Rutgers University-The State University of New Jersey

Newark

Capstone Project:

PREDICTION OF HEART DISEASE

By

Shivani Malandkar

Ruid:189006969

Email: Shivani.malandkar@rutgers.edu

Under the Guidance of:

Prof. Sergei Schreider

Acknowledgement

I would like to express my gratitude to Professor Sergei Schreider for guiding me in this project, without which it would not have been possible to complete. His contribution is sincerely appreciated and greatly acknowledged.

I would also like to thank the staff of Rutgers University and faculty of Information Technology department for all their help and learning resources during the completion of this project.

I would also like to take the opportunity to thank my family for their constant support as I worked through the timeline of the project.

Contents

Page No.

Abstract

Prediction using machine learning models in healthcare domain is increasing rapidly. People are getting diagnosed with various chronic disease like heart attack and cancer which require complicated treatments that are not only complex but also expensive in nature. Unfortunately, these treatments are not affordable for many individuals.

Healthcare sector is growing and coming up with new advanced technologies using sensors, artificial intelligence (AI) and DNA analysis. Adding predictive analysis to these technologies will furthermore increase the cost effectiveness and give a suitable accuracy for the result.

A predictive analysis approach can be used in order to achieve a solution for such problems. Using various machine learning models on huge database helps give pre-predictive analysis. This can help patients easily diagnose and can help reduce cost for their diagnosis. Using machine learning algorithm can help make a proper decision and solution for the problem and give a computer-based diagnosis and disease predictions. Along with this, we will come up with various challenges in order to achieve best model based on accuracy but using proper tools and algorithm can help us to overcome these hurdles.

The main aim of this project is to reduce a cost in healthcare domain in order to advancement in healthcare. The growth in the use of machine learning can further lead to valuable advancement in healthcare sector

Introduction

The data set consists of various factors such as age, cholesterol, depression, etc. According to Google, the cause of heart disease is often cholesterol, however, in this data set we are trying to analyze if there is another factor that also highly correlates to having heart disease. The main idea behind this project is to do predictive analysis using data of patients having heart disease or not and factors that can cause this disease.

This analysis will help find which factors can lead to heart disease. There is a target variable as well in dataset which states whether patient has heart disease or not.

Based on risk factors like age, cholesterol amount, whether a patient is depressed, we can analyze the likelihood of them getting a heart disease or not using machine learning algorithms.

Applying various machine learning models and analysis and based on their accuracy, we can get a pre-planned solution for a patient in the database who could have a disease in the future. For example, if a patient has higher cholesterol, doctors might be able to control it to avoid future diseases.

In this project, various machine learning algorithms are used, out of which we find the one with the highest accuracy. Through this, we can visualize the data and find out the important factors which cause heart disease or if the patient has a healthy heart.

People these days have an unhealthy lifestyle caused by habits such as smoking that could lead to high cholesterol and imbalances in vital stats. As per news articles, heart disease is a proven leading cause of death for both males and females. A lot of people die of heart diseases every year, which is almost 1 in every 4 deaths. According to 2015 data, 17.9 million deaths occurred mainly due to cardiovascular diseases. Every minute, one person in the United States dies due to cardiovascular or heart diseases.

In this data set, we are trying to find the variable that could be the most important when it comes to getting a heart disease. Also, on the basis of confusion matrix, we are trying to differentiate the number of healthy and unhealthy patients from the data set in various models.

Data Description

This database contains attributes, but all 76 published experiments refer to using a subset of 14 of them.

In particular, the Cleveland database is the only one that has been used by ML researchers to this date.

Link for the dataset: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

303 obesrvation and 14 variables

Attribute Information:

> 1. age

> 2. sex

> 3. chest pain type (4 values)

> 4. resting blood pressure

> 5. serum cholestoral in mg/dl

> 6. fasting blood sugar > 120 mg/dl

> 7. resting electrocardiographic results (values 0,1,2)

> 8. maximum heart rate achieved

> 9. exercise induced angina

> 10. oldpeak = ST depression induced by exercise relative to rest

> 11. the slope of the peak exercise ST segment

> 12. number of major vessels (0-3) colored by fluoroscopy

> 13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

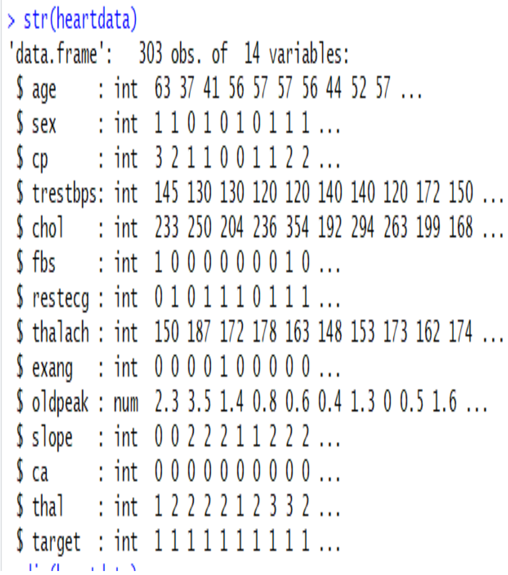
> 14. Target variable

Data Pre-processing:

There were no missing values found in data set but we performed following process in order to format the dataset.

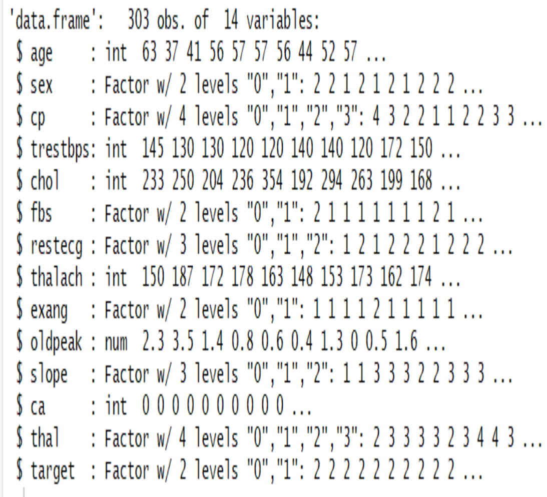
1.Converted integers to factors

Old dataset



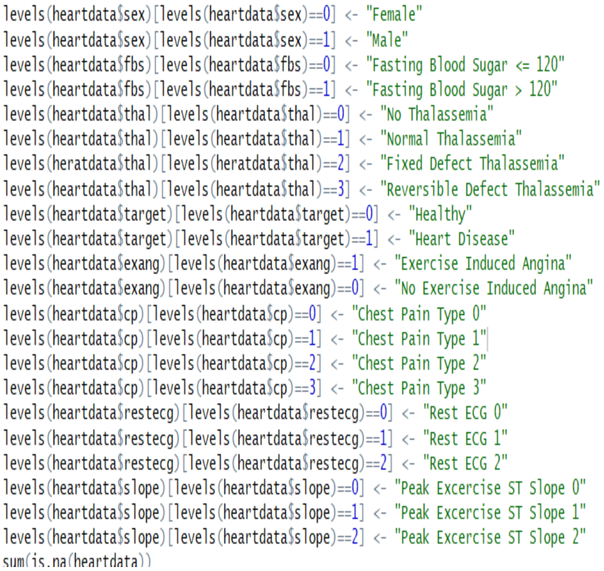
```
> str(heartdata)
'data.frame':   303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
 $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
 $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
 $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : int  0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : int  0 0 2 2 2 1 1 2 2 2 ...
 $ ca      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ thal    : int  1 2 2 2 2 1 2 3 3 2 ...
 $ target  : int  1 1 1 1 1 1 1 1 1 1 ...
```

