Heart Disease

Sofya Rbinovich

06/11/2021

Introduction

Cardiovascular disease (CVD) is a leading cause of deaths worldwide. According to the Wolrd Health Organisation (WHO) 17.9 million people died in 2019 and it was accounted as 32% of deaths worldwide. The most deaths (85%) from the CVD were due to heart attack and stroke ("Cardiovascular diseases (CVDs)", 2021). CVD also divided into 4 groups: coronary artery disease (CAD) which is also referred to as coronary heart disease (CHD), cerebrovascular disease, peripheral artery disease (PAD), and aortic atherosclerosis (Lopez, Ballard & Jan, 2021). Atherosclerosis occurs due to narrowing of the arteries due to paque build up. Atherosclerosis is characterized by the accumulation of lipids and fibrous elements in the large arteries and is the primary cause of heart disease and stroke (Lusis, 2000).

There are many risk factors that lead to the development of the CVD:

1. Age
2. Sex
3. Smoking
4. Genetics
5. Alcohol consumption
6. Physical inactivity
7. Diet
8. Environmental Factors

When a person come to doctor first step is the risk assessment, where a GP asks about your medical and family history, eck blood pressure and do a blood test to check the cholesterol levels. Any further tests include: ECG, exercise stress test, X-rays, echocardiogram, coronary angiography, radionuclide tests, MRI scans, CT scans.

Aims and Objectives

Aim of the project is to train different machine learning algorithms and assess their accuracy.

Objectives:

1. Investigate correlation between variables
2. Divide data into training and test sets
3. Train MLAs

Data Investigation

In the data set provided on the https://www.kaggle.com/ronitf/heart-disease-uci website contains 14 variables:

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 flourosopy

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

All patients names and identification numbers were removed due to Data Protection Act. First step is to upload data and all lubraries that will be used in the project.

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Using poppler version 20.12.1

##
## Attaching package: 'matrixStats'

## The following object is masked from 'package:dplyr':
##
##     count

## Rborist 0.2-3

## Type RboristNews() to see new features/changes/bug fixes.

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

##
## Attaching package: 'strucchange'

## The following object is masked from 'package:stringr':
##
##     boundary

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:modeltools':
##
##     Predict
```

```
## The following object is masked from 'package:Rborist':
##
##     Export

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

## Loading required package: bitops

## Rattle: A free graphical interface for data science with R.
## Version 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

##
## Attaching package: 'rattle'

## The following object is masked from 'package:randomForest':
##
##     importance
```

Data was downloaded as a csv file on the desktop and moved to the working reprository.

```
data <- read.csv("heart.csv")
```

Now we can transform the data set into more readable type.

```
data_tidy <- data %>%
  mutate(sex = ifelse(sex == 1, "Male", "Female"),
         fbs = ifelse(fbs == 1, ">120", "<= 120"),
         exang = ifelse(exang == 1, "Yes", "No"),
         restecg = ifelse(restecg == 0, "Normal", "Abnormal"),
         cp = ifelse(cp == 1, "Typical Angina",
                     ifelse(cp == 2, "Atypical Angina",
                            ifelse(cp == 3, "Non-Anginal Pain", "Asymptomatic"))),
         target = ifelse(target == 1, "Heart Disease", "No Heart Disease"),
         slope = as.factor(slope),
         ca = as.factor(ca),
         thal = as.factor(thal))
#rename Columns
data_tidy <- data_tidy %>%
  rename("Age" = age,
         "Sex" = sex,
         "Chest_Pain_type" = cp,
         "Resting_blood_pressure" = trestbps,
         "Cholesterol" =chol,
         "Fasting_blood_sugar" = fbs,
```

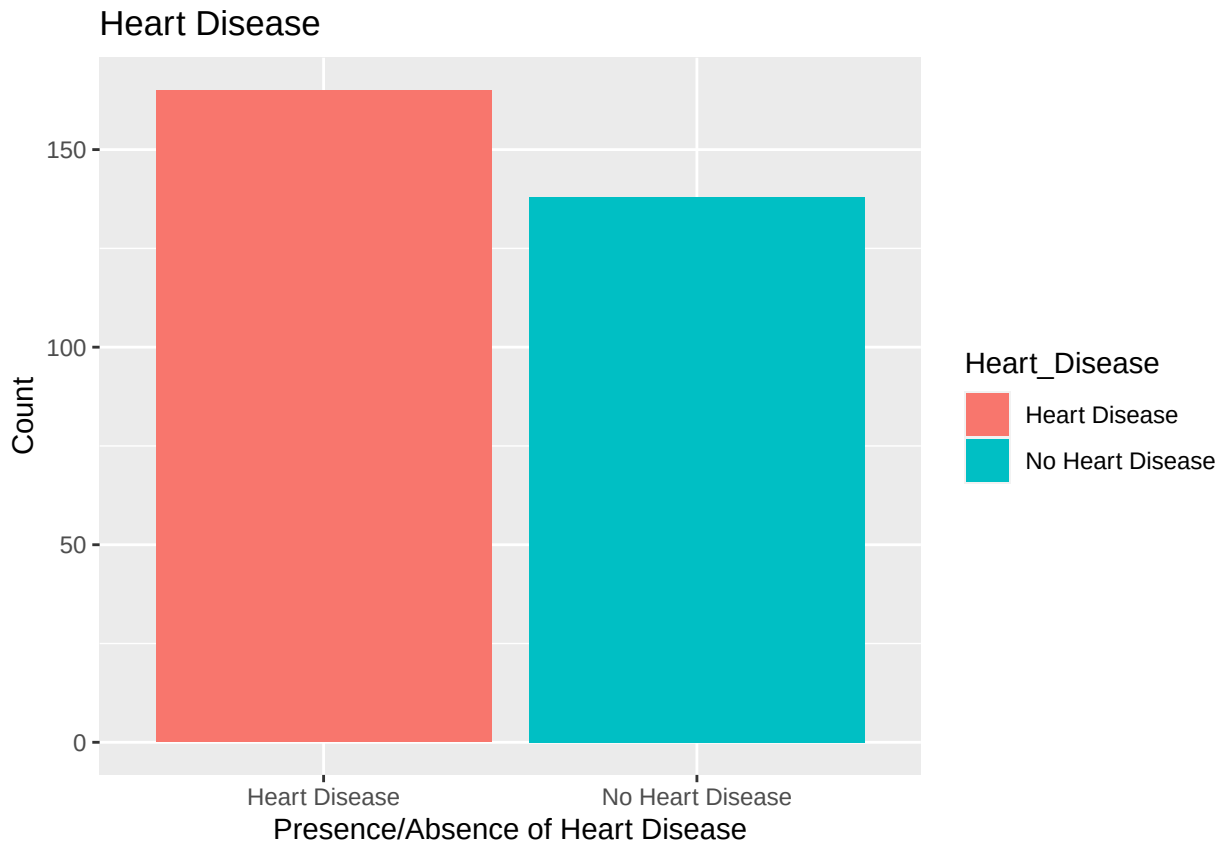
```
         "Resting_electrocardiographic_results" = restecg,
         "Maximum_heart_rate_achieved" = thalach,
         "Exercise_induced_angina " = exang,
         "ST_depression" = oldpeak,
         "ST_slope" = slope,
         "Number_of_major_vessels " = ca,
         "Thalassemia" = thal,
         "Heart_Disease" = target) %>%
  mutate_if(is.character, as.factor)
```

Data Exploration

```
##       Age            Sex              Chest_Pain_type Resting_blood_pressure
##  Min.   :29.00   Female: 96   Asymptomatic    :143    Min.   : 94.0
##  1st Qu.:47.50   Male  :207   Atypical Angina : 87    1st Qu.:120.0
##  Median :55.00                Non-Anginal Pain: 23    Median :130.0
##  Mean   :54.37                Typical Angina  : 50    Mean   :131.6
##  3rd Qu.:61.00                                        3rd Qu.:140.0
##  Max.   :77.00                                        Max.   :200.0
##   Cholesterol    Fasting_blood_sugar Resting_electrocardiographic_results
##  Min.   :126.0   <= 120:258          Abnormal:156
##  1st Qu.:211.0   >120  : 45          Normal  :147
##  Median :240.0
##  Mean   :246.3
##  3rd Qu.:274.5
##  Max.   :564.0
##  Maximum_heart_rate_achieved Exercise_induced_angina  ST_depression  ST_slope
##  Min.   : 71.0               No :204                  Min.   :0.00   0: 21
##  1st Qu.:133.5               Yes: 99                  1st Qu.:0.00   1:140
##  Median :153.0                                        Median :0.80   2:142
##  Mean   :149.6                                        Mean   :1.04
##  3rd Qu.:166.0                                        3rd Qu.:1.60
##  Max.   :202.0                                        Max.   :6.20
##  Number_of_major_vessels  Thalassemia          Heart_Disease
##  0:175                    0:  2       Heart Disease   :165
##  1: 65                    1: 18       No Heart Disease:138
##  2: 38                    2:166
##  3: 20                    3:117
##  4:  5
##
```

Now we are ready to investigate how different variable affect the occurence of the heart disease. Firtly, we will look how many people from the data set have heart disease.

