Applied Statistics STA4102 Project: Heart Disease Analysis

Thanh Pham

Instructor Houston Sanders

STA4102

2023-06-20

Abstract

In this project, I utilized a heart disease dataset sourced from Kaggle website and performed an exploratory analysis using R to discover whether the age of a person related to their gender, and other symptoms of heart rate disease or not. The dataset comprised various clinical features associated with heart disease, including age, sex, chest pain type, resting blood pressure, cholesterol levels, and presence of heart disease, etc. Initially, I performed data preprocessing, including cleaning missing values, removing duplicates, and standardizing variables. Then, I explored the data trends, correlation, characteristics and make a short conclusion based on observation. Furthermore, I visualized the data using various plots and charts to enhance the understanding of the relationships between categorical variables of heart disease. Next, I employed statistical techniques and machine learning algorithms to investigate potential relationships and predictive factors associated with heart disease. This includes employing linear regression, logistic regression to gain desired statistical results; and using decision tree, and random forest models to access the accuracies of the data.

Introduction

Being one of the main causes of death and morbidity, heart disease is a major public health concern in every country. For early detection, prevention, and efficient treatment options, it is essential to understand the risk factors and predictors connected with heart disease. In this work, I use the R programming language to analyze a heart illness dataset that I downloaded from Kaggle website and examine the correlations between several clinical variables and the presence of heart disease. Examining the heart disease dataset and identifying the relevant factors connected to the occurrence of heart disease is the main goal of this study. We want to find patterns, connections, and potential predictive models for heart disease using statistical analytic methods and machine learning algorithms.

Although there are 76 attributes in this database, all published experiments only mention using a portion of 14. The Cleveland database in particular is the only one that ML researchers have used up until this point. The "goal" field alludes to the patient's having heart disease. It has integer values ranging from zero (no presence) to four. The Cleveland database has been used for

experiments that have mostly focused on attempting to discern between presence (values 1, 2, 3, 4) and absence (value 0).

Here are the main variables explanation of the dataset:

- Age: age in years
- Sex: (1 = male; 0 = female)
- Cp: chest pain type (4 values)
- Trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- Chol: serum cholesterol in mg/dl
- Fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- Restecg: resting electrocardiographic results
- Thalach: maximum heart rate achieved
- Exang: exercise induced angina (1 = yes; 0 = no)
- Oldpeak: ST depression induced by exercise relative to rest
- Slope: the slope of the peak exercise ST segment
- Ca: number of major vessels (0-3) colored by fluoroscopy
- Thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
- Target: 0 absence heart disease

Data Description

First of all, I import the downloaded dataset as an .csv file into a heart variable using RStudio as an Integrated Development Environment for R programming. There are 1025 rows and 14 columns.

```
#import data

heart <- read.csv("~/Downloads/heart.csv")

print(head(heart))
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  52   1  0      125  212   0       1     168     0     1.0     2  2    3
## 2  53   1  0      140  203   1       0     155     1     3.1     0  0    3
## 3  70   1  0      145  174   0       1     125     1     2.6     0  0    3
## 4  61   1  0      148  203   0       1     161     0     0.0     2  1    3
## 5  62   0  0      138  294   1       1     106     0     1.9     1  3    2
## 6  58   0  0      100  248   0       0     122     0     1.0     1  0    2
##   target
## 1      0
## 2      0
## 3      0
## 4      0
## 5      0
## 6      1
```

```
# Total of rows and columns
nrow(heart)
## [1] 1025
ncol(heart)
## [1] 14
```

Data Cleaning

To clean the data, I need to find out if the data contains any missing NA values using the sum of is.na() onto the data. The result of this operation is '0' which means there are no missing values. Hence, no cleaning needed for this dataset.