Converted Dataset

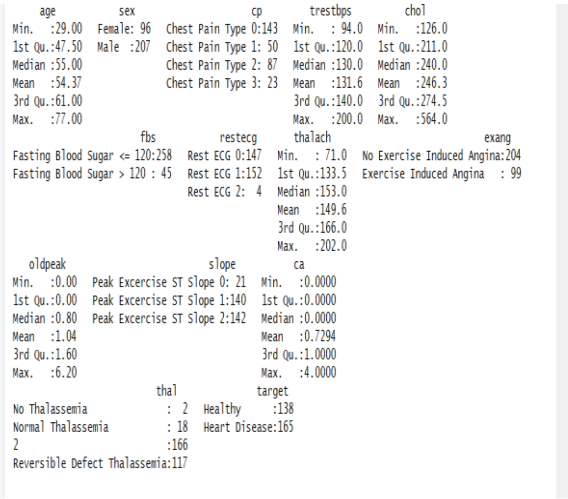


```
'data.frame':   303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
 $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
 $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
 $ restecg : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
 $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
 $ ca      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ thal    : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
 $ target  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

2.Creating levels and differentiating the values in dataset and assigning name for each



```
levels(heartdata$sex)[levels(heartdata$sex)==0] <- "Female"
levels(heartdata$sex)[levels(heartdata$sex)==1] <- "Male"
levels(heartdata$fbs)[levels(heartdata$fbs)==0] <- "Fasting Blood Sugar <= 120"
levels(heartdata$fbs)[levels(heartdata$fbs)==1] <- "Fasting Blood Sugar > 120"
levels(heartdata$thal)[levels(heartdata$thal)==0] <- "No Thalassemia"
levels(heartdata$thal)[levels(heartdata$thal)==1] <- "Normal Thalassemia"
levels(heratdata$thal)[levels(heratdata$thal)==2] <- "Fixed Defect Thalassemia"
levels(heartdata$thal)[levels(heartdata$thal)==3] <- "Reversible Defect Thalassemia"
levels(heartdata$target)[levels(heartdata$target)==0] <- "Healthy"
levels(heartdata$target)[levels(heartdata$target)==1] <- "Heart Disease"
levels(heartdata$exang)[levels(heartdata$exang)==1] <- "Exercise Induced Angina"
levels(heartdata$exang)[levels(heartdata$exang)==0] <- "No Exercise Induced Angina"
levels(heartdata$cp)[levels(heartdata$cp)==0] <- "Chest Pain Type 0"
levels(heartdata$cp)[levels(heartdata$cp)==1] <- "Chest Pain Type 1"
levels(heartdata$cp)[levels(heartdata$cp)==2] <- "Chest Pain Type 2"
levels(heartdata$cp)[levels(heartdata$cp)==3] <- "Chest Pain Type 3"
levels(heartdata$restecg)[levels(heartdata$restecg)==0] <- "Rest ECG 0"
levels(heartdata$restecg)[levels(heartdata$restecg)==1] <- "Rest ECG 1"
levels(heartdata$restecg)[levels(heartdata$restecg)==2] <- "Rest ECG 2"
levels(heartdata$slope)[levels(heartdata$slope)==0] <- "Peak Excercise ST Slope 0"
levels(heartdata$slope)[levels(heartdata$slope)==1] <- "Peak Excercise ST Slope 1"
levels(heartdata$slope)[levels(heartdata$slope)==2] <- "Peak Excercise ST Slope 2"
sum(is.na(heartdata))
```



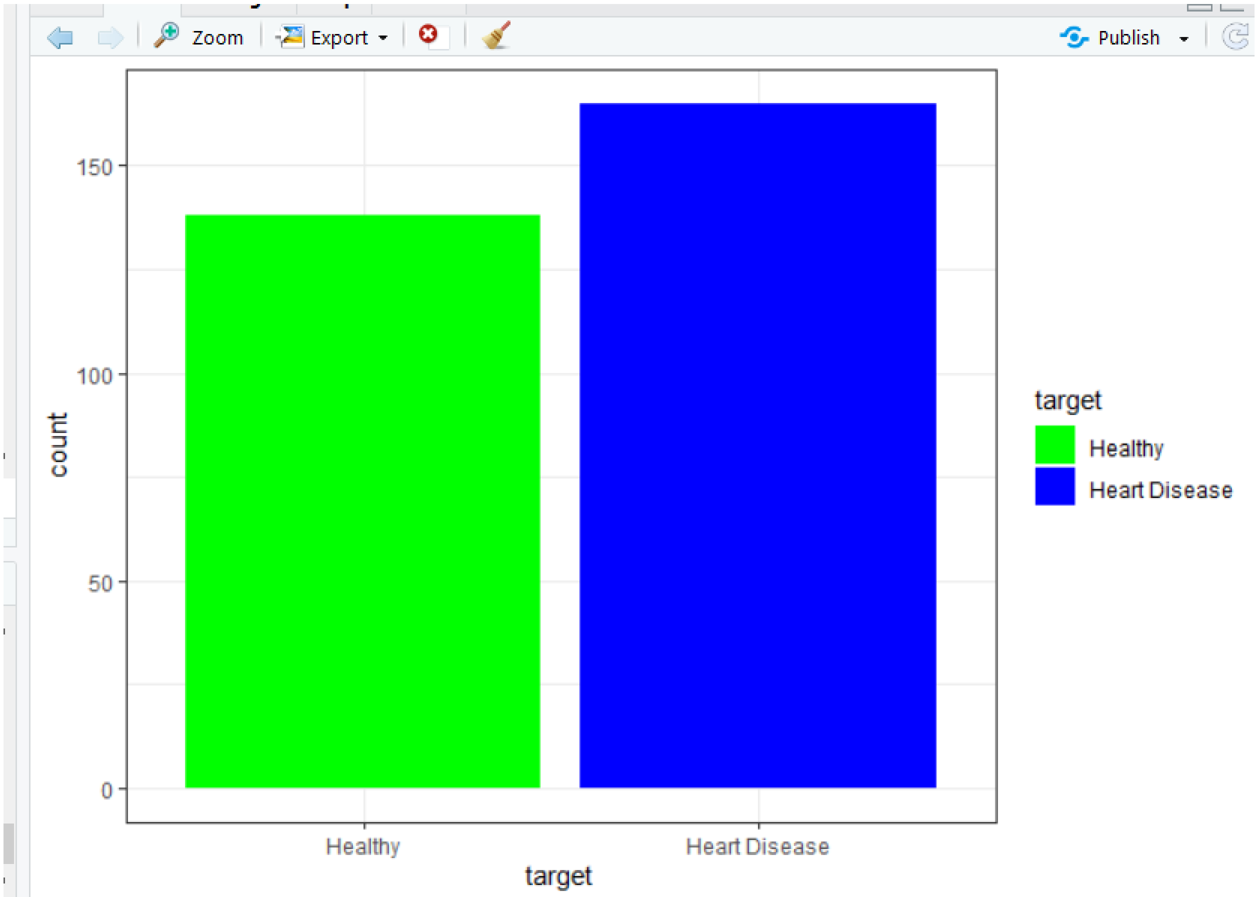
```
      age             sex                         cp          trestbps          chol      
 Min.   :29.00   Female: 96   Chest Pain Type 0:143   Min.   : 94.0   Min.   :126.0  
 1st Qu.:47.50   Male  :207   Chest Pain Type 1: 50   1st Qu.:120.0   1st Qu.:211.0  
 Median :55.00                Chest Pain Type 2: 87   Median :130.0   Median :240.0  
 Mean   :54.37                Chest Pain Type 3: 23   Mean   :131.6   Mean   :246.3  
 3rd Qu.:61.00                                        3rd Qu.:140.0   3rd Qu.:274.5  
 Max.   :77.00                                        Max.   :200.0   Max.   :564.0  
