# Heart Disease - Identification of Predictors and Prevention

Aditi Rajesh Deshpande, Kevin Fernandes, Daksh Alpesh Shah, Meng Hsuan Tsai

2020-11-25

## Contents

## 1. Objective

Heart Diseases are one of the major problems prevalent today due to health and lifestyle choices. Our main objective in the project is to predict the chances of Heart Disease and the major factors contributing to it. The data set we found helps us to evaluate deeper relationships between potential factors and perhaps reshape health care products.

## 2. Executive Summary

We used R, a statistical computing and graphics tool for building our models. At first, we explored the dataset to understand our dataset and develop a general idea for further analysis. We found some correlations among variables. Then, we used this dataset to build classification models, including the Decision Tree model and Logistic Regression model, to predict whether someone, with certain diagnostic measurements, has chances of getting Heart Disease or not. For the Decision Tree model, we sampled 80% of records as training data sets and 20% of records as validation data sets, plotted the Decision Tree, and evaluated it using confusion matrix and ROC. For the Logistic Regression model, also we sampled training data sets and validation data sets, built the Logistic Regression model, computed the odds ratios, and evaluated the Logistic Regression model using the confusion matrix and ROC. We evaluated all models and selected the best possible model.

## 3. Introduction

The motivation of this work comes from the fact that [1] although the death rates from Heart Diseases are declining; Heart Disease is still the major cause of death in the USA. An estimated 92.1 million US adults have at least one kind of Heart Disease and by 2030, around 44% of the US adult population is expected to have some form of Heart Disease.

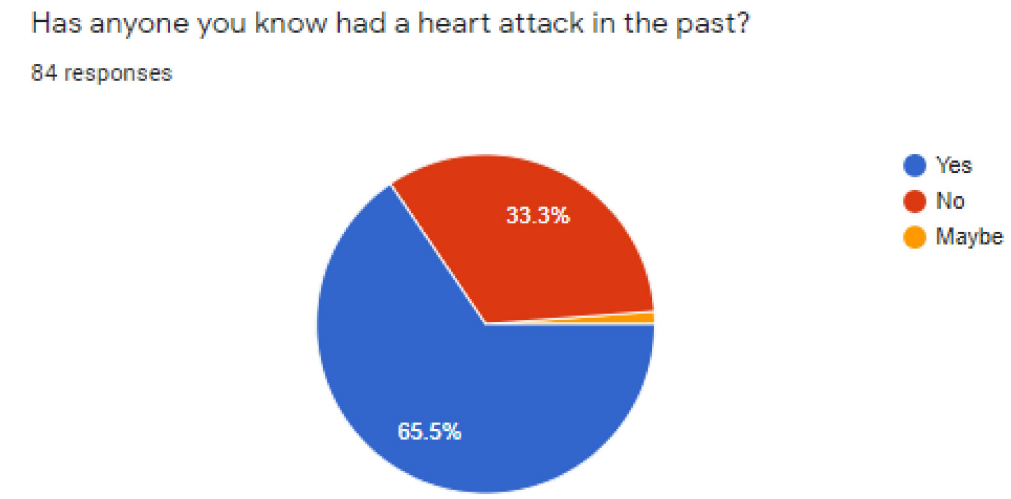
Heart Disease comes under many categories such as coronary artery disease, heart rhythm problems, chest pain (angina), or stroke. In [2] using data from the Global Burden of Disease Study, approximately 90% of the stroke risk could be attributed to modifiable risk factors. In our study, we tried identifying those risk factors that contribute the most to Heart Diseases. The estimated direct costs of Heart Diseases and stroke increased from $103.5 billion in 1996 to 1997 to $213.8 billion from 2014 to 2015.

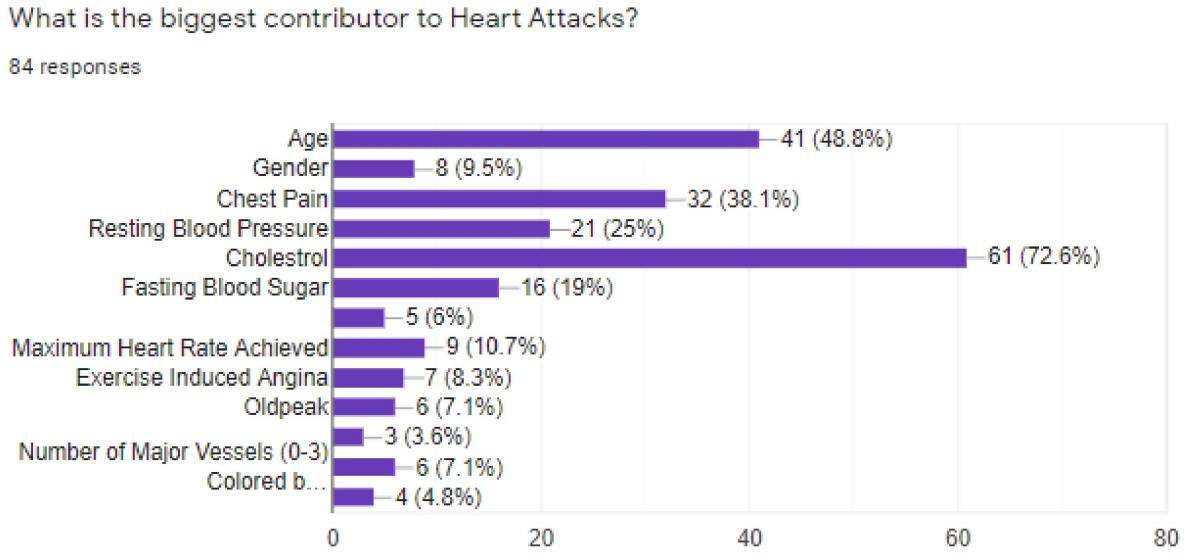
In our study we will consider the following business questions:

- Which factors contribute the most to Heart Diseases?
- Age - Are elderly people more prone to Heart Disease compared to younger people?
- Cholesterol - Does high cholesterol level lead to Heart Disease?

## 4. Research

[3] As there is a significant health impact of Heart Diseases, the awareness of the factors leading to Heart Disease and its symptoms should be common knowledge among everyone. But, in this study, it was identified that there was suboptimal knowledge about it. We conducted a survey using Google Forms to understand public awareness and to identify which variable they think could be the biggest contributor to having Heart Diseases.