However, I still transform the 'heart' data into 'heart_cleaned' for better observation and print out the structure of data.

```
# check if there are any missing NA values in the data
NA_values <- sum(is.na(heart))
print(NA_values)
## [1] 0
# Remove the missing values
heart_cleaned <- na.omit(heart)
str(heart_cleaned)
## 'data.frame':    1025 obs. of  14 variables:
##  $ age     : int  52 53 70 61 62 58 58 55 46 54 ...
##  $ sex     : int  1 1 1 1 0 0 1 1 1 1 ...
##  $ cp      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ trestbps: int  125 140 145 148 138 100 114 160 120 122 ...
##  $ chol    : int  212 203 174 203 294 248 318 289 249 286 ...
##  $ fbs     : int  0 1 0 0 1 0 0 0 0 0 ...
##  $ restecg : int  1 0 1 1 1 0 2 0 0 0 ...
##  $ thalach : int  168 155 125 161 106 122 140 145 144 116 ...
##  $ exang   : int  0 1 1 0 0 0 0 1 0 1 ...
##  $ oldpeak : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
##  $ slope   : int  2 0 0 2 1 1 0 1 2 1 ...
##  $ ca      : int  2 0 0 1 3 0 3 1 0 2 ...
##  $ thal    : int  3 3 3 3 2 2 1 3 3 2 ...
##  $ target  : int  0 0 0 0 0 1 0 0 0 0 ...
```

Data Exploratory

To explore the insights and trends of data, I import a 'ggplot2' library to plot the correlation between the categorical data.

Below is also the statistical summary of each category including the minimum value, 1st quantile, median, mean, 3rd quantile, and maximum value.

```
# import library
library(ggplot2)
```

```
# Statistical summary of heart data
summary(heart_cleaned)
##       age             sex               cp            trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :0.0000   Min.   : 94.0
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:120.0
##  Median :56.00   Median :1.0000   Median :1.0000   Median :130.0
##  Mean   :54.43   Mean   :0.6956   Mean   :0.9424   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.0000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :3.0000   Max.   :200.0
##       chol           fbs             restecg          thalach
##  Min.   :126   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:132.0
##  Median :240   Median :0.0000   Median :1.0000   Median :152.0
##  Mean   :246   Mean   :0.1493   Mean   :0.5298   Mean   :149.1
##  3rd Qu.:275   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##      exang           oldpeak          slope             ca
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   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 :1.000   Median :0.0000
##  Mean   :0.3366   Mean   :1.072   Mean   :1.385   Mean   :0.7541
##  3rd Qu.:1.0000   3rd Qu.:1.800   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.200   Max.   :2.000   Max.   :4.0000
##       thal           target
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000
##  Median :2.000   Median :1.0000
##  Mean   :2.324   Mean   :0.5132
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
# Explore the age
age <- heart_cleaned$age
hist(age)
```

