20th August, 2022

Heart Diseases Analysis R

Ocansey Kevin

Submitted to HarvardX (PH125.9X)

Abstract

For our final project we were asked to choose our own project. I decided to take on something simple but I also wanted to explore the medical sector, to be able to analyze a dataset that I might not completely understand in terms of terminology and technicality, which will prevent me from inferring any basic knowledge I have into my analysis. A leap into the unknown will make me focus only on patterns I find while analysing the data and also learn something new.

Signing up to Kaggle I chanced on the heart disease data set and decided to use it for my project.

INTRODUCTION

Heart disease describes numerous conditions that affect the heart, these include irregular heartbeats, heart problems that one is born with, diseases that affect blood vessels etc. From these conditions arise several types of heart diseases example coronary artery disease (CAD), heart arrhythmias. heart valve disease, pericardial disease, cardiomyopathy (Heart Muscle Disease), congenital heart disease. The most common is coronary artery disease.

The main blood veins that nourish the heart muscle are affected by coronary artery disease, a common cardiac disorder. The most common cause of coronary artery disease is cholesterol. Atherosclerosis, which refers to the development of these plaques (ath-ur-o-skluh-ROE-sis) occurs in the arteries of the heart. Because of this blood flow to the heart and other organs is decreased. A heart attack, angina, or a stroke may result from it.

Male and female symptoms of coronary artery disease may vary. Men are more prone to experience chest pain, for example. In addition to chest tightness, women are more prone to experience other symptoms such severe exhaustion, nausea, and shortness of breath.

Other heart diseases occur even while in a mother's womb. One month after conception, the baby's heart begins to form, and this is when a congenital heart abnormality develops. Heart blood flow is altered by congenital heart abnormalities. Congenital cardiac problems are more likely due to certain illnesses, drugs, and genetic factors. Moving on, there are some heart diseases that are caused by infections. This occurs when germs reach the heart or heart valves. They are usually bacteria or viruses, transmitted by parasites.

There are certain factors that increase the risk of heart disease. They include:

- **Age** : Growing older increases the risk of damaged and narrowed arteries and a weakened or thickened heart muscle.
- **Sex:** Men are at greater risk of having heart disease, where as for women, the risk increases after menopause
- **Diet:** Heart disease has been related to diets heavy in cholesterol, fat, salt, and sugar.
- **High Blood Pressure:** Uncontrolled high blood pressure can cause the arteries to become hard and thick. These changes interrupt blood flow to the heart and body.
- **Cholesterol:** High cholesterol increases risk of atherosclerosis,
- **Obesity**
- **Poor Dental health:** Unhealthy teeth and gums makes it easier for germs to enter the body
- **Stress:** Stress may damage the heart arteries
- **Diabetes**

1.1

OVERVIEW

The goal of this project is to develop an algorithm for predicting heart disease in patients.

Since there are several types of heart diseases, we can simply say our goal is to point out a healthy patient from an unhealthy patient, without specifically specifying which disease they might have.

To get started we take a look at the heart.cvs dataset I downloaded from Kaggle. (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset).

It is composed of 14 attributes, with the predicted attribute included.

1. Age
2. Sex
3. Cp : Chest pain type (4 values). Where 0 = Asymptomatic Pain; 1 = Typical Angina Pain; 2 = Atypical Angina Pain; 3 = Non-Angina Pain.
4. Trestbps : Resting blood pressure measured in mm Hg on admission to the hospital.
5. Chol : Serum cholesterol measured in mg/dl
6. Fbs : Fasting blood sugar measured in mg/dl
7. Restecg : Resting electrocardiographic results (values 0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes)
8. Thalach : Maximum heart rate achieved (bpm)
9. Exang : Exercise induced angina
10. Oldpeak : ST depression induced by exercise relative to rest
11. Slope : The slope of the peak exercise ST segment; 1 = up-slopping 2 = flat 3 = down-slopping
12. Ca : number of major vessels (0-3) coloured by fluoroscopy
13. Thal : Where 0 = normal; 1 = fixed defect; and 2 = reversable defect
14. Target : This represents the predicted attribute - diagnosis of heart disease (angiographic disease status) value 0 <= 50% diameter narrowing; and value 1 => 50% diameter narrowing

Overall, there were 1025 observation. 69.56% were Males, while 30.43% were Females.

2.1

EXPLORATORY ANALYSIS

Cleaning the Data Set

Also, according to the description of the data set, ‘Ca’ should have 4 different categories (0-3), however on further inspection of the dataset we notice a fifth category (4). This was the same for the ‘thal’ attribute (instead of distinct categories from 1 to 3, It had its levels beginning from 0 to 3). Thankfully the number of discrepancies were few hence, i decided to remove these errors all together.

Hence a total of 25 rows of data were filtered out.

Summary

After cleaning the data

- A total of 1000 patients
- We have a mean age of 54 years, with the minimum being 29 and maximum 77.
- For Sex we have 691 Males (69.1%) and 309 Females (30.1%).
- Also, our target shows that 508 (50.8 %) actually have heart disease whiles 492 patients (49.2%) do not

Before carrying on it is very important to point out the prevalence of Males in our study. I anticipate that it could have an influence on our algorithm.