                        fbs              restecg        thalach                                  exang    
 Fasting Blood Sugar <= 120:258   Rest ECG 0:147   Min.   : 71.0   No Exercise Induced Angina:204  
 Fasting Blood Sugar > 120 : 45   Rest ECG 1:152   1st Qu.:133.5   Exercise Induced Angina   : 99  
                                  Rest ECG 2:  4   Median :153.0                                   
                                                   Mean   :149.6                                   
                                                   3rd Qu.:166.0                                   
                                                   Max.   :202.0                                   
    oldpeak                            slope            ca        
 Min.   :0.00   Peak Excercise ST Slope 0: 21   Min.   :0.0000  
 1st Qu.:0.00   Peak Excercise ST Slope 1:140   1st Qu.:0.0000  
 Median :0.80   Peak Excercise ST Slope 2:142   Median :0.0000  
 Mean   :1.04                                   Mean   :0.7294  
 3rd Qu.:1.60                                   3rd Qu.:1.0000  
 Max.   :6.20                                   Max.   :4.0000  
                            thal                  target   
 No Thalassemia               :  2   Healthy      :138  
 Normal Thalassemia           : 18   Heart Disease:165  
 2                            :166                      
 Reversible Defect Thalassemia:117                      
```

Exploratory Analysis

Performing analysis both in R and Power Bi helps to visualize important variables and their indications.

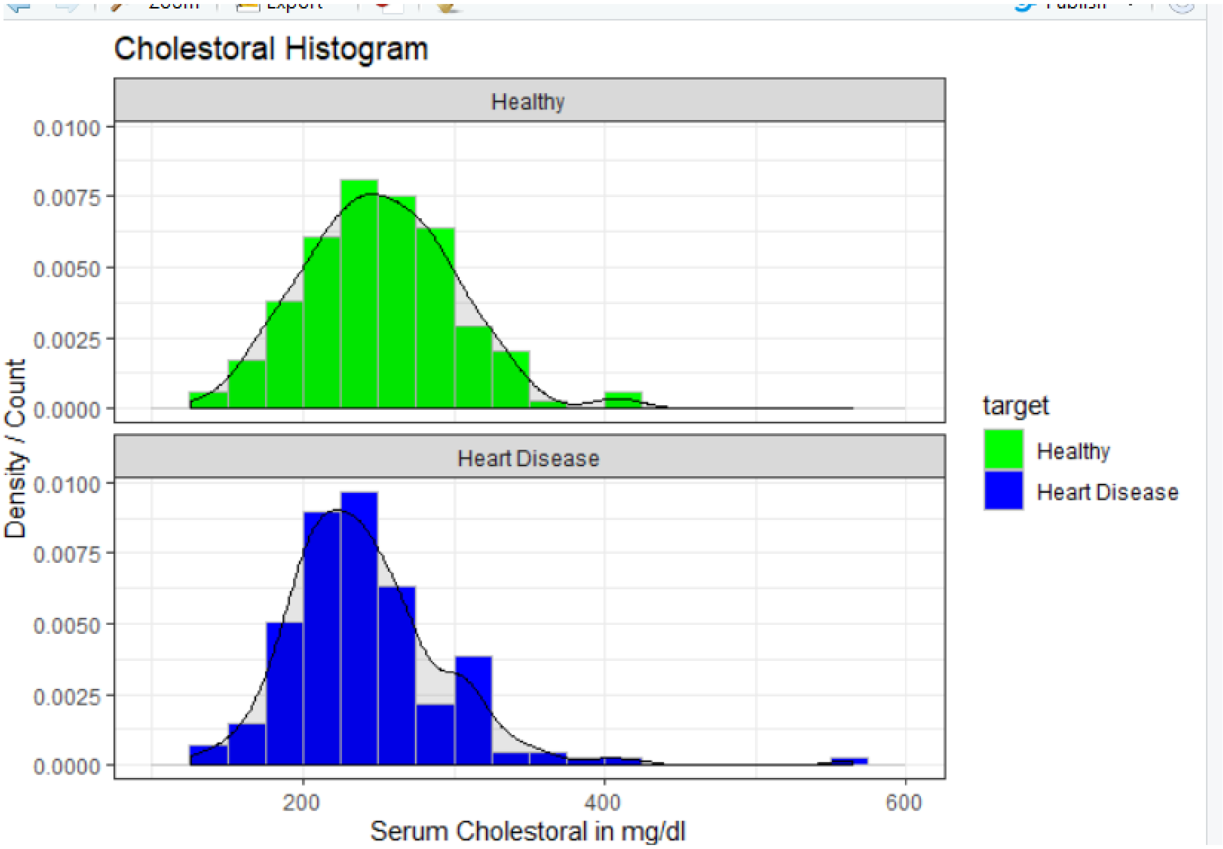
1.Plotting bar graph of target based on number of observations

Analysis: From the graph the dataset has more patients having heart disease



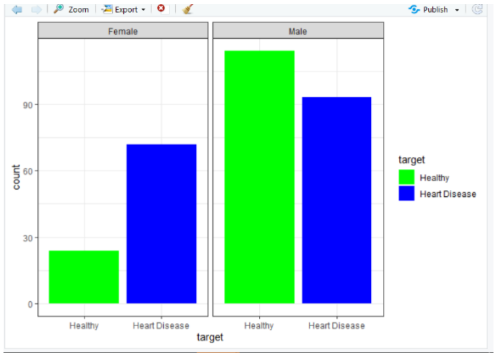
2.Histogram of Cholesterol v/s count

Analysis: Heart Disease patient seems to have more cholesterol between 200- 250mg/dl



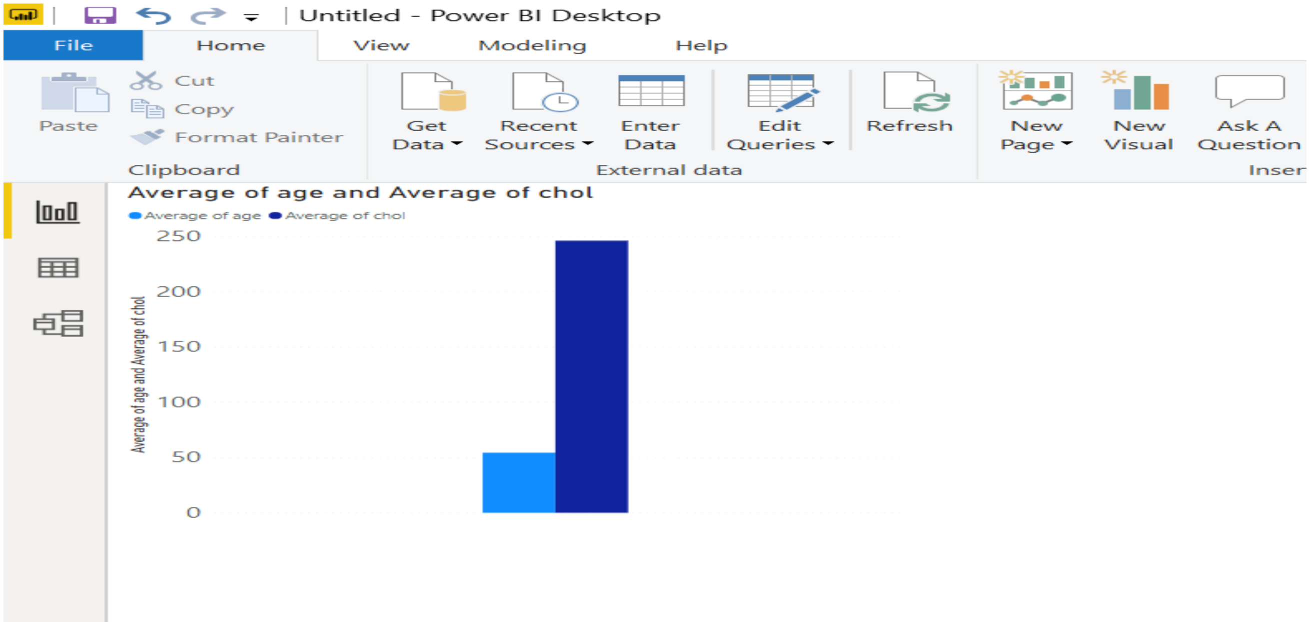
3. Plotting Gender based on Target variable

Analysis: Males are diagnosed with heart disease compared to females



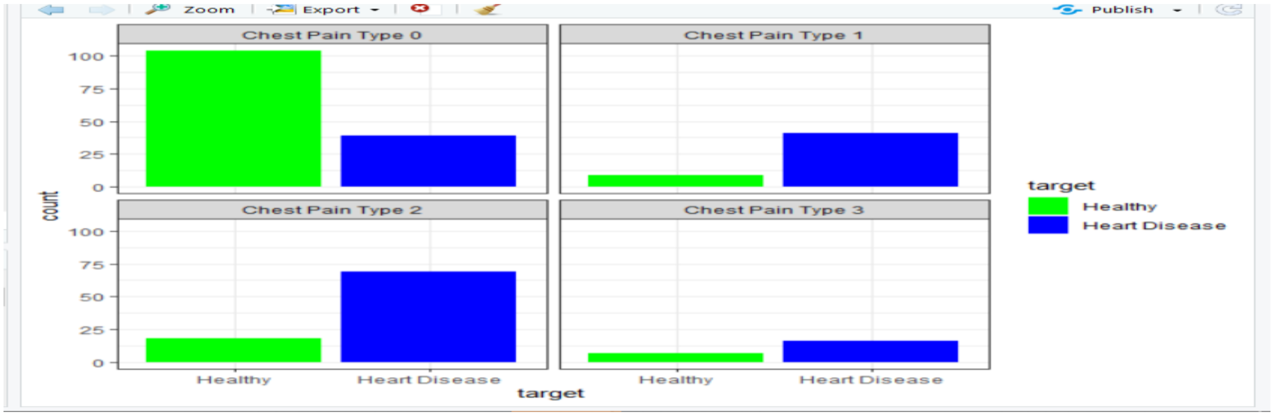
4. Average age v/s Cholesterol

Analysis: Here it gives count of average age which is above 40 and cholesterol which is 245 mg/dl in this dataset



5.Chest pain and target variable

Analysis: Chest Pain type 0 has great impact compared to other types. This can be a leading factor in occurrence of heart diseases

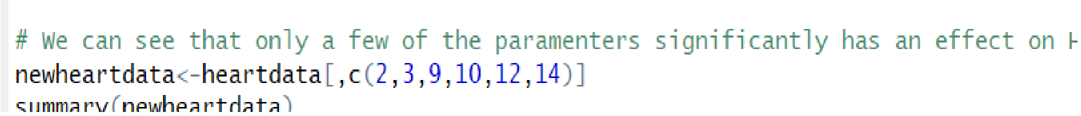


Machine Learning Models

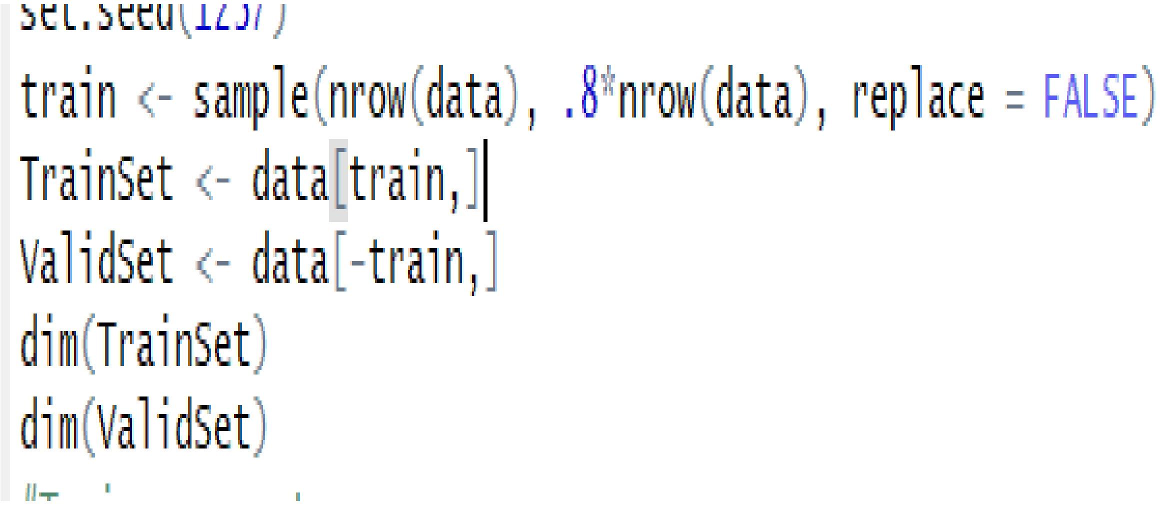
Using 4 Machine Learning Models and comparing the Accuracy

- Logistics Regression
- Random Forest
- Decision Tree
- Neural Network

Before this, based on exploratory analysis, only few variables were having huge impact on heart disease such as Chest pain, depression and hence we subset the dataset and divided the new dataset into train set and Valid set in 80:20 ratio.

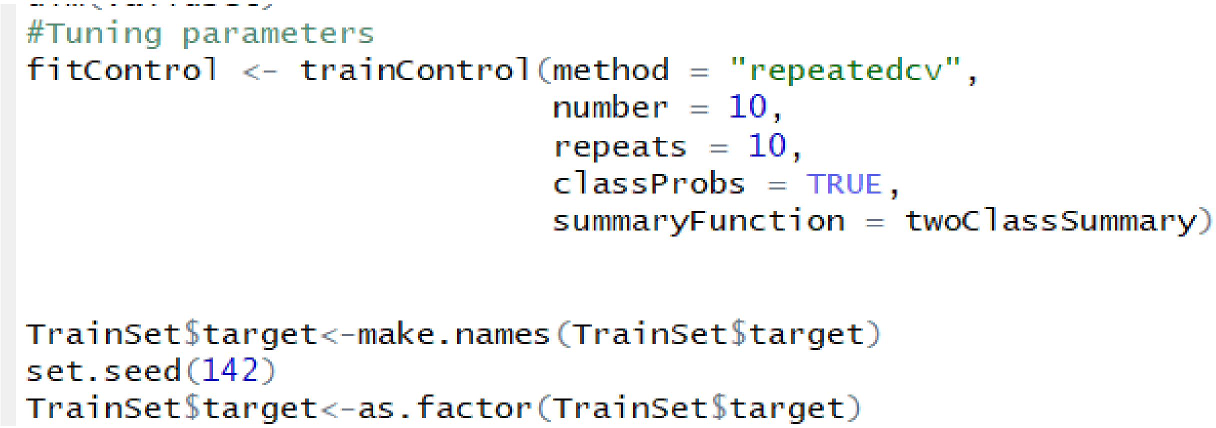