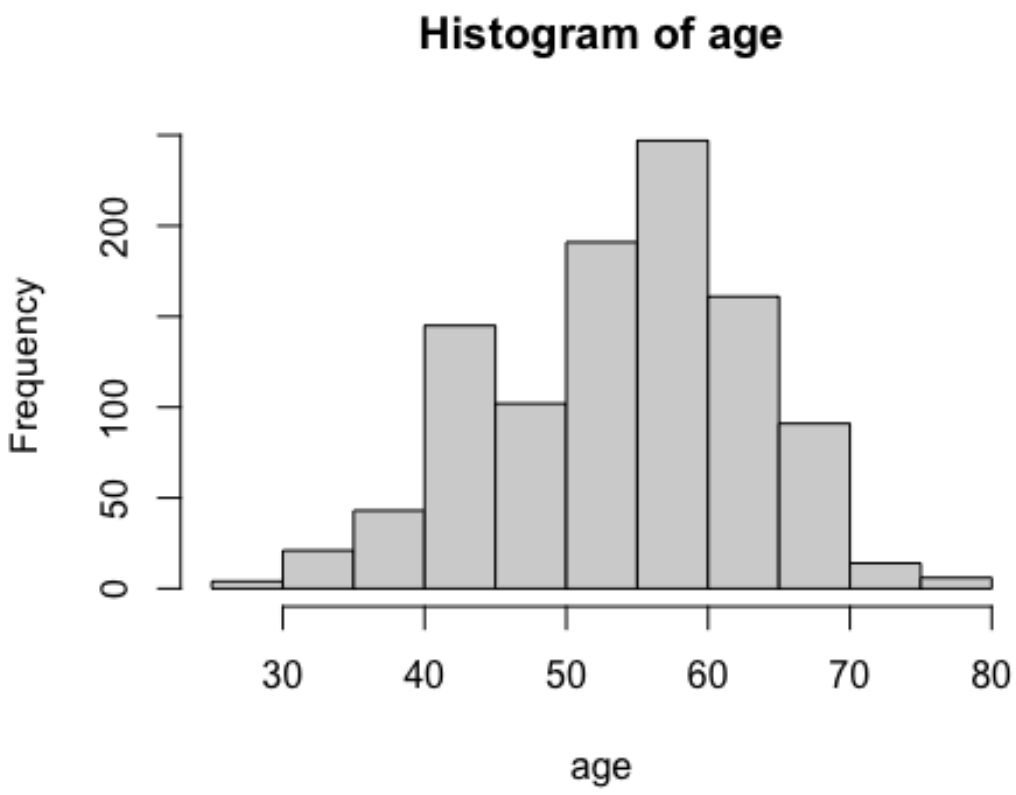


Figure 1. Frequency of Age

The figure shows the frequency of each specific age group and based on the histogram I observe that the distribution depicts a logical distribution for the subject; it is more common for people to experience cardiac problems as they get older, and there are also fewer elderly people than young people.

```
# Explore the sex
sex <- heart_cleaned$sex
male <- sum(sex == 1)
female <- sum(sex == 0)

barplot(c(male, female), names.arg = c("Male", "Female"), xlab = "Sex", ylab
= "Count", col = c("red", "blue"), main = "Distribution of Sex")
```