We can see here, from the 84 respondents, 61 or 72.6% think Cholesterol and 41 or 48.8% think Age was the biggest contributor to Heart Disease

A Machine Learning model was to predict the risk of a Heart Disease in the subjects and these predictions were compared to the actual experiences of the subjects over fifteen years [4]. The predicted machine learning scores aligned accurately with the actual distribution of observed events. Experimental results show 100% accurate prediction for the system using Neural Networks [5].

## 5. Data Description

Our primary dataset for this analysis is the "Heart Disease Data Set". The dataset is obtained from the University of California, Irvine's Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). No additional sources were used as the primary dataset was found to be sufficient enough. The dataset consists of 13 predictor variables and 1 target variable, 'target'. Here is the description of the variables within the dataset:

| Independent variabes | Type | Description |
|---|---|---|
| Age | Continuous | Age in years |
| Sex | Discrete | Male = 1, Female = 0 |
| CP | Discrete | Chest pain type, Typical angina = 0, Atypical angina = 1, Non angina pain = 2, Asymptomatic = 3 |
| Trestbps | Continuous | Resting blood pressure (in mmHg) |
| Chol | Continuous | Serum cholesterol in mg/dl |
| Fbs | Discrete | Fasting blood sugar > 120 mg/dl: True = 1, False = 0 |
| Restecg | Discrete | Resting electrocardiographic results; Normal = 0, Having ST-T wave abnormality = 1, Showing probable or define left = 2 |
| Thalach | Continuous | Maximum heart rate achieved |
| Exang | Discrete | Exercise induced angina, Yes = 1, No = 0 |
| Oldpeak | Continuous | ST depression induced by exercise relative to rest |
| Slope | Discrete | Slope of the peak exercise ST segment; upsloping = 0; Flat = 1, Down sloping = 2 |
| Ca | Discrete | Number of major colored vessels from fluoroscopy (0-3) |
| Thal | Discrete | A blood disorder called thalassemia; Normal = 0, Fixed defect = 1, Reversible defect = 2 |

| Dependent variabes | Type | Description |
|---|---|---|
| Target | Discrete | 0 = Heart Disease Absent, 1 = Heart Disease Present |

## 6. Preprocessing Data

- First we read the data file from the University of California, Irvine Machine Learning Repository into R. After reading the file into R, we converted the values into numeric
- We found a total of 6 null values in the dataset.The thal column contained 2 and the ca column contained 4 of the null values. We have removed these null values in order to make the dataset accurate

- The following shows the data and the summary statistics:

```
head(heart_data.dt)
```

```
##    age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1:  67   1  4      160  286   0       2     108     1     1.5     2  3    0
## 2:  67   1  4      120  229   0       2     129     1     2.6     2  2    2
## 3:  37   1  3      130  250   0       0     187     0     3.5     3  0    0
## 4:  41   0  2      130  204   0       2     172     0     1.4     1  0    0
## 5:  56   1  2      120  236   0       0     178     0     0.8     1  0    0
## 6:  62   0  4      140  268   0       2     160     0     3.6     3  2    0
##    target
## 1:      1
## 2:      1
## 3:      0
## 4:      0
## 5:      0
## 6:      1
```

```
summary(heart_data.dt)
```

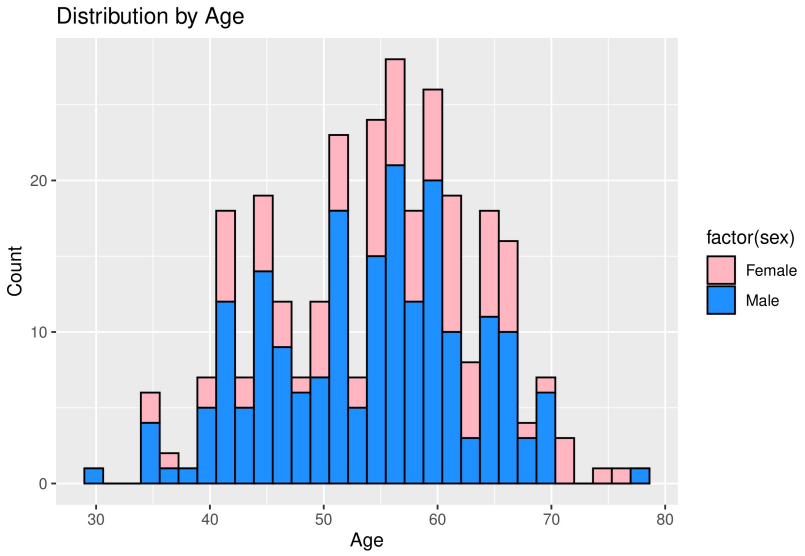
```
##       age             sex               cp           trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0
##  Median :56.00   Median :1.0000   Median :3.000   Median :130.0
##  Mean   :54.51   Mean   :0.6757   Mean   :3.166   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0
##       chol            fbs            restecg          thalach
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.0
##  Median :243.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :247.4   Mean   :0.1419   Mean   :0.9932   Mean   :149.6
##  3rd Qu.:276.2   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##      exang           oldpeak          slope             ca
##  Min.   :0.0000   Min.   :0.000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.800   Median :2.000   Median :0.0000
##  Mean   :0.3277   Mean   :1.051   Mean   :1.598   Mean   :0.6791
##  3rd Qu.:1.0000   3rd Qu.:1.600   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.200   Max.   :3.000   Max.   :3.0000
##       thal            target
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.8345   Mean   :0.4628
##  3rd Qu.:2.0000   3rd Qu.:1.0000
##  Max.   :2.0000   Max.   :1.0000
```

## 7. Exploratory Data Analysis(EDA)

- **Graphical Analysis**

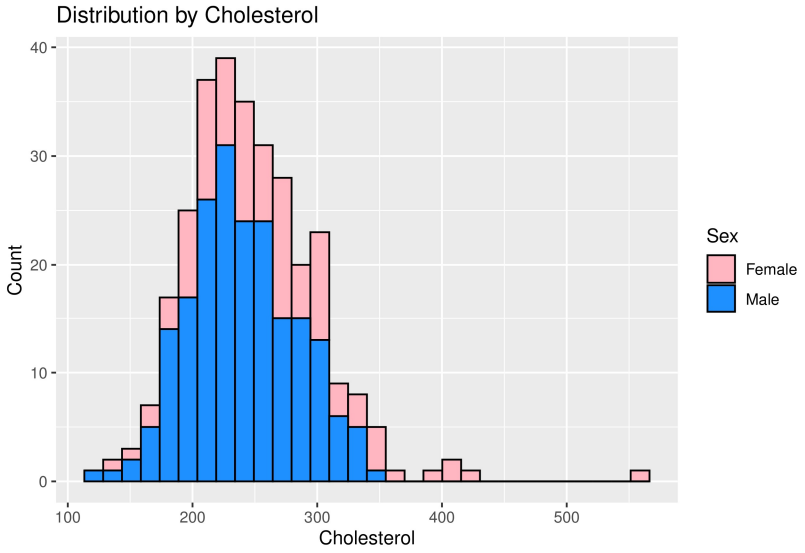
1. **Histogram for age:**

The minimum age is 29 and maximum age is 77. Average age of Population is 54.37. Maximum number of population lies between the age group 55 and 60 years.