```
# We can see that only a few of the paramenters significantly has an effect on H
newheartdata<-heartdata[,c(2,3,9,10,12,14)]
summary(newheartdata)
```



```
set.seed(1237)
train <- sample(nrow(data), .8*nrow(data), replace = FALSE)
TrainSet <- data[train,]
ValidSet <- data[-train,]
dim(TrainSet)
dim(ValidSet)
```

Tuning the Parameters :

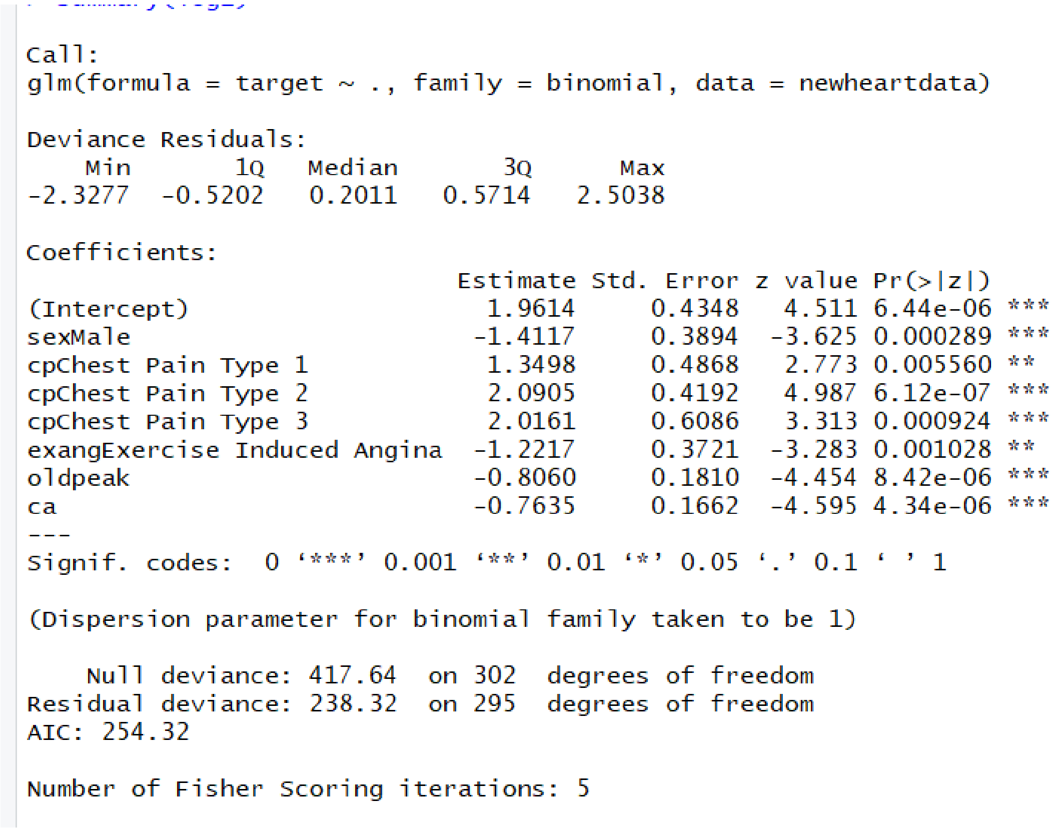