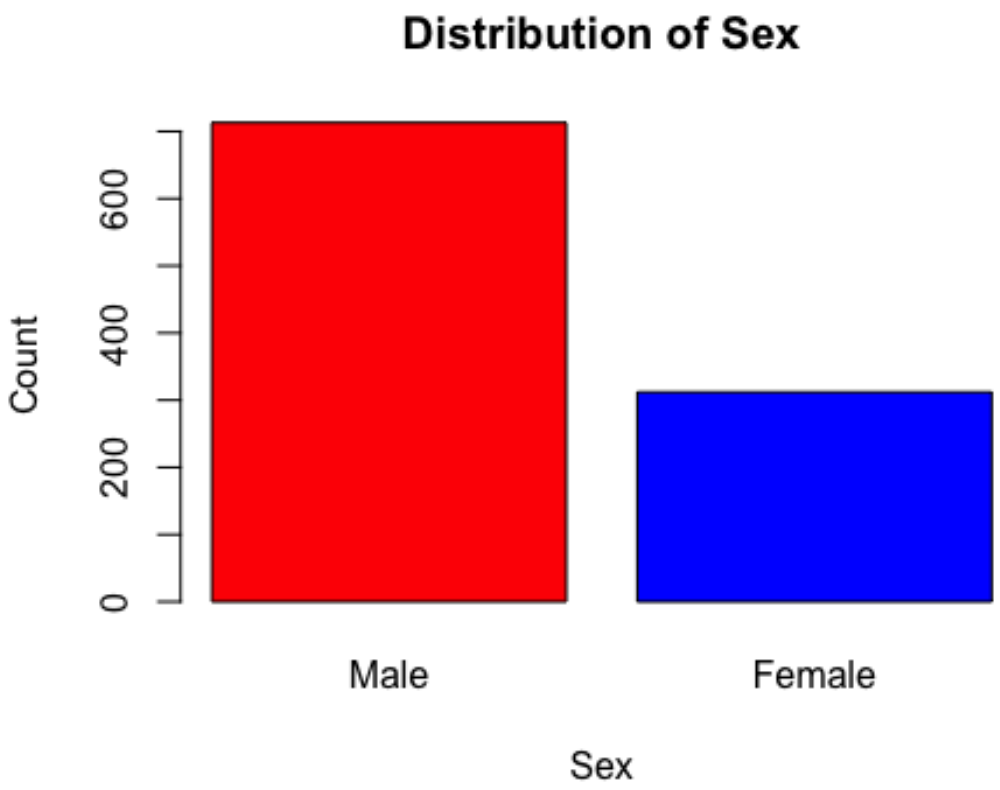


Figure 2. Distribution of Sex

Based on the bar plot, it is obvious that male is likely to have heart disease than female. Where male are 1's and female are 0's.

```
# Create a density plot to explore the biased relation between sex and age
ggplot(heart_cleaned, aes(x = age, fill = sex)) +
  geom_density(alpha = 0.5) +
  labs(x = "Age", y = "Density") +
  ggtitle("Density Estimate by Sex v.s Age")
## Warning: The following aesthetics were dropped during statistical transfor
mation: fill
## ℹ This can happen when ggplot fails to infer the correct grouping structur
e in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

Age vs sex

Then, I perform a ggplot() function to visualize the relation between sex and age of heart disease patients.

As we can see the distribution for male and female is balanced and following the same pattern in quantity of data. As the result, no bias about the relation sex and age.

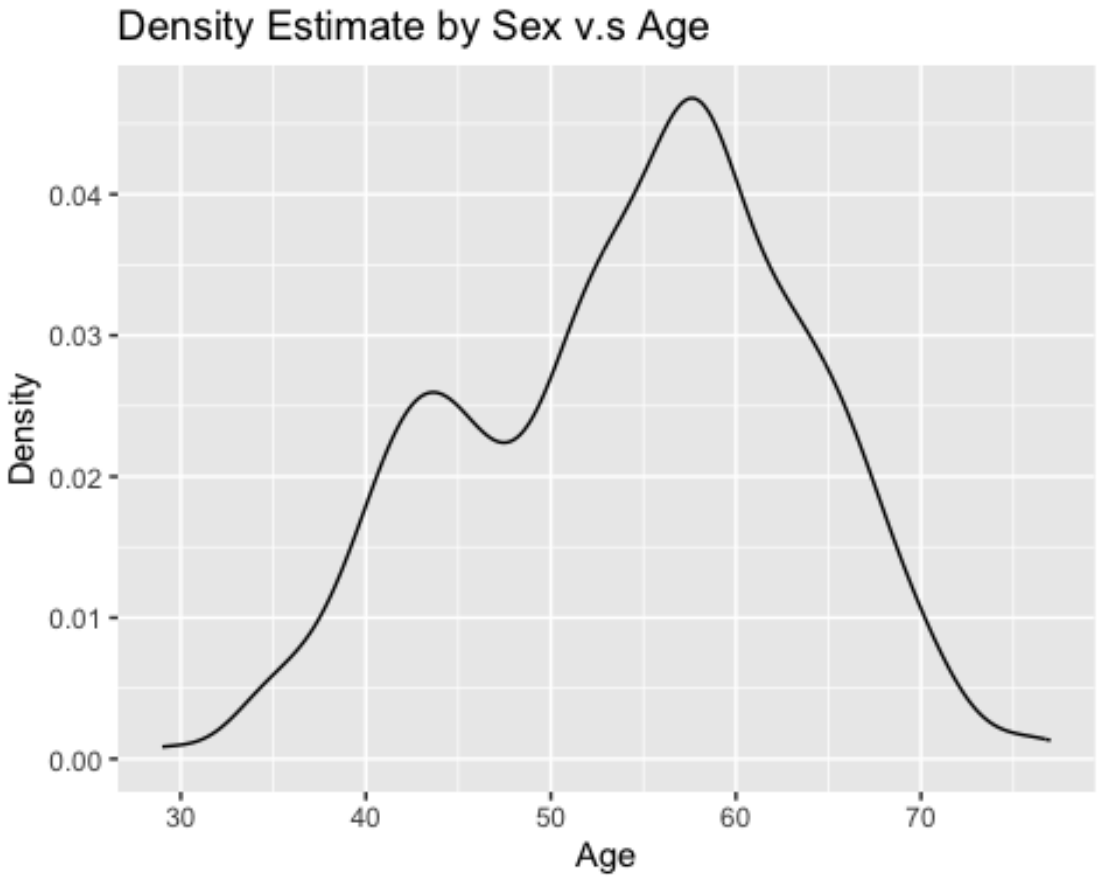


Figure 3. Age v.s Sex

```
# The relation between thalach (maximum heart rate achieved) and age
thalach <- heart_cleaned$thalach
ggplot(heart_cleaned, aes(x = age, y = thalach)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Age", y = "thalach", title = "Linear Regression Line between age
and maximum heart rate")
## `geom_smooth()` using formula = 'y ~ x'
```

Figure 4. Linear Regression Line between Age and maximum heart rate

Age vs Maximum Heart Rate

Based on my observations, individuals at the age of 30 can reach a maximum heart rate of 180 beats per minute. However, for individuals aged 70 or older, the highest recorded heart rate is 130 beats per minute.

