Final Project Report

Group 12

Student 1: Saketh V V

857-869-1728

venkata.s@husky.neu.edu

Student 2: Pranali Bhosale

857-800-4778

Bhosale.pr@husky.neu.edu

Percentage of Effort Contributed by Saketh: 50%

Percentage of Effort Contributed by Pranali: 50%

Signature of Saketh: saketh vommina venkata

Signature of Pranali:

Submission Date: 04/21/2019

TABLE OF CONTENTS

PROBLEM SETTING:

The American Heart Association recently published the 'Heart Disease and Stroke Statistics 2018', which states that about 1 out of every 3 deaths in the US is caused by Cardiovascular disease. According to this statistic report, approximately 2,300 Americans die of Cardiovascular disease each day, which means an average of 1 death every 38 seconds. The key lies in identifying the problem before it becomes fatal. The diagnosis of heart disease is a challenging task and is usually based on signs, symptoms and physical exam of the patient. Numerous factors such as cholesterol level, smoking habits, obesity, family history, blood pressure aggravate the risk of heart disease. These factors need to be examined by a Data Miner to find a relationship between them and the possibility of developing a heart disease.

PROJECT TOPIC:

The topic of this project is: **Prediction of Heart disease using the Cleveland Hospital dataset**

The Healthcare industry collects a deluge of patient data that contains hidden information, which is useful for making effective decisions. Vital information such as age, medical history, blood sugar levels, electrocardiographic results and so on can aid in predicting the presence of heart disease in patients. This will help us save thousands of lives and significantly reduce cost of health care services and medications.

AIMS & OBJECTIVES:

The objective of this project is to build a predictive model to detect the presence of heart disease in a patient. The focus will be distinguishing the presence value (1) from the absence value (0) of a heart disease. Exploratory Data Analysis of the impact of variables on a healthy heart versus an unhealthy one will be performed. Furthermore, this model will conduct performance evaluation to check for False positives.

DATA GATHERING:

The data for this project was collected from the following sources:

- Cleveland dataset: *https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data*

- Kaggle dataset for Heart Attack prediction: *https://www.kaggle.com/imnikhilanand/heart-attack-prediction*

Two datasets are being combined for this project due to limited availability of data. Both the datasets contain 14 variables. Confidential patient information such as names and social security numbers have been replaced with dummy variables.

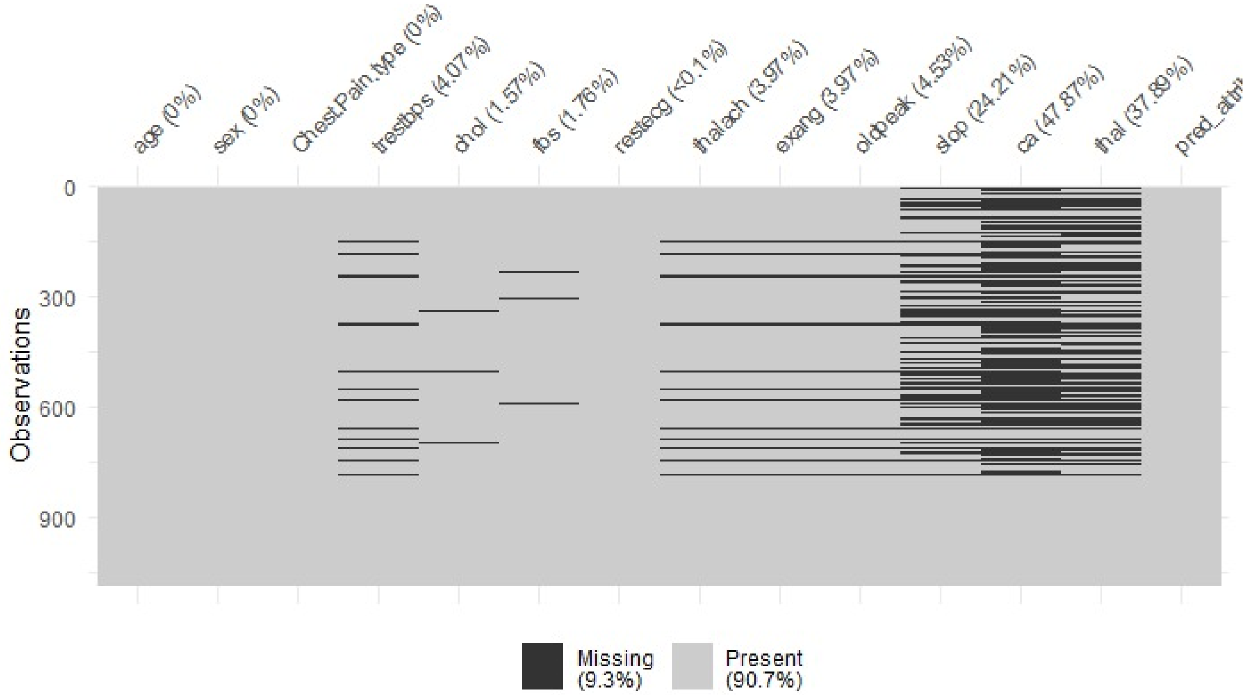
The 14 variables being used are as follows:

1. **age**: age in years
2. **sex**: 1 = male; 0 = female
3. **cp**: chest pain type - Value 1: typical angina; Value 2: atypical angina; Value 3: non-anginal pain; Value 4: asymptomatic
4. **trestbps**: resting blood pressure (in mm Hg on admission to the hospital)
5. **chol**: serum cholesterol in mg/dl
6. **fbs**: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. **restecg**: resting electrocardiographic results - Value 0: normal; Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. **thalach**: maximum heart rate achieved
9. **exang**: exercise induced angina (1 = yes; 0 = no)
10. **oldpeak**: ST depression induced by exercise relative to rest
11. **slope**: the slope of the peak exercise ST segment
12. **ca**: number of major vessels (0-3) colored by fluoroscopy
13. **thal**: 3 = normal; 6 = fixed defect; 7 = reversable defect

14. **num**: the predicted attribute

DATA EXPLORATION AND VISUALIZATION:

1. Missing Values:

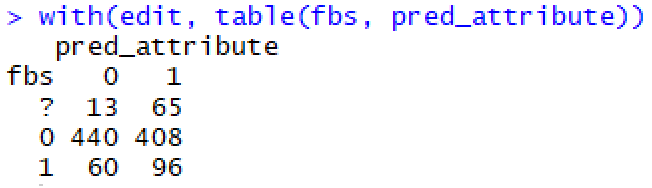


After analyzing the extracted datasets and the graph above, it was observed that there were missing values in 10 of the variables. These are listed as follows:

- **Categorical**
  - slop
  - ca
  - thal
  - fbs
  - exang
  - restecg

- **Continuous**
  - trestbps
  - chol
  - thalach
  - oldpeak

To handle the missing values in the categorical variables, first two-way tables were constructed to analyze the relationship between the variable and the target.



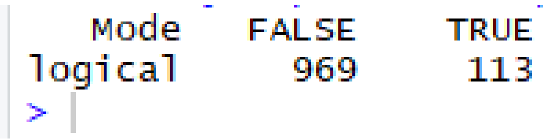
```
> with(edit, table(fbs, pred_attribute))
   pred_attribute
fbs   0   1
  ?  13  65
  0 440 408
  1  60  96
```

Hence, the missing values in the “fbs” column whose target value was 0 were replaced with 0’s since 0 had the most values (440). Missing values whose target value was 1 were replaced again with 0’s since that was the highest (408).

Similar imputation was performed for all the other categorical variables who had missing values.

To handle missing values in continuous variables, the R package MICE was then used. The MICE package implements a method – “PMM (Predictive Mean Matching)” to deal with the missing data. The package creates multiple imputations for multivariate missing data. The method is based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model. The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data. In addition, MICE can impute continuous two-level data, and maintain consistency between imputations by means of passive imputation. Many diagnostic plots are implemented to inspect the quality of the imputations. Checking to see if there are any NaN or invalid entries left to be dealt with using the is.null function in R proved negative. Thus, all the entries were valid.

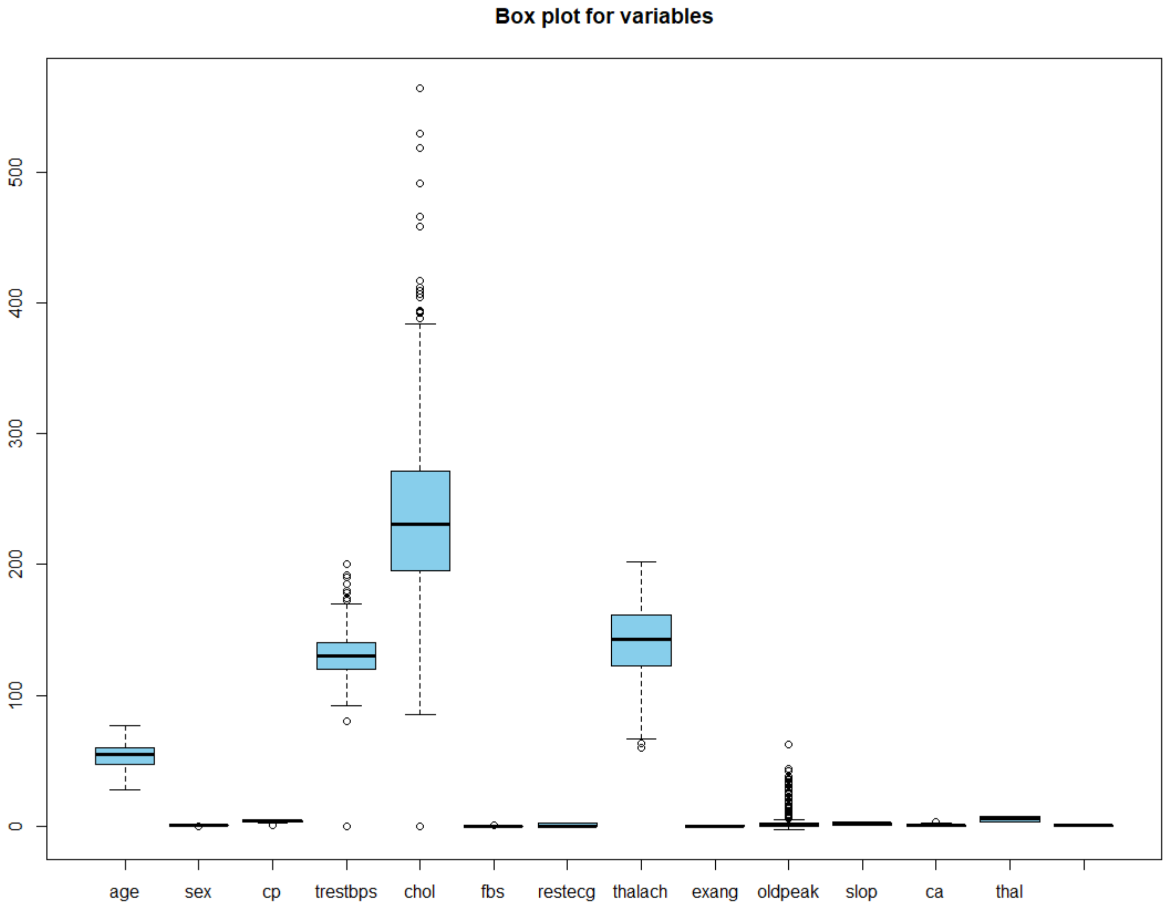
2. Checking for Duplicate records:



```
   Mode   FALSE    TRUE
logical     969     113
>
```

On further analysis of the data, it was found that there are 113 duplicate records in the dataset. Since the records don’t have any unique identifiers such as “name” or “patient ID”, we can’t know for sure if they are duplicate or just two patients with the same medical history. Therefore, we have decided to keep the records as it is for further analysis.