```
#Tuning parameters
fitControl <- trainControl(method = "repeatedcv",
                           number = 10,
                           repeats = 10,
                           classProbs = TRUE,
                           summaryFunction = twoClassSummary)


TrainSet$target<-make.names(TrainSet$target)
set.seed(142)
TrainSet$target<-as.factor(TrainSet$target)
```

1.Logistics Regression

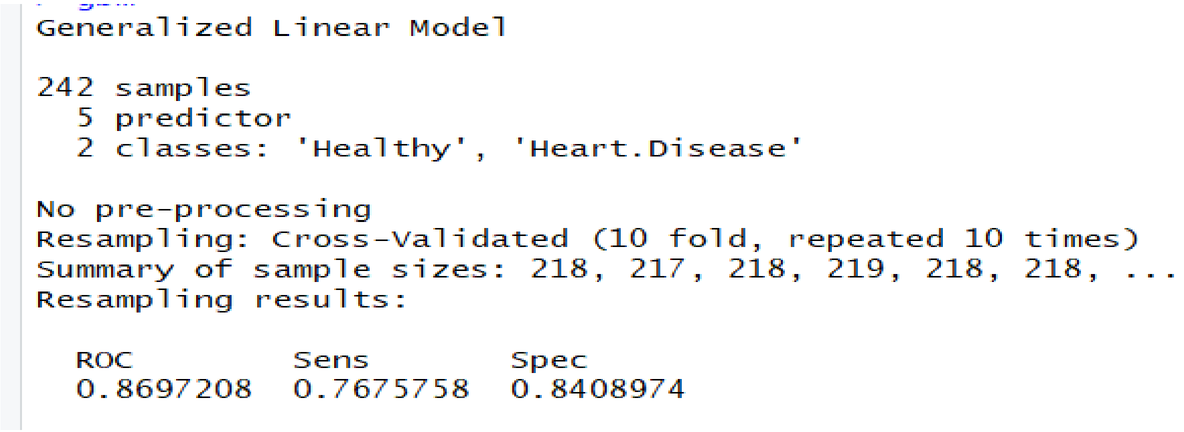
Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables

Running Logistic Regression Model using GLM function gives following summary:



```
Call:
glm(formula = target ~ ., family = binomial, data = newheartdata)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.3277  -0.5202   0.2011   0.5714   2.5038  

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                    1.9614     0.4348   4.511 6.44e-06 ***
sexMale                       -1.4117     0.3894  -3.625 0.000289 ***
cpChest Pain Type 1            1.3498     0.4868   2.773 0.005560 ** 
cpChest Pain Type 2            2.0905     0.4192   4.987 6.12e-07 ***
cpChest Pain Type 3            2.0161     0.6086   3.313 0.000924 ***
exangExercise Induced Angina  -1.2217     0.3721  -3.283 0.001028 ** 
oldpeak                       -0.8060     0.1810  -4.454 8.42e-06 ***
ca                            -0.7635     0.1662  -4.595 4.34e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 238.32  on 295  degrees of freedom
AIC: 254.32

Number of Fisher Scoring iterations: 5
```

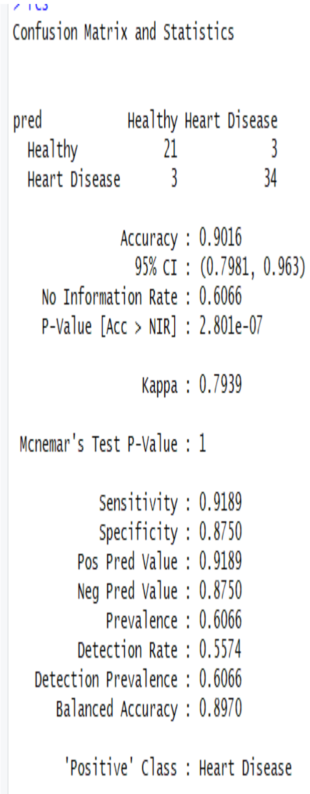
Then, performing Logistic Regression using Train Function In caret Library in order to Find ROC Sensitivity and Specifications:



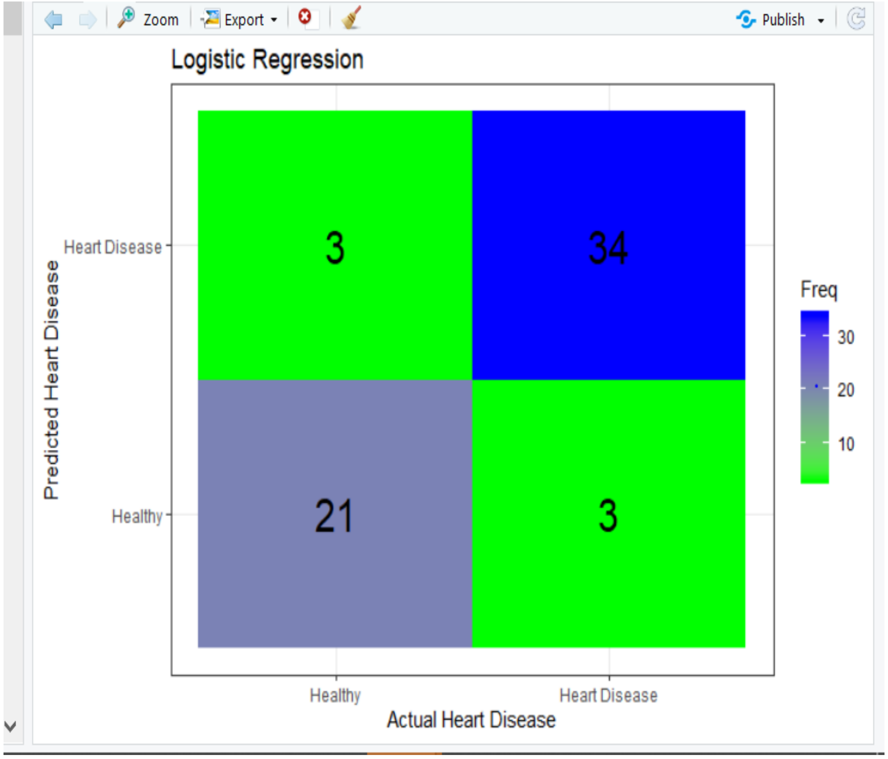
```
Generalized Linear Model 

242 samples
  5 predictor
  2 classes: 'Healthy', 'Heart.Disease' 

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times) 
Summary of sample sizes: 218, 217, 218, 219, 218, 218, ... 
Resampling results:

  ROC        Sens       Spec     
  0.8697208  0.7675758  0.8408974
```

Confusion Matrix Of Logistic Regression:

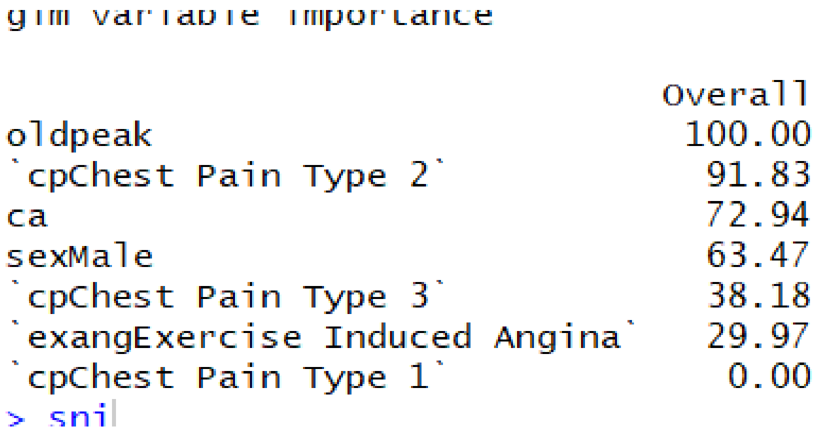
Accuracy was 90.16%



```
> res
Confusion Matrix and Statistics


pred           Healthy Heart Disease
  Healthy           21             3
  Heart Disease      3            34

               Accuracy : 0.9016
                 95% CI : (0.7981, 0.963)
    No Information Rate : 0.6066
    P-Value [Acc > NIR] : 2.801e-07

                  Kappa : 0.7939

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9189
            Specificity : 0.8750
         Pos Pred Value : 0.9189
         Neg Pred Value : 0.8750
             Prevalence : 0.6066
         Detection Rate : 0.5574
   Detection Prevalence : 0.6066
      Balanced Accuracy : 0.8970

       'Positive' Class : Heart Disease
```



Variable Importance in this Model:

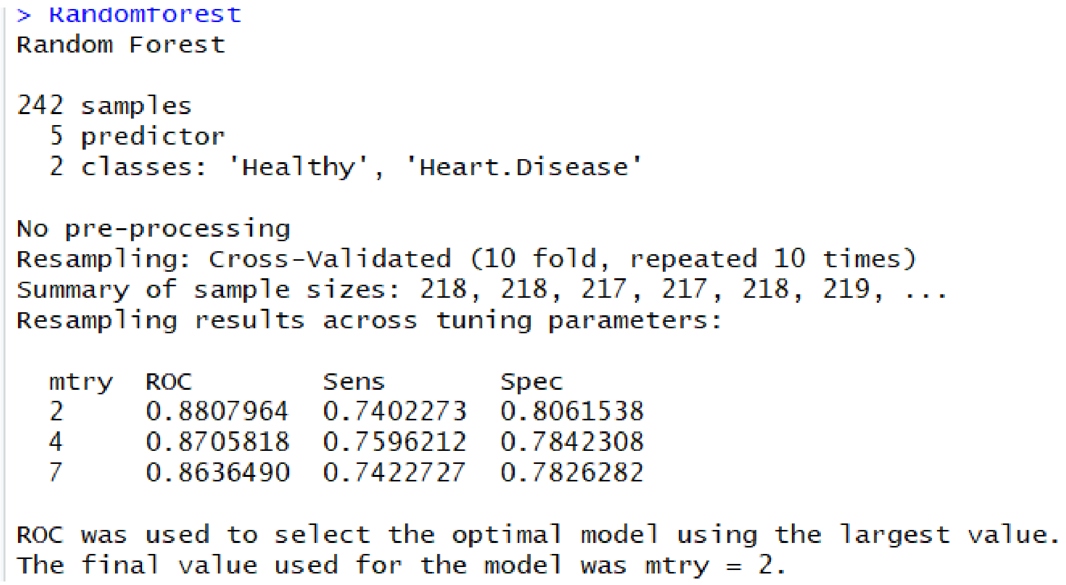


```
glm variable importance

                                   Overall
oldpeak                             100.00
`cpChest Pain Type 2`                91.83
ca                                   72.94
sexMale                              63.47
`cpChest Pain Type 3`                38.18
`exangExercise Induced Angina`       29.97
`cpChest Pain Type 1`                 0.00
> sni
```

2.Random Forest

Random forests is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees

Train Function in Caret Library

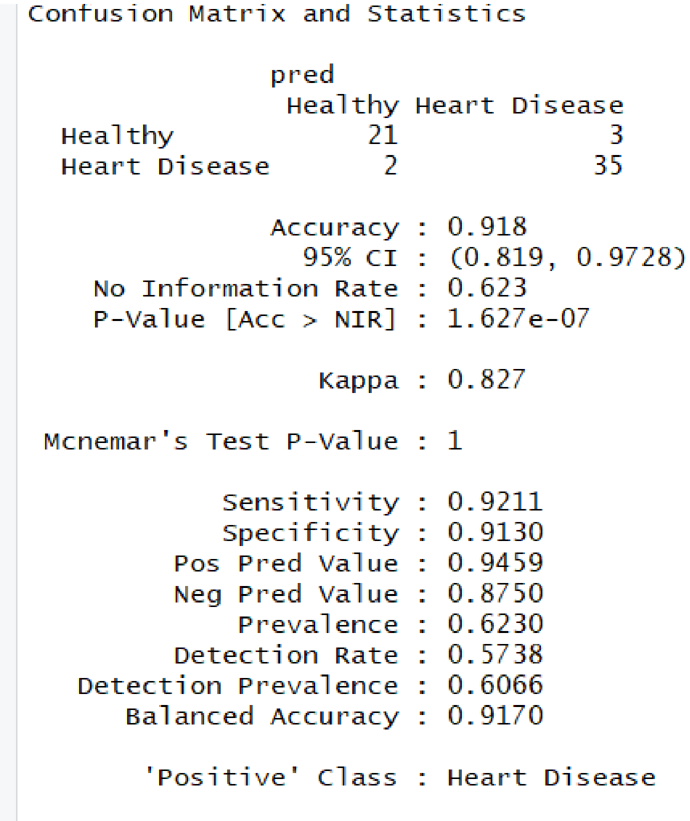


```
> Randomforest
Random Forest

242 samples
  5 predictor
  2 classes: 'Healthy', 'Heart.Disease'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 218, 218, 217, 217, 218, 219, ...
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
  2     0.8807964  0.7402273  0.8061538
  4     0.8705818  0.7596212  0.7842308
  7     0.8636490  0.7422727  0.7826282