2.2

VISUALIZATION

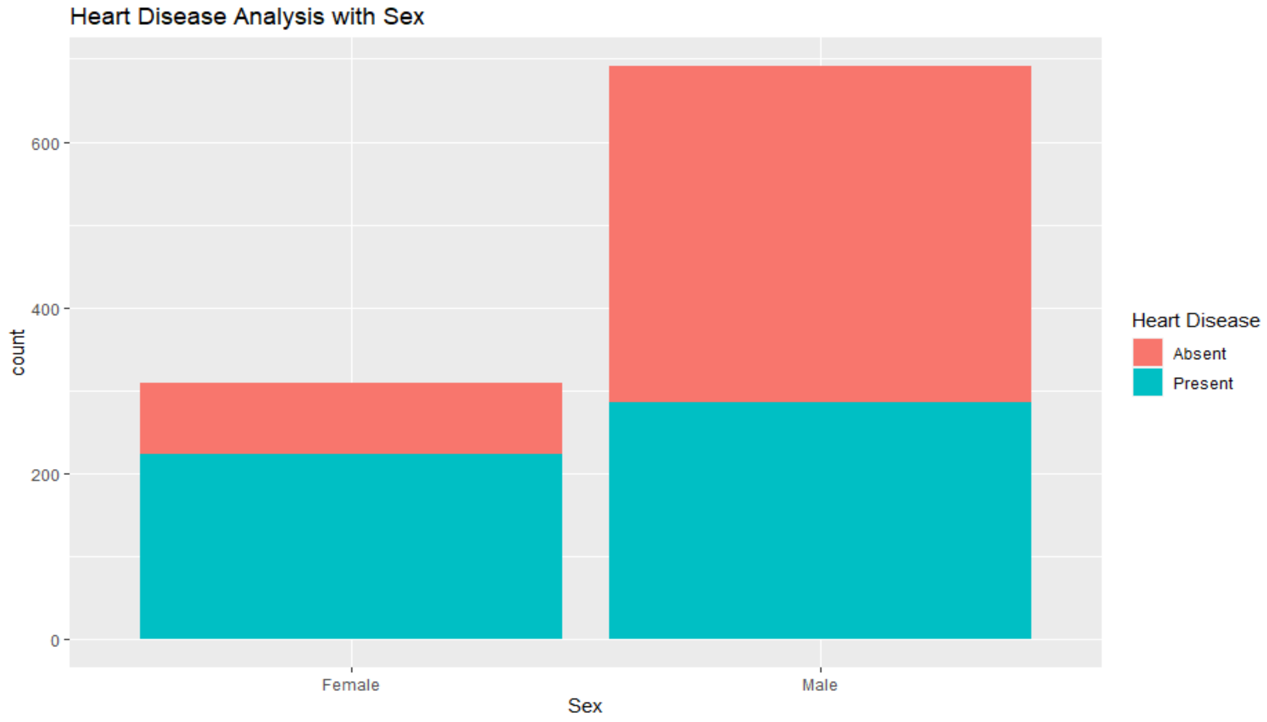


Fig a

For ‘fig b’ below, all patients between the ages of 24 and 64 fall under the category ‘Adult’ and those who are 64 and above are considered Elderly. Interesting to note that although Males are prevalent, according to this dataset Females are more likely to have heart diseases even across the groups of ages stratified.

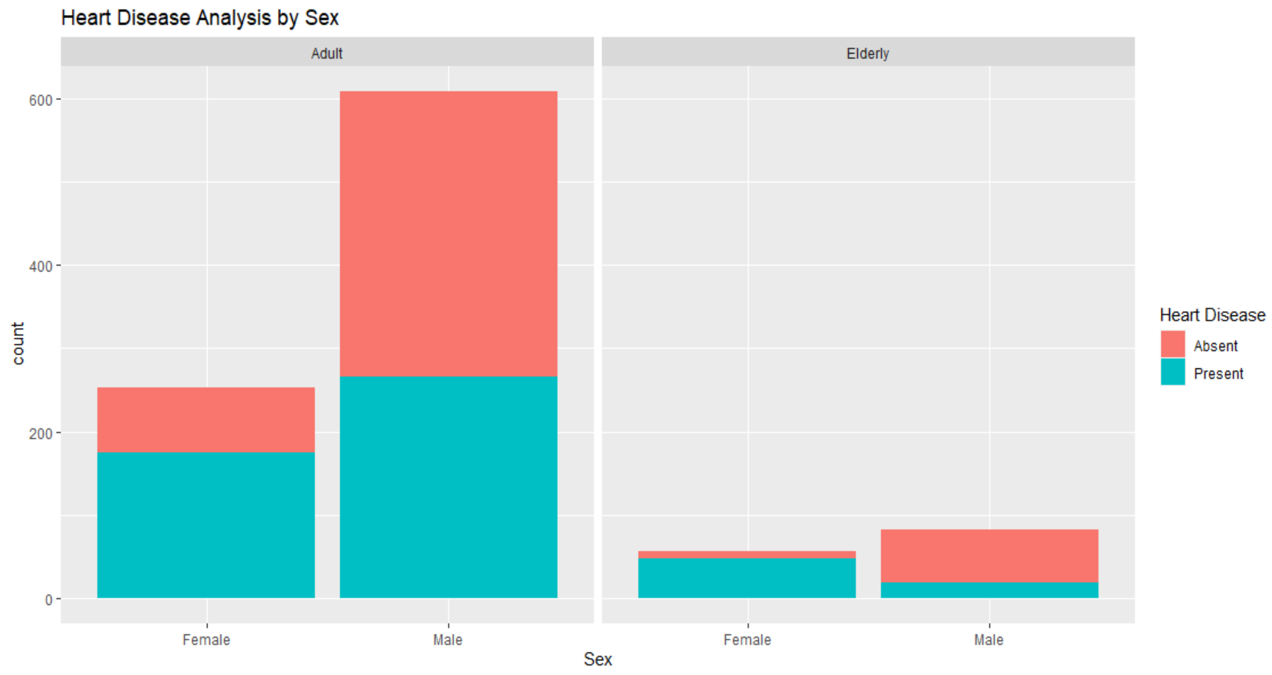


Fig b

With this graph we will later build a simple model to predict patients with heart disease using just their ages and/or sex.

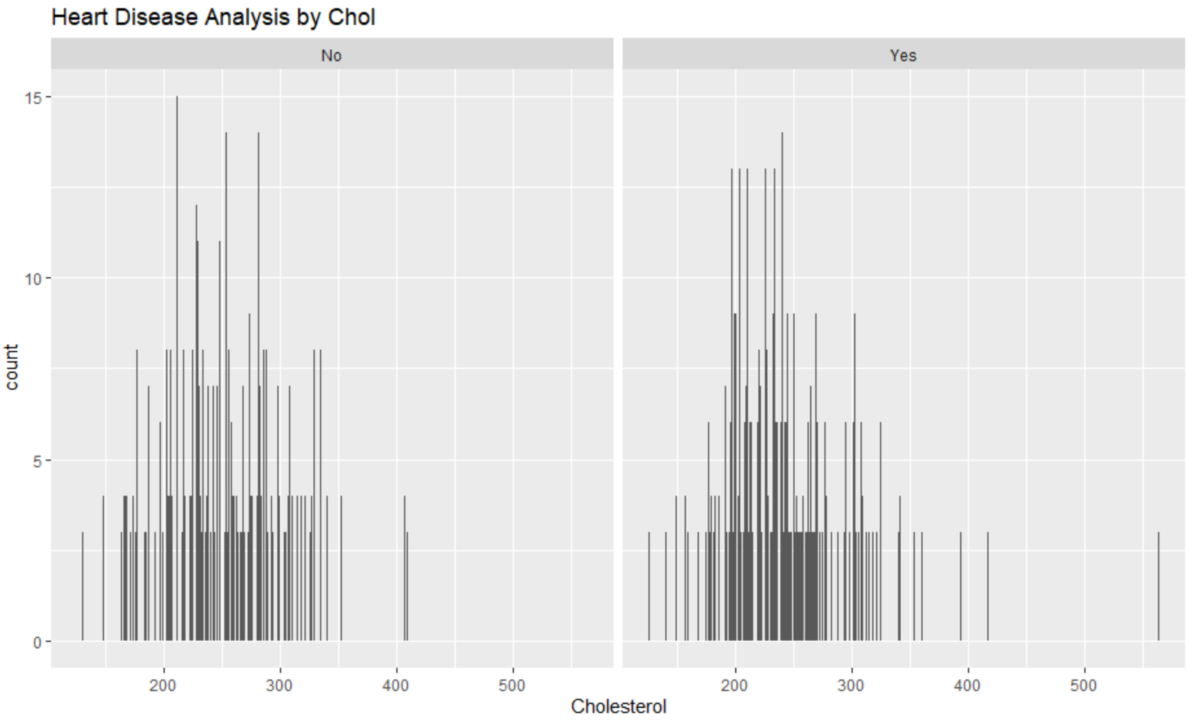

Heart Disease Analysis by Chol
No
Yes
count
15
10
5
0
200
300
400
500
Cholesterol