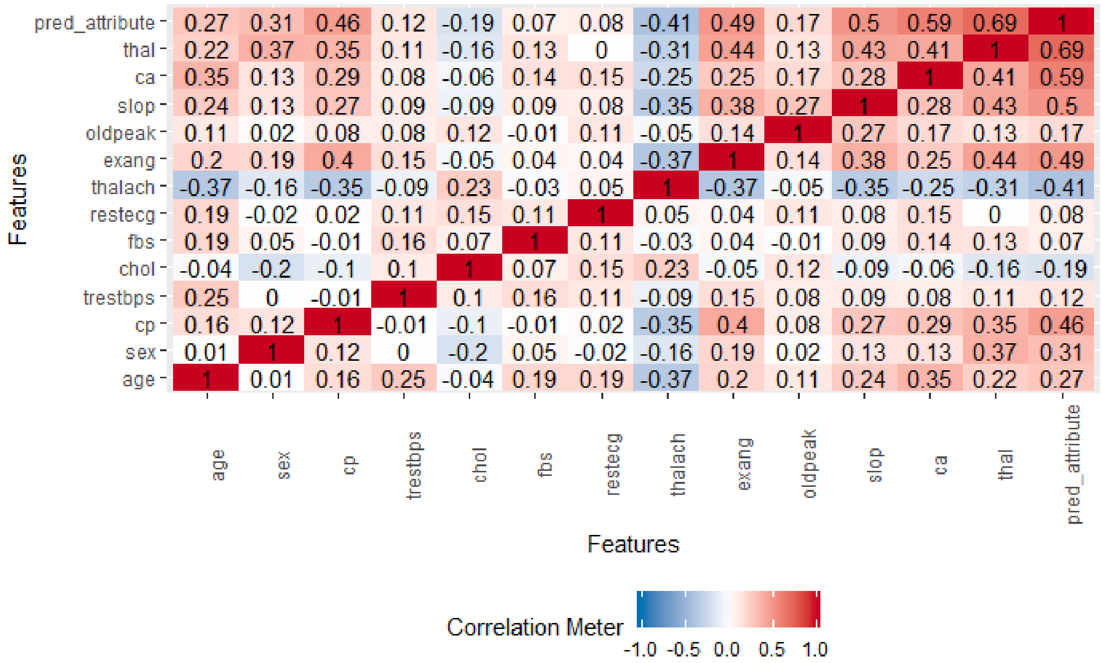
3. Checking for Outliers



The above graph is a box-plot for all variables. It is observed here that trestbps, chol, oldpeak have some outliers. There are various options when dealing with outliers such as dropping the outlier record, capping the outlier data, assigning a new value and so on. However, if we drop the outliers we will lose the data for the rest of the variables in that record. Altering the outlier data will affect the correlation between the variables and the target variable. It might even alter the final analysis as a result.

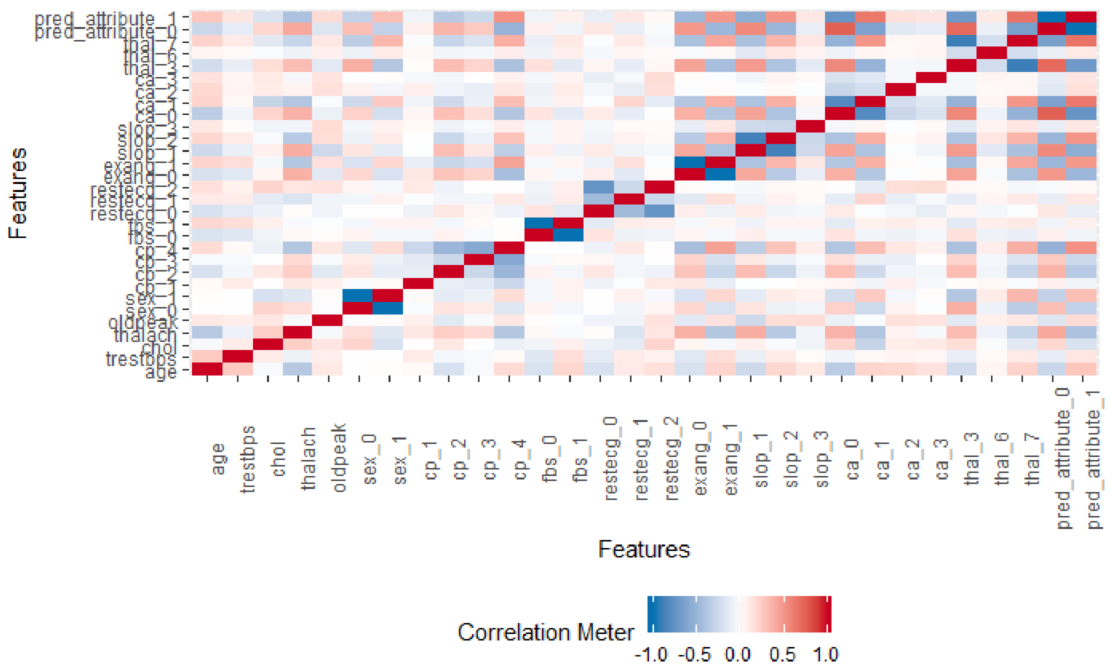
After research, some domain knowledge was obtained about the variables with outliers. The average cholesterol levels in an adult is in the range of 200 – 240 mg/dl. Some of the outliers shown in the box plot have the values of below 200 and above 400. Though these are rare, they are possible in the medical world. Therefore, we have decided to keep the outlier points as such.

4. Pearson Correlation Coefficient:



Features	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slop	ca	thal	pred_attribute
pred_attribute	0.27	0.31	0.46	0.12	-0.19	0.07	0.08	-0.41	0.49	0.17	0.5	0.59	0.69	1
thal	0.22	0.37	0.35	0.11	-0.16	0.13	0	-0.31	0.44	0.13	0.43	0.41	1	0.69
ca	0.35	0.13	0.29	0.08	-0.06	0.14	0.15	-0.25	0.25	0.17	0.28	1	0.41	0.59
slop	0.24	0.13	0.27	0.09	-0.09	0.09	0.08	-0.35	0.38	0.27	1	0.28	0.43	0.5
oldpeak	0.11	0.02	0.08	0.08	0.12	-0.01	0.11	-0.05	0.14	1	0.27	0.17	0.13	0.17
exang	0.2	0.19	0.4	0.15	-0.05	0.04	0.04	-0.37	1	0.14	0.38	0.25	0.44	0.49
thalach	-0.37	-0.16	-0.35	-0.09	0.23	-0.03	0.05	1	-0.37	-0.05	-0.35	-0.25	-0.31	-0.41
restecg	0.19	-0.02	0.02	0.11	0.15	0.11	1	0.05	0.04	0.11	0.08	0.15	0	0.08
fbs	0.19	0.05	-0.01	0.16	0.07	1	0.11	-0.03	0.04	-0.01	0.09	0.14	0.13	0.07
chol	-0.04	-0.2	-0.1	0.1	1	0.07	0.15	0.23	-0.05	0.12	-0.09	-0.06	-0.16	-0.19
trestbps	0.25	0	-0.01	1	0.1	0.16	0.11	-0.09	0.15	0.08	0.09	0.08	0.11	0.12
cp	0.16	0.12	1	-0.01	-0.1	-0.01	0.02	-0.35	0.4	0.08	0.27	0.29	0.35	0.46
sex	0.01	1	0.12	0	-0.2	0.05	-0.02	-0.16	0.19	0.02	0.13	0.13	0.37	0.31
age	1	0.01	0.16	0.25	-0.04	0.19	0.19	-0.37	0.2	0.11	0.24	0.35	0.22	0.27

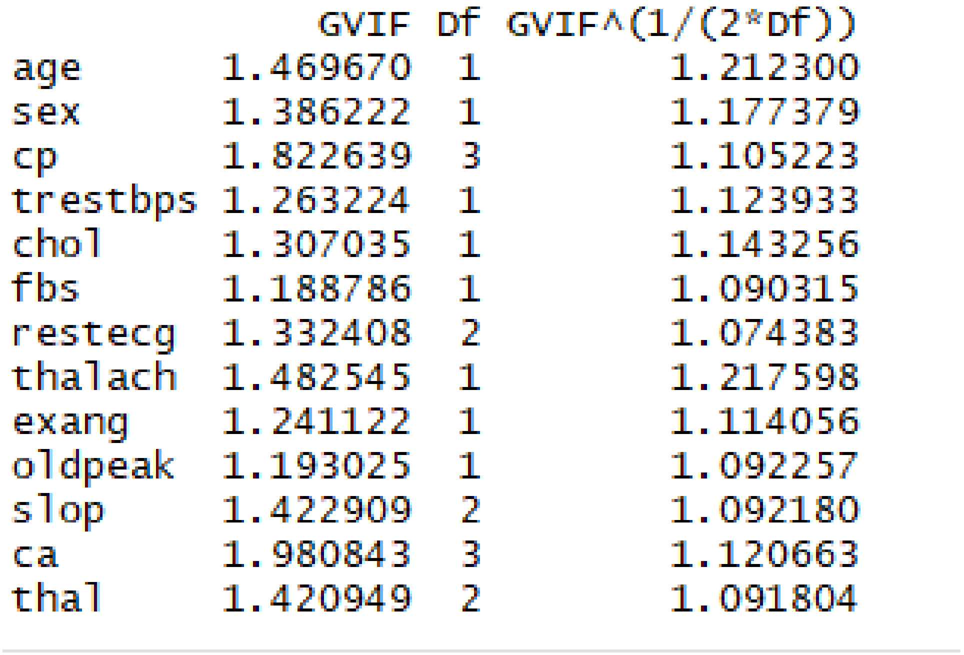
The Pearson Correlation graph clearly shows that there are no strongly correlated variables. This is a good sign since entering highly correlated variables into the model causes data redundancy.