2. **Histogram for Cholesterol:**

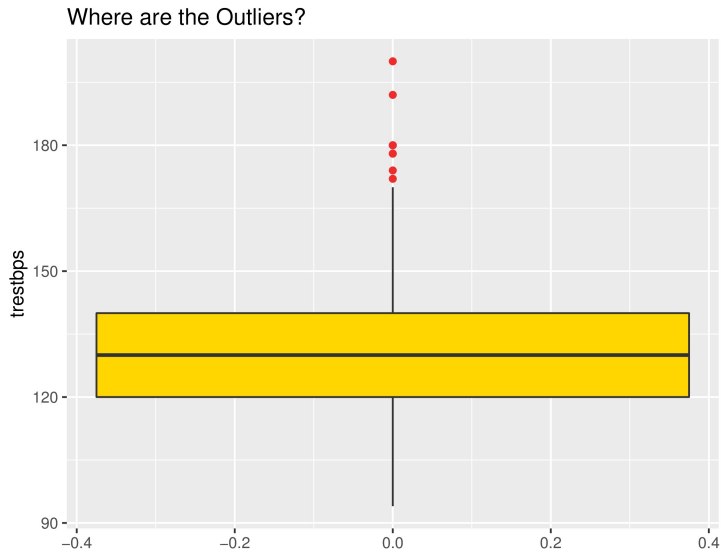
The minimum cholesterol among patients is 126, the maximum is 564 and the average cholesterol is 247.3. We can clearly see in Histogram that maximum population have cholesterol between 200 and 250 unit.



- **Finding outliers and understanding them**

1. trestbps

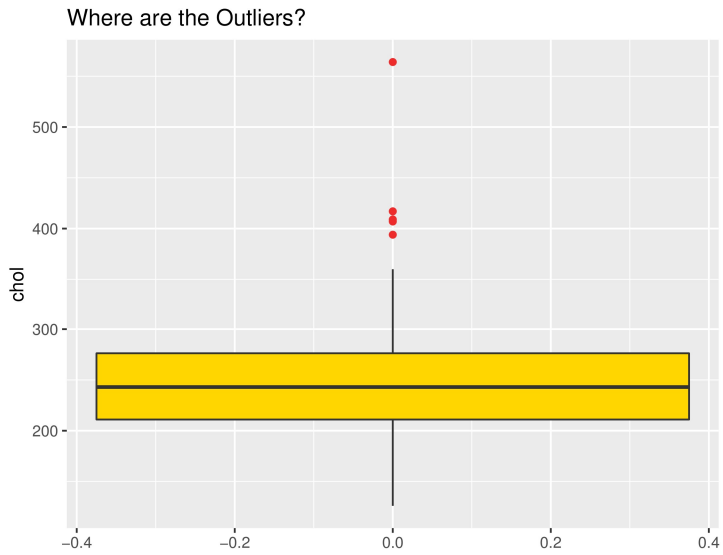
Resting Blood Pressure (in mm Hg on admission to the hospital)




- We do not remove outliers in the trestbps column because values above 180 indicate risk of critical medical conditions

2. chol

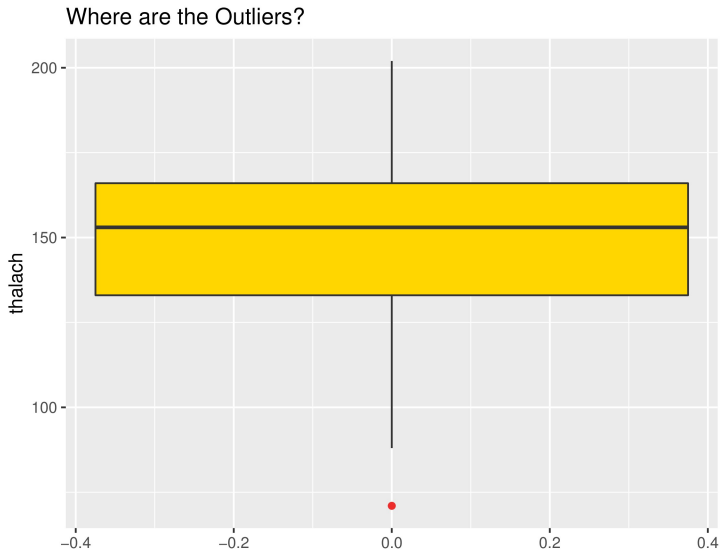
Serum Cholestoral in mg/dl



- We do not remove outliers in the chol column because patients can have unusally high values due to inherited conditions

3. thalach

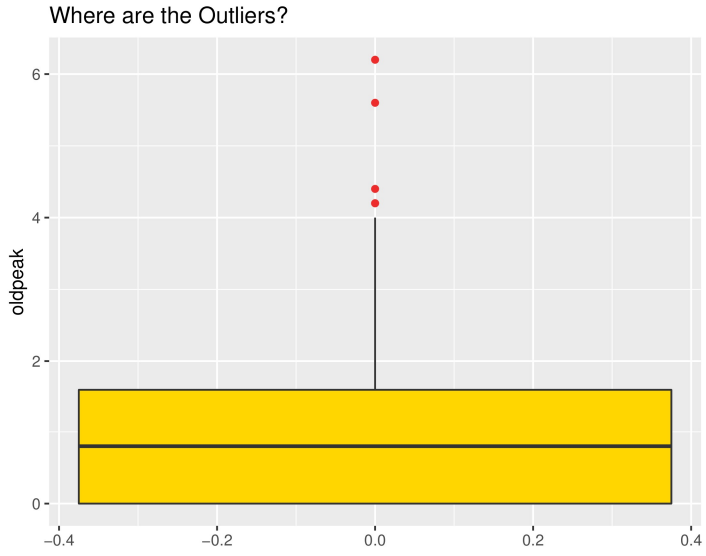
Maximum Heart Rate Achieved




- We do not remove outliers in the thalach column because it indicates a healthy individual