Fig c

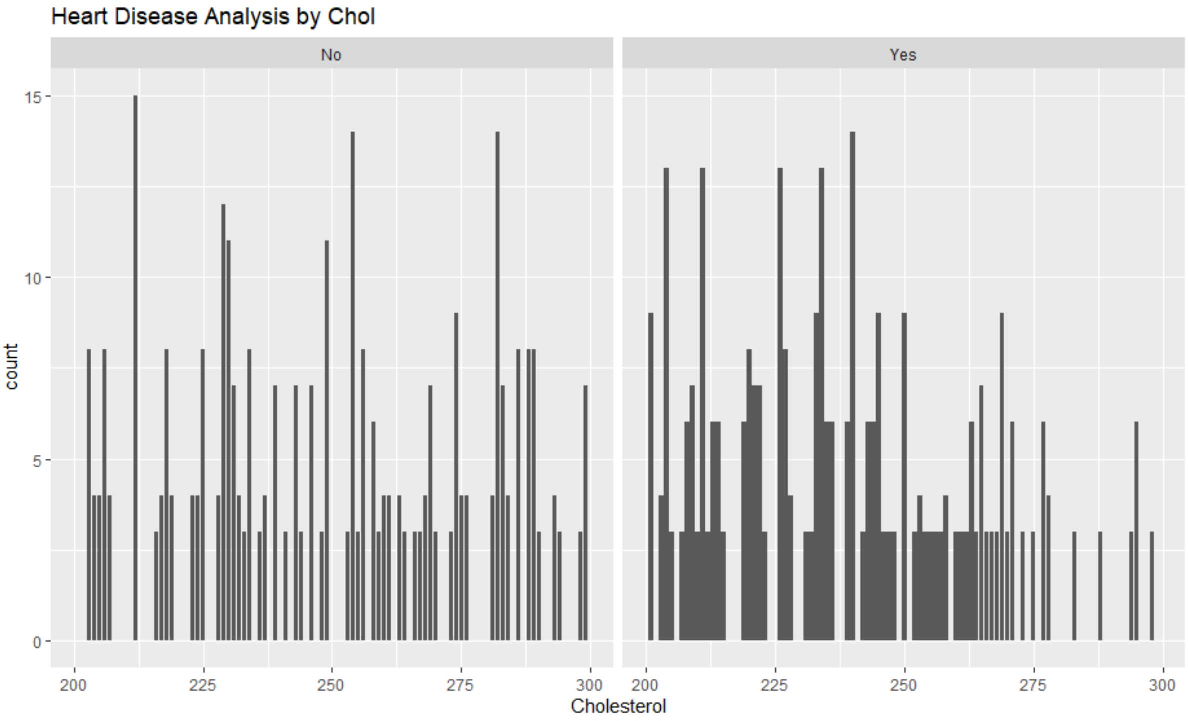

Heart Disease Analysis by Chol
No
Yes
count
15
10
5
0
200
225
250
275
300
Cholesterol

Fig d

In fig c and fig d there seems to be little to no correlation between cholesterol and heart disease. Also, when we zoom in, between 200 and 300 mg/dl, where there seems to be more heart disease patients between that range, cumulatively we learn that the difference isn’t that significant.

For the range 200 to 300 mg/dl cholesterol levels, the number of patients with heart disease are 346 where as those without are 340.

If you reduce the range, to let’s say 200 to 275 mg/dl we get 315 heart diseases patients and 259 patients without any heart complication. However, that would look like deliberately fishing for a correlation that is not there. According to research, cholesterol levels above 240 mg/dl is considered unhealthy. Maybe there is a correlation between high cholesterol and heart diseases but only if other factors are considered.