The heat map above is a more detailed graph. It displays the correlation for each of the factor of the variables. We can see from the first heat map that the variable "thal" has high correlation with the "pred_attribute". However, when split into its factors – that_3, thal_6, thal_7 we see that thal 3 is highly negatively correlated with pred_attribute 1 and highly positively correlated with pred_attribute 0. Conversely, that_7 is highly positively correlated with pred_attribute 1 and highly negatively correlated with pred_attribute 0. The variable that_6 is not correlated with the target variable at all.

Once the model is applied, the performance is going to be measured and compared if it can be improved by applying the models on the data set with the above shown dummy variables i.e. dataset where the variables are partitioned into its different levels.

5. Test for Multicollinearity:

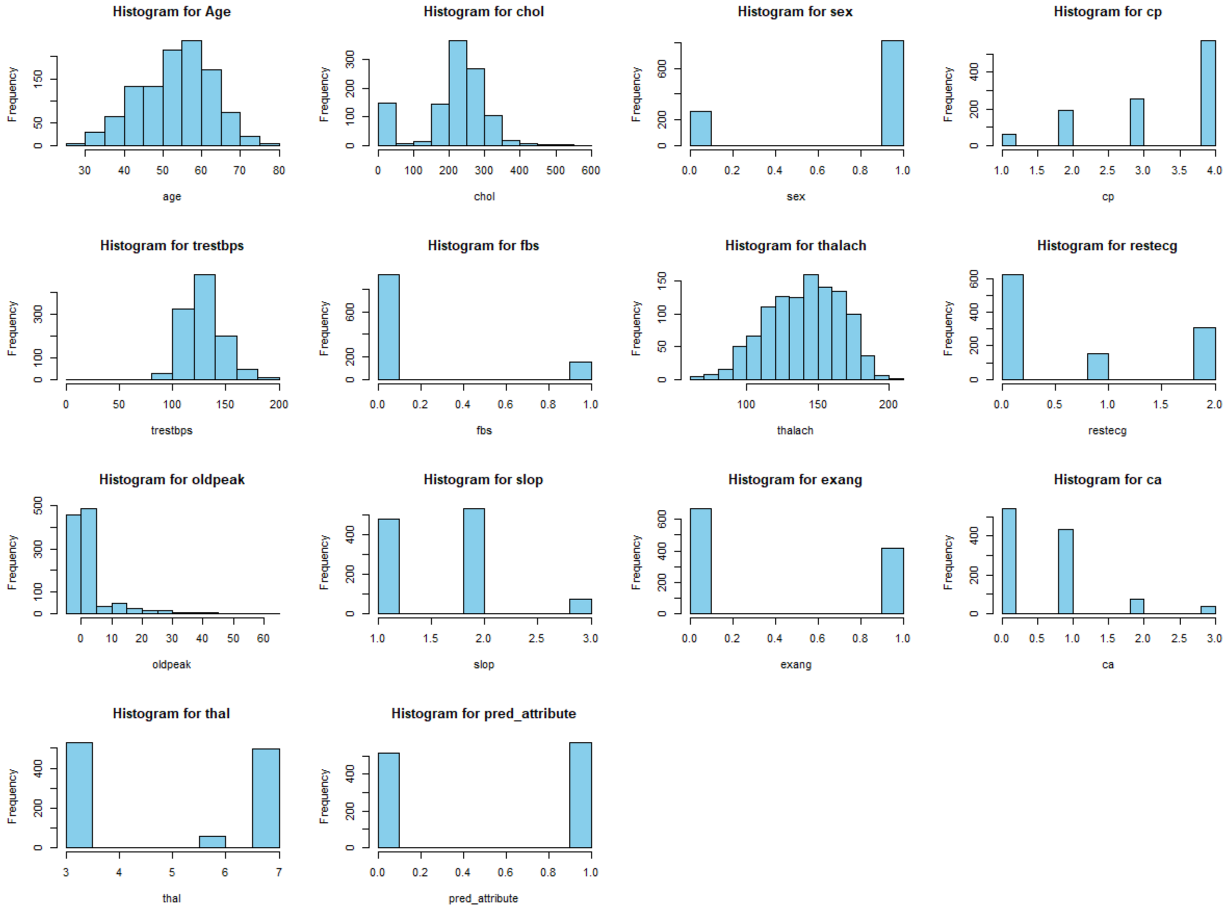


```
             GVIF Df GVIF^(1/(2*Df))
age      1.469670  1        1.212300
sex      1.386222  1        1.177379
cp       1.822639  3        1.105223
trestbps 1.263224  1        1.123933
chol     1.307035  1        1.143256
fbs      1.188786  1        1.090315
restecg  1.332408  2        1.074383
thalach  1.482545  1        1.217598
exang    1.241122  1        1.114056
oldpeak  1.193025  1        1.092257
slop     1.422909  2        1.092180
ca       1.980843  3        1.120663
thal     1.420949  2        1.091804
```

The study uses Variance Inflation Factor (VIF), the most widely used diagnostic to study the multicollinearity. The VIF estimates the how much the variance of a coefficient is “inflated” because of linear dependence with other predictors. For example, the VIF of “age” can be interpreted as, variance of age is 46% larger than it would be if it was completely uncorrelated with all the other predictors. The VIF has a lower bound of 1 and no upper bound.

The results here clearly show that the values for all the variables are between 1-2 which means that the collinearity existing within these predictors are acceptable.

6. Histograms for all Explanatory variables:

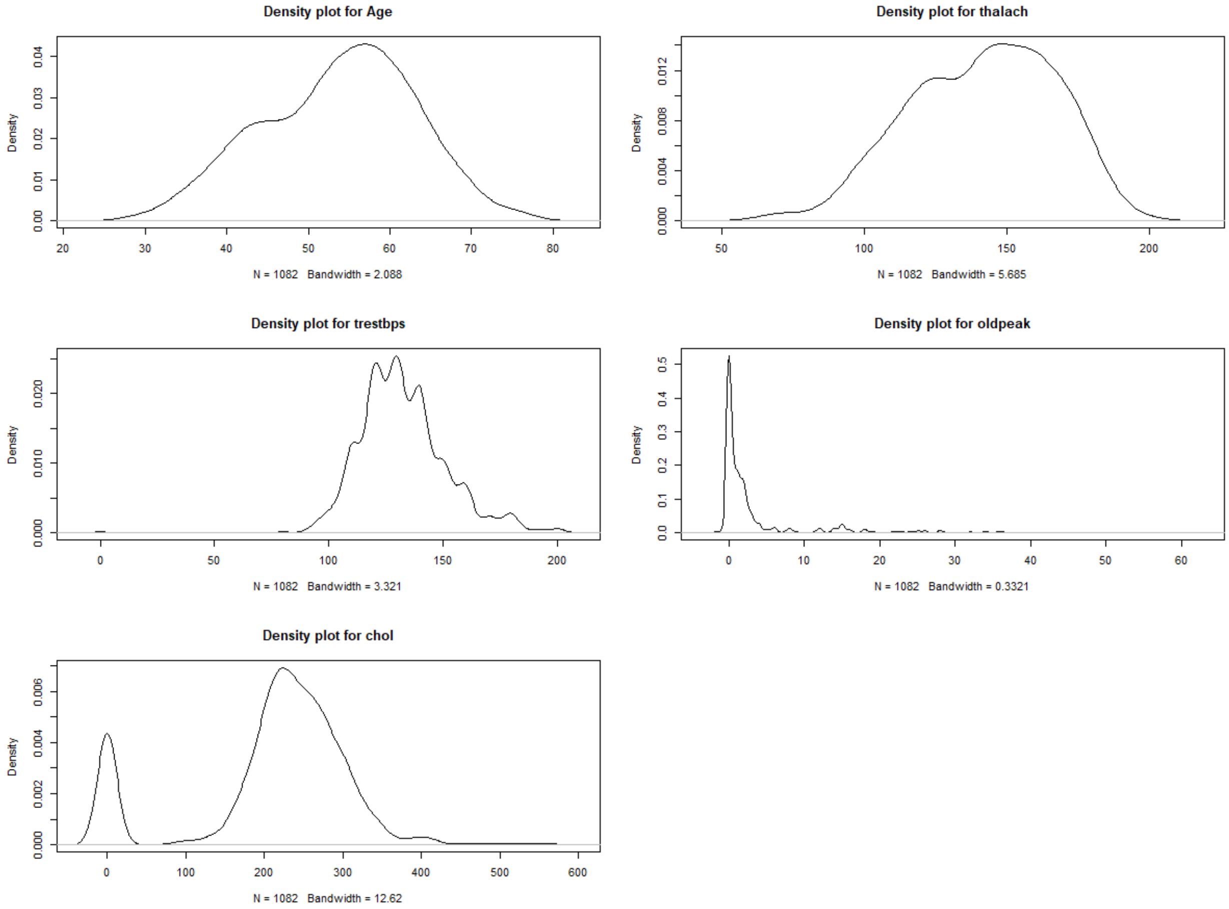


The above graphs clearly show the histograms of all variables in this dataset. These histograms display the entire distribution of the numerical variables.

The histogram for “age” shows that the number of records is the highest between the ages of 50-70. Similarly, the frequency is the most for the range of 200-300 “chol”. The histogram for “sex” has only 2 values as expected as this is a categorical variable which stands for “male/female”.

