Heart diseases features selection and exploratory data analysis

Student Name: Wangjun Shen

Student ID: 110248810

The University of South Australia

Table

Introduction

The objective of this assignment is to prepare a dataset that can be used to predict heart disease, a common and serious health issue that affects a significant portion of the population. The dataset we will be working on is a subset of a larger real-world dataset collected by multiple healthcare institutions. It contains various attributes related to patients' health, such as age, gender, blood pressure, and other clinical measurements. These attributes can be leveraged to predict the presence of heart disease.

Data mining plays a critical role in addressing challenges related to predicting disease spread and similar healthcare problems. By analyzing large and complex datasets, we can identify patterns and relationships that may not be immediately apparent. This, in turn, allows us to develop more accurate predictive models. In the context of heart disease, data mining techniques can help us extract insights and patterns from patient data. These insights can aid in identifying risk factors, predicting disease outcomes, and developing effective treatment strategies.

For example, by using data mining techniques on the given dataset, we can identify the most important predictors of heart disease. This knowledge can then be used to develop a classification model that accurately predicts the presence of the disease in patients. Moreover, data mining can also help us identify subpopulations that are more susceptible to the disease. We can then develop tailored prevention and treatment strategies for these subpopulations.

Overall, this assignment provides an opportunity to apply data mining techniques to a real-world dataset. It also allows us to gain hands-on experience with feature engineering, data exploration, and predictive modeling. The insights gained from this exploration can inform our work in subsequent assignments. This, in turn, can help us develop more accurate and effective predictive models for heart disease and other similar health challenges.

Related Work

The paper "Predictive data mining for medical diagnosis: An overview of heart disease prediction" by Subramanian and Krishnan (2013) aimed to provide an overview of predictive data mining techniques for the diagnosis of heart disease. The authors focused on identifying the key predictive features for heart disease and evaluating the effectiveness of various predictive models. The study used a dataset consisting of 303 patients with 14 features related to the diagnosis of heart disease, including age, sex, blood pressure, and cholesterol levels. The authors used various classification algorithms, including decision trees, k-nearest neighbors, and artificial neural networks, to build predictive models for heart disease diagnosis. The results showed that decision trees and artificial neural networks performed better than other models in predicting heart disease. The study provides insights into the potential of predictive data mining techniques in improving the accuracy of heart

disease diagnosis. This study informs my research on heart disease prediction by highlighting the importance of feature selection and the effectiveness of various predictive models. It provides guidance on the selection of appropriate predictive models and features that can improve the accuracy of heart disease diagnosis.

The study "Heart Disease Prediction Using Machine Learning Techniques: A Review" by Singh et al. (2020) aimed to provide an overview of various machine learning (ML) techniques used for predicting heart disease. The authors reviewed several studies that used ML algorithms to predict heart disease and identified the key features used in those studies. The study included features such as age, sex, blood pressure, cholesterol levels, and smoking habits. The authors compared the performance of various ML algorithms, including decision trees, support vector machines (SVM), and artificial neural networks (ANN), in predicting heart disease. The results showed that ANN performed better than other algorithms in predicting heart disease. The study provides insights into the potential of ML techniques in improving the accuracy of heart disease prediction. This study informs my research on heart disease prediction by providing a comprehensive review of various ML techniques used for heart disease prediction. It highlights the importance of feature selection and the effectiveness of various ML algorithms in predicting heart disease.

The paper "An Analysis of Heart Disease Prediction using Machine Learning Techniques" aimed to predict heart disease by using machine learning algorithms. The authors used several features including age, sex, chest pain, resting blood pressure, serum cholesterol, maximum heart rate, and other clinical parameters. The study utilized various algorithms, such as decision trees, Naive Bayes, and k-nearest neighbor (KNN). Among the algorithms, the authors found that the KNN algorithm was the most effective for heart disease prediction (Bhatla & Jyoti, 2012). The study informed my research by providing a baseline for feature selection and algorithms to use in predicting heart disease. Additionally, the study demonstrated that machine learning techniques could be effectively used to predict heart disease, supporting the potential of using such techniques in clinical decision-making.

Data Exploration

Features Selection

The details of the original dataset are as follows.

Table 1: Details for each variable of the original dataset.

Variable	Description	Type
id	A unique ID that identifies a participant in the study	Numerical
age	Age in years	Numerical
sex	Male and Female were recorded	Categorical
cp	Chest Pain type: typical angina; atypical angina; non-	Categorical

Variable	Description	Type
	anginal pain; and asymptomatic	
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Numerical
chol	Serum Cholestoral in mg/dl	Numerical
fbs	Fasting blood sugar > 120 mg/dl (True or False)	Boolean
restecg	Resting electrocardiographic results: normal; having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) or showing probable or definite left ventricular hypertrophy by Estes' criteria	Categorical
thalach	Maximum heart rate achieved	Numerical
exang	Exercise induced angina (True/False)	Boolean
oldpeak	ST depression induced by exercise relative to rest	Numerical
slope	The slope of the peak exercise ST segment: upsloping; flat; downsloping	Categorical
major_vessels	Number of major vessels (0-3) colored by flourosopy	Numerical
restwm	Rest wall motion abnormality: none; mild or moderate; moderate or severe; akinesis or dyskmem	Categorical
target	Heart disease diagnosed (disease/no disease)	Categorical

Import the dataset and view.

```
# load dataset first
heart.full <- read.csv("heart.csv")

# then check the dataset
str(heart.full)

## 'data.frame':    1025 obs. of  15 variables:
##  $ id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age          : int  52 53 70 61 62 58 58 55 46 54 ...
##  $ sex          : chr  "male" "male" "male" "male" ...
##  $ cp           : chr  "typical angina" "typical angina" "typical an
gina" "typical angina" ...
##  $ 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          : logi  FALSE TRUE FALSE FALSE TRUE FALSE ...
##  $ restecg      : chr  "ST-T wave abnormality" "normal" "ST-T wave a
bnormality" "ST-T wave abnormality" ...
##  $ thalach      : int  168 155 125 161 106 122 140 145 144 116 ...
##  $ exang        : logi  FALSE TRUE TRUE FALSE FALSE FALSE ...
##  $ oldpeak      : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
##  $ slope        : chr  "downsloping" "upsloping" "upsloping" "downsl
oping" ...
```

```
##  $ major_vessels: int  2 0 0 1 3 0 3 1 0 2 ...
##  $ restwm       : chr  "akinesis or dyskmem" "akinesis or dyskmem" "
akinesis or dyskmem" "akinesis or dyskmem" ...
##  $ target       : chr  "no disease" "no disease" "no disease" "no di
sease" ...
```

From the output results, it can be found that the dataset contains 15 variables and 1025 observation. The details of each variable are consistent with the description.

Then check the distribution of missing values in the dataset.

```
sum(is.na(heart.full))

## [1] 0
```

There are not missing values in this data.

