exang_{yes}$ (β = -5.2248):** Compared to the asymptomatic chest pain without exercise-induced angina, odds of having heart disease increase by a factor of exp(0.4328+6.4839-5.2248) = 5.429 for individuals those who are having exercise-induced angina with non-anginal pain when other variables are held constant.
- **$cp_{typical\ angina}$ * $exang_{yes}$ (β = -8.5163):** Compared to the asymptomatic chest pain without exercise-induced angina, odds of having heart disease decrease by a factor of exp(-1.2321+6.4839-8.5163)=0.0382 for individuals those who are having exercise-induced angina with typical angina when other variables are held constant.
- **oldpeak * $exang_{yes}$ (β = -3.1477):** For individuals with exercise-induced angina , each one-unit increase in oldpeak (ST depression induced by exercise relative to rest) is associated with a decrease in the odds of having heart disease by approximately 97.3% ($e^{-0.4608-3.1477}$= 0.027), compared to those without exercise-induced angina, holding other variables constant.
- **ca * $exang_{yes}$ (β = -4.8066):** For individuals with exercise-induced angina , each one-unit increase in ca (number of major vessels colored by fluoroscopy) is associated with a decrease in the odds of having heart disease by approximately 99.74% ($e^{-1.1482-4.8066}$ = 0.0026), compared to those without exercise-induced angina, holding other variables constant.
- **sex_{male} * $exang_{yes}$ (β = 4.2033):** Compared to the female individuals without exercise-induced angina, odds of having heart disease decrease by a factor of 0.0026 exp(-10.1638+6.4839+4.2033)= 1.6877 for male individuals those who are having exercise-induced angina when other variables are held constant.
- **sex_{male} * thalach (β = 0.0501):** For male individuals, each one-unit increase in thalach (maximum heart rate achieved) is associated with an increase in the odds of having heart disease by approximately 3.2% ($e^{-0.0188+0.0501}$= 1.032), compared to female individuals, holding other variables constant.

5. Discussion

This study aimed to analyzes a heart disease classification dataset to predict which patients are most likely to suffer from a heart disease in the near future and also to identify key predictors of heart disease. The analysis revealed several significant findings.

Firstly, all variables were carefully explored. According to that data, female population is more than twice of females. Further analysis showed that more than 47% of individuals experience typical angina, approximately 15% show potential signs of diabetes (based on fasting blood sugar) and more than 53% of individuals have a healthy heart. However, only 7% of individuals exhibit better heart rate with exercise (upsloping ST segment).

When considering quantitative variables, after treating for outliers, all predictors except oldpeak and thalach were approximately symmetrically distributed. Further analysis of the target variable indicated that more than 54% of individuals are suffering from heart disease.

Next, the association of each predictor variables with outcome variable was investigated and it revealing highlighted differences in the distribution of several variables between individuals with and without heart disease. Gender appears to be a notable factor, with approximately 75% of females having heart disease compared to around 45% of males. Additionally, approximately 86% of individuals with heart disease did not have exercise-induced angina, and more than 80% of individuals with heart disease had 0 major vessels colored by fluoroscopy. Furthermore, the age of individuals with heart disease tends to be lower than those without heart disease and individuals who has heart disease have considerably higher median max heart rate than individuals who has not heart disease while ST depression induced by exercise relative to rest shows an opposite trend to maximum heart rate.

Finally, a logistic regression model was constructed to predict the presence of heart disease. The forward selection process using AIC values identified chest pain type (cp), ST depression induced by exercise relative to rest (oldpeak), the number of major vessels colored by fluoroscopy (ca), sex, slope of the peak exercise ST segment (slope), exercise-induced angina (exang), serum cholesterol (chol), and maximum heart rate achieved (thalach) as potential predictors. However, serum cholesterol (chol) and maximum heart rate achieved (thalach) showed less significance at the 5% level.