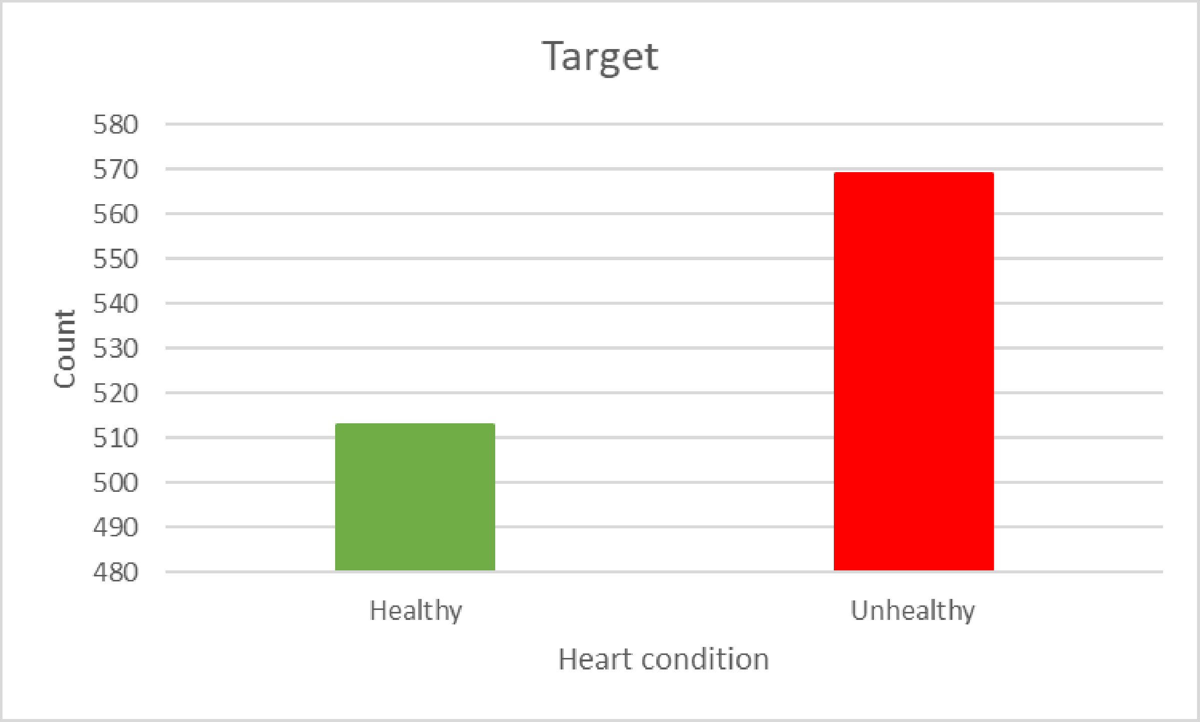
Similarly, for the categorical variable’s “cp”, “fbs”, “restecg”, “slop”, “exang”, “ca”, “thal” and “pred_attribute” are all categorical which is why we have the peaks at their corresponding values with gaps between the peaks.

7. Density Plots for all continuous variables:



The above graphs clearly illustrate that variables - cholesterol, age, and maximum heart rate (i.e. thalach) follow a normal distribution. Oldpeak variable seems to be left-skewed, but since this is a categorical variable the skewedness is expected and hence we need not do any transformation to the data such as applying log values to the data points. In the Chol variable the presence of outliers leads to two peaks.

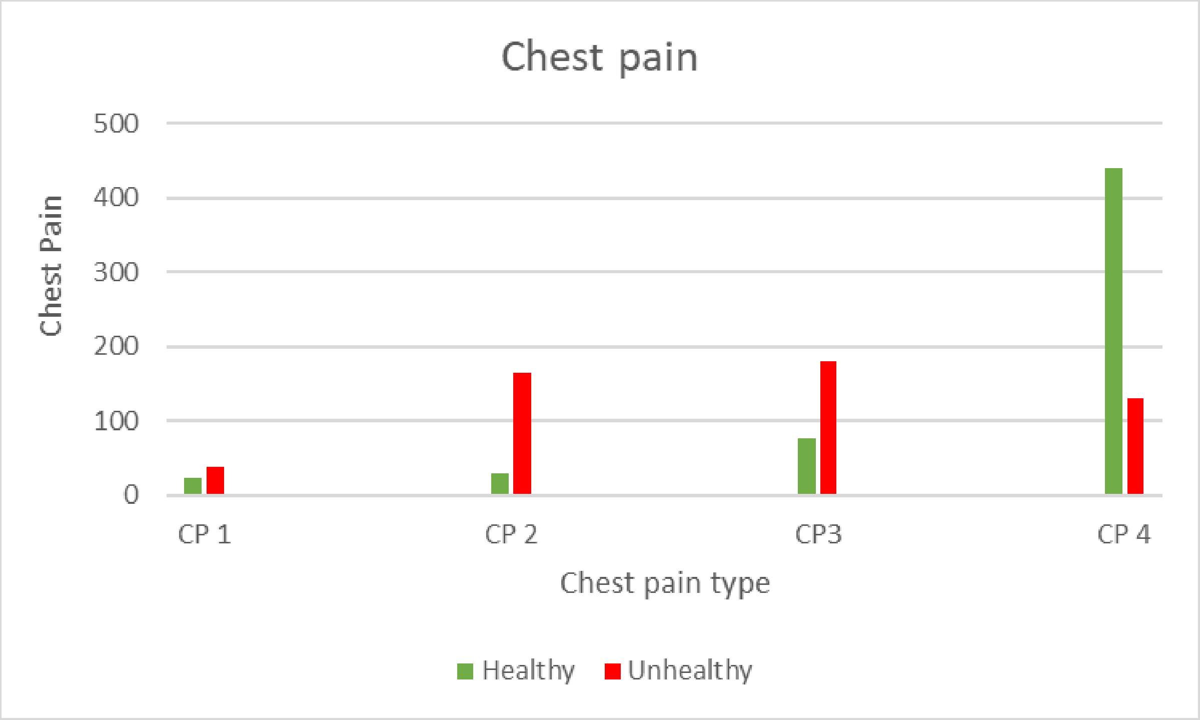
8. Count of healthy and unhealthy hearts in the target variable:



In the above graph we have charted the number of healthy and number of heart patients to get an idea of the difference between the two. We can see that there are 513 healthy hearts and 560 patients with heart disease.

The original dataset further bifurcated those affected with unhealthy heart diseases into multiple levels 1-4 all indicating a presence of an unhealthy heart. We have decided to make this project into a classification analysis project where we would identify an unhealthy heart, hence these levels have all been converted to 1’s which indicates the presence of an unhealthy heart.

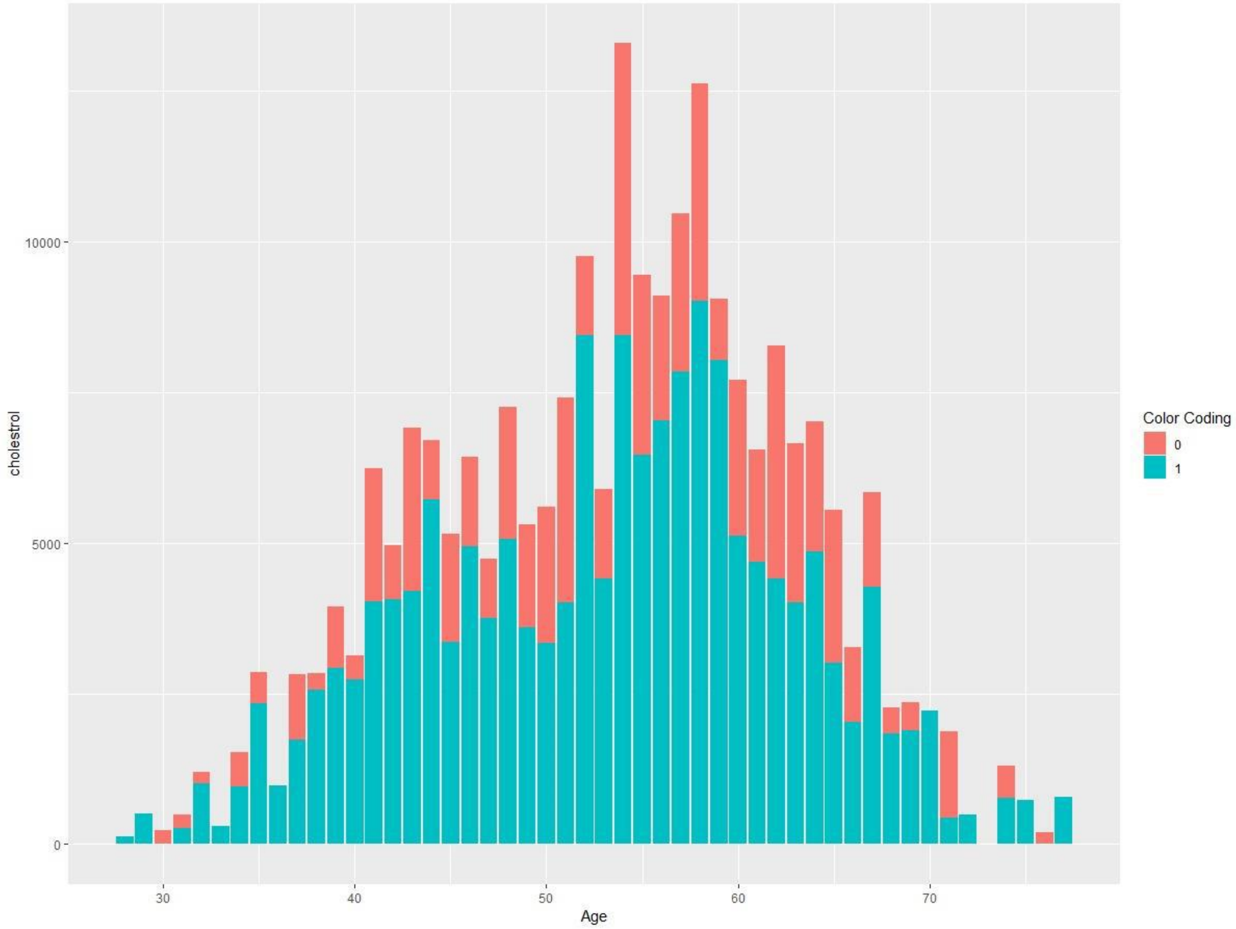
9. Chest pain in healthy and unhealthy patients by pain type:



In the above graph we have charted the number of healthy and number of heart patients who have different types of chest pain (CP 1 : Asymptomatic, CP 2: Atypical Angina , CP 3: Non-anginal pain, CP 4: Typical angina).