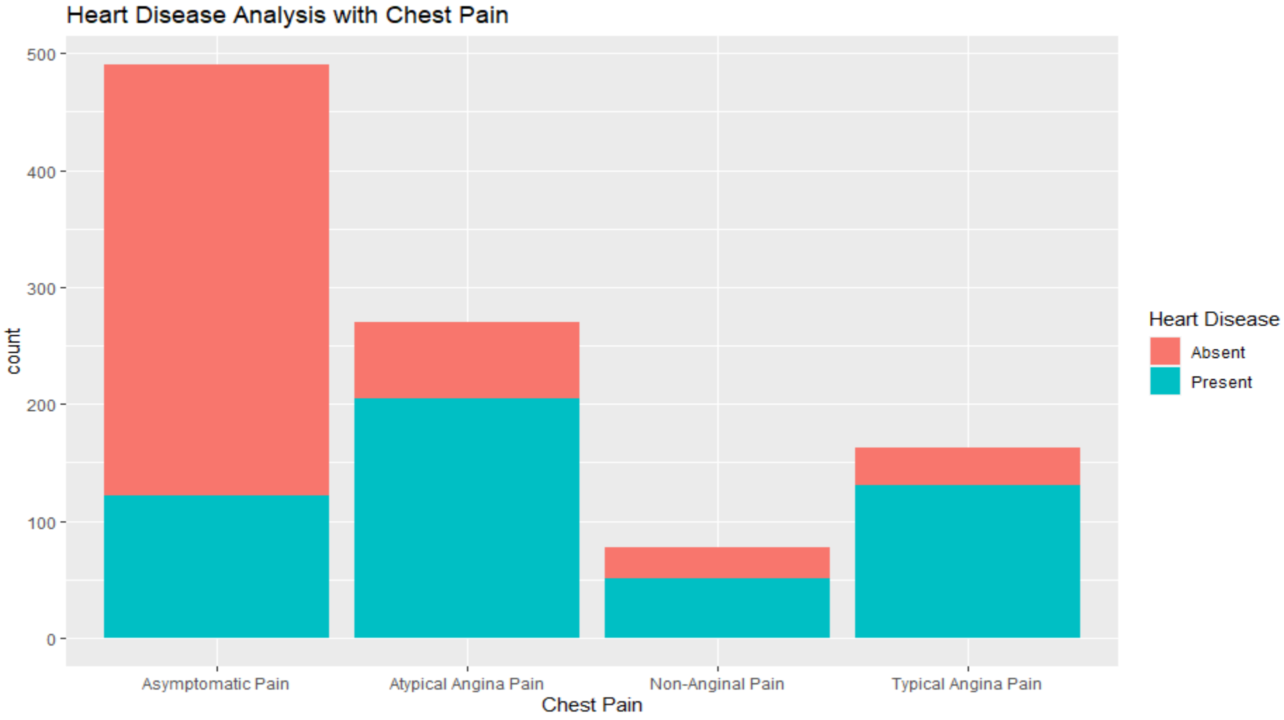


Fig e

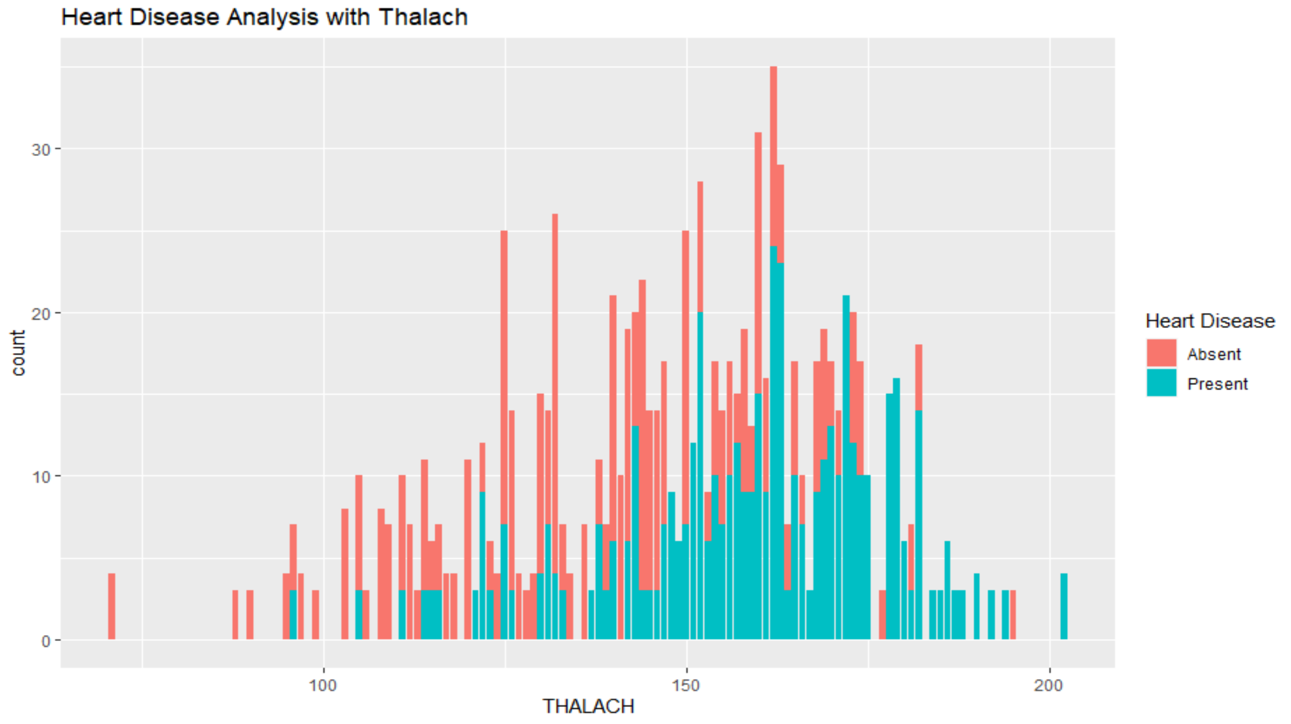

Heart Disease Analysis with Thalach
count
0
10
20
30
100
150
200
THALACH
Heart Disease
Absent
Present

Fig f

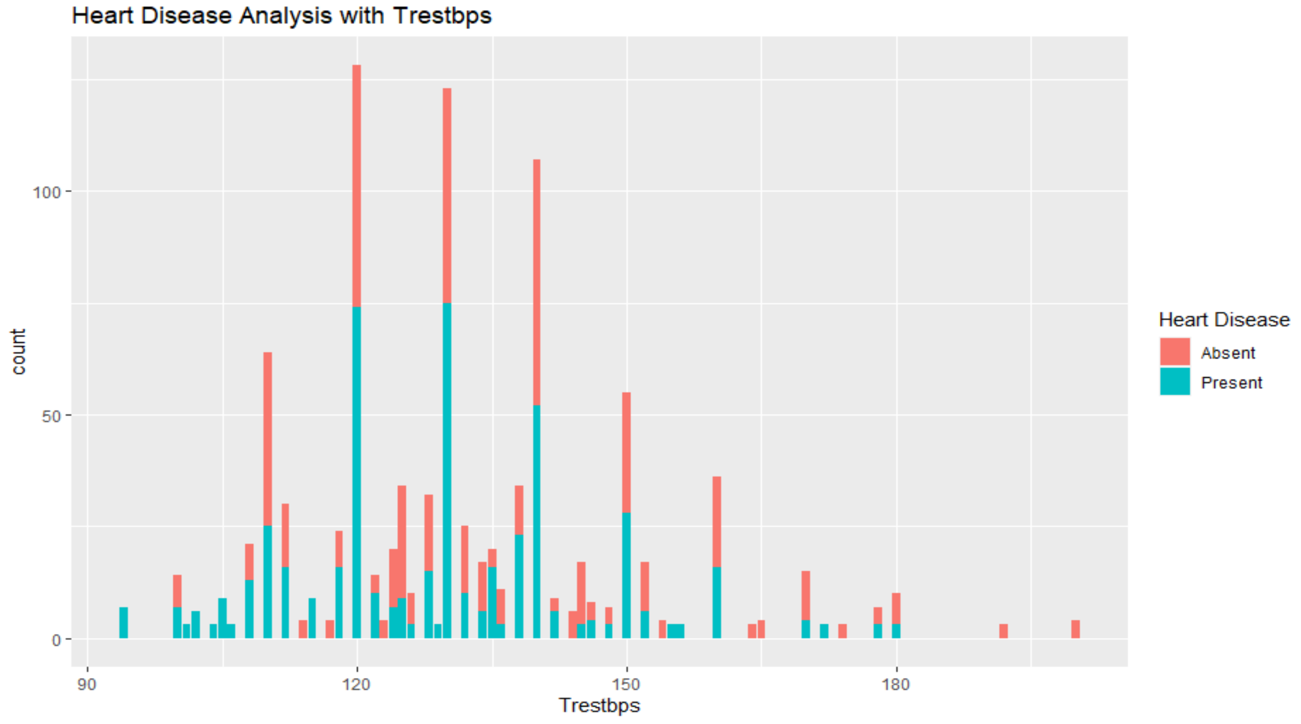

Heart Disease Analysis with Trestbps
count
0
50
100
90
120
150
180
Trestbps
Heart Disease
Absent
Present