The dataset is undergoing a transformation from its original data types to appropriate data types required for analysis. Originally, the variables age, trestbps, chol, thalach, and oldpeak were imported as character vectors representing age, resting blood pressure, serum cholesterol, maximum heart rate achieved, and ST depression induced by exercise relative to rest, respectively. These variables have been converted to a numeric data type since they represent numerical measurements. Similarly, the variables sex, cp, restecg, slope, restwm, and target were originally imported as character vectors representing sex, chest pain type, resting electrocardiographic results, slope of the peak exercise ST segment, presence of a major vessels colored by fluoroscopy, and heart disease status, respectively. These variables have been converted to factor data type since they represent categorical variables. This transformation allows for easier data manipulation and analysis, especially when exploring relationships between variables.

```
# convert age, trestbps, chol, thalach, and oldpeak to numeric
heart.full$age <- as.numeric(heart.full$age)
heart.full$trestbps <- as.numeric(heart.full$trestbps)
heart.full$chol <- as.numeric(heart.full$chol)
heart.full$thalach <- as.numeric(heart.full$thalach)
heart.full$oldpeak <- as.numeric(heart.full$oldpeak)


# convert sex, cp, restecg, slope, restwm, and target to factor
heart.full$sex <- as.factor(heart.full$sex)
heart.full$cp <- as.factor(heart.full$cp)
heart.full$restecg <- as.factor(heart.full$restecg)
heart.full$slope <- as.factor(heart.full$slope)
heart.full$restwm <- as.factor(heart.full$restwm)
heart.full$target <- as.factor(heart.full$target)
```

Then use random forest to detect the importance of each feature. In order to understand whether these features differ between genders, the data will firstly be created based on the sub-dataset.

```
# dataset for each gender
heart.male <- subset(heart.full, sex == "male")
heart.female <- subset(heart.full, sex == "female")
```

Then use those two sub datasets to build the random forest.

For male sub-dataset.

```
# Random forest for male subset
library(randomForest)

## Warning: package 'randomForest' was built under R version 4.2.2

## randomForest 4.7-1.1

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

set.seed(123)
rf_model_male <- randomForest(target ~ ., data = heart.male, importance
 = TRUE, ntree = 500)
importance(rf_model_male)

##                 disease no disease MeanDecreaseAccuracy MeanDecrease
Gini
## id             1.606825   3.213858            3.376221        11.15
4653
## age           33.928469  40.857084           42.838250        34.44
8119
## sex            0.000000   0.000000            0.000000         0.00
0000
## cp            38.745483  36.059468           41.121602        47.16
1819
## trestbps      33.903759  37.024285           41.751265        25.96
2476
## chol          34.122660  35.756098           39.121768        29.97
2875
## fbs           16.795603  16.203869           19.636756         3.76
1825
## restecg       19.083360  19.416802           21.509917         5.53
1215
## thalach       37.581333  37.899592           43.252118        51.57
2260
## exang         17.532815  19.401905           20.468760        11.93
3409
## oldpeak       32.589020  38.902000           40.817055        37.10
8368
## slope         24.318522  24.566740           27.686364        17.11
1266
```

```
## major_vessels 37.680805  40.115231            43.835239         42.18
4090
## restwm        31.201322  32.398437            34.984151         25.36
7047
```

According to the provided correlation matrix, the most important features for predicting the presence of heart disease are: cp, thalach, major_vessels, oldpeak, trestbps, and age. These features have the highest correlation with the target variable (disease/no disease), as well as high values for MeanDecreaseAccuracy and MeanDecreaseGini, indicating that they are crucial predictors for a machine learning model.

To facilitate observation, data can be visualized.

```
varImpPlot(rf_model_male, col = "red", pch = 20)
```

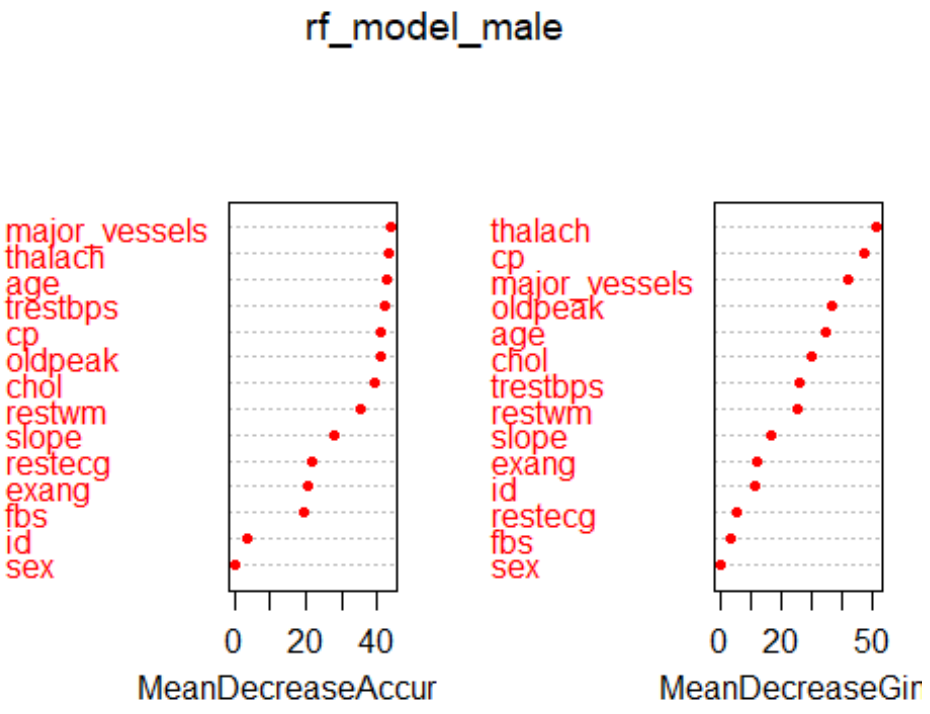

Figure 1 RandomForest Model for Male

The feature chest pain type (cp) has high values for both MeanDecreaseAccuracy and MeanDecreaseGini, suggesting that it is a strong predictor for heart disease. Maximum heart rate achieved (thalach) and the number of major vessels colored by fluoroscopy (major_vessels) also have high correlations and values for MeanDecreaseAccuracy and MeanDecreaseGini, making them strong predictors as well. Though ST depression induced by exercise relative to rest (oldpeak) and resting blood pressure (trestbps) have lower correlations, they still have high values for MeanDecreaseAccuracy and MeanDecreaseGini, indicating their importance in

predicting the presence of heart disease. Finally, age is also an important feature as it has a moderate correlation with the target variable and a relatively high value for MeanDecreaseAccuracy.

In conclusion, these six features can be considered the most important for predicting the presence of heart disease in males in this dataset.

For female sub-dataset.

```
# Random forest for female subset
set.seed(123)
rf_model_female <- randomForest(target ~ ., data = heart.female, import
ance = TRUE, ntree = 500)
importance(rf_model_female)
```

```
##                  disease no disease MeanDecreaseAccuracy MeanDecrease
Gini
## id              1.932703  -3.148355           -0.5202009         2.59
0325
## age            23.376089  24.724758           28.2364781        12.01
8921
## sex             0.000000   0.000000            0.0000000         0.00
0000
## cp             19.702221  20.271296           22.9715941        12.53
7614
## trestbps       18.409296  18.651176           22.1370717         8.42
9236
## chol           19.941082  22.898914           25.6314494         8.32
1454
## fbs             7.611649   9.737944           10.2944996         1.41
8907
## restecg        14.299858  15.552097           17.3754050         2.98
2791
## thalach        19.430050  20.969639           24.2566433         8.27
2732
## exang          18.303884  21.574578           22.6248431         9.16
7964
## oldpeak        20.677704  21.784792           24.3639560        16.48
6048
## slope          17.218824  19.973871           21.5111294         8.15
0365
## major_vessels 19.573486  20.510840           22.7582749        10.40
7631
## restwm         22.156482  24.907233           26.5256308        22.29
0165
```

According to the provided correlation matrix, the following features are significant in predicting the presence of heart disease in women:

- Age: It exhibits a high correlation with both disease and non-disease and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.