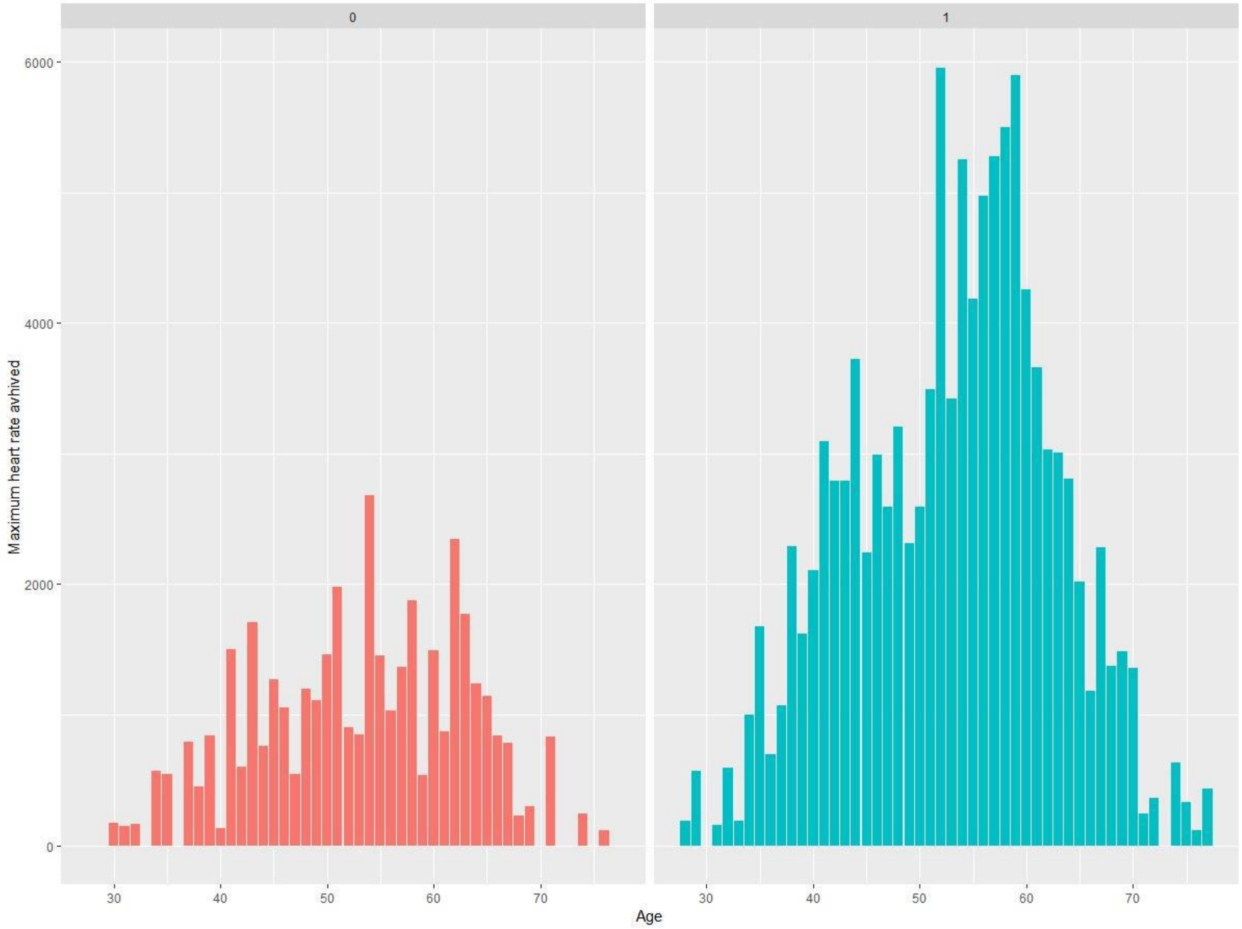
We can see that chest pain type 4 is highest in the healthy patients and chest pain type 1 the lowest. Whereas 180 patients with chest pain type 3 among the heart patients and lowest is chest pain type 1 (same as healthy heart patients).

10. Age vs Cholesterol:



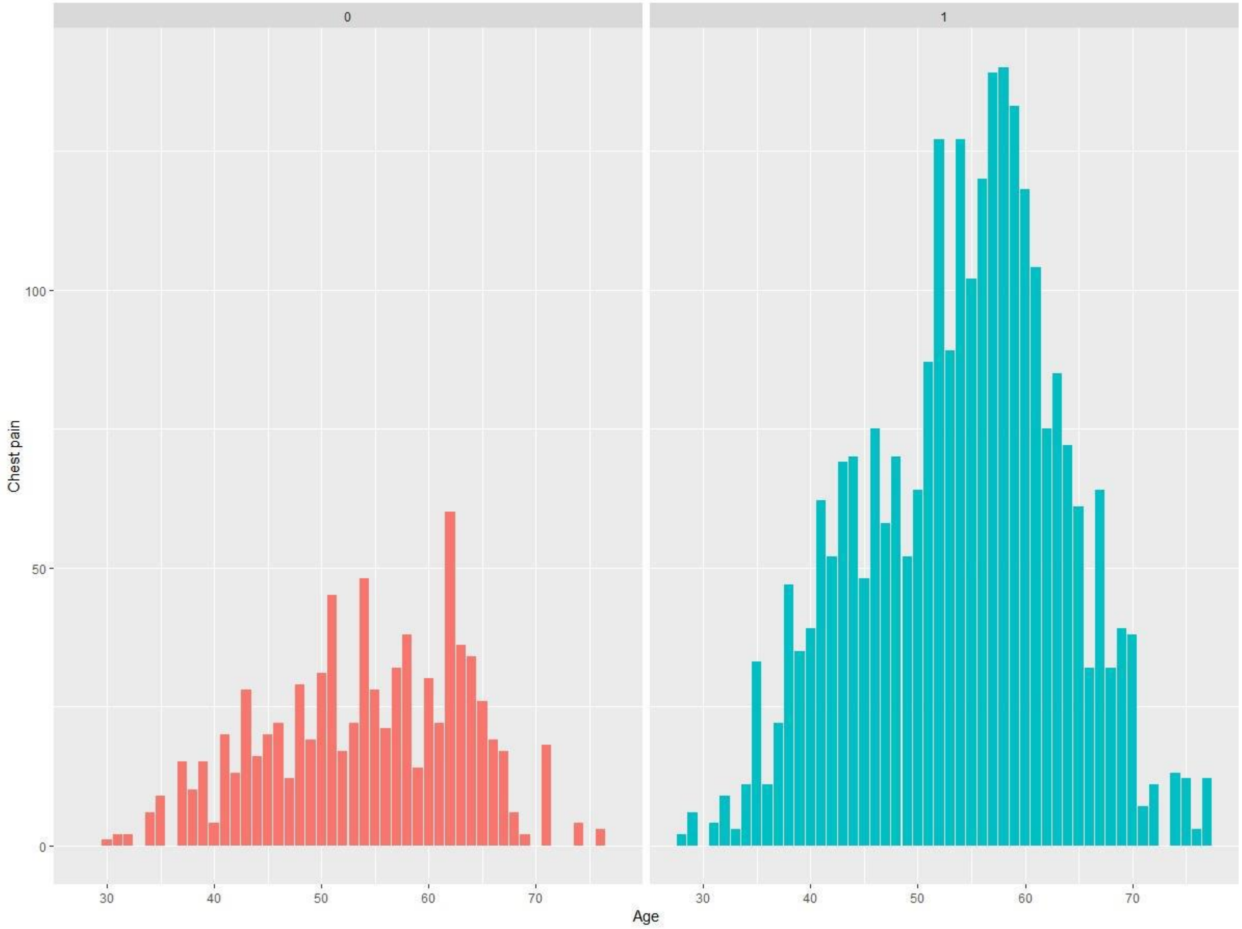
The graph above shows the comparison between age and cholesterol for males(blue) and females (pink). Some of the highest peaks are observed between the ages 50 and 60, indicating high stress levels at that age and possibility of heart disease. The graph has a bell shape.

11. Maximum heart rate achieved vs age by gender:



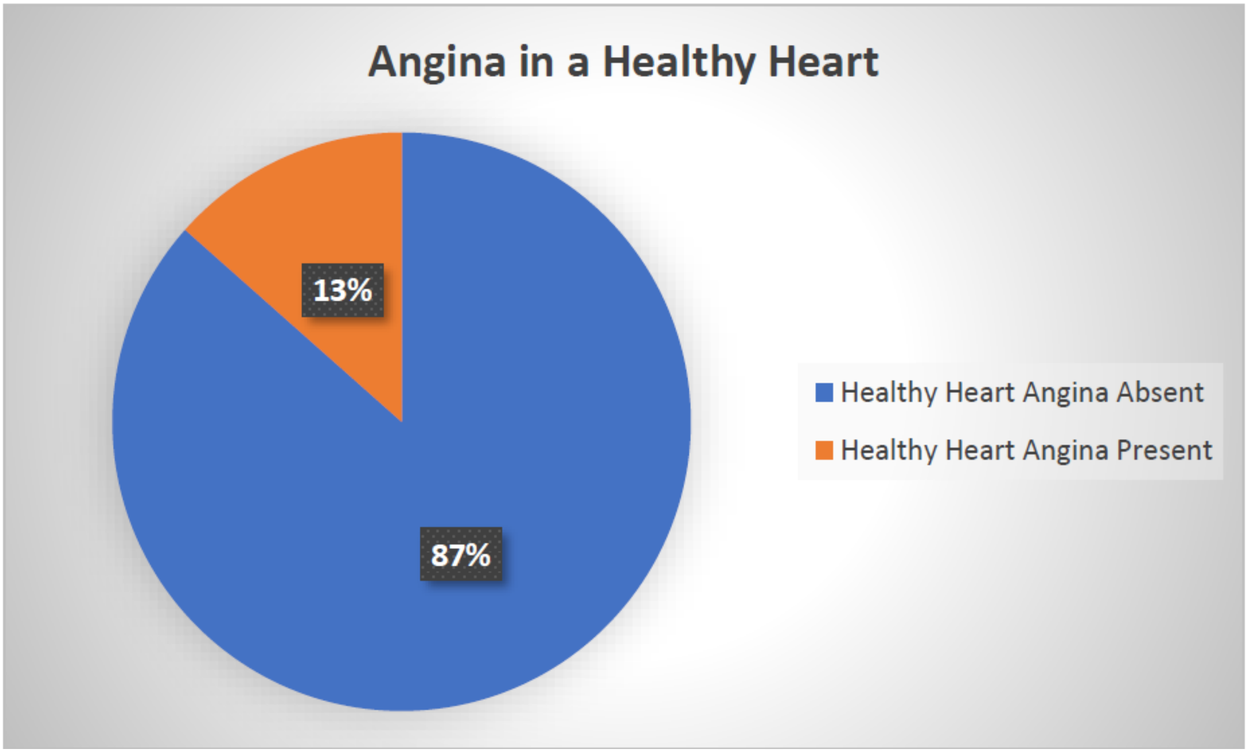
The graph above shows the comparison between age and maximum heart rate achieved for males(blue) and females (pink). Some of the highest peaks are observed between the ages 50 and 60, indicating high heart rate especially in the males. The graphs both have an approximate bell shape.