Fig g

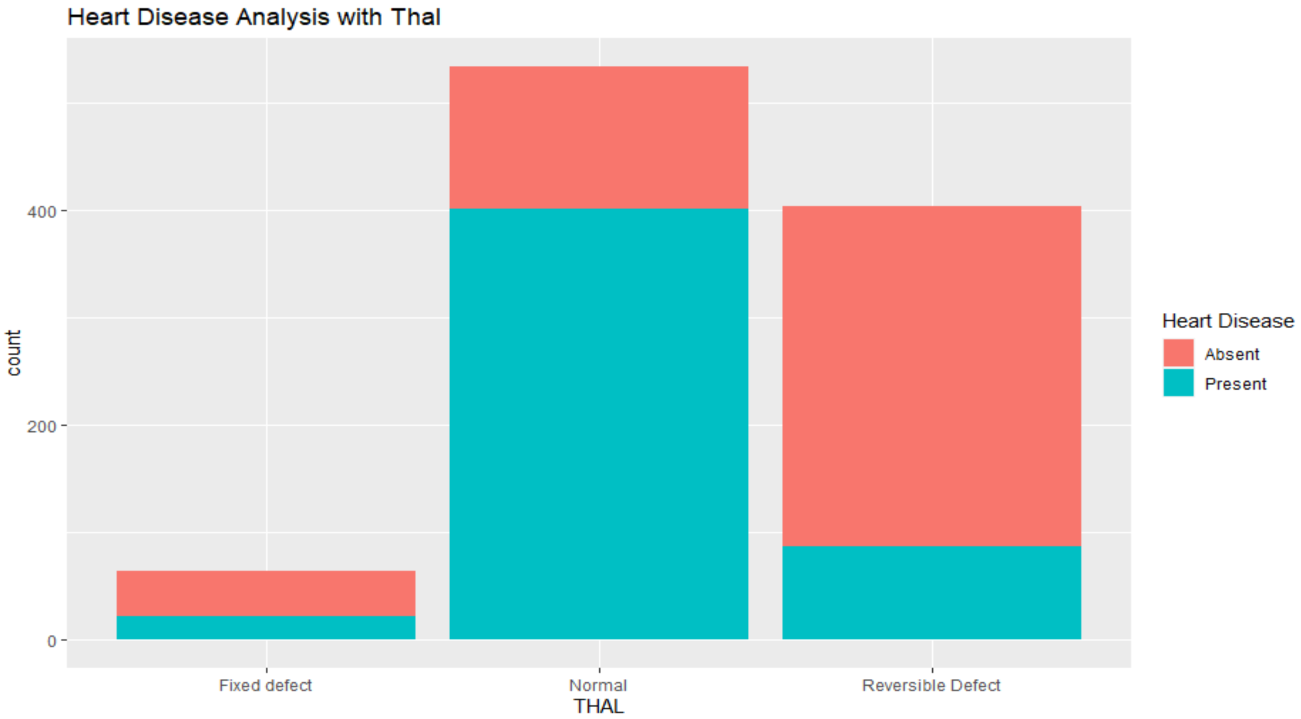

Heart Disease Analysis with Thal
count
0
200
400
Fixed defect
Normal
Reversible Defect
THAL
Heart Disease
Absent
Present

Fig h

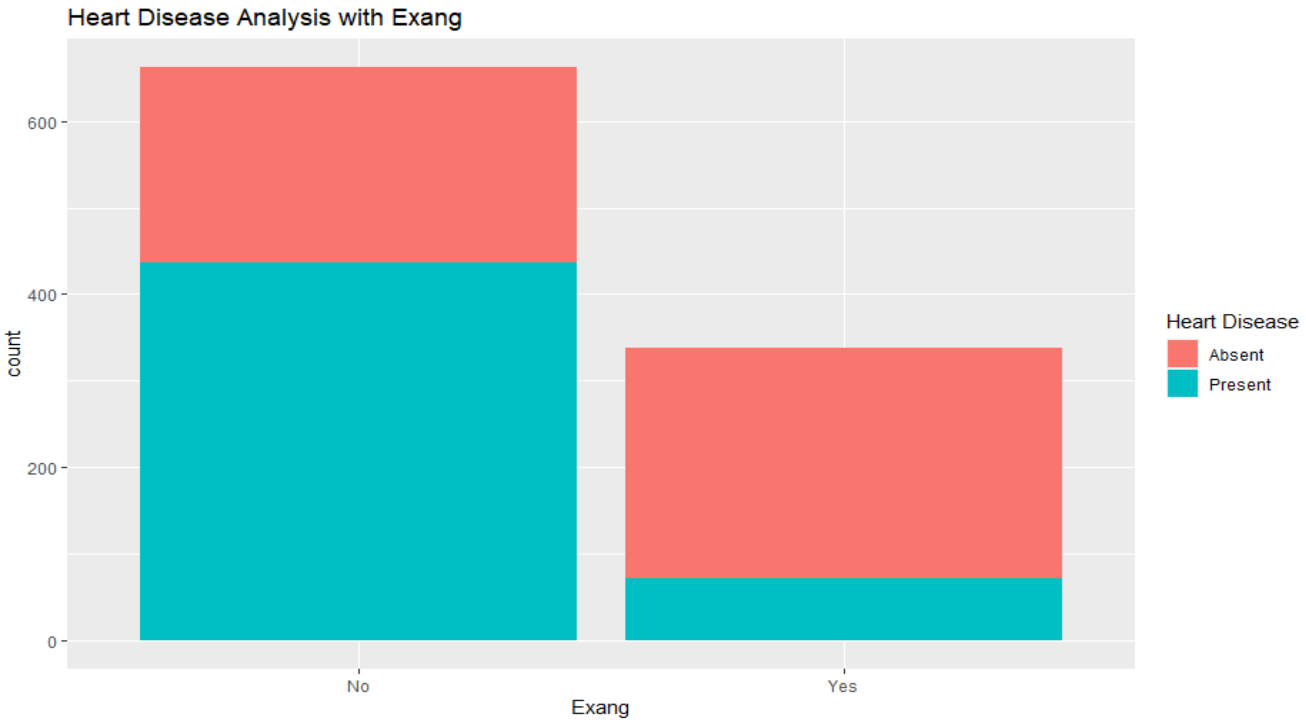

Heart Disease Analysis with Exang
count
0
200
400
600
No
Yes
Exang
Heart Disease
Absent
Present