- cp (Chest pain type): It exhibits a moderate correlation with disease and non-disease, and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
- thalach (Maximum heart rate achieved): It exhibits a moderate correlation with disease and non-disease and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
- oldpeak (ST depression induced by exercise relative to rest): It exhibits a moderate correlation with disease and non-disease and has a high value for MeanDecreaseAccuracy.
- major_vessels (Number of major vessels (0-3) colored by flourosopy): It exhibits a moderate correlation with disease and non-disease and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
- restwm (resting wall motion abnormalities): It exhibits a moderate correlation with disease and non-disease and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
- exang (exercise-induced angina): It exhibits a moderate correlation with both disease and non-disease and has a high value for MeanDecreaseGini.

On the other hand, “trestbps”, “chol”, “fbs”, and “restecg” display moderate correlations with disease and non-disease but have relatively lower values for MeanDecreaseAccuracy and MeanDecreaseGini. As a result, they are not included in the list of important features.

Overall, these seven features are the most significant in predicting the presence of heart disease in women using this dataset.

To facilitate observation, this is a visualization of the results.

```
varImpPlot(rf_model_female, col = "red", pch = 20)
```

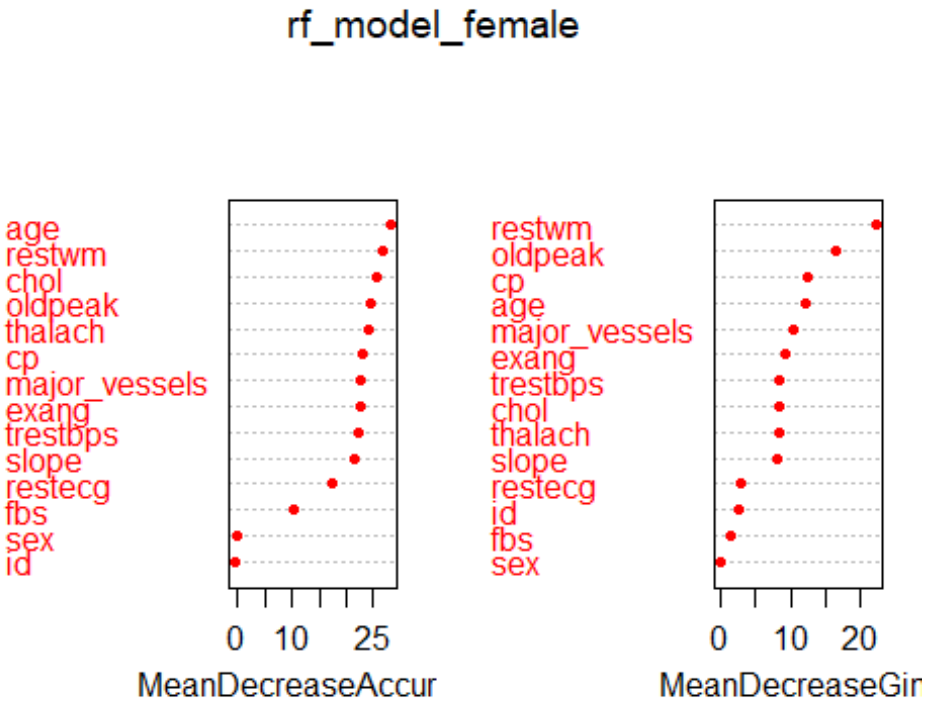

Figure 2 RandomForest Model for Female

Therefore, after comprehensive consideration, we have decided to retain the following features: age, cp, thalach, oldpeak, major_vessels, restwm, exang, and sex.

Now create a new data set with those selected features.

```
# Create new dataset with selected features and target variable
heart_features_selected <- heart.full[, c("id", "age", "sex", "cp", "th
alach", "exang", "oldpeak", "major_vessels", "restwm", "target")]

# Save new dataset as CSV file
write.csv(heart_features_selected, "heart_features_selected.csv", row.n
ames = FALSE)
```

Descriptive Statistics

Import the new data set and check the details of that new data set:

```
heart.selected <- read.csv("heart_features_selected.csv")

str(heart.selected)

## 'data.frame':    1025 obs. of  10 variables:
##  $ id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age          : int  52 53 70 61 62 58 58 55 46 54 ...
##  $ sex          : chr  "male" "male" "male" "male" ...
##  $ cp           : chr  "typical angina" "typical angina" "typical an
```

```
gina" "typical angina" ...
##  $ thalach      : int  168 155 125 161 106 122 140 145 144 116 ...
##  $ exang        : logi  FALSE TRUE TRUE FALSE FALSE FALSE ...
##  $ oldpeak      : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
##  $ major_vessels: int  2 0 0 1 3 0 3 1 0 2 ...
##  $ restwm       : chr  "akinesis or dyskmem" "akinesis or dyskmem" "
akinesis or dyskmem" "akinesis or dyskmem" ...
##  $ target       : chr  "no disease" "no disease" "no disease" "no di
sease" ...
```

Summarize the selected data set.

```
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.2

## Warning: package 'readr' was built under R version 4.2.2

## Warning: package 'purrr' was built under R version 4.2.2

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.2

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyve
rse 2.0.0 ──
## ✔ dplyr     1.1.1     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_co
nflicts() ──
## ✖ dplyr::combine()  masks randomForest::combine()
## ✖ dplyr::filter()   masks stats::filter()
## ✖ dplyr::lag()      masks stats::lag()
## ✖ ggplot2::margin() masks randomForest::margin()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to
force all conflicts to become errors

heart.selected$id <- as.numeric(heart.selected$id)
heart.selected$age <- as.numeric(heart.selected$age)
```

```
heart.selected$sex <- as.factor(heart.selected$sex)
heart.selected$cp <- as.factor(heart.selected$cp)
heart.selected$thalach <- as.numeric(heart.selected$thalach)
heart.selected$oldpeak <- as.numeric(heart.selected$oldpeak)
heart.selected$major_vessels <- as.numeric(heart.selected$major_vessels)
heart.selected$target <- as.factor(heart.selected$target)


summary(heart.selected[, -1])
```

```
##       age            sex                      cp          thalach     
##  Min.   :29.00   female:312   asymptomatic    : 77   Min.   : 71.0  
##  1st Qu.:48.00   male  :713   atypical angina :167   1st Qu.:132.0  
##  Median :56.00                non-anginal pain:284   Median :152.0  
##  Mean   :54.43                typical angina  :497   Mean   :149.1  
##  3rd Qu.:61.00                                       3rd Qu.:166.0  
##  Max.   :77.00                                       Max.   :202.0  
##    exang           oldpeak      major_vessels       restwm         
##  Mode :logical   Min.   :0.000   Min.   :0.0000   Length:1025       
##  FALSE:680       1st Qu.:0.000   1st Qu.:0.0000   Class :character  
##  TRUE :345       Median :0.800   Median :0.0000   Mode  :character  
##                  Mean   :1.072   Mean   :0.7541                     
##                  3rd Qu.:1.800   3rd Qu.:1.0000                     
##                  Max.   :6.200   Max.   :4.0000                     
##         target     
##  disease   :526  
##  no disease:499  
##                  
##                  
##                  
##
```

This result shows the statistical summary of a dataset related to heart disease. The dataset consists of 1025 observations and several variables, including age, sex, chest pain type, maximum heart rate, exercise-induced angina, ST depression induced by exercise relative to rest, number of major vessels colored by fluoroscopy, and the presence or absence of heart disease.