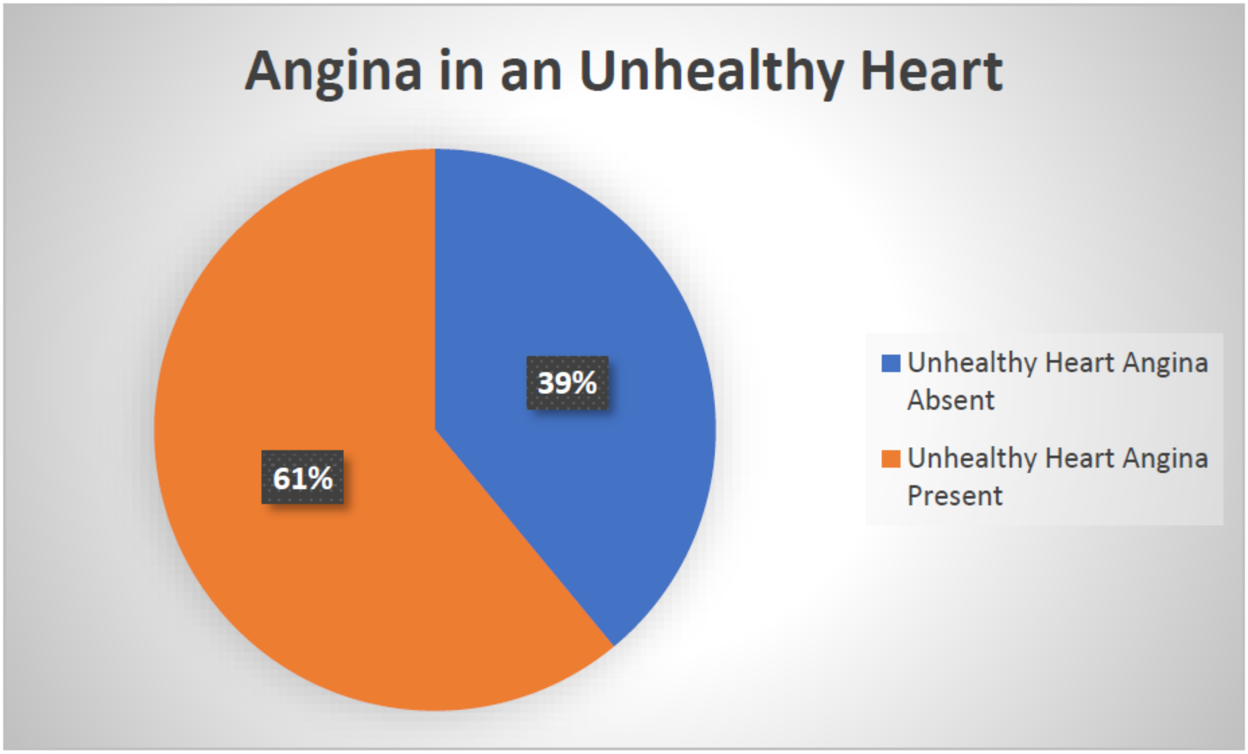
12. Chest pain with age gender wise:



The graph above shows the comparison between age and chest pain for males(blue) and females (pink). For females the highest peak occurs after the age of 60 while for males the highest peak appears to be right before 60. The graphs both have an approximate bell shape.

13. Exercise induced angina in healthy and unhealthy patients:





Angina also known as angina pectoris, a condition marked by severe pain in the chest, often also spreading to the shoulders, arms, and neck, caused by an inadequate blood supply to the heart. From the above pie charts, we can observe that the presence of angina in unhealthy hearts is almost 5 times than that of a healthy heart, therefore it clearly indicates that Angina is one of the important factors and its presence is not favored.

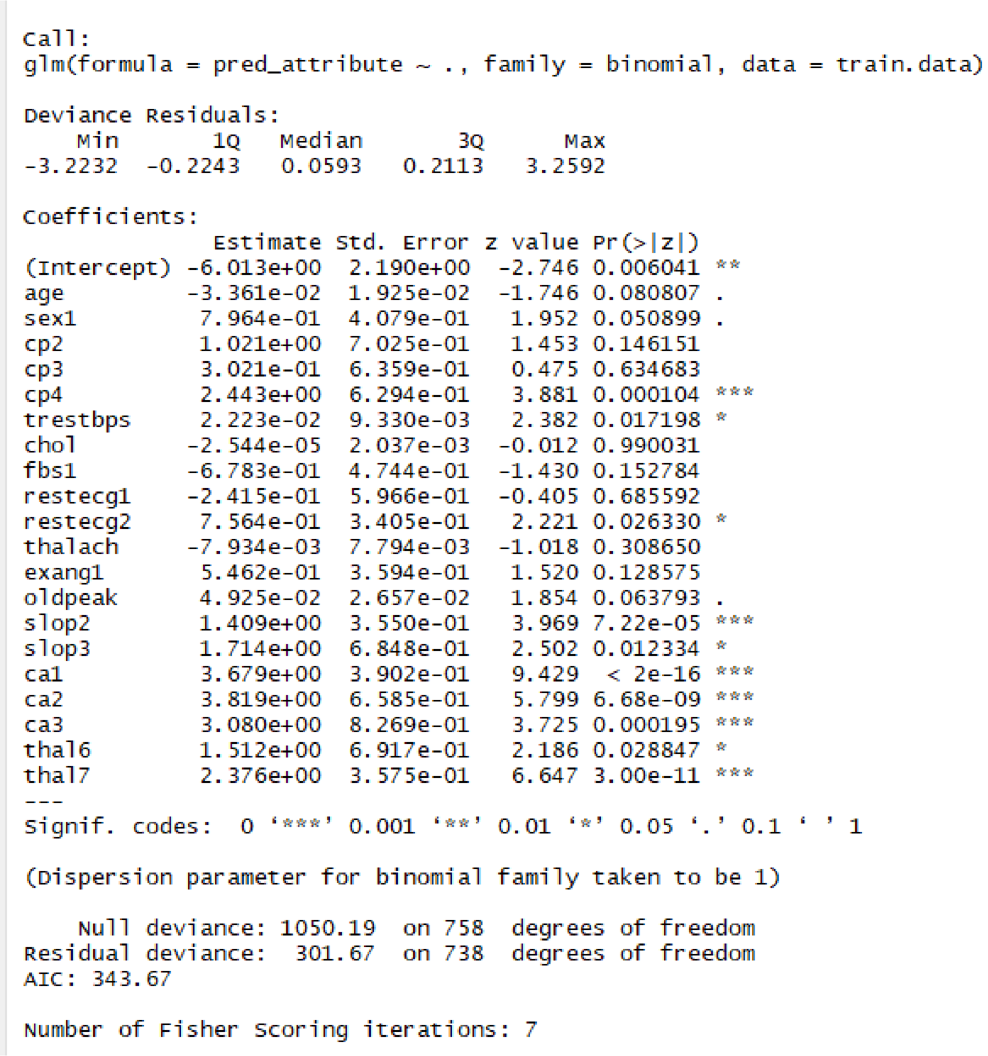
MODEL BUILDING:

Naïve's Rule: The data consists of 569 unhealthy heart records, which means that bifurcating the records according to the majority class should give an accuracy of 56%. Any model used should perform better than this.

To elevate the search for an accurate model this project implemented and compared the results of two algorithms.

1. Classification using Logistic Regression –

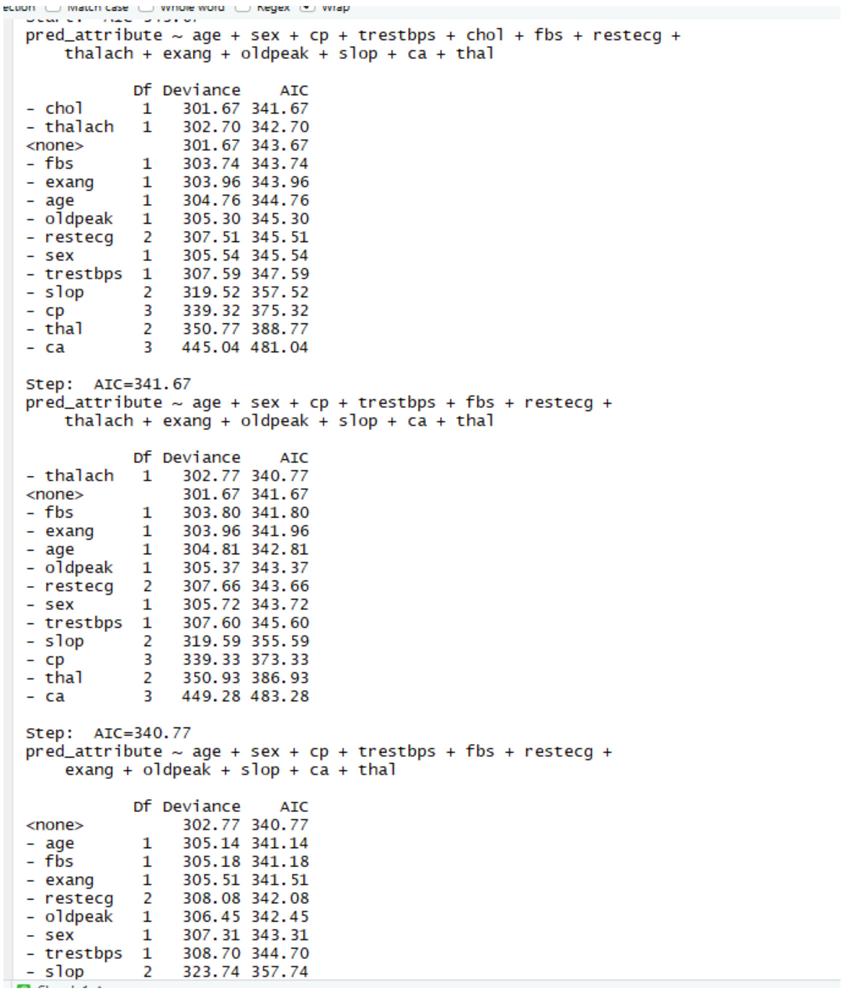
   It is one of the most popular algorithms used for classification problems. Logistic regression models the probability of the default class. The data was first divided into training and test data sets where 70% of the records were training and 30% of the records were the test data set. The model was then trained using the "gml()" function with all the predictors. The below screenshot shows the summary of the output.



```
Call:
glm(formula = pred_attribute ~ ., family = binomial, data = train.data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.2232  -0.2243   0.0593   0.2113   3.2592  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.013e+00  2.190e+00  -2.746 0.006041 ** 
age         -3.361e-02  1.925e-02  -1.746 0.080807 .  
sex1         7.964e-01  4.079e-01   1.952 0.050899 .  
cp2          1.021e+00  7.025e-01   1.453 0.146151    
cp3          3.021e-01  6.359e-01   0.475 0.634683    
cp4          2.443e+00  6.294e-01   3.881 0.000104 ***
trestbps     2.223e-02  9.330e-03   2.382 0.017198 *  
chol        -2.544e-05  2.037e-03  -0.012 0.990031    
fbs1        -6.783e-01  4.744e-01  -1.430 0.152784    
restecg1    -2.415e-01  5.966e-01  -0.405 0.685592    
restecg2     7.564e-01  3.405e-01   2.221 0.026330 *  
thalach     -7.934e-03  7.794e-03  -1.018 0.308650    
exang1       5.462e-01  3.594e-01   1.520 0.128575    
oldpeak      4.925e-02  2.657e-02   1.854 0.063793 .  
slop2        1.409e+00  3.550e-01   3.969 7.22e-05 ***
slop3        1.714e+00  6.848e-01   2.502 0.012334 *  
ca1          3.679e+00  3.902e-01   9.429  < 2e-16 ***
ca2          3.819e+00  6.585e-01   5.799 6.68e-09 ***
ca3          3.080e+00  8.269e-01   3.725 0.000195 ***
thal6        1.512e+00  6.917e-01   2.186 0.028847 *  
thal7        2.376e+00  3.575e-01   6.647 3.00e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1050.19  on 758  degrees of freedom
Residual deviance:  301.67  on 738  degrees of freedom
AIC: 343.67

Number of Fisher Scoring iterations: 7
```