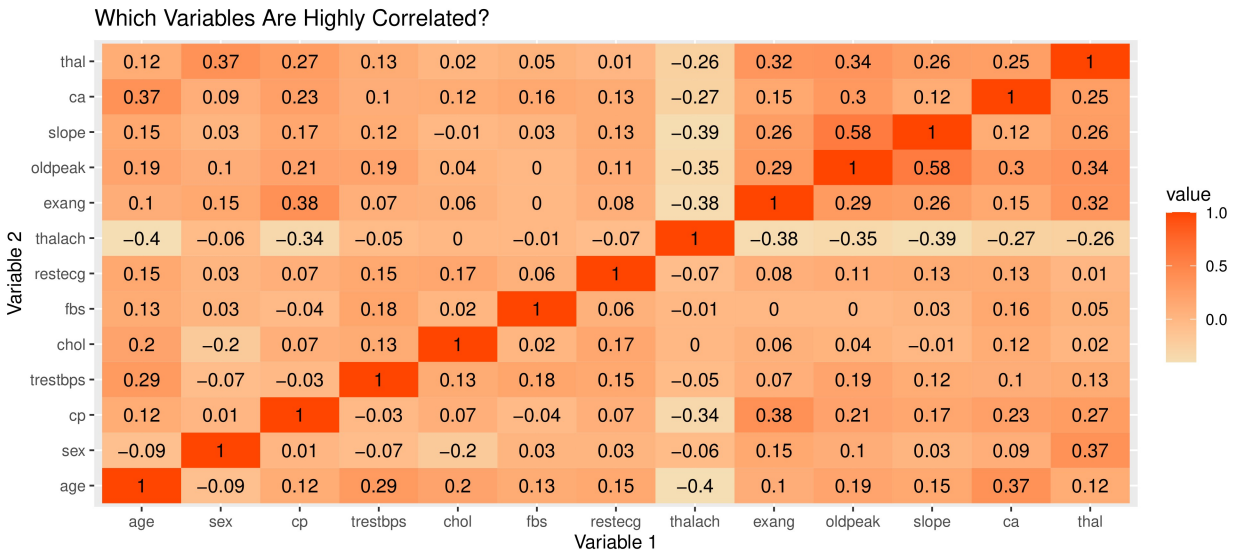
4. oldpeak

ST Depression Induced by Exercise Relative to Rest



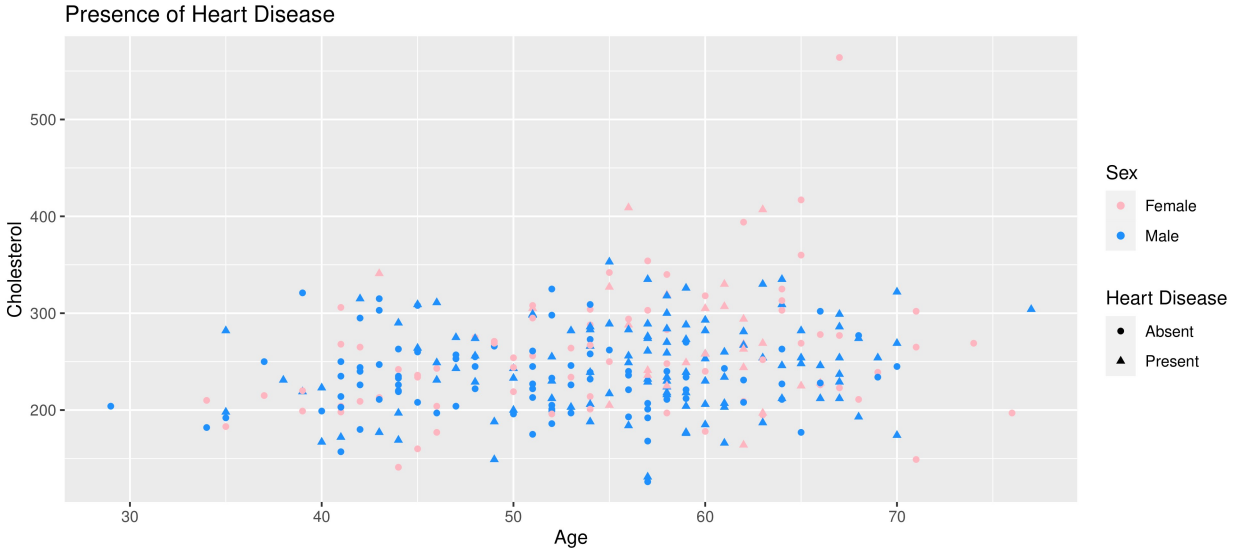

- We do not remove the outliers in the oldpeak column because high values of oldpeak(greater than 4) could indicate the presence of Heart Disease in patients.

- **Correlation heatmap between variables:**




We used a correlation heatmap to plot out the strongest positive and negative correlations. As indicated, "slope" and "oldpeak" have the strongest correlation, with a positive 0.58. "thalach" and "age" have the lowest correlation, with a negative 0.4.

- **Analyzing relationships between cholesterol and age:**




As we can see in the scatterplot, age and cholesterol do not appear to have a significant correlation with Heart Disease. In the top right corner, we can see an elderly female with high cholesterol, but who does not have Heart Disease. On the other hand, on the bottom left, we can see a young male with low cholesterol, who has Heart Disease. Hence we cannot derive any relationship between these variables and Heart Disease.

## 8. Data Modelling

1. Splitting the data:

First, we divided the dataset into two parts: training dataset and validation dataset. We allocated 80% of the dataset for the training dataset and the remaining 20% of the dataset for the validation dataset.

2. Logistic Regression:

It extends the idea of Linear Regression to the situation where the outcome variable is categorical. It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Sex, cp, trestbps, ca, thalach , exang, slope are significant variables.

The null deviance shows how well the response is predicted by the model with nothing but an intercept. The residual deviance shows how well the response is predicted by the model when the predictors are included. Residual deviance is the measure of error.Smaller the residual deviance , better the predictive power of the model.