The average age of the patients in the dataset is 54.43 years, with a minimum age of 29 years and a maximum age of 77 years. Out of the 1025 patients, 312 are female and 713 are male. Chest pain type is categorized into four types, namely asymptomatic, atypical angina, non-anginal pain, and typical angina. The most frequent type is typical angina, with 497 occurrences. The maximum heart rate (thalach) ranges from 71 to 202 beats per minute, with a mean of 149.1 beats per minute. Exercise-induced angina (exang) is a binary variable, with 345 patients experiencing it during exercise and 680 patients not experiencing it. ST depression induced by exercise relative to rest (oldpeak) ranges from 0 to 6.2, with an average value of 1.072. The number of major vessels colored by fluoroscopy (major_vessels) ranges from 0 to 4, with a mean of 0.7541. The dataset is labeled with the presence

or absence of heart disease (target). Out of the 1025 patients, 526 have heart disease and 499 do not have heart disease.

After examining the results, it is clear that there is room for further analysis. For example, while the report provides some information about the results, it does not provide a complete picture. In order to gain a deeper understanding, it would be helpful to explore additional metrics such as standard deviation, skewness, and kurtosis. This would allow us to better understand the distribution of the data and identify any outliers or patterns that may be present. By conducting a more thorough analysis, we can gain a more comprehensive understanding of the data and make more informed decisions based on the results.

```
library(psych)

## Warning: package 'psych' was built under R version 4.2.3

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

## The following object is masked from 'package:randomForest':
##
##     outlier

describe(heart.selected[, -1])

## Warning in FUN(newX[, i], ...): no non-missing arguments to min; ret
urning Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; ret
urning -Inf

##               vars    n   mean    sd median trimmed   mad min   max
range  skew
## age              1 1025  54.43  9.07   56.0   54.66  8.90  29  77.0
 48.0 -0.25
## sex*             2 1025   1.70  0.46    2.0    1.74  0.00   1   2.0
  1.0 -0.85
## cp*              3 1025   3.17  0.96    3.0    3.31  1.48   1   4.0
  3.0 -0.86
## thalach          4 1025 149.11 23.01  152.0  150.40 23.72  71 202.0
131.0 -0.51
## exang            5 1025    NaN    NA     NA     NaN    NA Inf  -Inf
 -Inf    NA
## oldpeak          6 1025   1.07  1.18    0.8    0.89  1.19   0   6.2
  6.2  1.21
## major_vessels    7 1025   0.75  1.03    0.0    0.57  0.00   0   4.0
  4.0  1.26
```

```
## restwm*          8 1025   2.14  0.97    3.0    2.17  0.00   1   4.0
  3.0 -0.25
## target*          9 1025   1.49  0.50    1.0    1.48  0.00   1   2.0
  1.0  0.05
##                kurtosis   se
## age               -0.53 0.28
## sex*              -1.28 0.01
## cp*               -0.39 0.03
## thalach           -0.10 0.72
## exang                NA   NA
## oldpeak            1.29 0.04
## major_vessels      0.68 0.03
## restwm*           -1.81 0.03
## target*           -2.00 0.02
```

This study provides a summary of the results obtained from a sample of subjects with different variables related to heart disease. The subjects had an average age of 54.43 years, with a standard deviation of 9.07. The majority of subjects were male, with an average value of 1.70 and a standard deviation of 0.46. The average level of chest pain experienced by the subjects was 3.17, with a standard deviation of 0.96. The average maximum heart rate achieved by the subjects was 149.11 bpm, with a standard deviation of 23.01 bpm. The average ST depression induced by exercise was 1.07 mm, with a standard deviation of 1.18 mm. The average number of major vessels colored by fluoroscopy was 0.75, with a standard deviation of 1.03. The average resting wall motion score index was 2.14, with a standard deviation of 0.97. The majority of subjects did not have heart disease, with a mean value of 1.49 and a standard deviation of 0.50. These findings provide important insights into the characteristics of the sample and may inform future research on heart disease.

```
par(mfrow=c(1,2)) # To plot the histograms side by side
hist(heart.selected$age[heart.selected$sex == "male"], main = "Age Dist
ribution for Males", xlab = "Age for male")
hist(heart.selected$age[heart.selected$sex == "female"], main = "Age Di
stribution for Females", xlab = "Age for female")
```

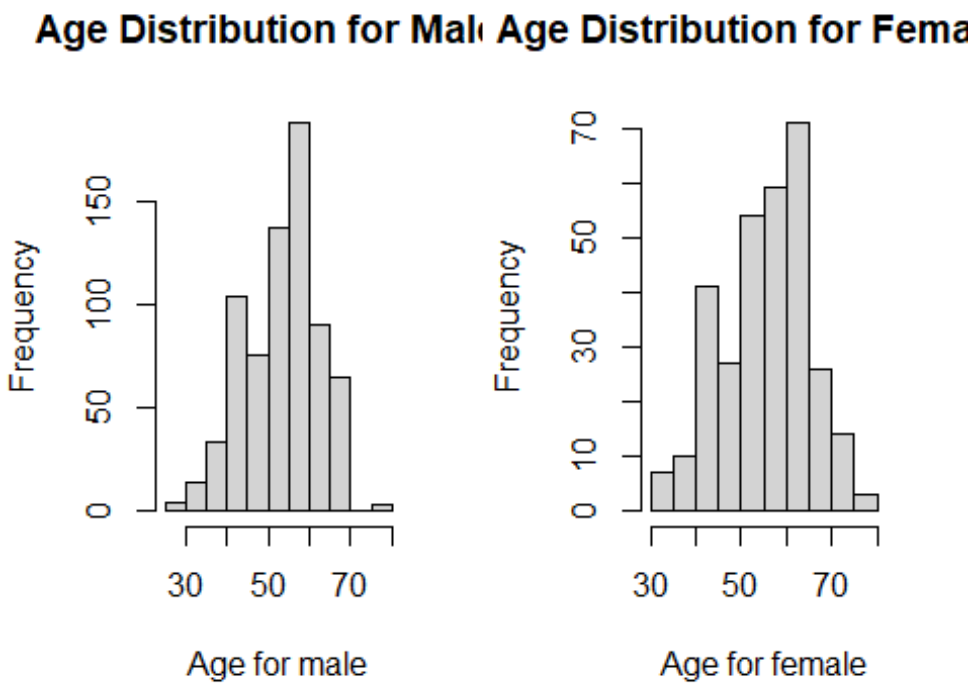

Figure 3 Age Distribution for Male and Female

In terms of male age, the distribution is generally symmetric, or slightly left-skewed. When compared to the male age distribution, the female age distribution is more left-skewed. Additionally, it's worth noting that there are more females over the age of 70 than males.

```
boxplot(heart.selected$age ~ heart.selected$target,
        main="Heart disease diagnosis distribution by Age",
         ylab="Age",xlab="Heart disease diagnosed")
```