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

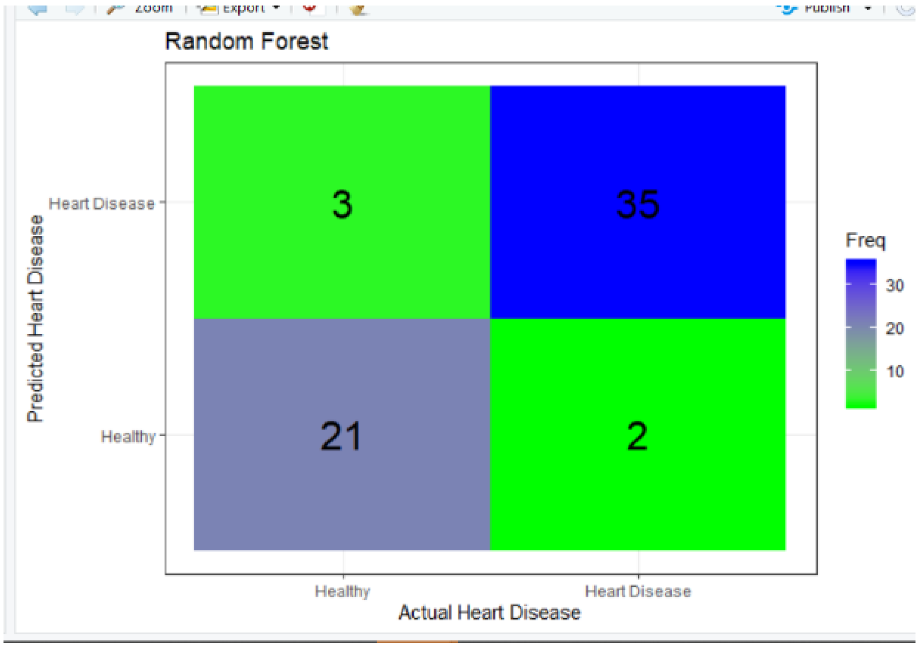
Confusion Matrix

The accuracy we got 91.8%

p-Value is highly significant

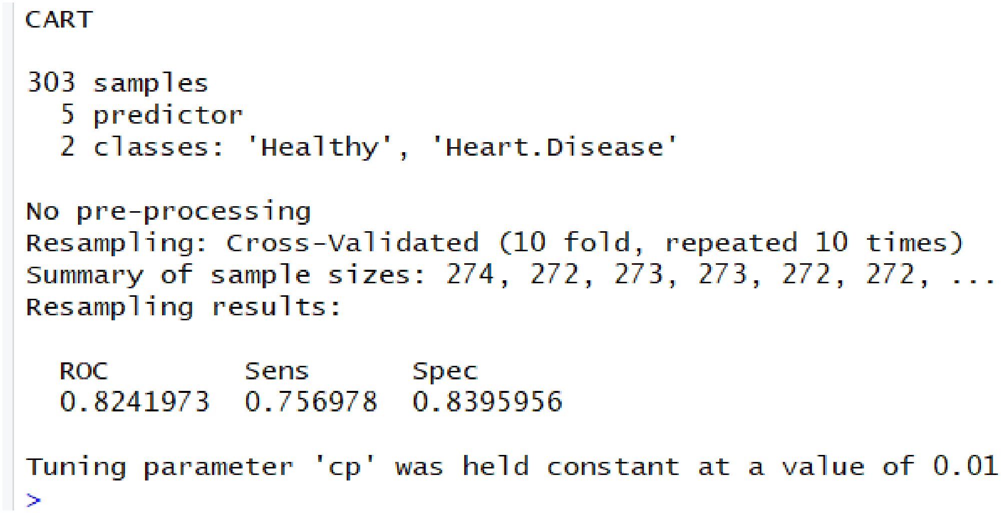


```
Confusion Matrix and Statistics

               pred
                Healthy Heart Disease
  Healthy            21             3
  Heart Disease       2            35

               Accuracy : 0.918
                 95% CI : (0.819, 0.9728)
    No Information Rate : 0.623
    P-Value [Acc > NIR] : 1.627e-07

                  Kappa : 0.827

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9211
            Specificity : 0.9130
         Pos Pred Value : 0.9459
         Neg Pred Value : 0.8750
             Prevalence : 0.6230
         Detection Rate : 0.5738
   Detection Prevalence : 0.6066
      Balanced Accuracy : 0.9170

       'Positive' Class : Heart Disease
```



3.Decision Tree

The Decision Tree Analysis is a schematic representation of several decisions followed by different chances of the occurrence., a tree-shaped graphical representation of decisions related to the investments and the chance points that help to investigate the possible outcomes is called as a decision tree analysis.
Running Decision Tree Roc comes out to be 82%

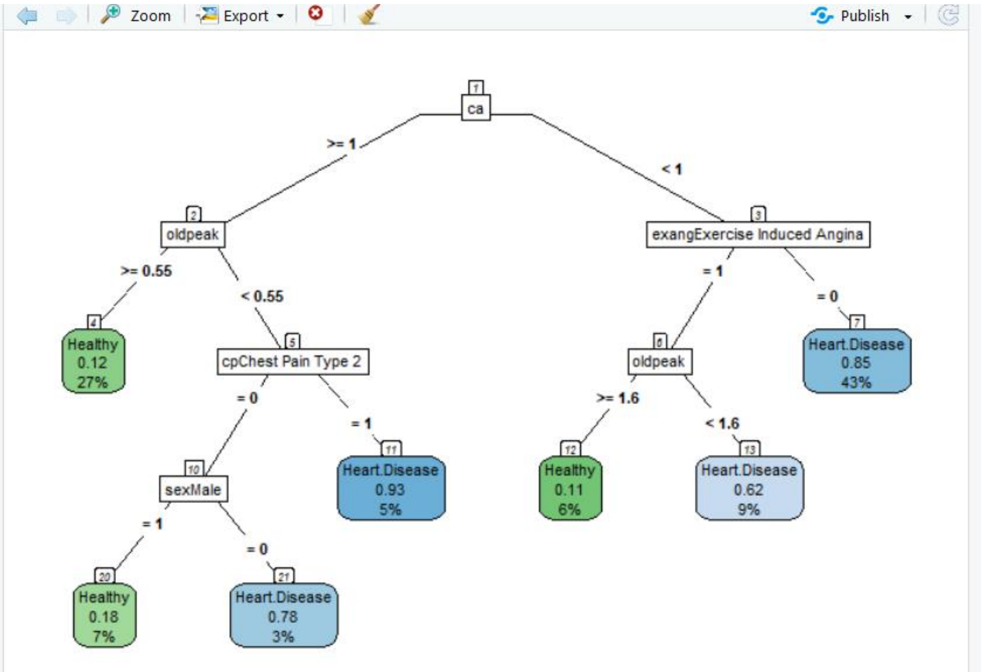


```
CART

303 samples
  5 predictor
  2 classes: 'Healthy', 'Heart.Disease'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 274, 272, 273, 273, 272, 272, ...
Resampling results:

  ROC        Sens      Spec
  0.8241973  0.756978  0.8395956

Tuning parameter 'cp' was held constant at a value of 0.01
>
```

Plotted the decision tree

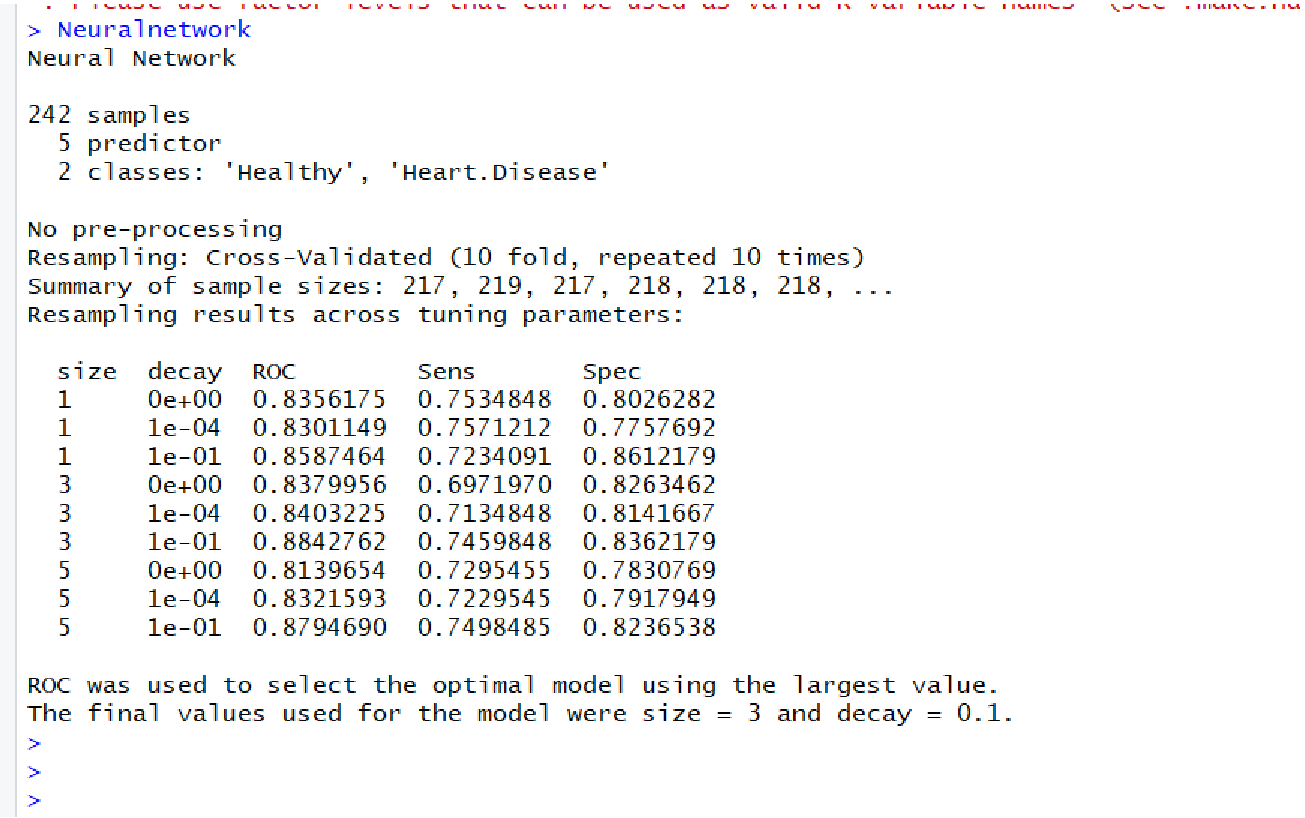
- ca: Number of major vessels (0-3) colored by fluoroscopy
- oldpeak: ST depression induced by exercise relative to rest
- Deeper the red, higher the probability of Heart Disease
- Deeper the green, more the chances of being healthy
- If no. of vessels >= 1 AND ST depression < 0.55 AND Chest Pain Type = 2, then there is a 93% chance of Heart Disease