```
# The relation between cp (chest pain type: 0,1,2,3) and age
cp <- heart_cleaned$cp
ggplot(heart_cleaned, aes(x = age, y = cp)) +
  geom_density_2d() +
  labs(x = "Age", y = "cp", title = "2D Kernel Density Plot")
```

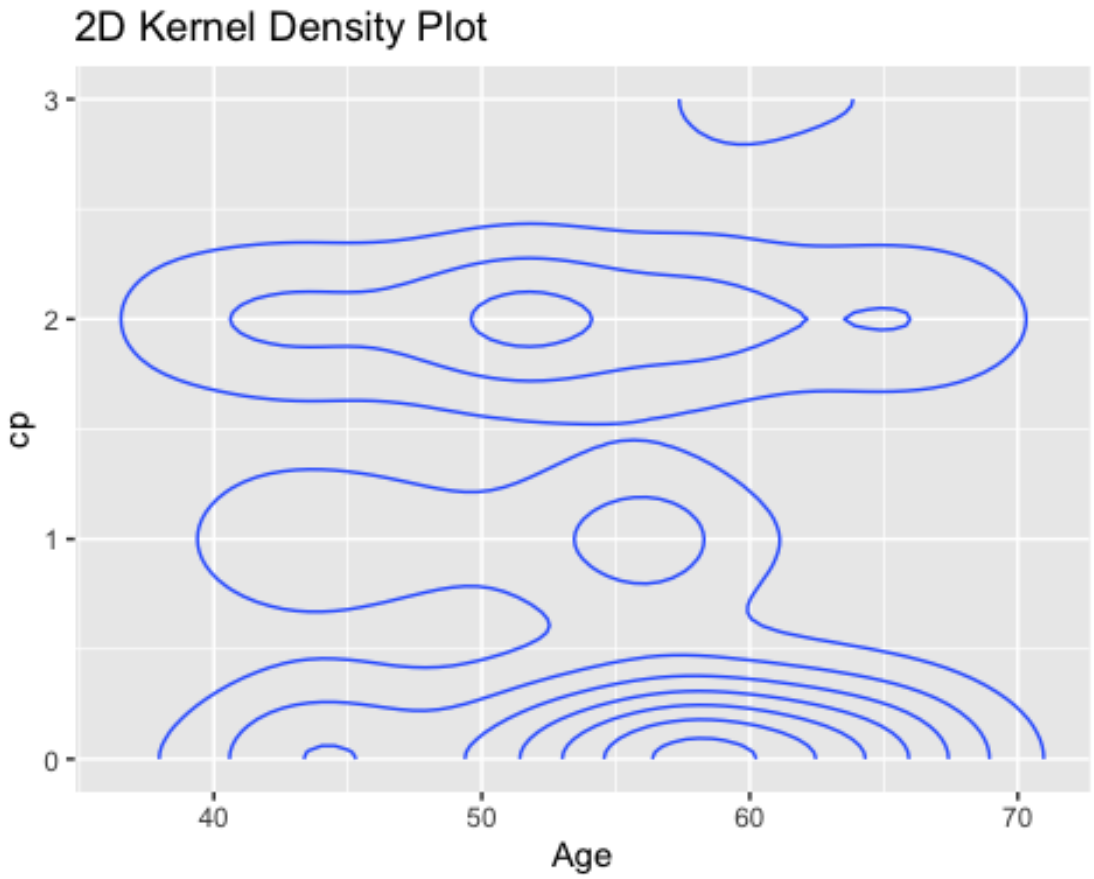


Figure 5. Age vs Chest Pain Types

Age vs Chest Pain Types

I observe that chest pain is more prevalent in the age range of 50 to 60.

Methodology

Linear Regression

I perform linear regression model for the relations of Age vs Maximum Heart Rate and Age vs Chest Pain Types. The first observed p-value is significantly low, that describes how likely we are to have found a particular set of observations if the null hypothesis were true.

```
# Linear regression model between age and maximum heart rate
model1 <- lm(age ~ thalach, data=heart_cleaned)
summary(model1)
##
## Call:
## lm(formula = age ~ thalach, data = heart_cleaned)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3755  -6.6071   0.6255   6.3954  24.5488
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 77.38069    1.71274   45.18   <2e-16 ***
## thalach     -0.15389    0.01135  -13.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.357 on 1023 degrees of freedom
## Multiple R-squared:  0.1523, Adjusted R-squared:  0.1514
## F-statistic: 183.8 on 1 and 1023 DF,  p-value: < 2.2e-16
# Linear regression model between age and 4 chest pain types
model2 <- lm(age ~ cp, data=heart_cleaned)
summary(model2)
##
## Call:
## lm(formula = age ~ cp, data = heart_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.3976  -6.7635   0.9682   6.2365  22.2365
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.0318     0.3834 143.526   <2e-16 ***
## cp           -0.6341     0.2748  -2.308   0.0212 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.053 on 1023 degrees of freedom
## Multiple R-squared:  0.005179,   Adjusted R-squared:  0.004207
## F-statistic: 5.326 on 1 and 1023 DF,  p-value: 0.02121
```

Logistic Regression

In this model, I observe that the predicted score indicates a likelihood of heart disease. However, a good cutoff point must be established from which it is simple to discern between having heart disease and not.