```
data_tidy %>%
  ggplot(aes(Heart_Disease, fill = Heart_Disease)) +
  geom_bar() +
  xlab("Presence/Absence of Heart Disease") +
  ylab("Count") +
  ggtitle("Heart Disease")
```

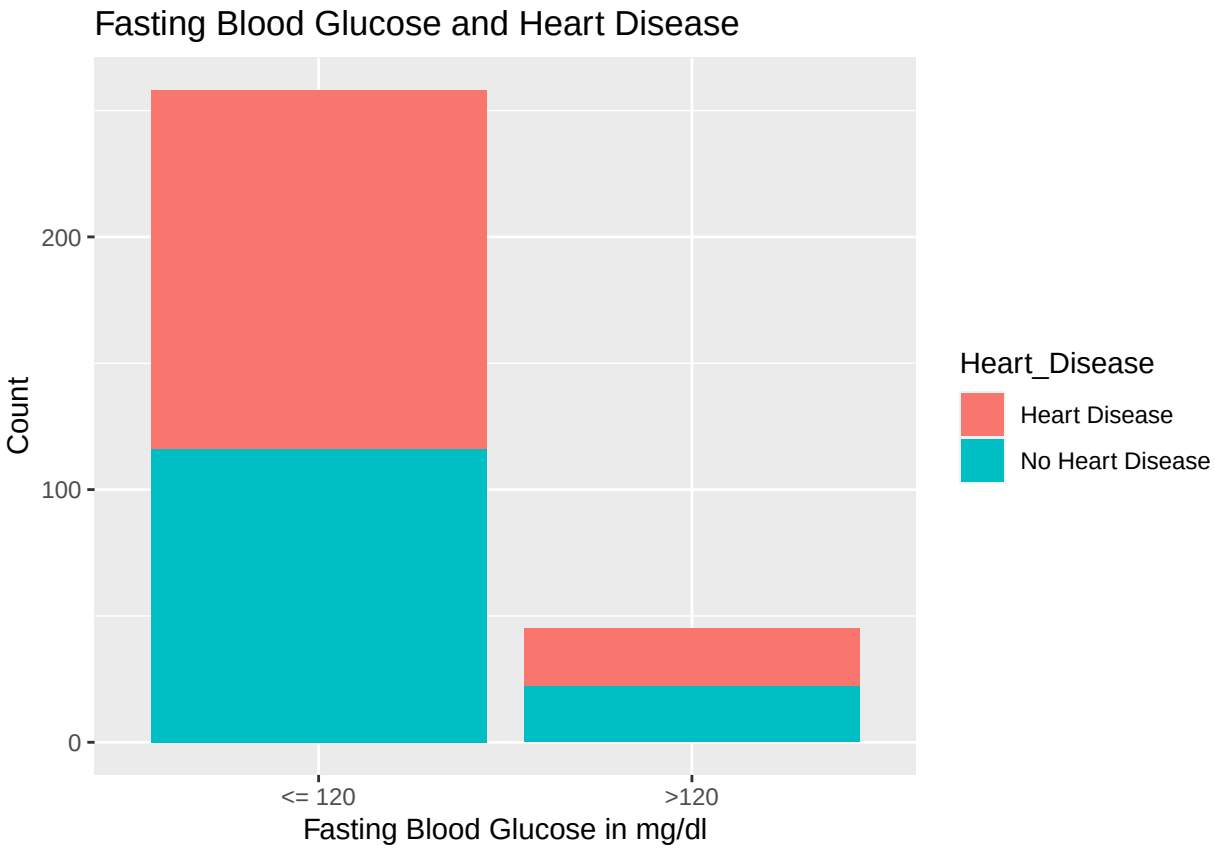