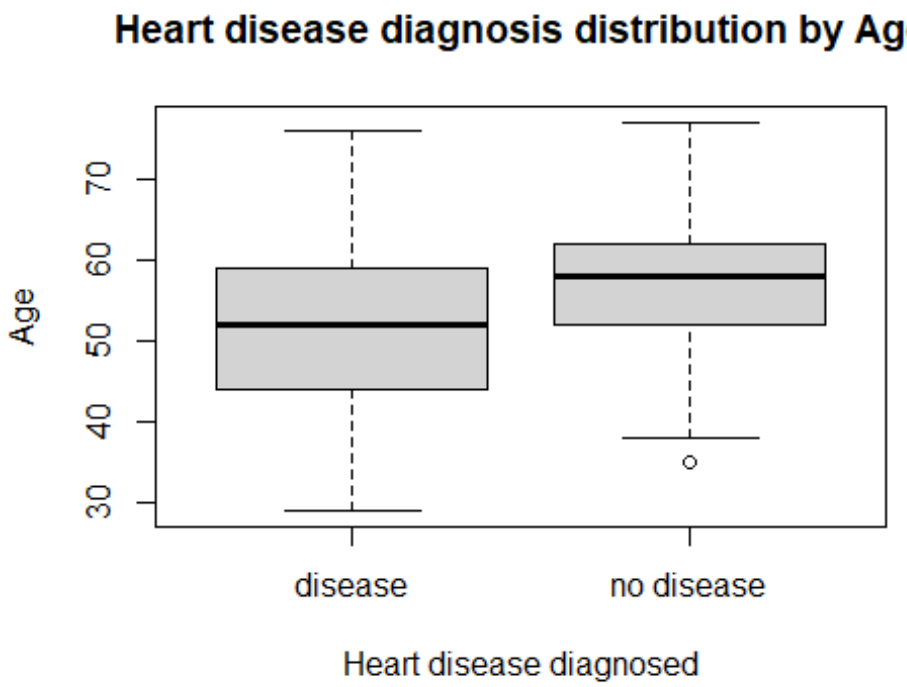

Figure 4 Heart Disease Diagnosis Distribution by Age

According to our findings, the median age of individuals without disease is higher than that of those with disease. Moreover, the range in age is relatively smaller for those without disease. Though there is an outlier among the no disease group, it is not considered an extreme outlier and therefore can be retained rather than removed.

```
ggplot(heart.selected, aes(x = target, y = age, fill = sex)) +
  geom_boxplot() +
  labs(title = "Heart disease diagnosis distribution by Age and Gender",
       x = "Heart disease diagnosed",
       y = "Age") +
  scale_fill_discrete(name = "Sex",
                      labels = c("Female", "Male")) +
  theme_minimal()
```

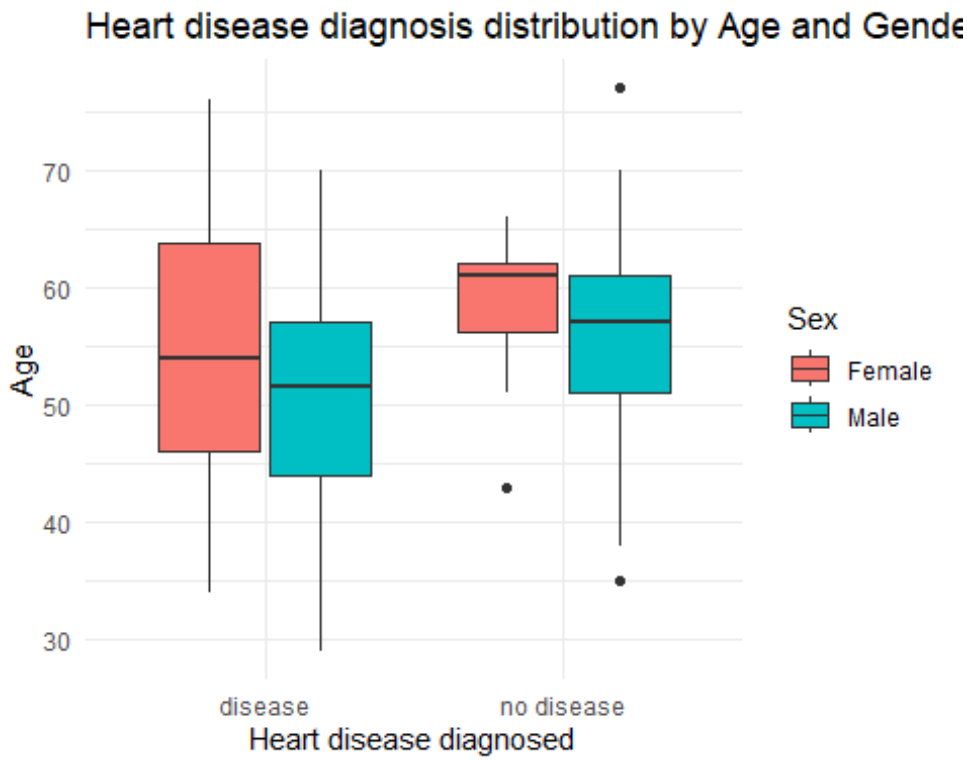

Figure 5 Heart Disease Diagnosis Distribution by Age and Gender

Females have higher median ages than males, regardless of disease status. Additionally, it's worth noting that individuals without disease have a higher average age than those with disease. This may be due to the fact that individuals with heart disease tend to have shorter lifespans.

```
ggplot(data = heart.selected, aes(x = target, fill = cp)) +
  geom_bar(position = "fill") +
  labs(title = "Heart disease diagnosis Distributions by Chest pain",
       x = "Heart disease diagnosis",
       y = "chest pain") +
  theme_test()
```

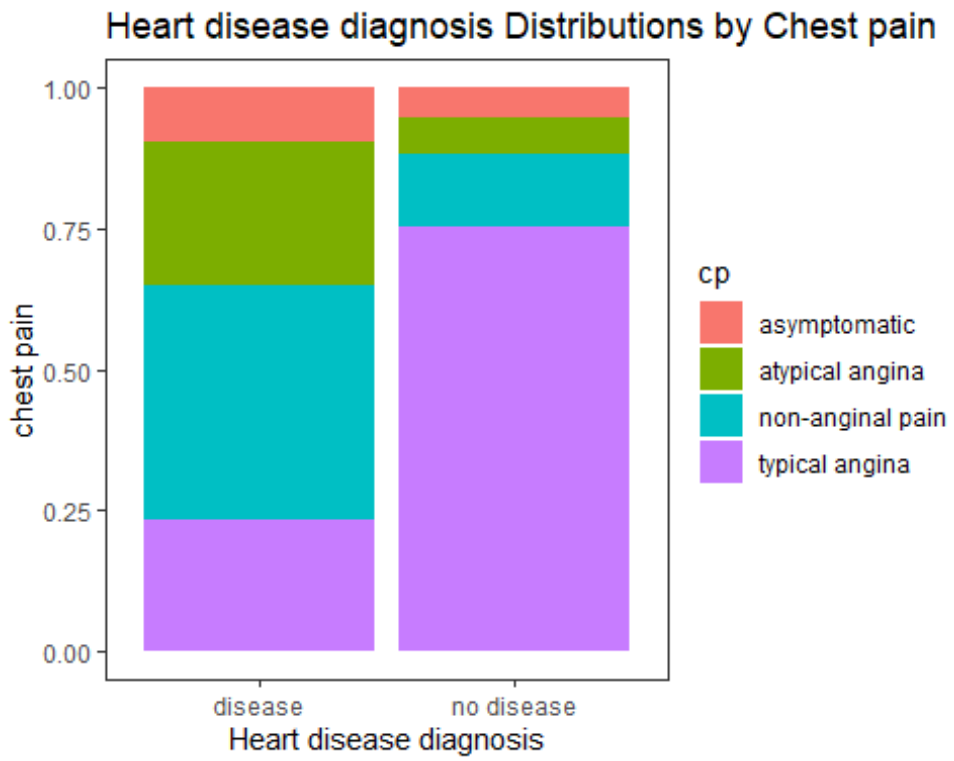

Figure 6 Heart Disease Diagnosis Distributions by Chest Pain

There are four possible outcomes for chest pain (CP), with varying proportions between individuals with and without the disease. For individuals with the disease, non-anginal pain has the highest proportion, followed by atypical angina, while asymptomatic has the lowest proportion. Conversely, for individuals without the disease, typical angina has the highest proportion, which is significantly higher than the proportion in those with the disease, while the other three types have smaller proportions.