Fig i

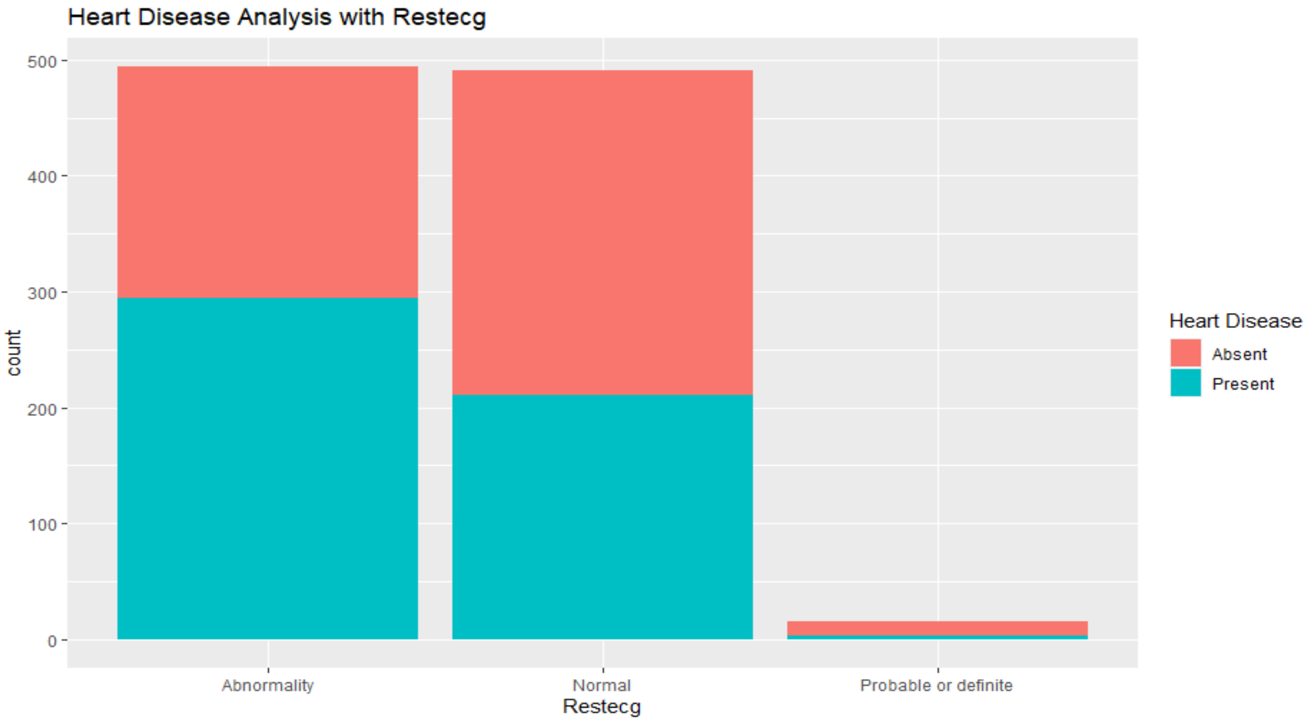

Heart Disease Analysis with Restecg
500
400
300
200
100
0
count
Abnormality
Normal
Probable or definite
Restecg
Heart Disease
Absent
Present

Fig j

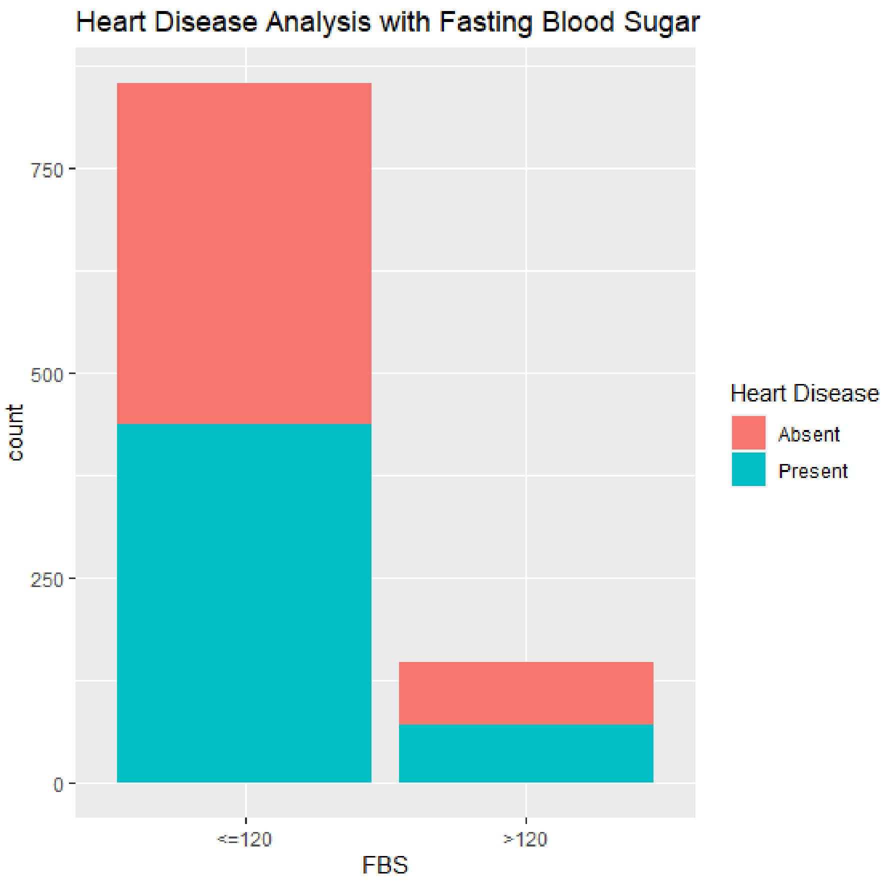

Heart Disease Analysis with Fasting Blood Sugar
count
750
500
250
0
<=120
>120
FBS
Heart Disease
Absent
Present