```
set.seed(100)

index <- sample(nrow(heart_cleaned),0.75*nrow(heart_cleaned))
train <- heart[index,]
test <- heart[-index,]
```

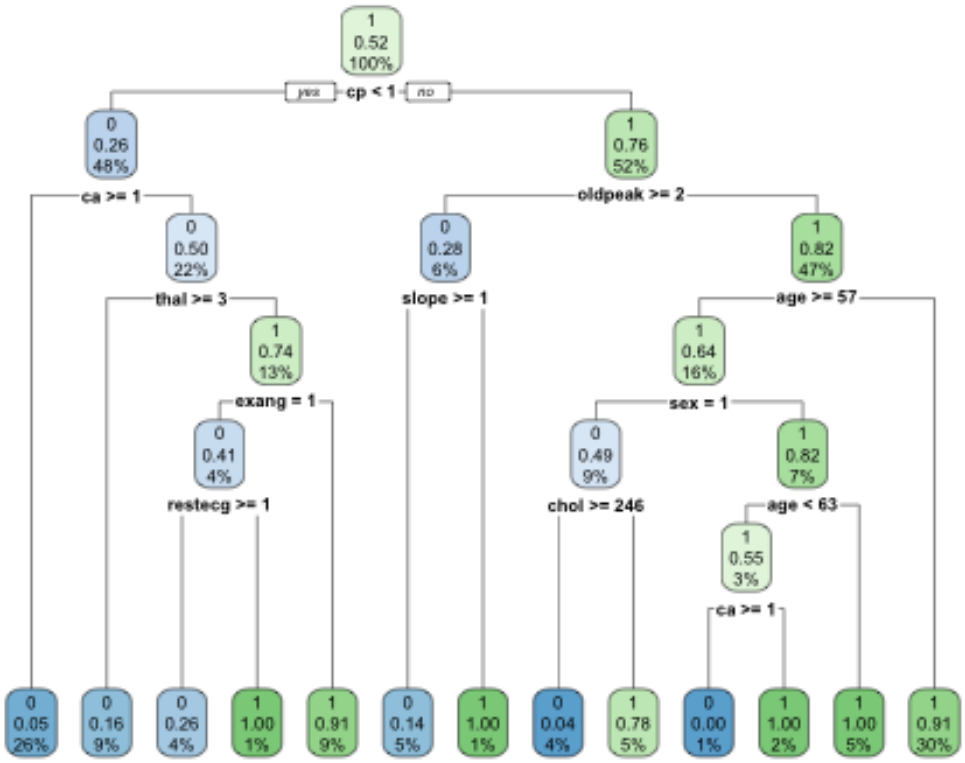
```
model <- glm(target~.,data = train,family = "binomial")
model
##
## Call:  glm(formula = target ~ ., family = "binomial", data = train)
##
## Coefficients:
## (Intercept)          age          sex           cp     trestbps         ch
ol
##    3.573217    -0.011343    -1.968592     0.793969    -0.018794    -0.0057
13
##         fbs      restecg      thalach        exang      oldpeak        slo
pe
##    0.061758     0.252697     0.026677    -1.197404    -0.647040     0.5637
45
##          ca         thal
##   -0.828620    -0.784109
##
## Degrees of Freedom: 767 Total (i.e. Null);  754 Residual
## Null Deviance:       1063
## Residual Deviance: 519.2     AIC: 547.2
```

Decision Tree

For this model, I import 'rpart' and 'rpart.plot' libraries to implement a tree. Recursive partitioning and regression trees are referred to as 'rpart'.When both the independent and dependent variables are continuous or categorical, 'rpart' is employed.