```
# calculate probability
prop.table(table(data_tidy$Heart_Disease))
```

```
##
##    Heart Disease No Heart Disease
##        0.5445545        0.4554455
```

We can see that the probability of having a heart disease is higher by approximately 84%. Therefore, we should understand what factors and why put people in the risk group of developing Heart Diseases. First of all, when you are coming to a GP, he or she will probably meassure your fasting blood glucose.

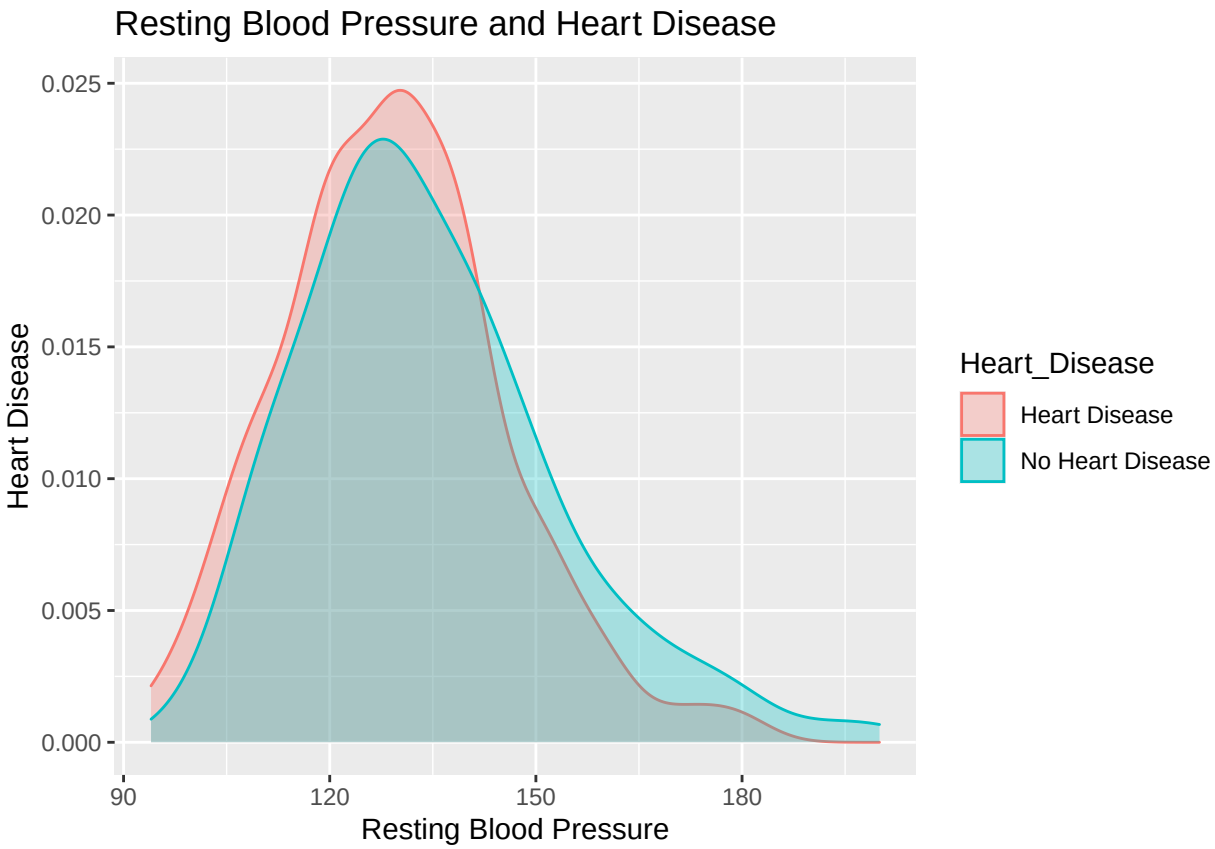
```
data_tidy %>%
  ggplot(aes(Fasting_blood_sugar, fill = Heart_Disease)) +
  geom_bar(stat = "count") +
  xlab("Fasting Blood Glucose in mg/dl") +
  ylab("Count") +
  ggtitle("Fasting Blood Glucose and Heart Disease")
```



We see that having fasting blood glucose $\leq$ 120 mg/dL cause more people to develope heart diseases. A fasting blood sugar level below 100 milligrams per deciliter (mg/dL) is considered normal. A fasting blood sugar level from 100 to 125 mg/dL is considered prediabetes. This result is sometimes called impaired fasting glucose ("Prediabetes - Diagnosis and treatment - Mayo Clinic", 2021). Long-lastging high blood glucose causes the damage of the blood vessels due to high blood pressure, and increase risk of developing atherosclerosis. People that are diagnosed with diabetes are in higher risk of developing CVD.

Blood pressure is the pressure of blood in your arteries – the vessels that carry your blood from your heart to your brain and the rest of your body. Normally, blood pressure changes over time. Some people have a condition called hypertension. Hypertension is a common condition in which the long-term force of the blood against your artery walls is high enough that it may eventually cause health problems, such as heart disease. Hypertension can be primary (unknown cause) and secondary, which is caused by another medical condition, such as renal parenchymal disease and coarctation of the aorta. From the graph below we can see that the higher blood pressure are more likely to cause heart disease.

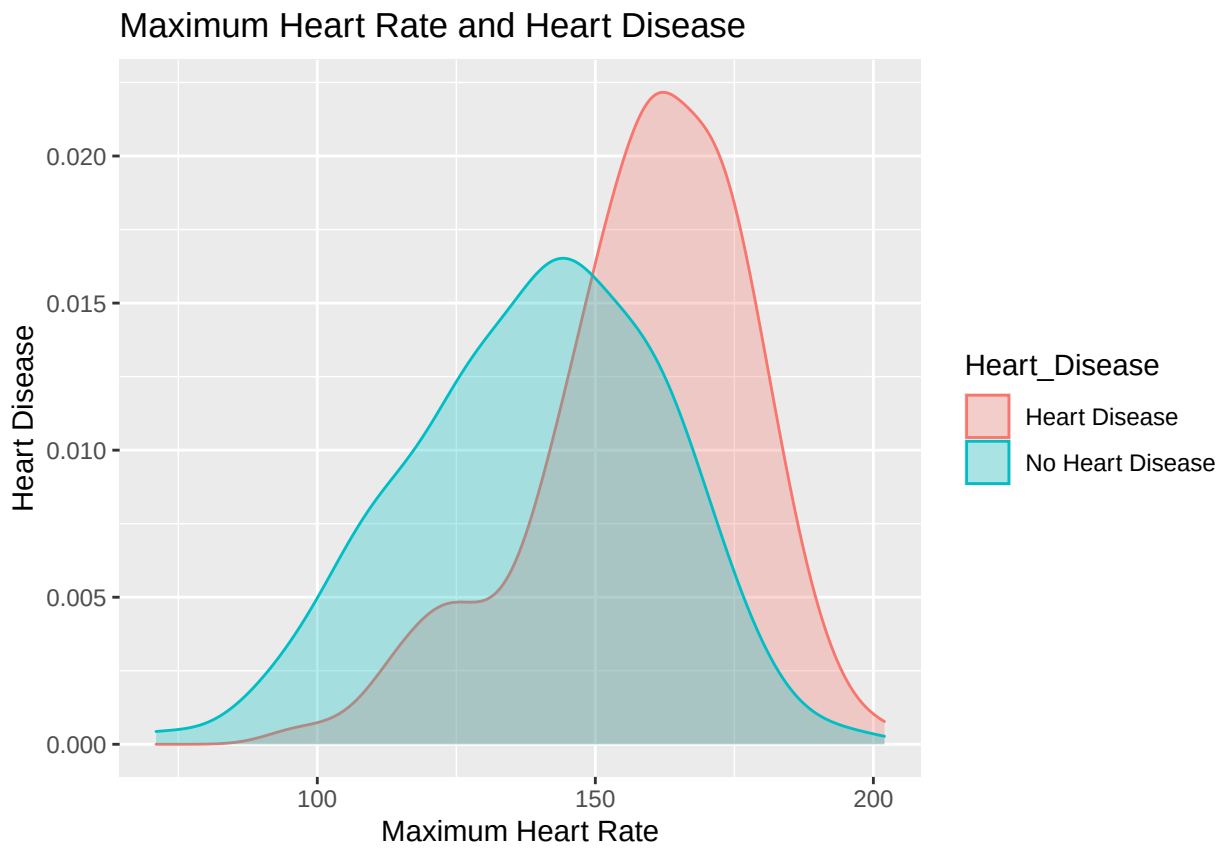
```
data_tidy %>%
  ggplot(aes(Resting_blood_pressure, fill = Heart_Disease, color = Heart_Disease)) +
  geom_density(alpha = 0.3) +
  xlab("Resting Blood Pressure") +
  ylab('Heart Disease') +
  ggtitle("Resting Blood Pressure and Heart Disease")
```



Some people think that heart rate and blood pressure increase simultaneously, however, they are two different measurements. Heart rate, also called pulse, is the number of times your heart beats per minute. Heart rate can change based on activity level, age, medication, and other factors throughout life. For most adults, a resting heart rate of 50 to 100 beats per minute is considered normal. People who exercise regularly often have lower resting heart rates.

In some situations, such as periods of acute stress or danger, blood pressure and heart rate may both increase at the same time, but that's not always the case. Your heart rate can increase without any change occurring in your blood pressure. Increased heart rate is strongly associated with coronary heart disease. The graph represents that the high HR increases the proportion of development of heart disease.

```
data_tidy %>%
  ggplot(aes(Maximum_heart_rate_achieved, fill = Heart_Disease, color = Heart_Disease)) +
  geom_density(alpha = 0.3) +
  xlab("Maximum Heart Rate") +
  ylab('Heart Disease') +
  ggtitle("Maximum Heart Rate and Heart Disease")
```



Cholesterol is essential and is needed for the production of healthyh cells and hormones, however, high cholesterol levels can develop fatty deposits in the blood vessels, which will grow and narrow arteries. Due to this the blood pressure increases and leads to the development of atherosclerosis, consenquently increasing the risk of CVD.

```
cholesterol_cat <-data.frame(Total_cholesterol = c("<200 mg/dL",
                                                   "200-239 mg/dL",
                                                   ">240 mg/dL"),
                             Category = c("Desirable", "Borderline High",
                                          "High"))
cholesterol_cat <- cholesterol_cat %>%
  rename("Total Cholesterol Levels" = Total_cholesterol,
         "Category" = Category)
knitr::kable(cholesterol_cat)
```

Total Cholesterol Levels	Category
<200 mg/dL	Desirable
200-239 mg/dL	Borderline High
>240 mg/dL	High