This finding is intriguing because individuals with typical angina should logically have a higher likelihood of having the disease. Therefore, further exploration is necessary in subsequent research to fully understand this relationship.

```
mosaicplot(heart.selected$sex ~ heart.selected$target,
           main="Heart disease outcome by Gender", shade=FALSE,color=TR
UE,
           xlab="Gender", ylab="Heart disease")
```

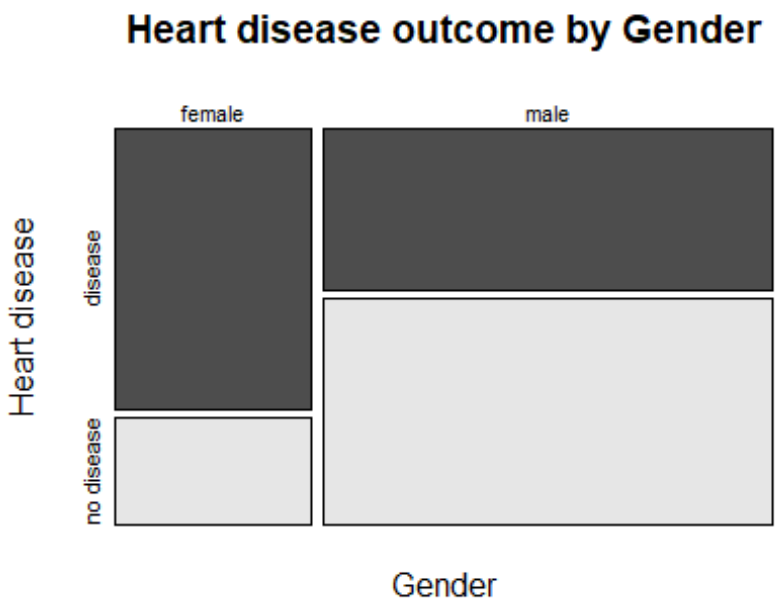

Figure 7 Heart Disease Outcome by Gender

Based on the figure above, it is evident that the proportion of disease is higher among females, while it is relatively low among males. Hence, it can be inferred that the probability of an observed individual being classified as having the disease is higher if they are female.

The relationship between CP and disease was examined in different genders.

```
# Create barplot by chest pain and gender
ggplot(heart.selected, aes(x = target, fill = cp)) +
  geom_bar(position = "fill") +
  facet_wrap(~sex) +
  labs(title = "Heart disease diagnosis Distributions by Chest pain",
       x = "Heart disease diagnosis",
       y = "chest pain") +
  scale_fill_discrete(name = "Chest Pain") +
  theme_test()
```

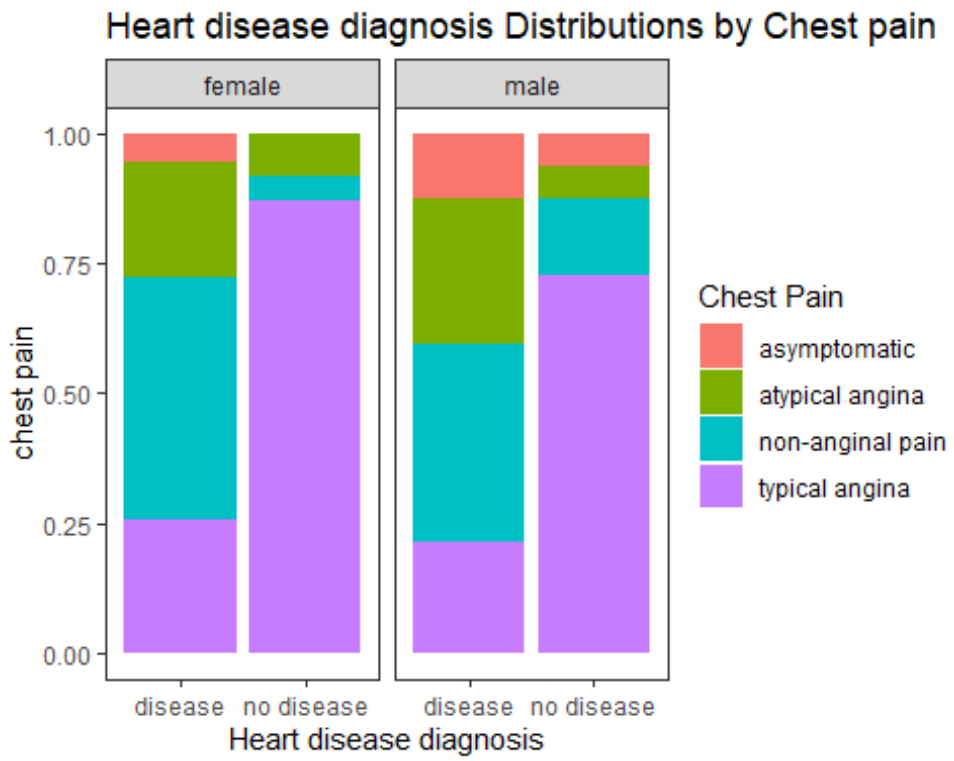

Figure 8 Heart Disease Diagnosis Distributions by Chest Pain

The figure above shows the situation of four types of chest pain under different genders, which is basically consistent with the overall results. Therefore, we can conclude that during classification, regardless of gender, if the observed individual has typical angina, they are more likely to be classified as having no disease.

The next step involves exploring the Thalach variable visually.

```
# Exploratory data analysis of thalach
ggplot(data = heart.selected, aes(x = thalach)) +
  geom_histogram(bins = 30, fill = "purple", color = "black") +
  labs(title = "Distribution of Maximum Heart Rate Achieved",
       x = "Maximum Heart Rate Achieved", y = "Frequency")
```