Fig k

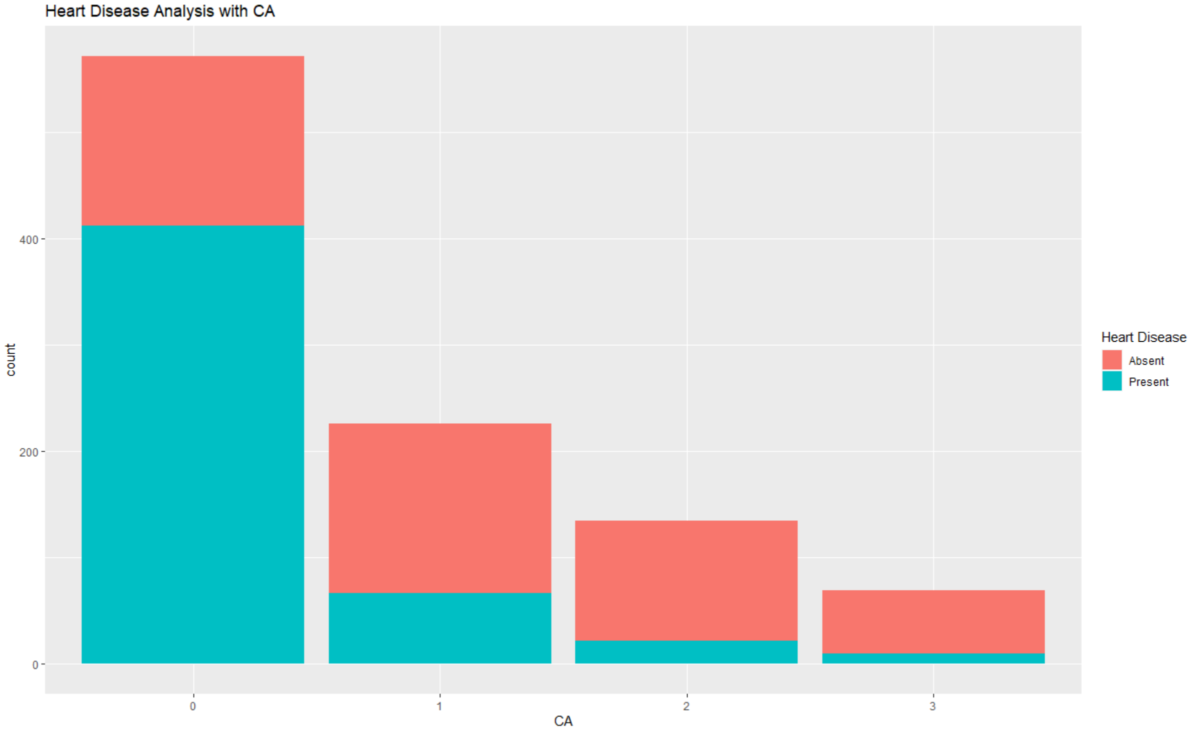

Heart Disease Analysis with CA
count
400
200
0
0
1
2
3
CA
Heart Disease
Absent
Present

Fig l

3.0

MACHINE LEARNING

*The data set '*heart_clean' *was partitioned into 80% training set and 20% test set.*

Looking at the fig a. We notice how a large number of females have heart diseases in our data set. In fact, 73% of the females have heart disease. Here is my first model that predicts target based on the sex of our patient.

```
y_hat_sex <- factor(ifelse(heart_clean$sex == "Female", 1, 0))
confusionMatrix(y_hat_sex, heart_clean$target)$overall["Accuracy"].

```

I had an accuracy of 0.62.

```
#Applying the model on the test set
y_hat_sex_test <- factor(ifelse(test_set$sex == 1, 0, 1))
confusionMatrix(y_hat_sex_test, as.factor(test_set$target))$overall["Accuracy"]

```

We had an accuracy of 0.6582915 on the test set

Using GLM (Logistic Regression)

```
ps <- seq(0.3, 0.7, 0.1)
accuracy_general <- map_df(ps, function(p){

fit <- glm(target ~ exang + trestbps + chol + age + sex + thal + restecg + cp +
fbs + ca + oldpeak  + slope + thalach, data = train_set, family = binomial)
p_hat <- predict(fit, test_set, type = "response")
y_hat <-  factor(ifelse(p_hat > p, 1, 0))
cm <- confusionMatrix(y_hat, as.factor(test_set$target))
acc <- cm$overall["Accuracy"]

f_meas <- F_meas(y_hat, test_set$target)
cm_sensitivity <- sensitivity(y_hat, test_set$target)
cm_specificity <- specificity(y_hat, test_set$target)


tibble( acc, f_meas, cm_sensitivity, cm_specificity)
})

accuracy_results <- cbind( ps, accuracy_general)
accuracy_results

```

GLM model

	ps	acc	F_meas	cm_sensitivity	cm_specificity
1	0.3	0.8442211	0.8287293	0.7653061	0.9207921
2	0.4	0.8442211	0.8287293	0.7653061	0.9207921
3	0.5	0.8391960	0.8279570	0.7857143	0.891089
4	0.6	0.8391960	0.8383838	0.8469388	0.8316832
5	0.7	0.8090452	0.8190476	0.8775510	0.7425743

Table 1 (accuracy_results)

Table above shows the results of the function mapped with a sequence of p-values. Looking at the results we can be satisfied with p = 0.6 as the best p value to provide accurate results for our model.

Because of how delicate our goal is (i.e predicting patients with heart disease) we must pick the result with the best F_meas and also the best sensitivity.

Improving our Glm model with K-Fold Cross Validation

When preparing this model, I decided to omit some attributes i.e chol and fbs and see what happens to the accuracy. This is because looking a Fig d and Fig k, I questioned their correlation with our target attribute.

First of all, I ran the model with all predictors and had an accuracy of 0.839196 and an F_Meas of 0.827957.

```r
train_control  <- trainControl(method = 'cv', number = 10)

model_glm <- train(target ~ exang + trestbps + chol + age + sex + thal +
restecg + cp + fbs + ca + oldpeak + slope + thalach, data = train_set,
family = binomial, method = "glm", trControl = train_control)

y_hat_glm <- predict(model_glm, test_set)
cm_glm <-  confusionMatrix(y_hat_glm, test_set$target)
cm_glm$overall["Accuracy"]
F_meas(y_hat_glm, test_set$target)

im_glm <- varImp(model_glm)
im_glm

```

Glm model with all predictor attributes

```r
 model_glm <- train(target ~  thal + restecg + cp  + exang + trestbps + chol + age +

sex + ca + thal + oldpeak + slope  + thalach, data = train_set,

family = binomial, method = "glm",  trControl = train_control)


 y_hat_glm <- predict(model_glm, test_set)

 cm_glm <-  confusionMatrix(y_hat_glm,test_set$target)

 cm_glm$overall["Accuracy"]

 # F_meas slightly better than our initial glm model as well.

 F_meas(y_hat_glm,test_set$target)

```

Model without fbs attribute

Next, I omitted only the ‘fbs’ attribute and ran the model again. The accuracy increased to 0.8844221. Then, I removed only the ‘chol attribute and brought back fbs the accuracy remained the same at 0.8844. Also, when I omitted both it remained the same, just like the first accuracy from the model which factors in all the predictor attributes provided in the dataset. Looking at the varImp of that model. I decided to stick with the chol attribute because it had a higher overall importance (31.388) compared to fbs1 (20.998).

Variable Importance of glm model

glm variable	**Variable importance**
chol	32.3620
trestpbs	30.258
oldpeak	23.911
cp1	23.644
restecg1	22.572
age	17.803
slope1	9.985
slope2	8.912
thal2	3.404
restecg2	0.00

This table represent the bottom 10 important variables in our model. We can see that ‘age’ is the lowest, cumulatively, because the others below it represents a fragment/factor of a specific attribute instead of the attribute as a whole.