However this dataset is influenced by outliers and exhibits a gender imbalance. The very large coefficient of sex might indicate multicollinearity. Additionally, the dataset includes only a specific set of risk factors, and other potentially relevant variables, such as lifestyle factors (e.g., smoking, diet, physical activity), were not available for inclusion in the model. The absence of

these variables may limit the model's predictive power and its ability to fully capture the complex interplay of factors contributing to heart disease.

6. Conclusions

In conclusion, this analysis of the heart disease dataset successfully identified several key predictors and associations relevant to heart disease.

The investigation into the associations between individual predictors and the target variable revealed significant relationships with age, chest pain type (cp), ST depression induced by exercise relative to rest (oldpeak), number of major vessels colored by fluoroscopy (ca), slope of the peak exercise ST segment (slope), resting ECG (restecg), exercise-induced angina (exang), and sex. In contrast, resting blood pressure (trestbps) and fasting blood sugar (fbs) did not show statistically significant associations with the presence of heart disease in this dataset.

The logistic regression model, developed using forward selection based on AIC and considering both main effects and interaction terms, identified chest pain type (cp), ST depression induced by exercise relative to rest (oldpeak), number of major vessels colored by fluoroscopy (ca), sex (specifically male), slope of the peak exercise ST segment (slope), exercise-induced angina (exang), serum cholesterol (chol), and maximum heart rate achieved (thalach) as potential predictors, along with significant interaction effects between cp and exang, oldpeak and exang (for exang = yes), ca and exang (for exang = yes), sex (male) and exang (for exang = yes), and sex (male) and thalach. It's important to note that while included in the final model, serum cholesterol and maximum heart rate achieved exhibited lower levels of statistical significance ($p > 0.05$).

While the study provides valuable insights, it is important to acknowledge its limitations, particularly the sample imbalance and the absence of lifestyle-related risk factors.

Future research should aim to expand the dataset with a more balanced representation of classes and include a broader range of relevant variables, such as smoking status, diet, and physical activity. Furthermore, exploring more advanced machine learning techniques could potentially enhance the accuracy and generalizability of heart disease prediction models.

Appendix

Appendix 01

Step 1.2

```
Start:  AIC=354.64
target ~ 1

           Df Deviance    AIC
+ cp        3   280.17 288.17
+ oldpeak   1   285.74 289.74
+ thalach   1   287.10 291.10
+ ca        1   294.40 298.40
+ exang     1   300.46 304.46
+ slope     2   309.06 315.06
+ sex       1   327.79 331.79
+ age       1   334.76 338.76
+ restecg   2   342.29 348.29
+ trestbps  1   348.62 352.62
+ chol      1   349.17 353.17
<none>          352.64 354.64
+ fbs       1   352.58 356.58
```

Step 1.3

```
Step:  AIC=288.17
target ~ cp

           Df Deviance    AIC
+ oldpeak   1   237.26 247.26
+ ca        1   240.63 250.63
+ thalach   1   249.29 259.29

+ slope     2   253.47 265.47
+ sex       1   258.32 268.32
+ exang     1   261.16 271.16
+ age       1   270.19 280.19
+ trestbps  1   276.13 286.13
+ restecg   2   275.58 287.58
+ chol      1   277.68 287.68
<none>          280.17 288.17
+ fbs       1   279.66 289.66
```

Step 1.4

```
Step:  AIC=247.26
target ~ cp + oldpeak

           Df Deviance    AIC
+ ca        1   208.01 220.01
+ sex       1   220.50 232.50
+ thalach   1   222.91 234.91
+ exang     1   227.30 239.30
+ slope     2   229.93 243.93
+ age       1   232.56 244.56
+ chol      1   234.54 246.54
<none>          237.26 247.26
+ restecg   2   233.34 247.34
+ trestbps  1   235.51 247.51
+ fbs       1   236.93 248.93
```

Step 1.5

```
Step:  AIC=220.01
target ~ cp + oldpeak + ca

           Df Deviance    AIC
+ sex       1   193.93 207.93
+ exang     1   198.04 212.04
+ slope     2   198.22 214.22
+ thalach   1   200.26 214.26
<none>          208.01 220.01
+ restecg   2   204.68 220.68
+ chol      1   206.69 220.69
+ trestbps  1   206.78 220.78
+ age       1   207.75 221.75
+ fbs       1   208.00 222.00
```