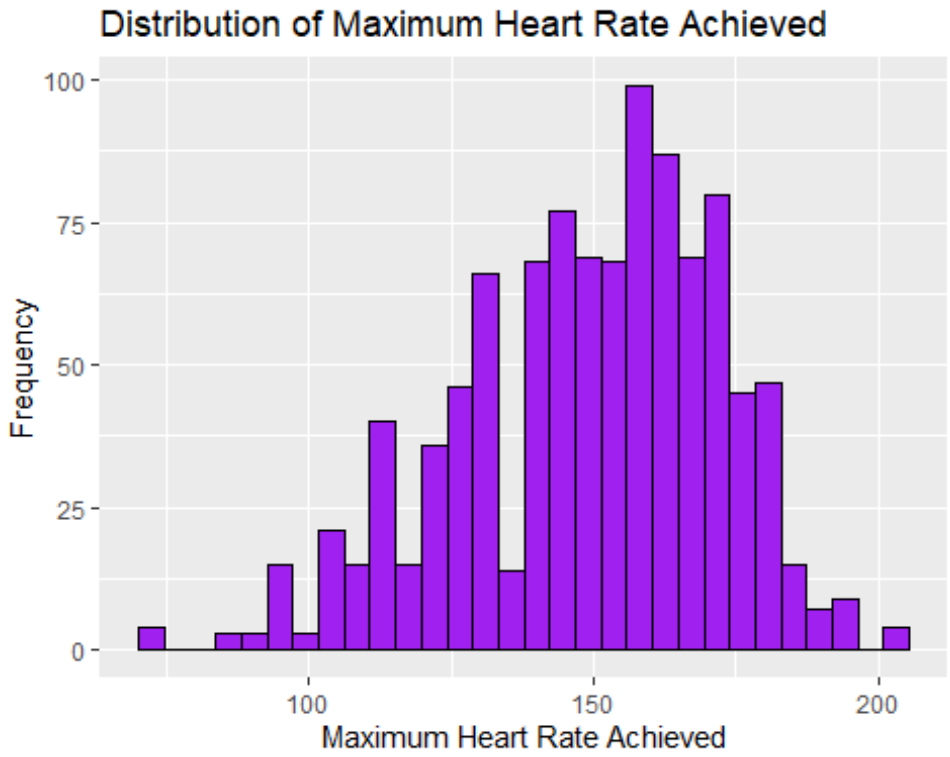

Figure 9 Distribution of Maximum Heart Rate Achieved

According to the results, the distribution of Thalach exhibits a slightly skewed left distribution, indicating the possible presence of an outlier on the left side. However, further analysis is required to confirm this observation.

```
boxplot(heart.selected$thalach ~ heart.selected$target,
        main="Heart disease diagnosis distribution by Thalach",
         ylab="Thalach",xlab="Heart disease diagnosed")
```

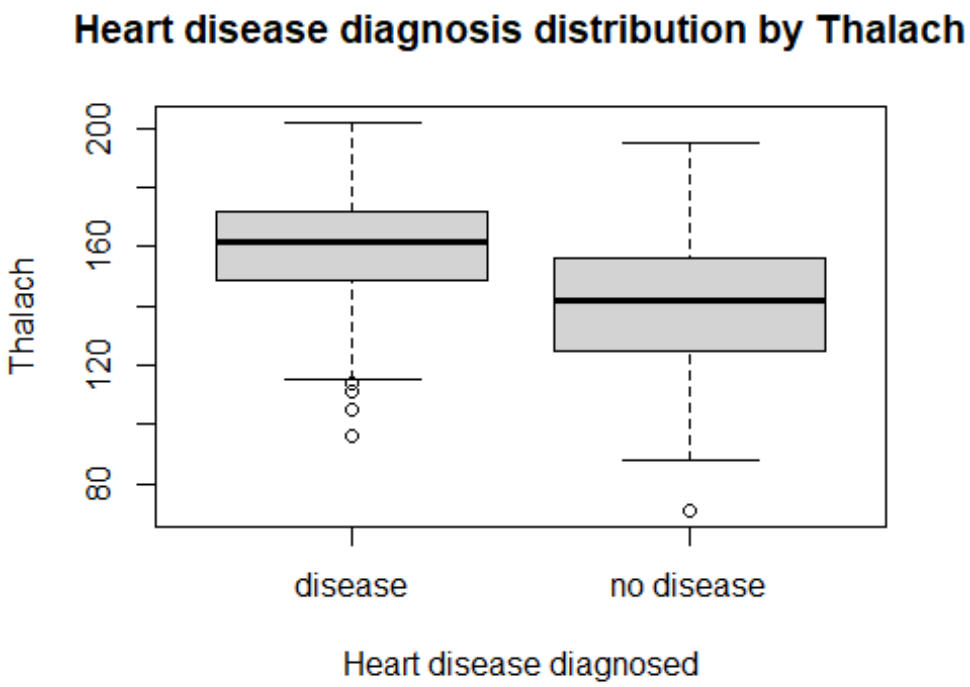

Figure 10 Heart Disease Diagnosis Distribution by Thalach

Based on the figure above, it is evident that the proportion of disease is higher among females, while it is relatively low among males. Therefore, it can be inferred that the probability of an observed individual being classified as having the disease is higher if they are female.

Moreover, the distribution of thalach for disease appears to be higher overall than that of no disease, which aligns with common knowledge. Additionally, there are more outliers for thalach in disease. Thus, it can be assumed that, in the case of an individual having a relatively high value of thalach, the probability of them having the disease is higher, making it easier for them to be classified into the disease category during classification.

The density diagram below illustrates the distribution of maximum heart rate achieved for individuals with and without heart disease.

```
# Relationship between thalach and target
ggplot(data = heart.selected, aes(x = thalach, fill = target)) +
  geom_density(alpha = 0.5) +
  labs(title = "Relationship between Maximum Heart Rate Achieved and He
art Disease Diagnosis",
       x = "Maximum Heart Rate Achieved", y = "Density", fill = "Diagno
sis") +
  scale_fill_manual(values = c("red", "green"))
```

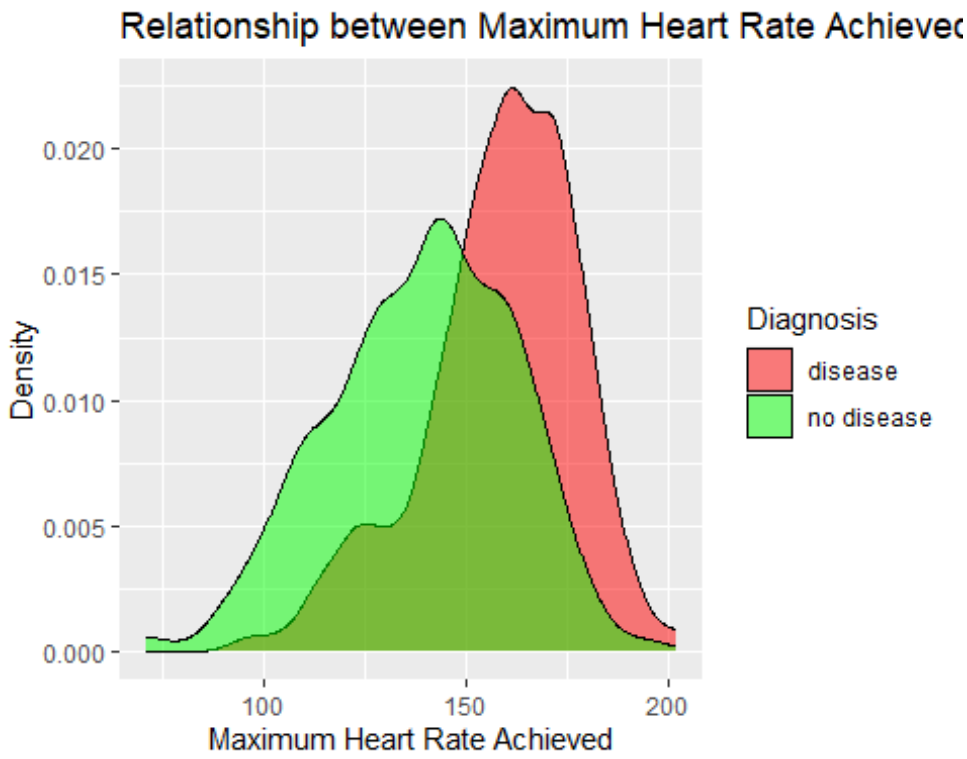

Figure 11 Relationship Between Maximum Heart Rate Achieved

The density diagram reveals that, under the condition of having the disease, the maximum heart rate achieved has higher values and proportion, which is significantly higher than that of the condition of no disease. It is suggested that individuals with a higher maximum heart rate achieved value are more likely to be classified as having the disease during classification.

```
# Relationship between thalach and target divided into gender
ggplot(data = heart.selected, aes(x = thalach, fill = target)) +
  geom_density(alpha = 0.5) +
  labs(title = "Relationship between Maximum Heart Rate Achieved and He
art Disease Diagnosis by Gender",
       x = "Maximum Heart Rate Achieved", y = "Density", fill = "Diagno
sis") +
  scale_fill_manual(values = c("red", "green")) +
  facet_wrap(~ sex)
```