```
data_tidy %>%
  ggplot(aes(Cholesterol, fill = Heart_Disease, color = Heart_Disease)) +
  geom_density(alpha = 0.3) +
  xlab("Cholesterol") +
```

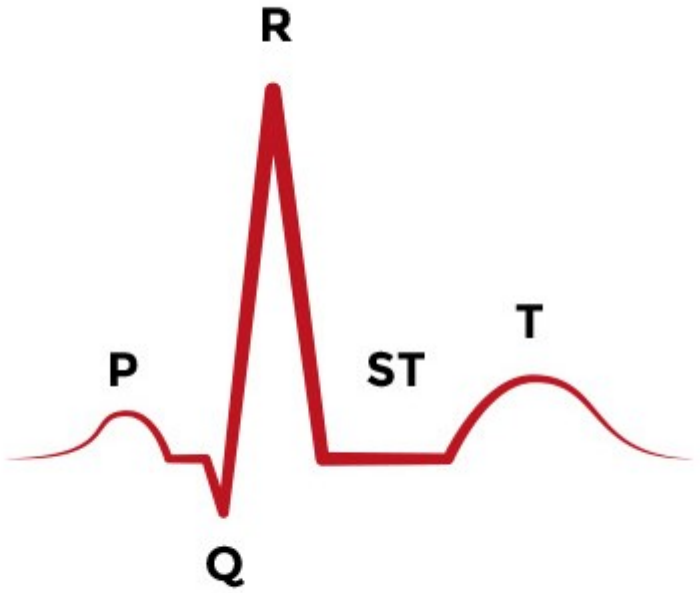
```
  ylab('Heart Disease') +
  ggtitle("Cholesterol and Heart Disease")
```

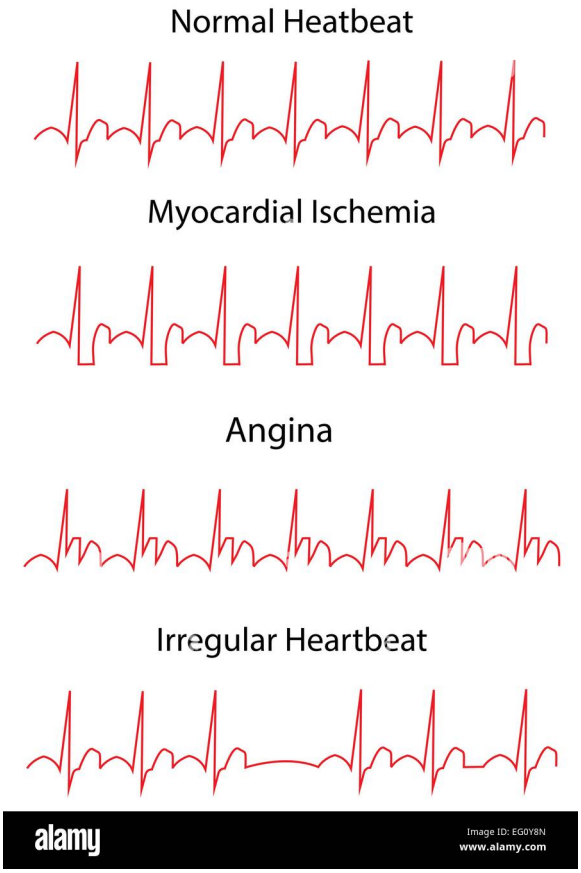
As it was said previously age and gender have an effect on the development of the CVD. Men are usually develop CVD in the younger age, meanwhile, women have higher stroke risk at older age. The difference between the risk of CVD development is also explained by hormones. Sex hormone, estrogen, has a protective actions by suppressing obesity and dyslypedimia. However, diabetes dcreases the protective actions of estrogen (Wakabayashi, 2017).

Age plays an important role in increasing the risk of developing of CVD. The American Heart Association (AHA) reports that the incidence of CVD in US men and women is ~40% from 40–59 years, ~75% from 60–79 years, and ~86% in those above the age of 80 (Rodgers et al., 2019).

```
Sex <- data_tidy %>%
  ggplot(aes(Sex, fill = Heart_Disease)) +
  geom_bar(stat = "count") +
  xlab("Sex") +
  ylab("Count") +
  ggtitle("Sex and Heart Disease")
Age <- data_tidy %>%
  ggplot(aes(Age, fill = Heart_Disease, color = Heart_Disease)) +
  geom_density(alpha = 0.3) +
  xlab("Age") +
  ylab('Heart Disease') +
  ggtitle("Age and Heart Disease")
ggarrange(Age, Sex)
```

Previously, I said that one of the diagnostic method is ECG, it detects electrical activity in th heart. It is quick and non-invasive test. ECG can detect various CVD conditions, such as coronary heart disease. ST segment on the ECG represents the interval between ventricular depolarisation and repolarisation. ST depression on the ECG indicates severe coronary lesions and large benefits of an early invasive treatment strategy in unstable coronary artery disease.


R
T
P
ST
Q



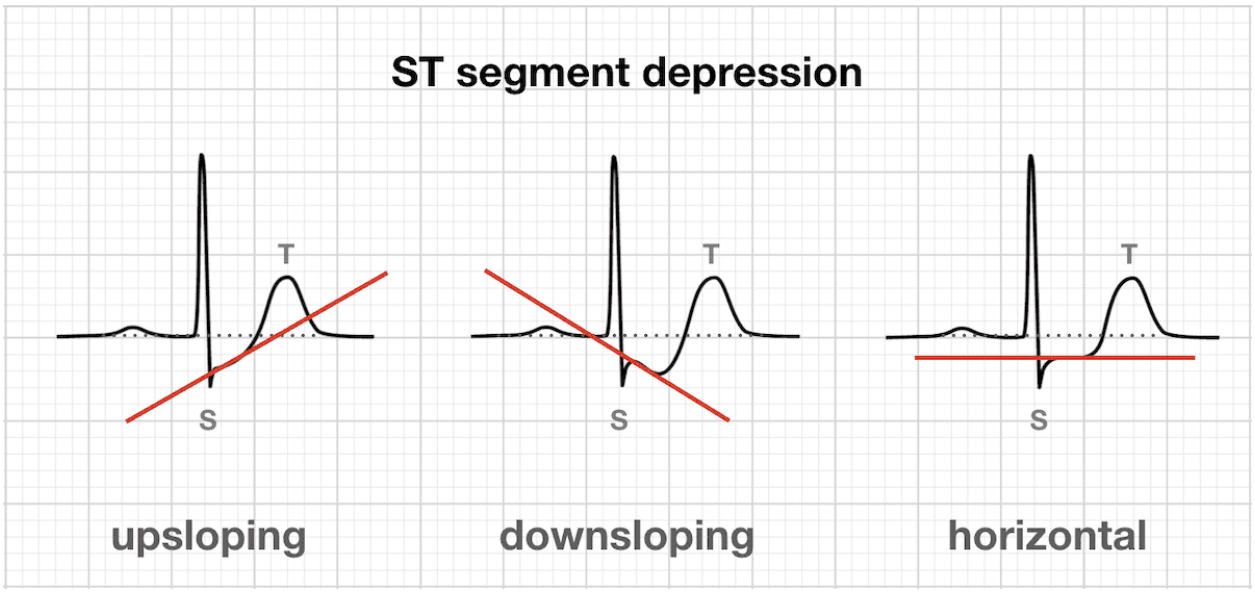
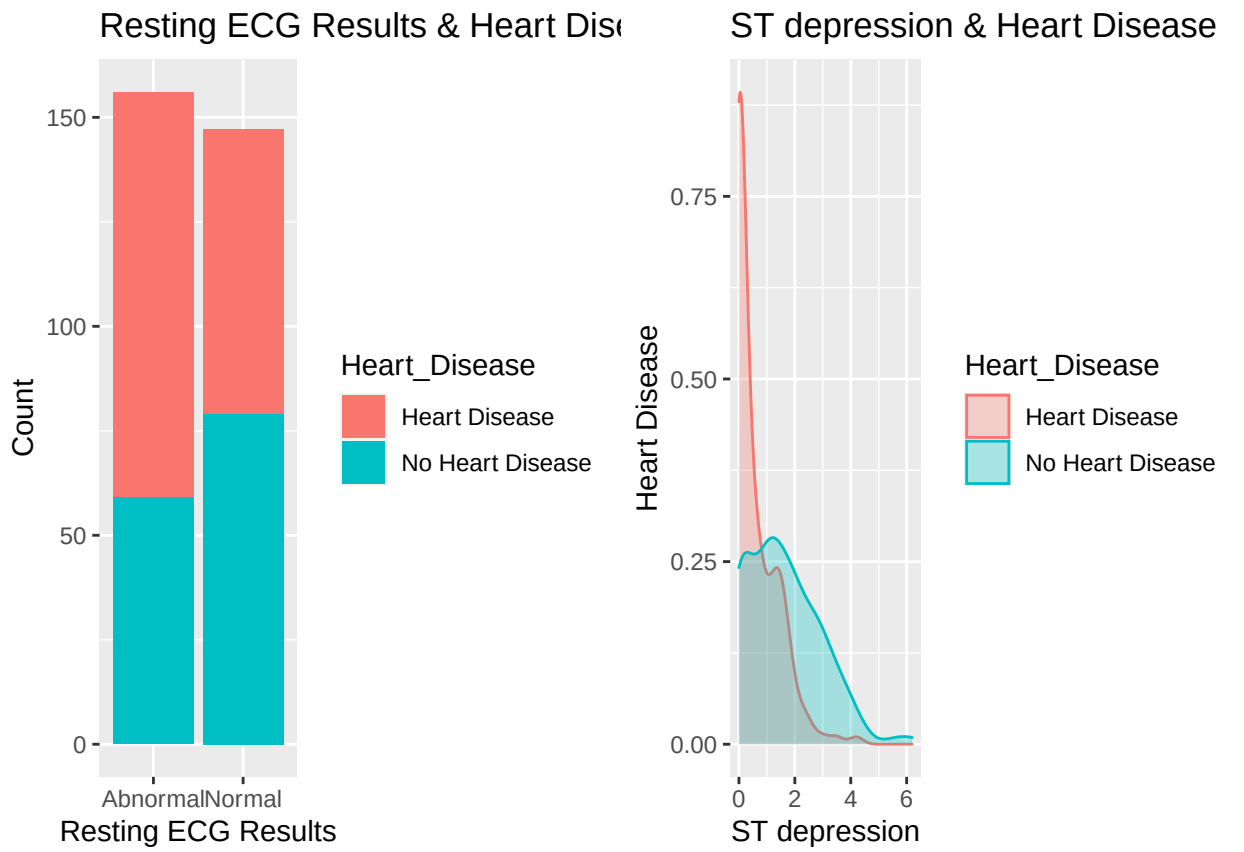


Figure 1: ST segment depression