The model has in built function to create dummy variables for the categorical data and perform the function. The Residual deviance as shown is much lesser than the Null Deviance which is a good indicator that the model used has good predictive usage. Further calculating the accuracy, Logistic regression with a cut off 0.5 gave an accuracy of 0.91640866873065 which is around 91.65%.

Stepwise Regression:-The study then performs Stepwise regression using Backward elimination method to check if the presence of all variables was of a good choice or if the model can be improved further by removing any predictor. The output of this is shown below:



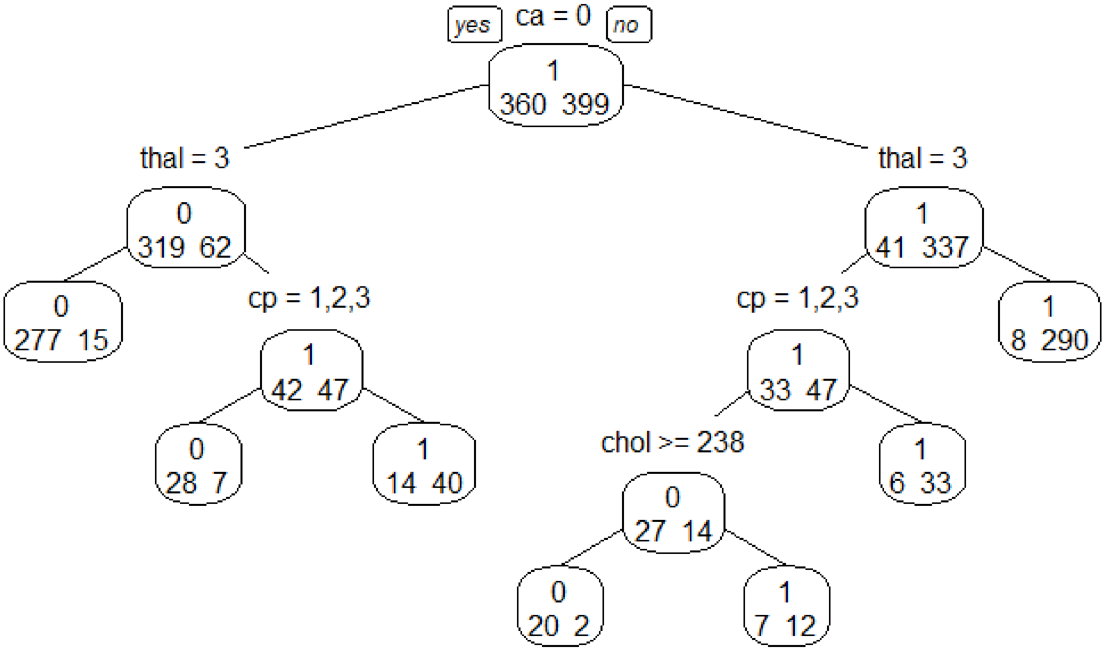
```
Start:  AIC=343.67
pred_attribute ~ age + sex + cp + trestbps + chol + fbs + restecg + 
    thalach + exang + oldpeak + slop + ca + thal

           Df Deviance    AIC
- chol      1   301.67 341.67
- thalach   1   302.70 342.70
<none>          301.67 343.67
- fbs       1   303.74 343.74
- exang     1   303.96 343.96
- age       1   304.76 344.76
- oldpeak   1   305.30 345.30
- restecg   2   307.51 345.51
- sex       1   305.54 345.54
- trestbps  1   307.59 347.59
- slop      2   319.52 357.52
- cp        3   339.32 375.32
- thal      2   350.77 388.77
- ca        3   445.04 481.04

Step:  AIC=341.67
pred_attribute ~ age + sex + cp + trestbps + fbs + restecg + 
    thalach + exang + oldpeak + slop + ca + thal

           Df Deviance    AIC
- thalach   1   302.77 340.77
<none>          301.67 341.67
- fbs       1   303.80 341.80
- exang     1   303.96 341.96
- age       1   304.81 342.81
- oldpeak   1   305.37 343.37
- restecg   2   307.66 343.66
- sex       1   305.72 343.72
- trestbps  1   307.60 345.60
- slop      2   319.59 355.59
- cp        3   339.33 373.33
- thal      2   350.93 386.93
- ca        3   449.28 483.28

Step:  AIC=340.77
pred_attribute ~ age + sex + cp + trestbps + fbs + restecg + 
    exang + oldpeak + slop + ca + thal

           Df Deviance    AIC
<none>          302.77 340.77
- age       1   305.14 341.14
- fbs       1   305.18 341.18
- exang     1   305.51 341.51
- restecg   2   308.08 342.08
- oldpeak   1   306.45 342.45
- sex       1   307.31 343.31
- trestbps  1   308.70 344.70
- slop      2   323.74 357.74
```

The output showed that, the model would perform better by removing the variables “chol” and “thalach”. These variables were then removed, and the model was trained once again. This gave an accuracy close to 91.95% which is not much of an improvement.

Cross Validation:- This study also performs cross validation primarily for two reasons. Since the number of records are only 1000, in order to create multiple sets of training and tests data sets cross validation has been used. Secondly, it has been used to check for the accuracy. On performing cross validation, we achieved a accuracy of 91.64% with all the predictors, which also proves that the accuracy achieved by the initial model was indeed a good model.

2. Classification using Decision Trees –

A Classification model is built in the form of a tree structure using Decision trees. The dataset is divided into smaller and smaller subsets while incrementing an associated decision tree. Decision trees can provide clear results as to which fields are most important for classification.



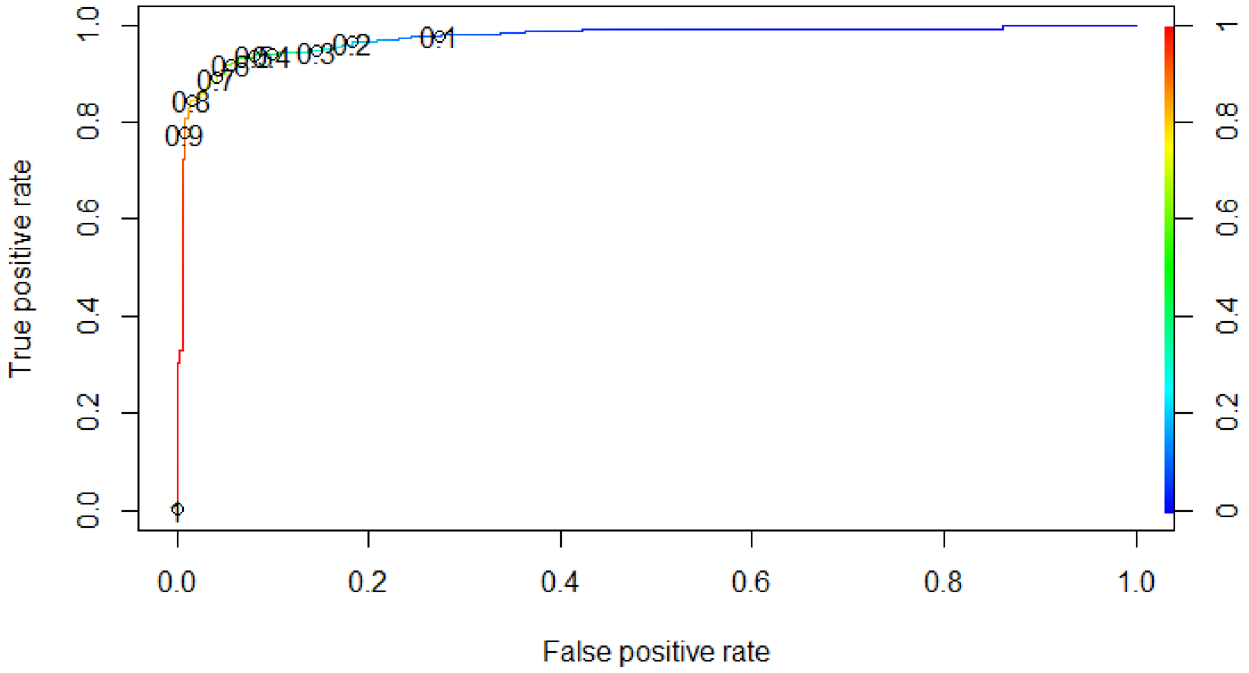
From the above tree we can see that the most important predictors are “ca”, “thal”, “cp” and “chol”. The advantage of decision trees is that clear rules can be written which are easy to interpret. For example, we can say that For all those patients where “ca” is no and “thal” not equal to 3, they have a good chance for the presence of a heart disease. Similarly, for all those patients with absence of ca, presence of thal equal to 3 and chest pain type not equal to 1,2 or 3 also indicated a presence of a heart disease for most of these records. An accuracy of 0.9133127 or 91.3% was achieved using the decision trees which is 0.3 less than that of Logistic regression model.

PERFORMANCE EVALUATION:

Various performance evaluation methods were implemented in this project.