Step 1.6

```
Step:  AIC=207.92
target ~ cp + oldpeak + ca + sex

           Df Deviance    AIC
+ slope     2   180.38 198.38
+ exang     1   183.93 199.93
+ thalach   1   184.49 200.49
+ chol      1   190.81 206.81
+ restecg   2   189.89 207.89
<none>          193.93 207.93
+ trestbps  1   192.26 208.26
+ age       1   192.79 208.79
+ fbs       1   193.80 209.80
```

Step 1.7

```
Step:  AIC=198.38
target ~ cp + oldpeak + ca + sex + sl
ope

           Df Deviance    AIC
+ exang     1   173.01 193.01
+ thalach   1   176.48 196.48
+ chol      1   177.96 197.96
+ trestbps  1   178.22 198.22
<none>          180.38 198.38
+ fbs       1   179.86 199.86
+ age       1   180.09 200.09
+ restecg   2   178.10 200.10
```

Step 1.8

```
Step:  AIC=193.01
target ~ cp + oldpeak + ca + sex + sl
ope + exang

           Df Deviance    AIC
+ chol      1   171.00 193.00
<none>          173.01 193.01
+ trestbps  1   171.07 193.07
+ thalach   1   171.10 193.10
+ fbs       1   172.13 194.13
+ restecg   2   170.50 194.50
+ age       1   172.91 194.91
```

Step 1.9

```
Step:  AIC=193
target ~ cp + oldpeak + ca + sex + sl
ope + exang + chol

           Df Deviance    AIC
+ thalach   1   168.61 192.61
<none>          171.00 193.00
+ trestbps  1   169.42 193.42
+ fbs       1   169.72 193.72
+ age       1   170.92 194.92
+ restecg   2   168.98 194.98
```

Step 1.10

```
Step:  AIC=192.61
target ~ cp + oldpeak + ca + sex + sl
ope + exang + chol + thalach

           Df Deviance    AIC
<none>          168.61 192.61
+ trestbps  1   166.70 192.70
+ fbs       1   167.41 193.41
+ age       1   168.59 194.59
+ restecg   2   166.67 194.67
```

Appendix 02

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.122377   2.388134   1.726  0.08431 .
cpatypical angina   -0.813935   0.779302  -1.044  0.29628
cpnon - anginal pain  0.287456   0.676101   0.425  0.67071
cptypical angina    -1.951764   0.659122  -2.961  0.00306 **
oldpeak             -0.594390   0.242439  -2.452  0.01422 *
ca                  -1.186501   0.273767  -4.334 1.46e-05 ***
sexmale             -2.147135   0.536581  -4.002 6.29e-05 ***
slopeflat           -1.248532   0.493354  -2.531  0.01138 *
slopeupsloping      -0.853422   0.941124  -0.907  0.36451
exangyes            -1.026537   0.465620  -2.205  0.02748 *
chol                -0.008060   0.005181  -1.556  0.11982
thalach              0.018488   0.012108   1.527  0.12678
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 352.64  on 255  degrees of freedom
Residual deviance: 168.61  on 244  degrees of freedom
AIC: 192.61
```

Appendix 03

Step 2.2

```
Start:  AIC=192.61
target ~ cp + oldpeak + ca + sex + sl
ope + exang + chol + thalach