```
train$pred<-NULL
test$pred<-NULL

library(rpart)
tree<-rpart(target~.,method = "class",data = train)
library(rpart.plot)
rpart.plot(tree)
```




```
acc_tr_tree<-(21+37)/76
acc_tr_tree
## [1] 0.7631579
```

The most important variables, according to the decision tree are cp, ca, thal, and oldpeak. The final predictive result states that the decision tree's accuracy rate is 76.32%, which means its misclassification rate is 23.68%.

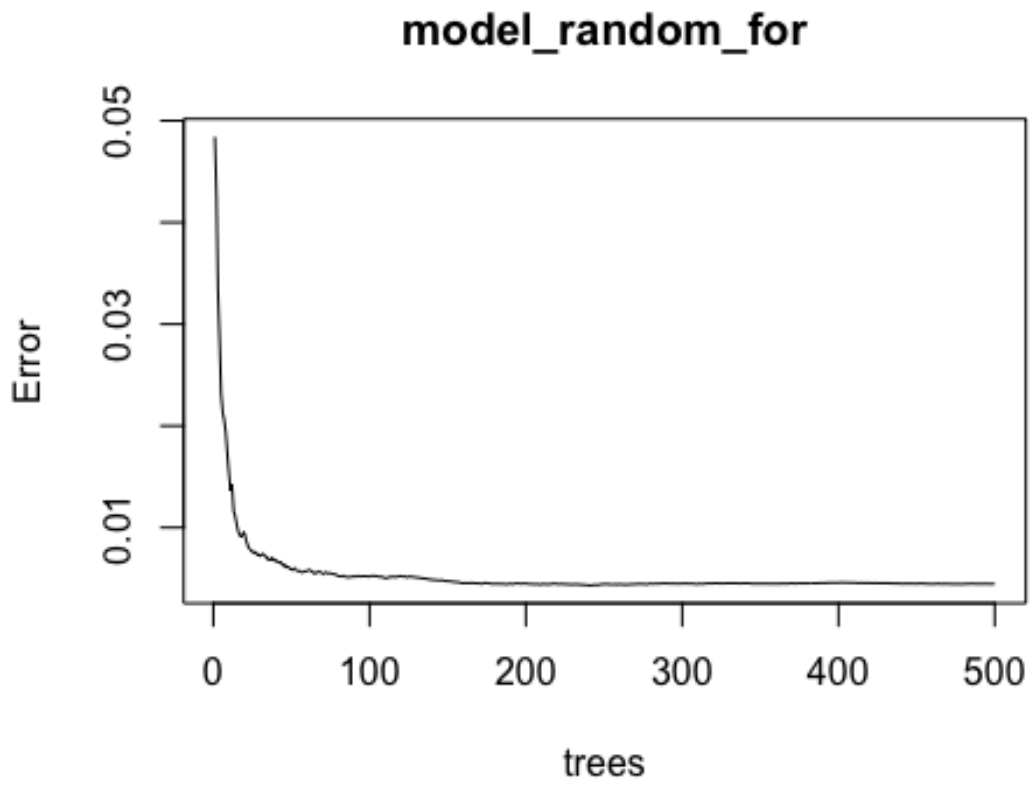
Random Forest

For the Random Forest model, I have to install a ‘randomForest’ package and upload its library to clarify desirable results. Then, I plot random forest on graph with respect to class error.

```
install.packages("randomForest")
## Error in contrib.url(repos, "source"): trying to use CRAN without setting
a mirror
library(randomForest)
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
test$pred<-NULL
set.seed(100)

target <- heart_cleaned$target
model_random_for <- randomForest(target ~ ., data = heart)
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values.  Are you sure you want to do regression?
model_random_for
##
## Call:
##  randomForest(formula = target ~ ., data = heart)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 0.004469573
##                     % Var explained: 98.21
plot(model_random_for)
```



Conclusion

Overall, my research has demonstrated that a higher risk of heart disease was specifically linked to older age, male gender, higher blood pressure, and elevated cholesterol levels. These findings support prior research and emphasize the significance of these elements in determining the likelihood of developing heart disease. Additionally, our investigation showed how well machine learning algorithms like decision trees and logistic regression can be used to forecast the occurrence of heart disease using the clinical data that is now available. Most categorical data have a trend of increasing when compared to increasing age. Sex is balanced and following the same pattern in quantity of data, hence, no bias about the relation sex and age. After performing alternate machine learning classification, methods, and access to their accuracies, all the models are concluded to have high accuracies from 75% to 85%.

Reference

https://www.kaggle.com/datasets/ineubytes/heart-disease-dataset/code

https://archive.ics.uci.edu/dataset/45/heart+disease