```
restecg <- data_tidy %>%
  ggplot(aes(Resting_electrocardiographic_results, fill = Heart_Disease)) +
```

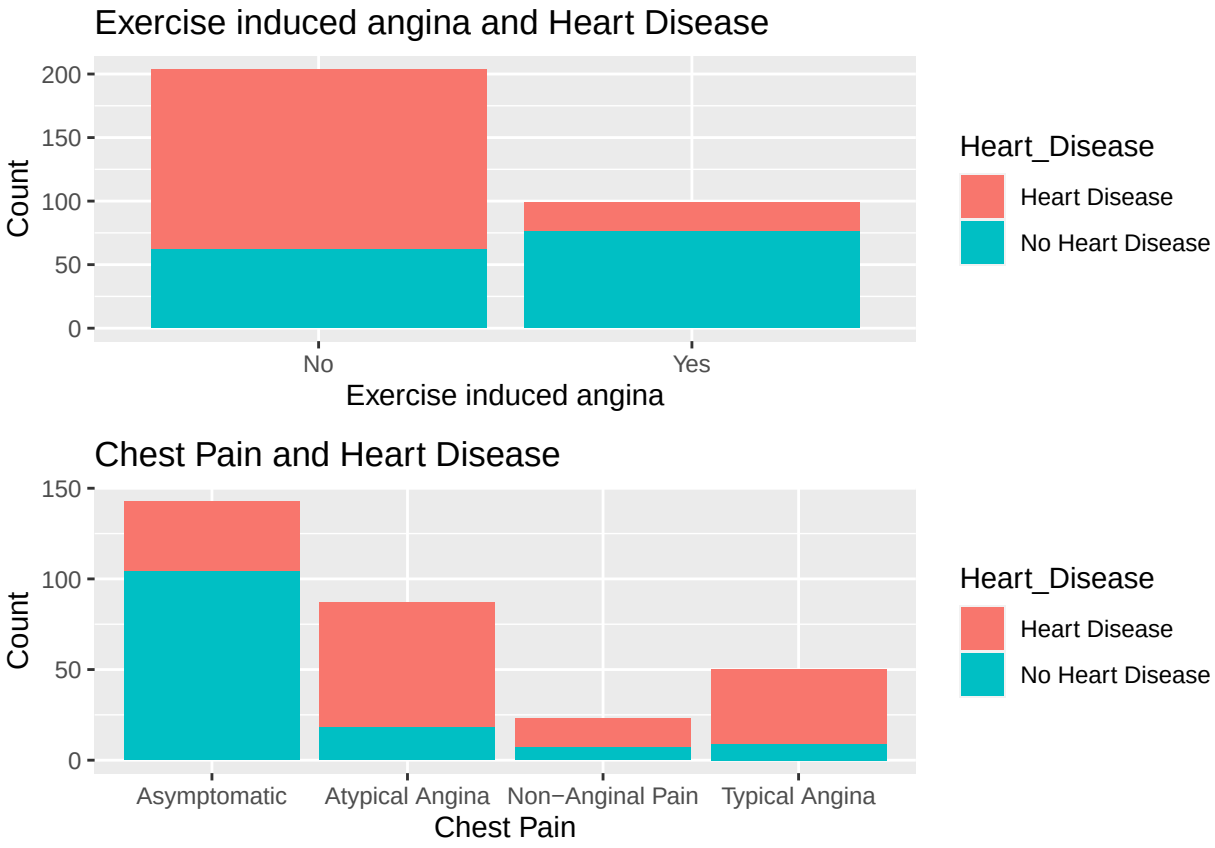
```
  geom_bar(stat = "count") +
  xlab("Resting ECG Results") +
  ylab("Count") +
  ggtitle("Resting ECG Results & Heart Disease")
ST_depr <- data_tidy %>%
  ggplot(aes(ST_depression, fill = Heart_Disease, color = Heart_Disease)) +
  geom_density(alpha = 0.3) +
  xlab("ST depression") +
  ylab('Heart Disease') +
  ggtitle("ST depression & Heart Disease")
ggarrange(restecg, ST_depr)
```



Angina is the chest pain that is caused by the reduced blood flow to the heart muscles. Angina tends to appear during physical activity, emotional stress, or exposure to cold temperatures, or after big meals. Symptoms of angina include: pressure, aching, or burning in the middle of the chest. pressure, aching, or burning in the neck, jaw, and shoulders (usually the left shoulder) and even down the arm. Anginal history is an important predictive factor for exercise-induced angina, and also of coronary artery disease disease morbidity and mortality. From the graph below we can see that exercise induced angina is not the major predictive fator of heart disease.

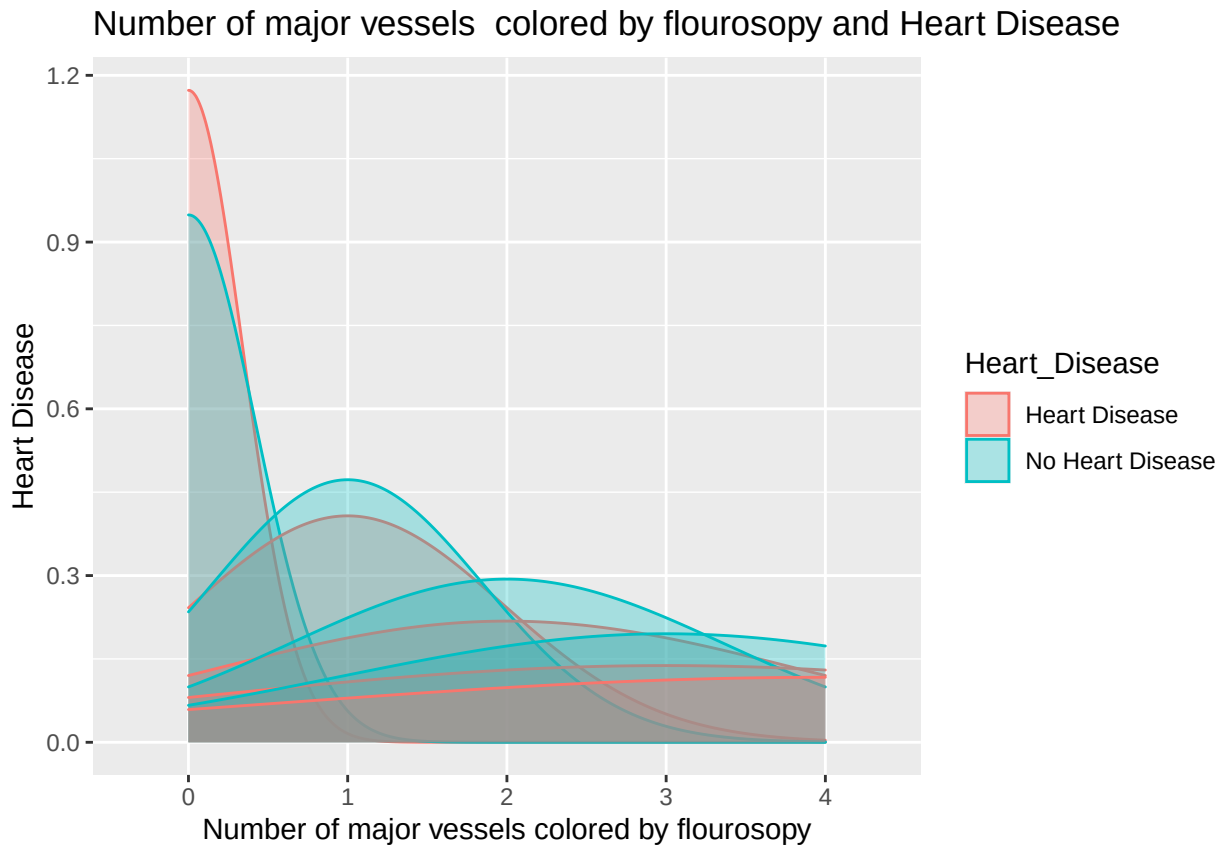
```
angina <- data_tidy %>%
  ggplot(aes(`Exercise_induced_angina `, fill = Heart_Disease)) +
  geom_bar(stat = "count") +
  xlab("Exercise induced angina") +
  ylab("Count") +
  ggtitle("Exercise induced angina and Heart Disease")
```