                 Df Deviance    AIC
+ cp:exang        3   160.89 190.89
+ sex:exang       1   165.66 191.66
+ sex:thalach     1   165.74 191.74
+ cp:chol         3   161.94 191.94
<none>                168.61 192.61
+ ca:thalach      1   166.75 192.75
+ ca:chol         1   166.80 192.80
+ exang:chol      1   167.30 193.30
+ chol:thalach    1   167.78 193.78
+ sex:chol        1   167.79 193.79
+ oldpeak:exang   1   167.81 193.81
+ oldpeak:sex     1   167.96 193.96
+ oldpeak:thalach 1   168.00 194.00
+ oldpeak:slope   2   166.11 194.11
+ ca:exang        1   168.12 194.12
+ ca:sex          1   168.14 194.14
+ oldpeak:chol    1   168.22 194.22
+ exang:thalach   1   168.24 194.24
+ oldpeak:ca      1   168.58 194.58
+ slope:exang     2   167.40 195.40
+ slope:thalach   2   167.85 195.85
+ cp:ca           3   166.06 196.06
+ cp:oldpeak      3   166.19 196.19
+ slope:chol      2   168.39 196.39
+ sex:slope       2   168.48 196.48
+ ca:slope        2   168.50 196.50
+ cp:sex          3   166.82 196.82
+ cp:thalach      3   167.76 197.76
+ cp:slope        6   162.28 198.28
```

Step 2.3

```
Step:  AIC=190.89
target ~ cp + oldpeak + ca + sex + sl
ope + exang + chol + thalach +
    cp:exang

                 Df Deviance    AIC
+ oldpeak:exang   1   157.19 189.19
+ sex:thalach     1   158.09 190.09
<none>                160.89 190.89
+ exang:chol      1   159.08 191.08
+ ca:exang        1   159.50 191.50
+ oldpeak:slope   2   157.51 191.51
+ sex:exang       1   159.59 191.59
+ chol:thalach    1   159.64 191.64
+ ca:thalach      1   159.74 191.74
+ ca:chol         1   159.83 191.83
+ oldpeak:thalach 1   159.84 191.84
+ sex:chol        1   160.01 192.01
+ oldpeak:sex     1   160.11 192.11
+ oldpeak:chol    1   160.27 192.27
+ exang:thalach   1   160.46 192.46
+ cp:ca           3   156.50 192.50
+ oldpeak:ca      1   160.69 192.69
+ cp:chol         3   156.73 192.73
+ ca:sex          1   160.83 192.83
+ slope:exang     2   159.27 193.27
+ slope:thalach   2   159.51 193.51
+ cp:thalach      3   158.44 194.44
+ cp:sex          3   158.44 194.44
+ sex:slope       2   160.56 194.56
+ slope:chol      2   160.74 194.74
+ cp:slope        6   152.77 194.77
+ ca:slope        2   160.84 194.84
+ cp:oldpeak      3   159.68 195.68
```

Step 2.4

```
Step:  AIC=189.19
target ~ cp + oldpeak + ca + sex + sl
ope + exang + chol + thalach +
    cp:exang + oldpeak:exang

                 Df Deviance    AIC
+ ca:exang        1   153.68 187.68
+ sex:thalach     1   154.56 188.56
<none>                157.19 189.19
+ sex:exang       1   155.91 189.91
+ ca:thalach      1   156.10 190.10
+ oldpeak:thalach 1   156.13 190.13
+ oldpeak:sex     1   156.23 190.23
+ exang:chol      1   156.25 190.25
+ oldpeak:chol    1   156.28 190.28
+ chol:thalach    1   156.32 190.32
+ ca:chol         1   156.37 190.37
+ sex:chol        1   156.52 190.52
```

Step 2.5

```
Step:  AIC=187.68
target ~ cp + oldpeak + ca + sex + sl
ope + exang + chol + thalach +
    cp:exang + oldpeak:exang + ca:exa
ng

                 Df Deviance    AIC
+ sex:exang       1   149.34 185.34
<none>                153.68 187.68
+ sex:thalach     1   151.69 187.69
+ exang:chol      1   152.20 188.20
+ oldpeak:sex     1   152.25 188.25
+ ca:thalach      1   152.29 188.29
+ oldpeak:chol    1   152.71 188.71
+ oldpeak:thalach 1   152.76 188.76
+ ca:chol         1   152.90 188.90
+ chol:thalach    1   152.93 188.93
+ sex:chol        1   153.16 189.16
```