Similarly doctors can take a decision based on these parameters whether there is a chance of Heart Disease in the future



4.Neural Network

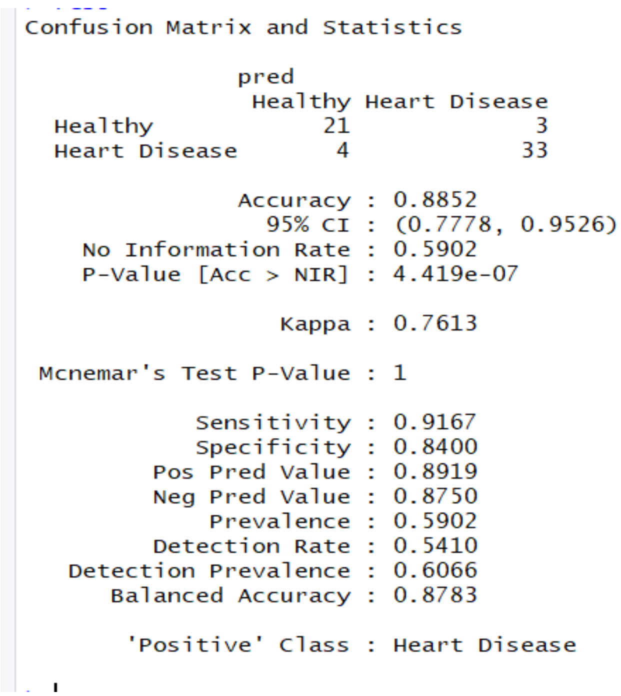
Train set in Caret Function



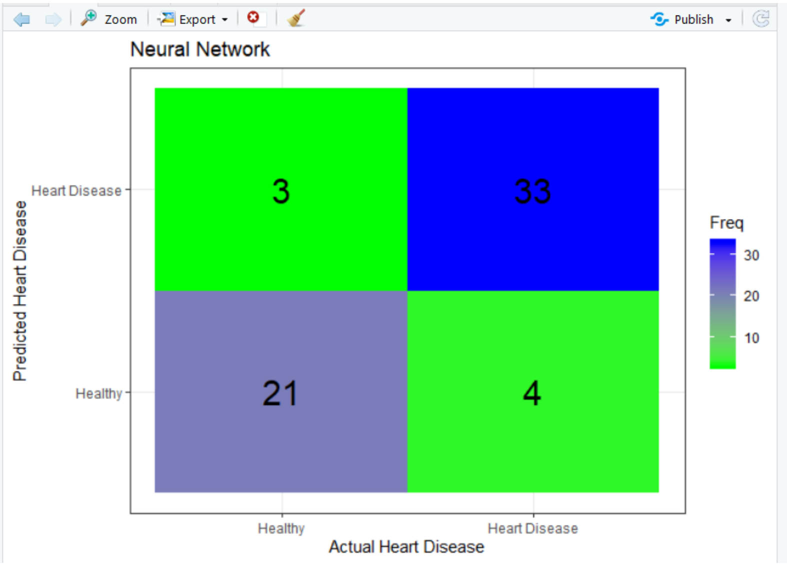
```
 : Please use factor levels that can be used as valid R variable names  (see ?make.na
> Neuralnetwork
Neural Network 

242 samples
  5 predictor
  2 classes: 'Healthy', 'Heart.Disease' 

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times) 
Summary of sample sizes: 217, 219, 217, 218, 218, 218, ... 
Resampling results across tuning parameters:

  size  decay  ROC        Sens       Spec     
  1     0e+00  0.8356175  0.7534848  0.8026282
  1     1e-04  0.8301149  0.7571212  0.7757692
  1     1e-01  0.8587464  0.7234091  0.8612179
  3     0e+00  0.8379956  0.6971970  0.8263462
  3     1e-04  0.8403225  0.7134848  0.8141667
  3     1e-01  0.8842762  0.7459848  0.8362179
  5     0e+00  0.8139654  0.7295455  0.7830769
  5     1e-04  0.8321593  0.7229545  0.7917949
  5     1e-01  0.8794690  0.7498485  0.8236538

ROC was used to select the optimal model using the largest value.
The final values used for the model were size = 3 and decay = 0.1.
> 
> 
> 
```

Confusion Matrix

Accuracy:88.5%



```
Confusion Matrix and Statistics

                pred
                 Healthy Heart Disease
  Healthy             21             3
  Heart Disease        4            33

               Accuracy : 0.8852
                 95% CI : (0.7778, 0.9526)
    No Information Rate : 0.5902
    P-Value [Acc > NIR] : 4.419e-07

                  Kappa : 0.7613

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9167
            Specificity : 0.8400
         Pos Pred Value : 0.8919
         Neg Pred Value : 0.8750
             Prevalence : 0.5902
         Detection Rate : 0.5410
   Detection Prevalence : 0.6066
      Balanced Accuracy : 0.8783

       'Positive' Class : Heart Disease
```



Conclusions

- According to the above analysis Random Forest (92% accuracy based on confusion matrix) is the best models and gives better predictive accuracy compared to the other 3 models.
- Confusion Matrix also stated that Positive Class: Heart Disease Patients to be more than healthy patients (Target condition).
- Variable importance function states that following variables such as oldpeak, chest pain Gender (Male) and exercise induced Angina plays a vital role in analysis and can be a major impact on analysis.

Challenges

- Neural Network I used nnet in caret function
- I tried using Neural Network library but while plotting it took a longer execution time as indicated that data frame should be in matrix form it was bit challenging to achieve this I was not able to plot the Neural Network.
- Instead I used NNs via the Caret Package for this analysis
- There are 3 types you can use neural Network package in R
    - Using package neuralnet
    - Using package nnet
    - Using NNs via the Caret package

Future Scope

- In future, Machine learning Algorithms should be created with more advanced feature in predicting accuracy with more accurate results which will allow the patient to take immediate care in low cost.
- Machine Learning algorithm can have a great impact on huge database available in health care sector and make an immediate decision and care in favor of patients, delay in treatments can be avoided and can help save many lives.

References

- **https://s3.amazonaws.com/academia.edu.documents/34837848/V1I8_IJERTV1IS8282.pdf?response-content-disposition=inline%3B%20filename%3DV1I8_IJERTV1IS8282.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20191201%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20191201T223805Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=61147d2db91f6c2524e204ea0dd6422cf73f888475df444578a5e923d79977bb**
- **https://en.wikipedia.org/wiki/Cardiovascular_disease**
- **https://en.wikipedia.org/wiki/Category:Heart_diseases**
- **http://www.di.fc.ul.pt/~jpn/r/neuralnets/neuralnets.html**
- **https://datascienceplus.com/neuralnet-train-and-test-neural-networks-using-r/**
- http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/
- **https://www.datacamp.com/community/tutorials/decision-trees-R**
- **https://www.tutorialspoint.com/r/r_functions.htm**
- https://archive.ics.uci.edu/ml/datasets/Heart+Disease