In the output we get the residual deviance smaller than the Null deviance, so our Logistic Model has some predictive power. The variables will have some explanatory power.

In logistic regression the odds ratio represents the constant effect of a predictor X, on the likelihood that one outcome will occur.

When a binary outcome variable is modeled using logistic regression, it is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables. This makes the interpretation of the regression coefficients tricky. So we make use of Odd's Ratio.

```
## [1] "Summary of  the Logistic Regression Model"
```

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8171  -0.5359  -0.1926   0.4157   2.2818
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.807764   3.321557  -1.146  0.25164
## age         -0.020362   0.027361  -0.744  0.45675
## sex          1.327736   0.549625   2.416  0.01570 *
## cp           0.525923   0.208812   2.519  0.01178 *
## trestbps     0.021971   0.011453   1.918  0.05506 .
## chol         0.003446   0.004294   0.803  0.42220
## fbs         -0.799407   0.640307  -1.248  0.21186
## restecg      0.235885   0.207367   1.138  0.25532
## thalach     -0.027953   0.011914  -2.346  0.01896 *
## exang        1.114523   0.479347   2.325  0.02007 *
## oldpeak      0.264658   0.233532   1.133  0.25709
## slope        0.405958   0.403248   1.007  0.31407
## ca           1.300778   0.297202   4.377 0.000012 ***
## thal         0.675913   0.229062   2.951  0.00317 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 325.79  on 235  degrees of freedom
## Residual deviance: 162.81  on 222  degrees of freedom
## AIC: 190.81
##
## Number of Fisher Scoring iterations: 6
```

```
## [1] "Exponents of the coefficients"
```

```
## (Intercept)         age         sex          cp    trestbps        chol
##  0.02219775  0.97984350  3.77249450  1.69201963  1.02221395  1.00345209
##         fbs     restecg     thalach       exang     oldpeak       slope
##  0.44959540  1.26602915  0.97243403  3.04811522  1.30298586  1.50073981
##          ca        thal
##  3.67215218  1.96582604
```

```
## [1] "Summary of  the Logistic Regression Model with stepAIC"
```

```
##
## Call:
## glm(formula = target ~ sex + cp + trestbps + thalach + exang +
##     oldpeak + ca + thal, family = "binomial", data = train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7496  -0.5628  -0.2116   0.4545   2.4868
##
## Coefficients:
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -3.29975    2.36529  -1.395   0.16299
## sex          1.18352    0.48265   2.452   0.01420 *
## cp           0.52635    0.20523   2.565   0.01033 *
## trestbps     0.01900    0.01071   1.773   0.07618 .
## thalach     -0.02525    0.01005  -2.513   0.01197 *
## exang        1.17624    0.46160   2.548   0.01083 *
## oldpeak      0.39812    0.19613   2.030   0.04237 *
## ca           1.12581    0.25511   4.413 0.0000102 ***
## thal         0.70873    0.21894   3.237   0.00121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 325.79  on 235  degrees of freedom
## Residual deviance: 167.90  on 227  degrees of freedom
## AIC: 185.9
##
## Number of Fisher Scoring iterations: 6
```

Odds ratio:

From the output , we can conclude that ca is the most significant variable among all the other variables. The way we interpret these coeffcients is as follows. Considering thalach (maximum heart rate achieved)as an example, if it goes up by one unit, the odds of having Heart Disease goes down by 2.8%.

AIC:

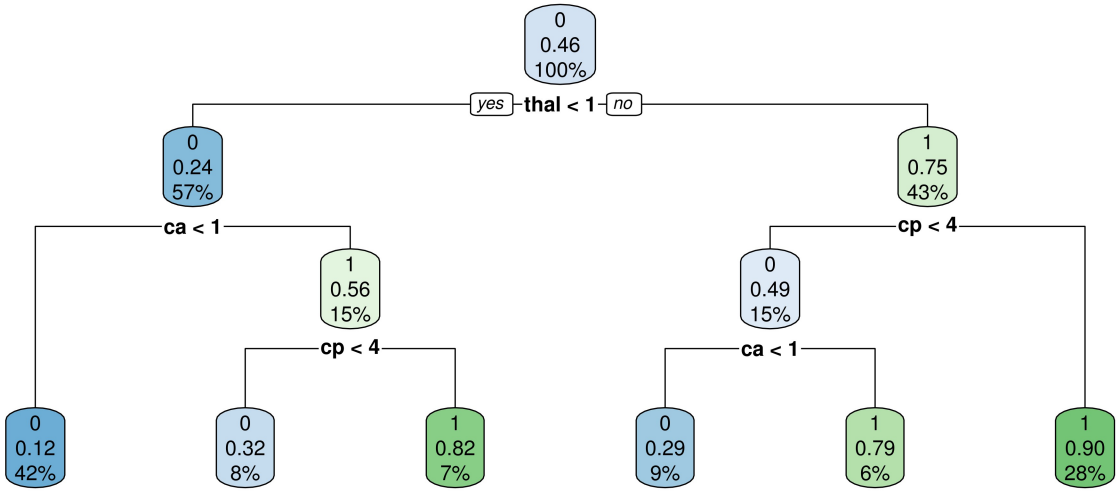
Using AIC , we select the best possible model available to us with all the significant variables . Even in the AIC model, we get residual deviance smaller than the null deviance, so our model has some predictive power.

3. Decision Tree:

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision Tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

From the Decision Tree , we get the following Rule with the most percentage cover of cases.

When thal $< 1$ & ca $< 1$ THEN CLASS $= 0$ and this rule covers 42% of cases.



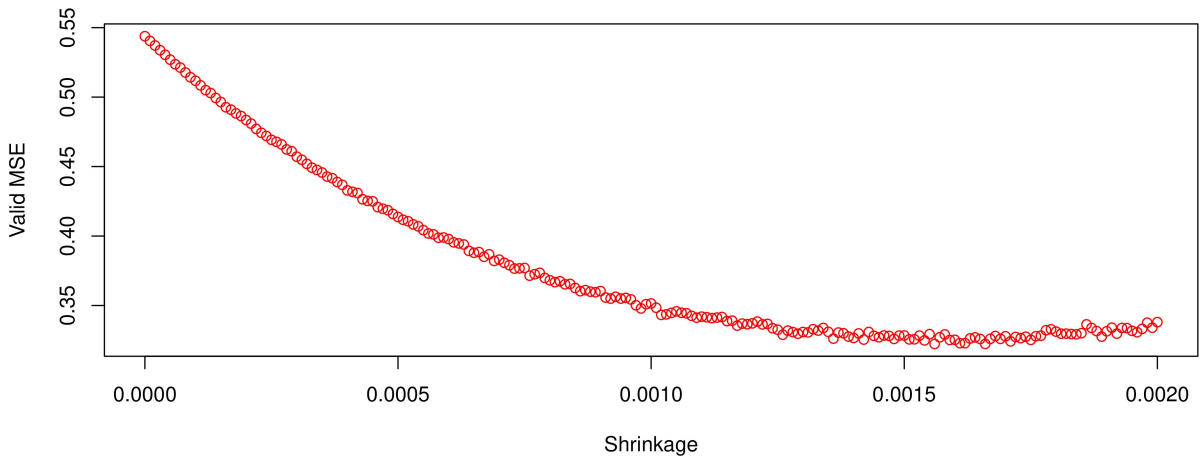