```
+ oldpeak:slope    2   154.69 190.69
+ exang:thalach    1   156.77 190.77
+ cp:ca            3   152.94 190.94
+ cp:chol          3   152.98 190.98
+ ca:sex           1   157.14 191.14
+ oldpeak:ca       1   157.19 191.19
+ slope:thalach    2   155.59 191.59
+ cp:oldpeak       3   154.25 192.25
+ cp:sex           3   154.55 192.55
+ cp:thalach       3   154.60 192.60
+ sex:slope        2   156.66 192.66
+ cp:slope         6   148.67 192.67
+ slope:exang      2   156.92 192.92
+ ca:slope         2   157.05 193.05
+ slope:chol       2   157.15 193.15
```

```
+ oldpeak:slope    2   151.27 189.27
+ exang:thalach    1   153.46 189.46
+ cp:chol          3   149.58 189.58
+ oldpeak:ca       1   153.66 189.66
+ ca:sex           1   153.66 189.66
+ cp:ca            3   149.73 189.73
+ slope:exang      2   152.28 190.28
+ slope:thalach    2   152.41 190.41
+ cp:oldpeak       3   150.42 190.42
+ cp:sex           3   150.77 190.77
+ cp:thalach       3   150.86 190.86
+ sex:slope        2   153.12 191.12
+ ca:slope         2   153.60 191.60
+ slope:chol       2   153.61 191.61
+ cp:slope         6   145.83 191.83
```

Step 2.6

```
Step:  AIC=185.34
target ~ cp + oldpeak + ca + sex + sl
ope + exang + chol + thalach +
    cp:exang + oldpeak:exang + ca:exa
ng + sex:exang

                  Df Deviance    AIC
+ sex:thalach      1   146.66 184.66
<none>                 149.34 185.34
+ cp:ca            3   144.12 186.12
+ oldpeak:slope    2   146.13 186.13
+ ca:thalach       1   148.14 186.14
+ sex:chol         1   148.31 186.31
+ oldpeak:sex      1   148.39 186.39
+ exang:chol       1   148.41 186.41
+ oldpeak:chol     1   148.45 186.45
+ chol:thalach     1   148.53 186.53
+ ca:chol          1   148.77 186.77
+ oldpeak:thalach  1   148.85 186.85
+ cp:slope         6   139.01 187.01
+ ca:sex           1   149.09 187.09
+ oldpeak:ca       1   149.29 187.29
+ exang:thalach    1   149.30 187.30
+ cp:chol          3   145.48 187.48
+ cp:sex           3   145.70 187.70
+ slope:thalach    2   148.11 188.11
+ cp:thalach       3   146.17 188.17
+ cp:oldpeak       3   146.54 188.54
+ sex:slope        2   148.72 188.72
+ slope:exang      2   149.02 189.02
+ ca:slope         2   149.08 189.08
+ slope:chol       2   149.08 189.08
```

Step 2.7

```
Step:  AIC=184.66
target ~ cp + oldpeak + ca + sex + sl
ope + exang + chol + thalach +
    cp:exang + oldpeak:exang + ca:exa
ng + sex:exang + sex:thalach

                  Df Deviance    AIC
<none>                 146.66 184.66
+ oldpeak:slope    2   143.33 185.33
+ chol:thalach     1   145.39 185.39
+ exang:chol       1   145.78 185.78
+ sex:chol         1   145.84 185.84
+ oldpeak:chol     1   145.97 185.97
+ ca:sex           1   146.00 186.00
+ oldpeak:sex      1   146.01 186.01
+ oldpeak:thalach  1   146.21 186.21
+ cp:ca            3   142.24 186.24
+ ca:thalach       1   146.27 186.27
+ ca:chol          1   146.33 186.33
+ oldpeak:ca       1   146.56 186.56
+ exang:thalach    1   146.62 186.62
+ cp:chol          3   143.01 187.01
+ cp:slope         6   137.25 187.25
+ cp:oldpeak       3   143.68 187.68
+ cp:sex           3   143.71 187.71
+ slope:thalach    2   145.87 187.87
+ slope:exang      2   146.13 188.13
+ ca:slope         2   146.37 188.37
+ slope:chol       2   146.47 188.47
+ sex:slope        2   146.64 188.64
+ cp:thalach       3   144.78 188.78
```