After removing the age attribute from our glm model our final accuracy increased to 0.8542714 with an F_meas of 0.8465608.

```
   # Final GLM Model
   # selecting the final predictors based on the Variable Importance

   model_glm <- train(target ~  thal +  cp  + exang + trestbps +  sex + chol + ca + thal
 + oldpeak +slope+ thalach, data = train_set, family = binomial, method = "glm",
 trControl = train_control)

   y_hat_glm <- predict(model_glm, test_set)
   cm_glm <-  confusionMatrix(y_hat_glm,test_set$target)
   cm_glm$overall["Accuracy"]
   F_meas(y_hat_glm,test_set$target)
```

Final Glm Model

Using KNN (K Nearest Neighbour)

```
model_knn <- train(target ~ thal + restecg + cp  + exang + trestbps + age +
sex + fbs + ca + oldpeak + slope + thalach, data = train_set,
method = "knn", trControl = train_control)

 y_hat_knn <- predict(model_knn, test_set)
 cm_knn <- confusionMatrix(y_hat_knn,test_set$target)
 cm_knn$overall["Accuracy"]
 im_knn <- varImp(model_knn)
 im_knn

 #I noticed fbs had 0.00 varImportance hence we run the Knn model again without it

```

Knn model with all initial predictors

```
model_knn <- train(target ~  oldpeak + thalach + ca + thal, data = train_set,  method = "knn", trControl = train_control)
  y_hat_knn <- predict(model_knn, test_set)
  cm_knn <-  confusionMatrix(y_hat_knn,test_set$target)
  cm_knn$overall["Accuracy"]

  im_knn <- varImp(model_knn)
  im_knn

```

Final Knn Model with stratified predictor attributes

The code above shows our final model after filtering out the predictors with very low variable importance. In our final knn model, the ‘thal’ predictor has a varImp of 0, however if we omit the ‘slope’ attribute our accuracy drops from 0.8241206 to 0.8190955.

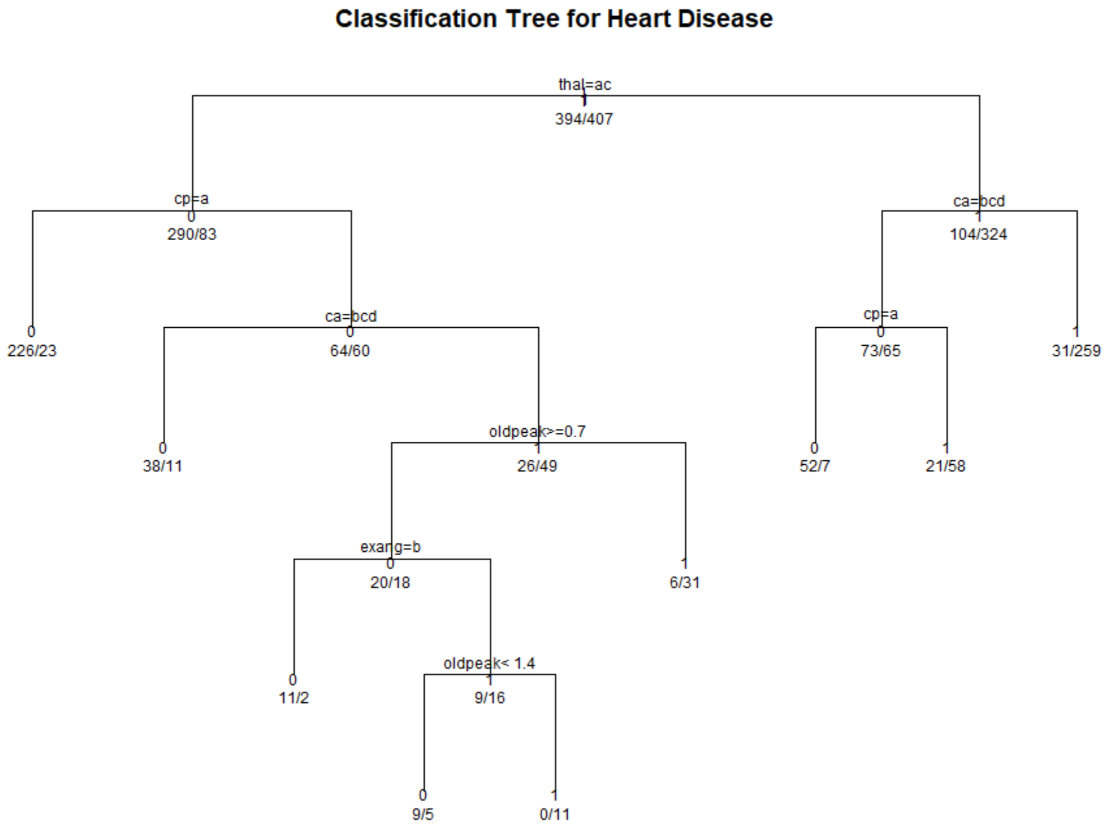
Using classification Tree (RPART)

```r
library(rpart)

tree_model <- rpart(target ~ thal + restecg + cp  + exang + trestbps +age + sex  + restecg +  fbs + ca + oldpeak + slope + thalach, data = train_set, method = "class")
  varImp(tree_model)

# Rerun the model without fbs, slope, age, sex, trestbps and restecg
tree_model <- rpart(target ~  exang + thal +  cp  + ca + oldpeak +  thalach, data = train_set, method = "class")

varImp(tree_model)

plotcp(tree_model)
plot(tree_model, uniform=TRUE,   main="Classification Tree for Heart Disease")
text(tree_model, use.n = TRUE, all=TRUE, cex=.8)

y_hat_tree <- predict(tree_model, test_set, type = 'class')
cm_tree <- confusionMatrix(y_hat_tree, test_set$target)

```

Classification Tree Model

With an accuracy of 0.8543 it is our best scoring model yet. The table below shows the variable Importance for the model

	Overall
ca	160.1036
cp	180.0138
oldpeak	148.5071
thal	112.6889
thalach	145.5428
exang	131.00704



Using Random Forest

```
library(randomForest)
model_forest <- randomForest(target ~ exang + trestbps + chol + age+
 sex + thal + restecg + cp + fbs +  ca + oldpeak + slope + thalach,
 data = train_set, ntree = 50)

y_hat_forest <- predict(model_forest, test_set, type = 'class')
cm_forest <- confusionMatrix(y_hat_forest, test_set$target)
varImp(model_forest)

```

I am more than satisfied with an accuracy of 0.9849.

CONCLUSION

At the end of the project Random Forest came out as the best model for my analysis. I was more than satisfied with an accuracy of 0.9849. A sensitivity of 1.0 and specificity of 0.9703.

It is important to note that this data set was generated in the 1980s, and I am sure there are better ways and new technology for detecting heart diseases now. However, it could be a stepping stone and or reference for any future analysis remotely close to the subject of heart diseases.

Furthermore, judging from the project it seems the best predictors across all models seems to be ‘chest pain’, ‘ca’, ‘thal’, and ‘oldpeak’. My goal now is to try this model with new data sets hopefully something recent with the aim of having consistent and high accuracy predictions.

References

1. Mayo Clinic Staff. “Heart Disease - Symptoms and Causes.” *Mayo Clinic*, www.mayoclinic.org, 13 Aug. 2022, https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118#:~:text=Heart%20disease%20describes%20a%20range,born%20with%20(congenital%20heart%20defects).