```
##  target                                          cover
##    0.12 when thal <  1 & ca <  1                  42%
##    0.29 when thal >= 1 & ca <  1 & cp <  4         9%
##    0.32 when thal <  1 & ca >= 1 & cp <  4         8%
##    0.79 when thal >= 1 & ca >= 1 & cp <  4         6%
##    0.82 when thal <  1 & ca >= 1 & cp >= 4         7%
##    0.90 when thal >= 1           & cp >= 4        28%
```

```
## [1] "Mean Valid MSE for Adaboost"
```

```
## [1] 0.3787996
```

```
## [1] "Min Valid MSE for Adaboost"
```

```
## [1] 0.3223143
```

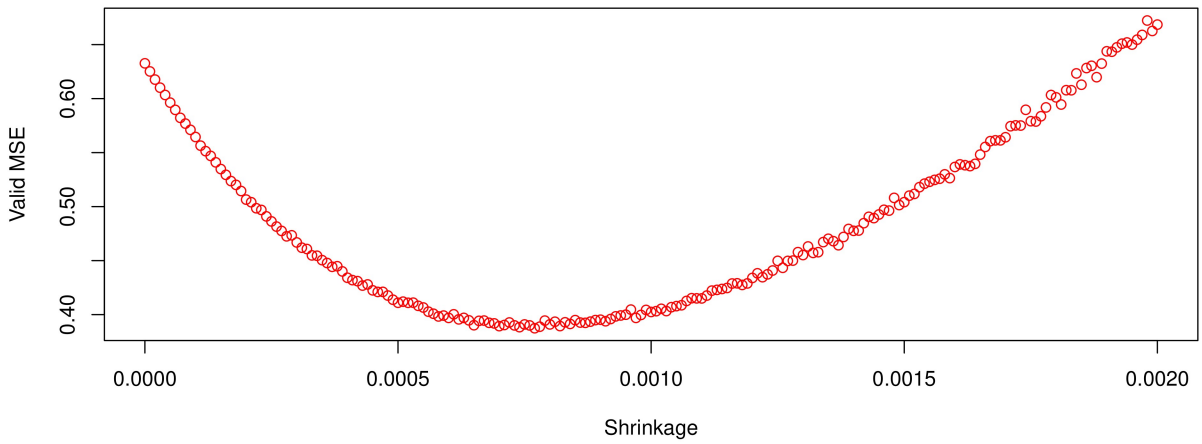


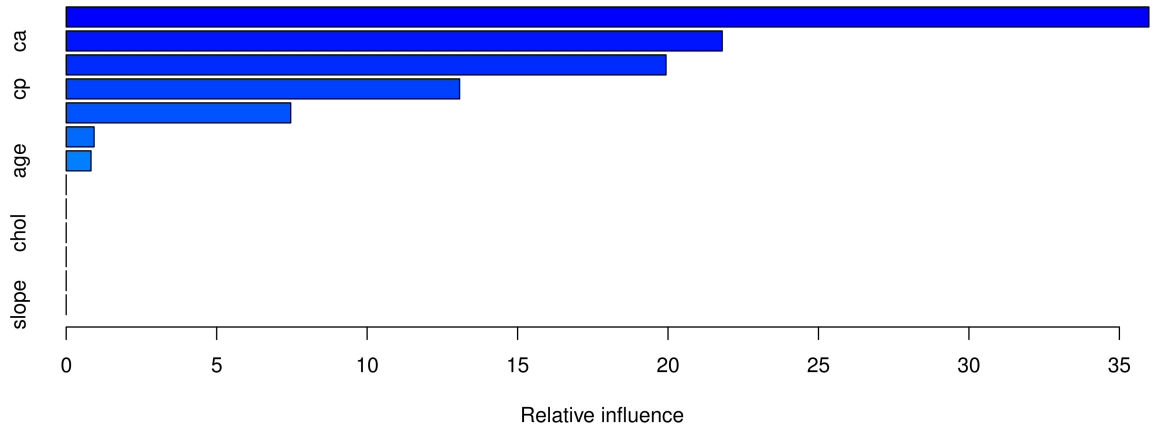

```
## [1] "Mean Valid MSE for Bernoulli"
```

```
## [1] 0.4828616
```

```
## [1] "Min Valid MSE for Bernoulli"
```

```
## [1] 0.3874044
```






```
##                var    rel.inf
## thal         thal 35.9797241
## ca             ca 21.8011640
## thalach   thalach 19.9336106
## cp             cp 13.0744268
## exang       exang  7.4605795
## oldpeak   oldpeak  0.9263317
## age           age  0.8241632
## sex           sex  0.0000000
## trestbps trestbps  0.0000000
## chol         chol  0.0000000
## fbs           fbs  0.0000000
## restecg   restecg  0.0000000
## slope       slope  0.0000000
```

## 9. Performance evaluation

The performance of a regression model can be understood by knowing the error rate of the predictions made by the model. You can also measure the performance by knowing how well your regression line fit the dataset and knowing the accuracy of such models.

Confusion Matrix:

Confusion matrix is a measurement that used to represent the performance of a classification model by recording the sources of errors: false positives and false negatives. We use confusion matrix to depict the accuracy of the training data.