```
chest_pain <- data_tidy %>%
  ggplot(aes(Chest_Pain_type, fill = Heart_Disease)) +
  geom_bar(stat = "count") +
  xlab("Chest Pain") +
  ylab("Count") +
  ggtitle("Chest Pain and Heart Disease")
ggarrange(angina, chest_pain,
          nrow = 2)
```

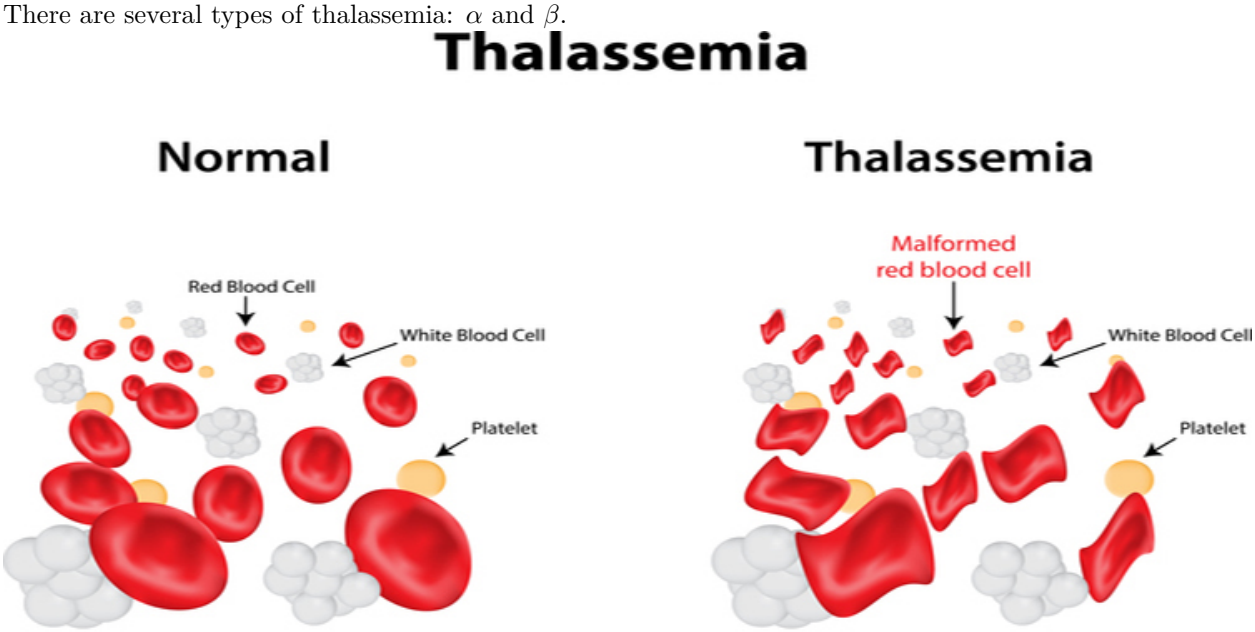


Another diagnostic method is X-ray fluoroscopy, whcih is considered as a modern real-time imaging. X-ray fluoroscopy, however, is limited for defining soft tissue and obtaining functional information.

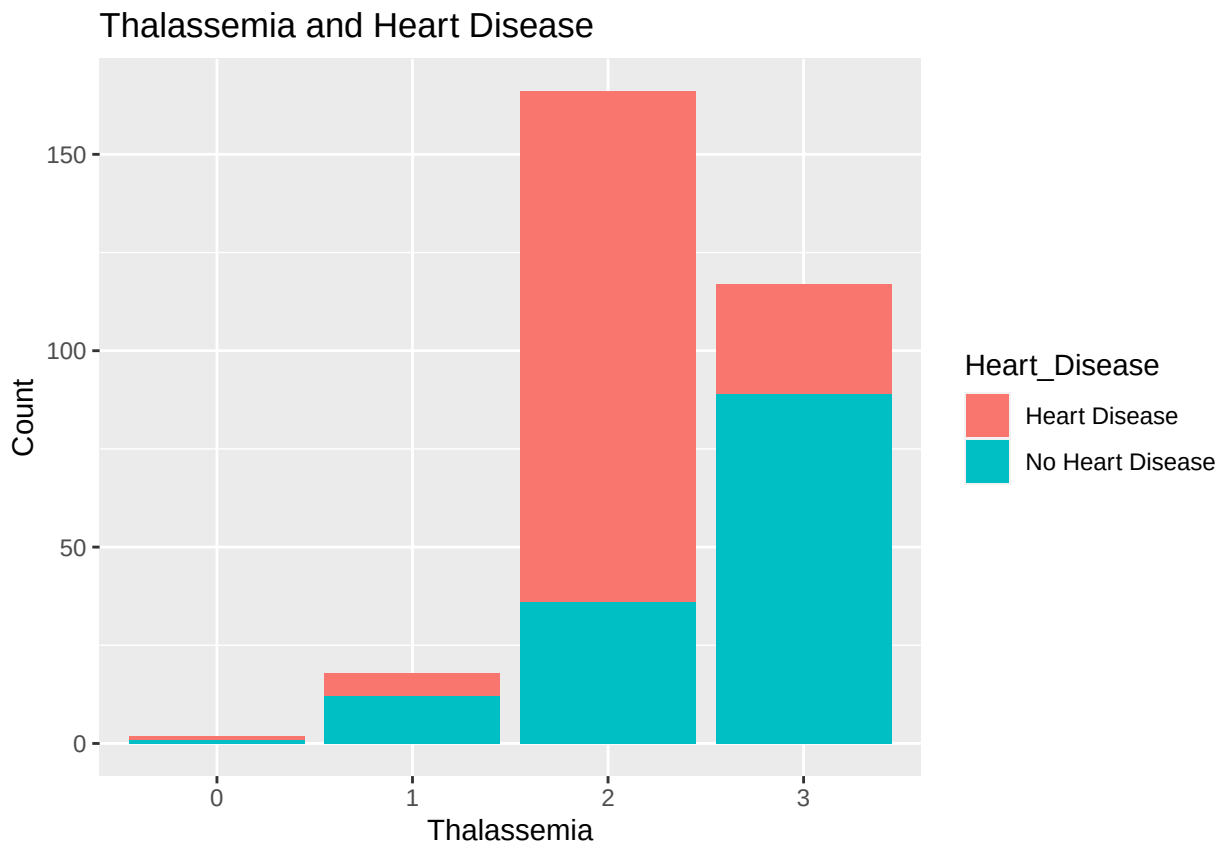
```
data_tidy %>%
  ggplot(aes(`Number_of_major_vessels `, fill = Heart_Disease, color = Heart_Disease)) +
  geom_density(alpha = 0.3) +
  xlab("Number of major vessels colored by flourosopy") +
  ylab('Heart Disease') +
  ggtitle("Number of major vessels  colored by flourosopy and Heart Disease")
```



Thalassemia is a inherited blood condition where Hemoglobin is abnormally synthesises, and therefore, affect the function of red blood cells, causing anemia. In patients with thalassemia have impaired endothelial relaxation, intimal thickening, abnormal vascular stiffening, and degeneration of elastic arteries.Patients with long-lasting anemia have a increased cardiac output, resulting in higher heart rate (Wood, 2009). There are several types of thalassemia: α and β.

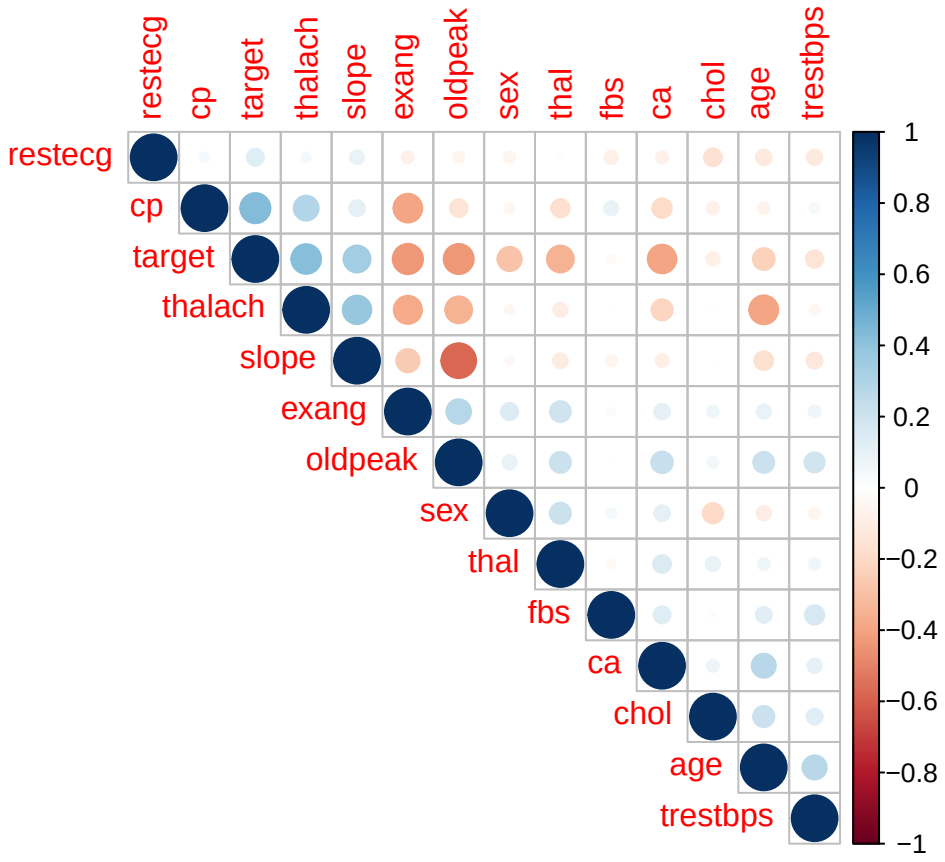