1. AUROC

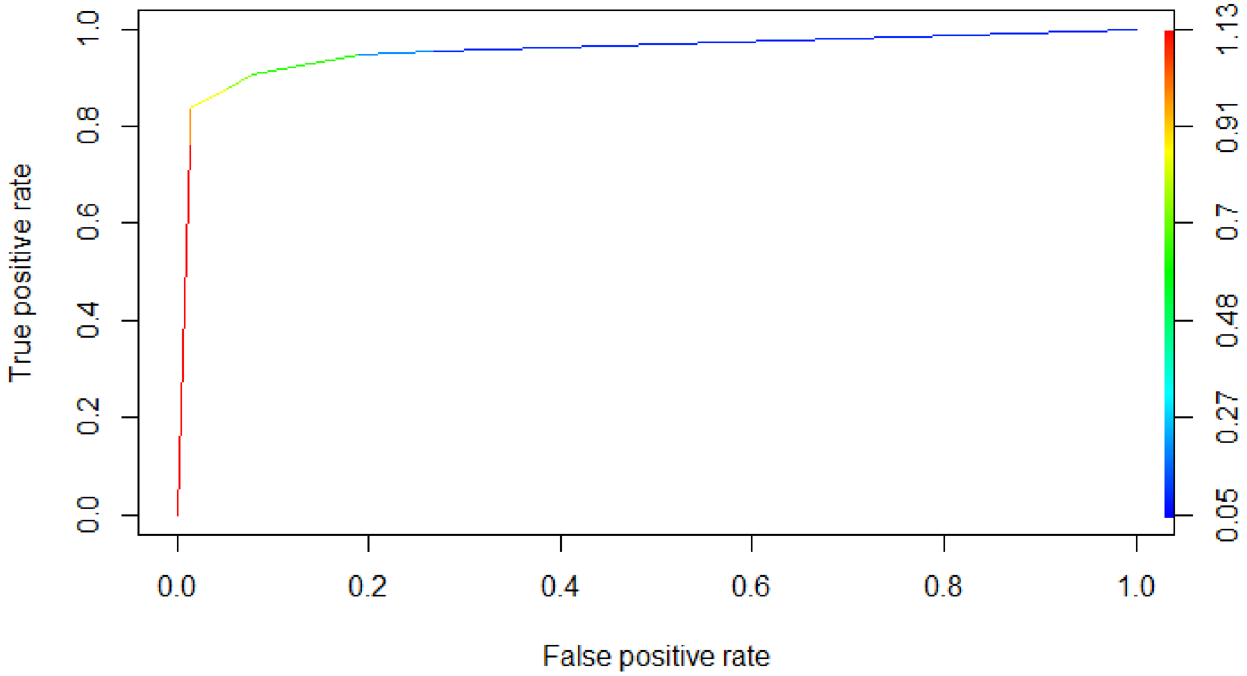
I. For Logistic Regression model:



The AOC is 0.9737538 which confirms the high predictive capability of the model, we can also see that the ideal cutoff value is around 0.3 and therefore we now classify the predicted values with this cutoff.

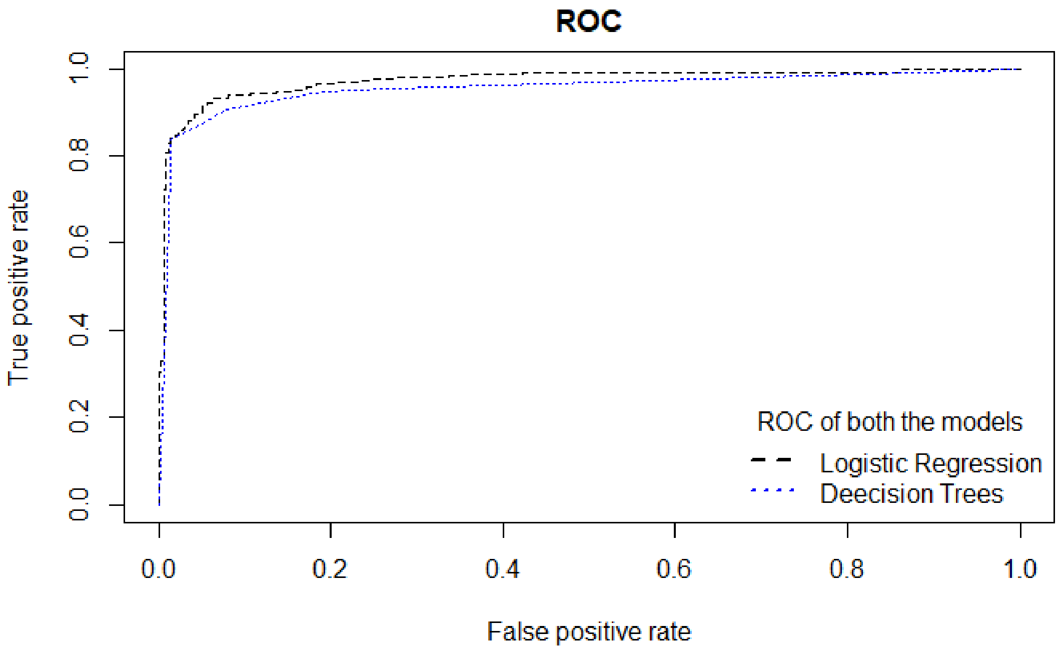
Using the cutoff 0.3, and then building the confusion matrix now gave an accuracy of 0.919504643962848 which is around 91.95%. This is close to 0.3% improvement on the original model.

II. For Tree model:



The AOC is 0.9541138, though a very good score it is still less compared to the AOC achieved by Logistic regression. Also, we can see that a cutoff of 0.2 might lead to better results, therefore we classify once again using this cutoff.

Overlapping the ROC for both models, we can see that the curve obtained by logistic regression is slightly better than that achieved by the decision trees. Therefore, we can safely conclude that Logistic Regression is the right model to go ahead for this study.



2. F1 scores - For Logistic Regression model

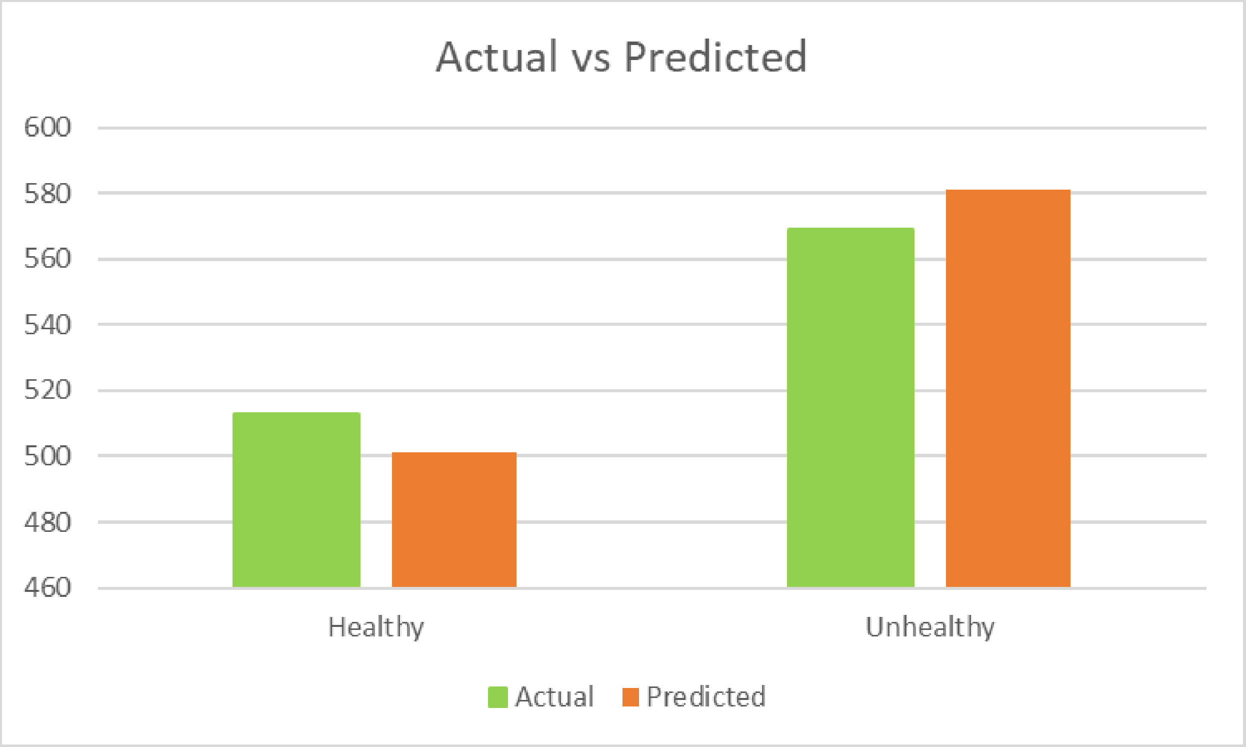
F1 scores for both the classification groups is around 91 to 92% which is a good score.

3. R Squared Value – For Logistic Regression model

R Squared value of 0.7759015 is achieved for the model and the adj R Squared is 0.7698284 which are quite close to each other thereby proving the number of predictors considered are good. However, the overall value achieved for R Squared is low. This is because the R Squared metric is primarily for linear models and not for non-linear models such as Logistic Regression.

The deviance test shown above, proves that this model has a good-fit.

FINAL RESULT:



The above graph illustrates the differences in actual vs the predicted values (of Logistic Regression model) of healthy and unhealthy hearts. The following table shows the exact values:

	Healthy	Unhealthy
Actual	513	569
Predicted	501	581

It is clear from the matrix above, that the error rate is quite low (as mentioned above in this report), and so it can be concluded that the Logistic Regression model is quite accurate.

CONCLUSION:

About 1 in every 4 deaths in the United States is due to a heart disease. Knowing the warning signs and symptoms of a heart attack so that one can act fast if they are having a heart attack is vital to prevent a fatality due to a heart disease. The chances of survival are greater when treatment begins quickly. The study hopes to be an aid in helping people around the world in early detection of a heart disease and to take precautionary measures.

The study also shows that the model developed can successfully predict the presence or absence of a heart disease with an accuracy close to 92%. It was also proven that the Logistic Regression is a better algorithm for the classification of the records for the given data set.

The following insights were drawn from the Logistic Regression model:

- Cholesterol and Maximum Heart Rate achieved are surprisingly not important variables for Logistic Regression model. However, Tree model considers Cholesterol an important predictor.
- The most important predictors are: **CP4** - Asymptomatic, **SLOPE2** – stress test is flat, **CA**: number of major vessels (0-3) colored by fluoroscopy, **THAL 7** – reversible defect.

However, the Tree model is preferred when we need generate clear rules for medical purposes, especially for doctors. Tree models require larger datasets to develop accurate rules. Since this project only had 1082 records, the tree model was not preferred.