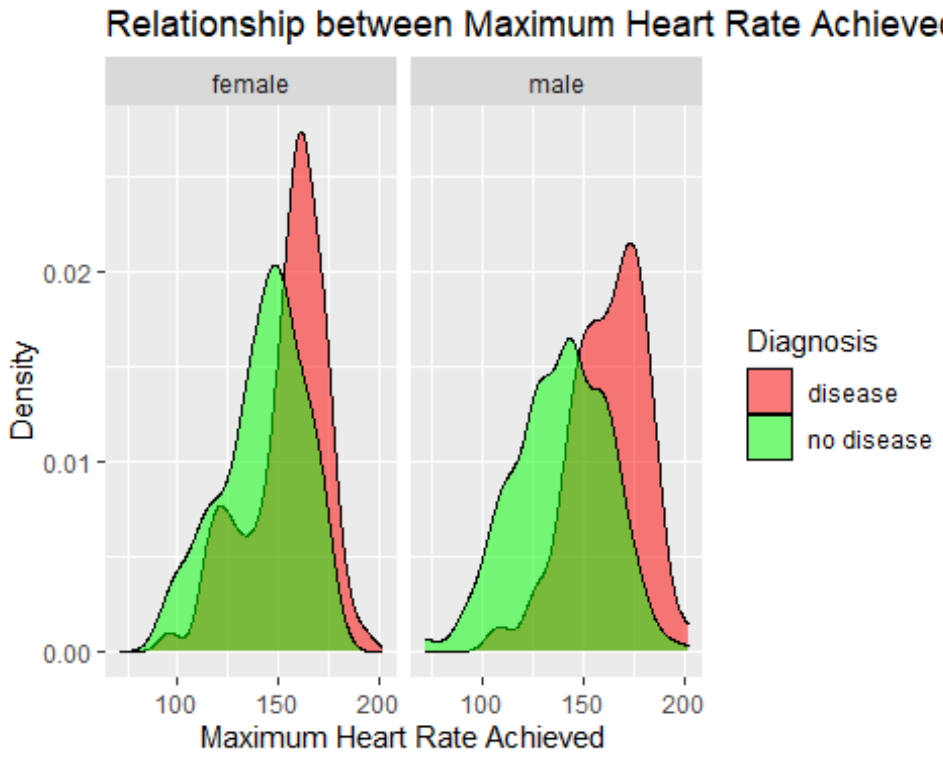

Figure 12 Relationship Between Maximum Heart Rate Achieved for Each Gender

Regardless of gender, the density of maximum heart rate achieved is very similar to the overall situation. However, among females with disease, the proportion of higher maximum heart rate achieved values is very significant.

Overall, no single variable can make highly accurate predictions, so combining all variables is necessary for classification.

Correlation Between Variables

To examine the relationship between variables, you can use the cor function to perform calculations and display the results with a corplot.

```
# calculate the correlation matrix
library(corrplot)

## Warning: package 'corrplot' was built under R version 4.2.3

## corrplot 0.92 loaded

heart.selected$age <- as.numeric(heart.selected$age)
heart.selected$thalach <- as.numeric(heart.selected$thalach)
heart.selected$oldpeak <- as.numeric(heart.selected$oldpeak)
heart.selected$major_vessels <- as.numeric(heart.selected$major_vessels)


cor.matrix <- cor(heart.selected[, c("age",  "thalach", "oldpeak", "maj
```

```
or_vessels")])
cor.matrix

##                      age    thalach    oldpeak major_vessels
## age            1.0000000 -0.3902271  0.2081367     0.2715505
## thalach       -0.3902271  1.0000000 -0.3497962    -0.2078884
## oldpeak        0.2081367 -0.3497962  1.0000000     0.2218160
## major_vessels  0.2715505 -0.2078884  0.2218160     1.0000000
```

Age has a positive correlation with major_vessels (0.2715) and a weak positive correlation with oldpeak (0.2081). Thalach has a negative correlation with oldpeak (-0.3498). Finally, oldpeak and major_vessels have a moderate positive correlation (0.2218).

```
corrplot(cor.matrix)
```

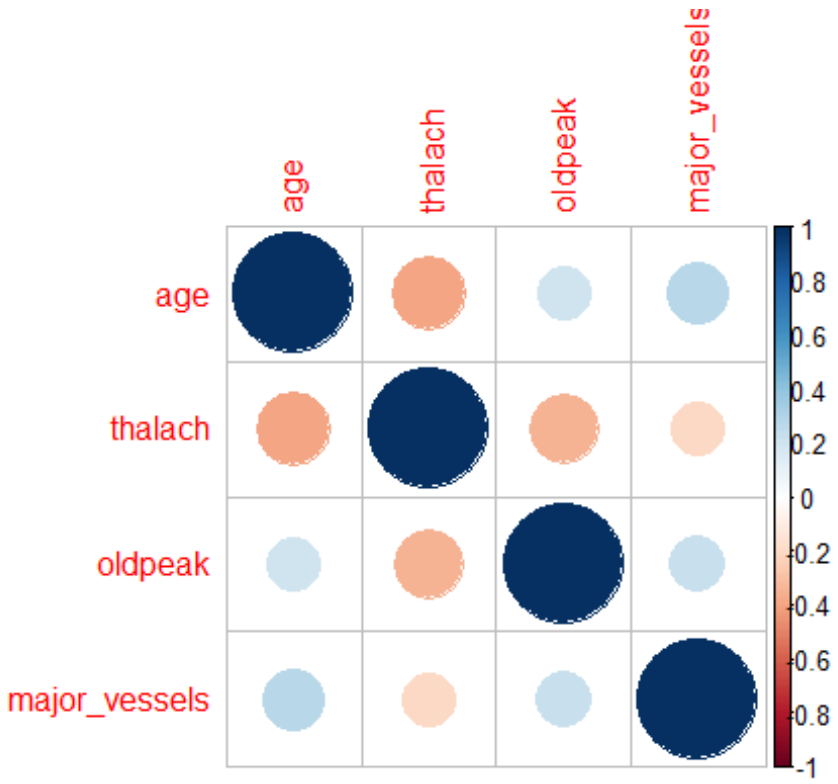

Figure 13 Correlation Visualization for Numeric Features

From the figure, it appears that thalach is negatively correlated with both age and oldpeak. In terms of positive correlation, major_vessels and age appear to be positively correlated.

Conclusion

Based on the results of data exploration, there is a clear difference between men and women. The other variables and target essentially show a positive correlation. The

larger the value, the higher the likelihood of the target being a disease. However, except for chest pain, the data is somewhat abnormal, and further exploration is necessary.

The selected features from the original dataset include "age", "sex", "cp", "thalach", "exang", "oldpeak", "major_vessels", "restwm", and "target". "Id" is not an independent variable and "target" is a dependent variable. In correlation research of numeric features, it was found that although some features have positive or negative correlations, the correlation coefficient is not high. Therefore, it can be concluded that one feature cannot replace another feature. The feature selection was successful, and new datasets have been created and uploaded.

References

Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart Disease Prediction using Machine Learning Techniques. *SN Computer Science*, *1*(6). https://doi.org/10.1007/s42979-020-00365-y

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, *17*(8), 43–48. https://doi.org/10.5120/2237-2860