```
data_tidy %>%
  ggplot(aes(Thalassemia, fill = Heart_Disease)) +
  geom_bar(stat = "count") +
  xlab("Thalassemia") +
  ylab("Count") +
  ggtitle("Thalassemia and Heart Disease")
```



Now we can check the correlation between all variables.

```
correlation <- cor(data)
corrplot::corrplot(correlation, type="upper",
                   order="hclust")
```



Methods

In this research we will use several machine learning algorithms (MLA), such as linear regression, random forests, naive Bayes, decision tree and k-nearest neighbors with cross-validation. All MLAs have their advantages and disadvantages.

Logistic Regression is one of the most common used algorithms. Logistic regression is an extension of linear regression that assures that the estimate of conditional probability $Pr(Y = 1|X = x)$ between 1 and 0 ("Yes" or "No"). It brings the logistic transformation $g(p) = log\frac{p}{1-p}$, therefore, we model the conditional probability as this:

$$g[Pr(Y = 1|X = x)] = \beta_0 + \beta_1 x$$

Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting. Logistic models can be updated easily with new data using stochastic gradient descent. However, Logistic regression tends to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships.

Random Forests combine predictions from many individual trees. Decision trees can learn non-linear relationships, and are fairly robust to outliers. Ensembles perform very well in practice. Unconstrained, individual trees are prone to overfitting because they can keep branching until they memorize the training data. However, this can be alleviated by using ensembles.

Naive Bayes is a very simple algorithm based around conditional probability and counting.It's called "naive" because its core assumption of conditional independence rarely holds true in the real world. Even though the conditional independence assumption rarely holds true, naive Bayes models actually perform surprisingly

well in practice, especially for how simple they are. Due to their sheer simplicity, naive Bayes models are often beaten by models properly trained and tuned using the previous algorithms listed.

Decision Tree is a supervised machine learning technique that builds a decision tree from a set of class labeled training samples during the machine learning process. Decision Trees are very simple and fast with a good accuracy. However, it has a long training time. Lack of available memory, when dealing with large databases.

K-nearest neighbors follows a method for classifying objects based on the closest training examples in the feature space and is used in many applications in the field of data mining, statistical pattern recognition, and many others.

Analysis

Firstly, we shoudl split our data into training and test sets.

```
test_index <- createDataPartition(data$target, times = 1, p = 0.3, list = FALSE)
test_set <- data[test_index, ]
training_set <- data[-test_index, ]
```

Now we can strat to train our models and calculate the accuracy of each. The first MLA would be a logistic regression.

```
fit_glm <- glm(target ~., data = training_set, family = 'binomial')
y_hat <- predict(fit_glm, type = 'response',
                 newdata = test_set)
p_hat <- ifelse(y_hat > 0.5, 1, 0)
accuracy_glm <- mean(p_hat == test_set$target) #accuracy of the model
```

We obtained quite good accuracy which is 83.5%. However, we should check other MLAs. The next algorithm we will train would be Random Forests.