```
## [1] "Confusion Matrix for Logistic Regression"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 26  4
##          1  6 24
```

```
##
##                Accuracy : 0.8333
##                  95% CI : (0.7148, 0.9171)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 0.000001056
##
##                   Kappa : 0.6667
##
##  Mcnemar's Test P-Value : 0.7518
##
##             Sensitivity : 0.8571
##             Specificity : 0.8125
##          Pos Pred Value : 0.8000
##          Neg Pred Value : 0.8667
##              Prevalence : 0.4667
##          Detection Rate : 0.4000
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.8348
##
##        'Positive' Class : 1
##
```

```
## [1] "Confusion Matrix for Logistic Regression with StepAIC"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 25  5
##          1  7 23
##
##                Accuracy : 0.8
##                  95% CI : (0.6767, 0.8922)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 0.00001609
##
##                   Kappa : 0.6
##
##  Mcnemar's Test P-Value : 0.7728
##
##             Sensitivity : 0.8214
##             Specificity : 0.7812
##          Pos Pred Value : 0.7667
##          Neg Pred Value : 0.8333
##              Prevalence : 0.4667
##          Detection Rate : 0.3833
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.8013
##
##        'Positive' Class : 1
##
```

```
## [1] "Confusion Matrix for Decision Tree"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 28  4
##          1  4 24
##
##                Accuracy : 0.8667
##                  95% CI : (0.7541, 0.9406)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 0.00000004403
##
##                   Kappa : 0.7321
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8571
##             Specificity : 0.8750
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.8750
##              Prevalence : 0.4667
##          Detection Rate : 0.4000
##    Detection Prevalence : 0.4667
##       Balanced Accuracy : 0.8661
##
##        'Positive' Class : 1
##
```

```
## [1] "Confusion Matrix for Decision Tree with Adaboost"
```

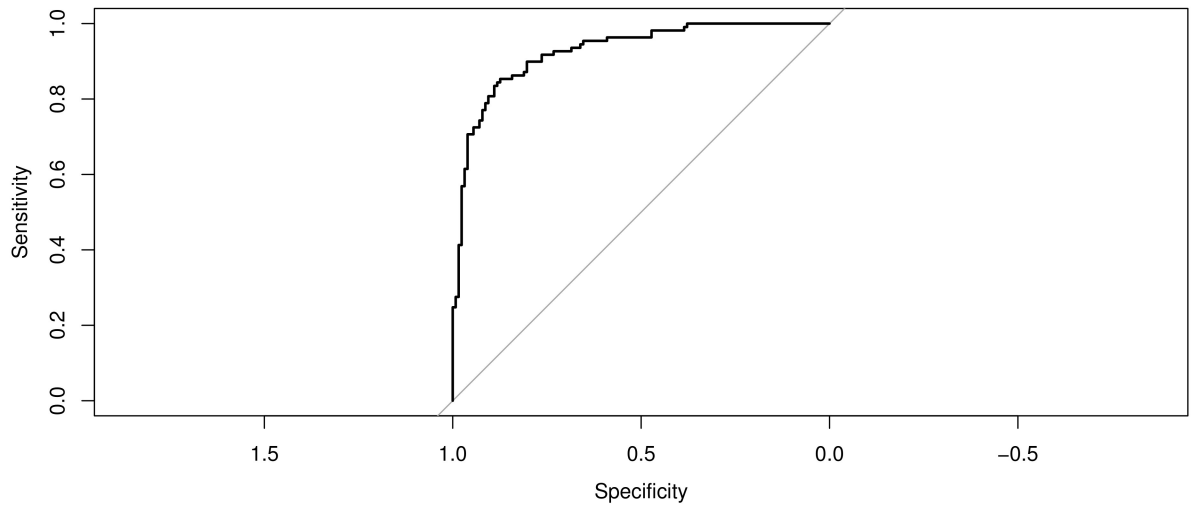
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 32 18
##          1  0 10
##
##                Accuracy : 0.7
##                  95% CI : (0.5679, 0.8115)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 0.006353
##
##                   Kappa : 0.3721
##
##  Mcnemar's Test P-Value : 0.00006151
##
##             Sensitivity : 0.3571
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.6400
##              Prevalence : 0.4667
##          Detection Rate : 0.1667
##    Detection Prevalence : 0.1667
##       Balanced Accuracy : 0.6786
```

```
##
##        'Positive' Class : 1
##
```

## 10. ROC curve

For the classification problem to check or visualize the performance of the classification problem, we use AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model's performance. It is also written as AUROC (Area Under the Receiver Operating Characteristics). ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better is the model in distinguishing. An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

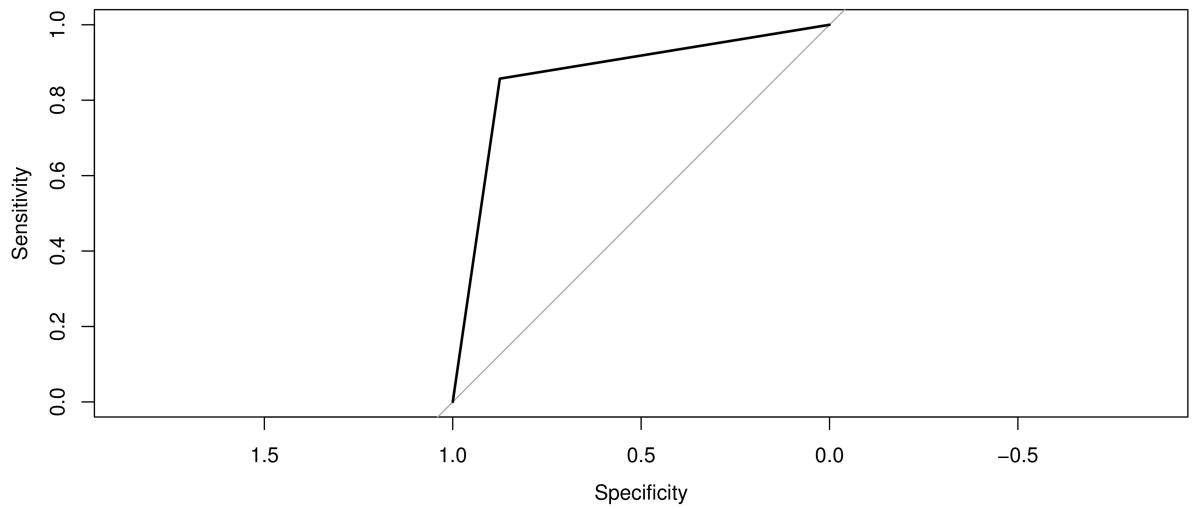
```
## [1] "ROC for Logistic Regression"
```




```
## [1] "Area Under the Curve for Decision Tree"
```

```
## Area under the curve: 0.9265
```

```
## [1] "ROC for Descision Tree"
```