```
fit_rf <- randomForest(target ~., data = training_set, ntree = 1000, mtry = 1)
y_hat_rf <- predict(fit_rf, test_set)
p_hat_rf <- ifelse(y_hat_rf > 0.5, 1, 0)
accuracy_rf <- mean(p_hat_rf == test_set$target) #accuracy of the model
```

The Random forests algorithm shows a little bit higher accuracy, however it is not quite significant. Third model we will train would be Naive Bayes. However, I am expecting low a‘ccuracy levels.

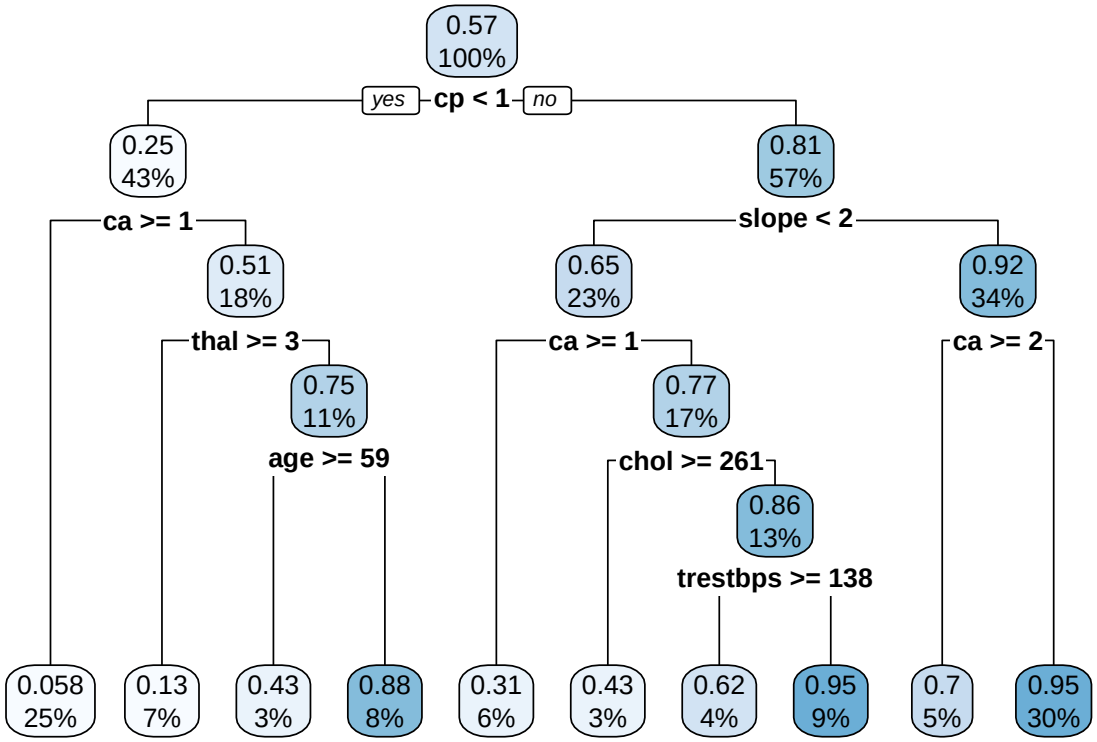
```
fit_nb <- train(as.factor(target)~., data = training_set,
                method = 'nb')
y_hat_nb <- predict(fit_nb, test_set)
p_hat_nb <- ifelse(as.numeric(y_hat_nb) > 0.5, 1, 0)
accuracy_nb <- mean(p_hat_nb == test_set$target) #accuracy of the model
```

The accuracy of the Naive Bayes model is really low and use of such MLA in this situation is inappropriate. Fourth model will be decision tree.

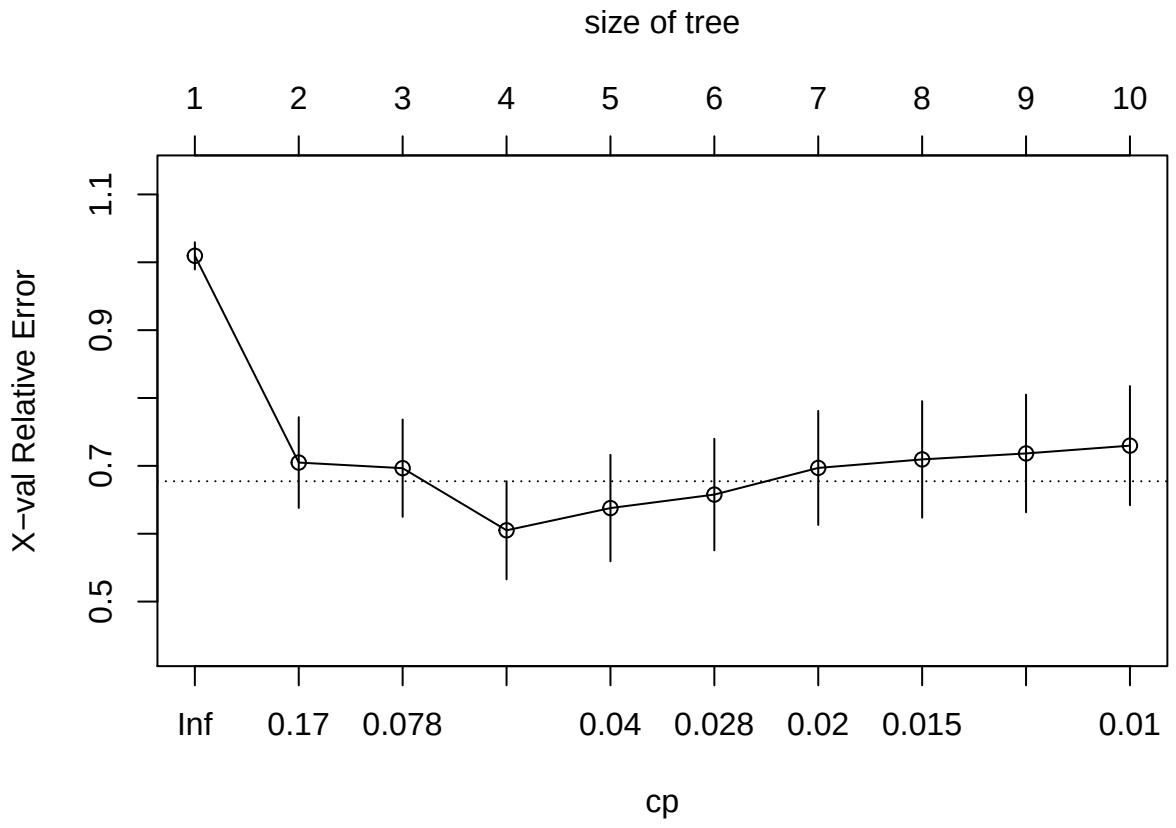
```
fit_rpart <- rpart(target ~., training_set)
```

We can visualise and choose optimal cp value to re-train model to get higher accuracy. The lower the cp value the bigger the tree, however very low cp values result in overfitting.

```
rpart.plot::rpart.plot(fit_rpart) #visualise
```




```
plotcp(fit_rpart)
```




```
printcp(fit_rpart)
```

```
##
## Regression tree:
## rpart(formula = target ~ ., data = training_set)
##
## Variables actually used in tree construction:
## [1] age      ca       chol     cp       slope    thal     trestbps
##
## Root node error: 51.939/212 = 0.24499
##
## n= 212
##
##          CP nsplit rel error  xerror     xstd
## 1  0.310439      0   1.00000 1.00953 0.020052
## 2  0.088880      1   0.68956 0.70489 0.066869
## 3  0.067585      2   0.60068 0.69659 0.071702
## 4  0.041243      3   0.53310 0.60503 0.072312
## 5  0.039248      4   0.49185 0.63778 0.078480
## 6  0.019804      5   0.45261 0.65768 0.082231
## 7  0.019658      6   0.43280 0.69696 0.084081
## 8  0.011621      7   0.41314 0.70953 0.085867
## 9  0.010584      8   0.40152 0.71831 0.086744
## 10 0.010000      9   0.39094 0.72979 0.087827
```

I decided to choose cp = 0.099284.

```
fit_rpart_2 <- rpart(target ~., training_set, cp = 0.099284) #train model with optimal cp value
y_hat_rpart <- predict(fit_rpart_2, test_set) #predictive model
p_hat_rpart <- ifelse(y_hat_rpart > 0.5, 1, 0)
accuracy_rpart <- mean(p_hat_rpart == test_set$target)
```

The decision tree model gives us good accuracy which is 79.12088%. The final MLA is a KNN model. First of all we should create train control as following:

```
control <- trainControl(method = 'repeatedcv',
                          number = 10,
                        repeats = 3)
```

Everything is ready to train Knn model.

```
fit_knn <- train(target ~.,
                 data = training_set,
                 method = 'knn',
                 trControl = control)
fit_knn
```

```
## k-Nearest Neighbors
##
## 212 samples
##  13 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 191, 191, 191, 190, 191, 191, ...
## Resampling results across tuning parameters:
##
##   k  RMSE       Rsquared   MAE
##   5  0.4949226  0.1235801  0.4161279
##   7  0.4913088  0.1091824  0.4280045
##   9  0.4798762  0.1264121  0.4235939
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.
```

The final K value was 9, so we can use tuneGrid parameter in the train function.

```
fit_knn_2 <- train(target ~.,
                 data = training_set,
                 method = 'knn',
                 trControl = control,
                 tuneGrid = data.frame(k=9))
y_hat_knn <- predict(fit_knn_2, test_set)
p_hat_knn <- ifelse(y_hat_knn > 0.5, 1, 0)
accuracy_knn <- mean(p_hat_knn == test_set$target) #accuracy of the model
```

The Knn model gave us low value of accuracy. Now we can visualise accuracies of different MLAs and decide which model suits this case the most.

```
accuracy_model <- data.frame(Model = c("Linear Regression",
                                       "Random Forest",
                                       "Naive Bayes",
                                       "Decision Tree",
                                       "Knn"),
                             Accuracy = c(accuracy_glm*100,
                                          accuracy_rf*100,
                                          accuracy_nb*100,
                                          accuracy_rpart*100,
                                          accuracy_knn*100)) %>%
  arrange(desc(Accuracy))
knitr::kable(accuracy_model)
```

Model	Accuracy
Random Forest	83.51648
Linear Regression	82.41758
Decision Tree	70.32967
Knn	61.53846
Naive Bayes	48.35165

Results

We have trained different models and found out that Random forest has the highest accuracy (84.6%), meanwhile, Naive Bayes model has the lowest accuracy (48.3%). The further assessment of accuracy can be done using more parameters.

References

1.Cardiovascular diseases (CVDs). (2021). Retrieved 6 November 2021, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2.Lopez, E., Ballard, B., & Jan, A. (2021). Cardiovascular Disease. Statpearls.

3.Lusis, A. (2000). Atherosclerosis. Nature, 407(6801), 233-241. doi: 10.1038/35025203

4.Prediabetes - Diagnosis and treatment - Mayo Clinic. (2021). Retrieved 6 November 2021, from https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284

5.Rodgers, J., Jones, J., Bolleddu, S., Vanthenapalli, S., Rodgers, L., & Shah, K. et al. (2019). Cardiovascular Risks Associated with Gender and Aging. Journal Of Cardiovascular Development And Disease, 6(2), 19. doi: 10.3390/jcdd6020019

6.Wakabayashi, I. (2017). Gender differences in cardiovascular risk factors in patients with coronary artery disease and those with type 2 diabetes. Journal Of Thoracic Disease, 9(5), E503-E506. doi: 10.21037/jtd.2017.04.30

7.Wood, J. (2009). Cardiac Complications in Thalassemia Major. Hemoglobin, 33(sup1), S81-S86. doi: 10.3109/03630260903347526