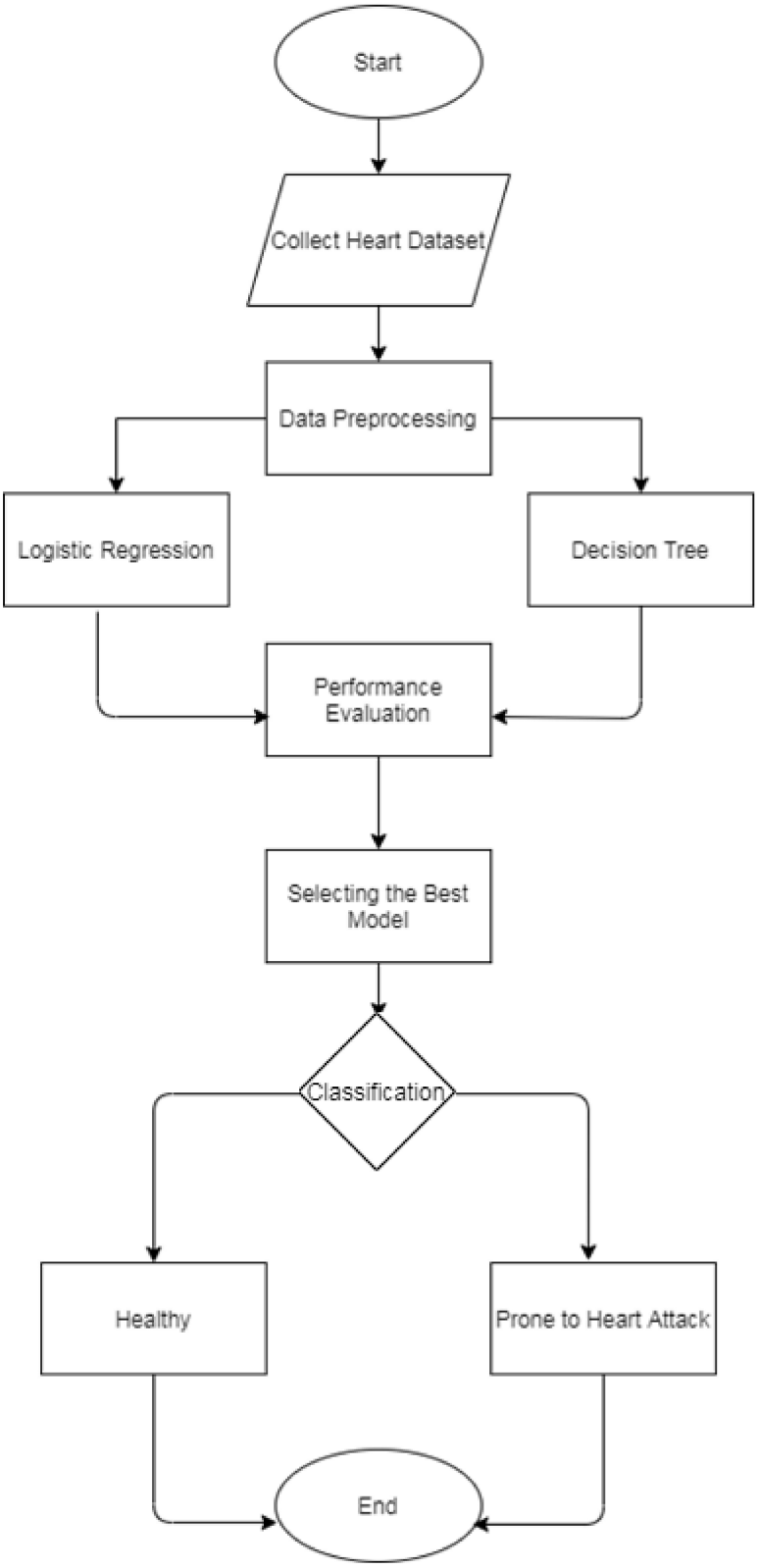



```
## [1] "Area Under the Curve for Decision Tree "
```

```
## Area under the curve: 0.8661
```

As we see here, the AUC for the training data is 0.9265 and the AUC for the validation dataset is 0.8661

## 11. Flow Diagram



## 12. Conclusion:

We have built and compared the Logistic Regression model and the Decision Tree model and have captured the results and the performance metrics for the same. With the Logistic Regression model we achieved an accuracy of 83.33% and for the Decision Tree we achieved an accuracy of 86.67%. Therefore, the Decision Tree model is better for our project analysis. They also are easy to implement and interpret, and they display higher accuracy than Logistic Regression model here. The Decision Tree model we built helped us identify thal, ca, thalach and cp as the most important predictors of Heart Disease. This proves that age and chol are not major contributors to Heart Disease.

Using this model, one can employ it to deduce if a particular patient, with a certain medical profile, is likely to have Heart Disease or not. This model can be used to serve bigger populations and provide predictions that are highly accurate enough with minimum error. The model's predictions can be looked upon as the basis for improvement of measures to study various medical factors that could help in preventing a Heart Disease. Similar models can be built to study other prone populations and help in serving the society better with analytics and various classification techniques.

## 13. References:

[1] Moonesinghe R, Yang Q, Zhang Z, Khoury MJ. Prevalence and cardiovascular health impact of family history of premature Heart Disease in the United States: Analysis of the National Health and Nutrition Examination Survey, 2007-2014

[2] Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Virani, S. S. (2019). Heart disease and stroke statistics-2019 update: A report from the American Heart Association.

[3] Fang J, Luncheon C, Ayala C, Odom E, Loustalot F. Awareness of heart attack symptoms and response among adults—United States, 2008, 2014, and 2017. MMWR. 2019;68(5):101–6.

[4] AI can better predict risk of heart attack, cardiac death, study published in Journal of Cardiovascular Research, https://health.economictimes.indiatimes.com/news/diagnostics/ai-can-better-predict-risk-of-heart-attack-cardiac-death-study/72899878

[5] Singh P, Singh S, Pandi-Jain GS. "Effective heart disease prediction system using data mining techniques".in International Journal of Nanomedicine, 13(T-NANO 2014 Abstracts):121-124, 2018