Heart diseases features selection and exploratory data analysis

Wangjun Shen

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

Data Exploration

Features Selection

The details of the original dataset are as follows.

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

Variable	Description	Type
cp	Chest Pain type: typical angina; atypical angina; non-anginal pain; and asymptomatic	Categorical
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.

```r
# 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 angina" "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 abnormality" "ST-T wave abnormali
##  $ 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" "downsloping" ...
##  $ 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
##  $ target       : chr  "no disease" "no disease" "no disease" "no disease" ...
```

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.

```r
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.

```r
# 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.

```r
# 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.
```

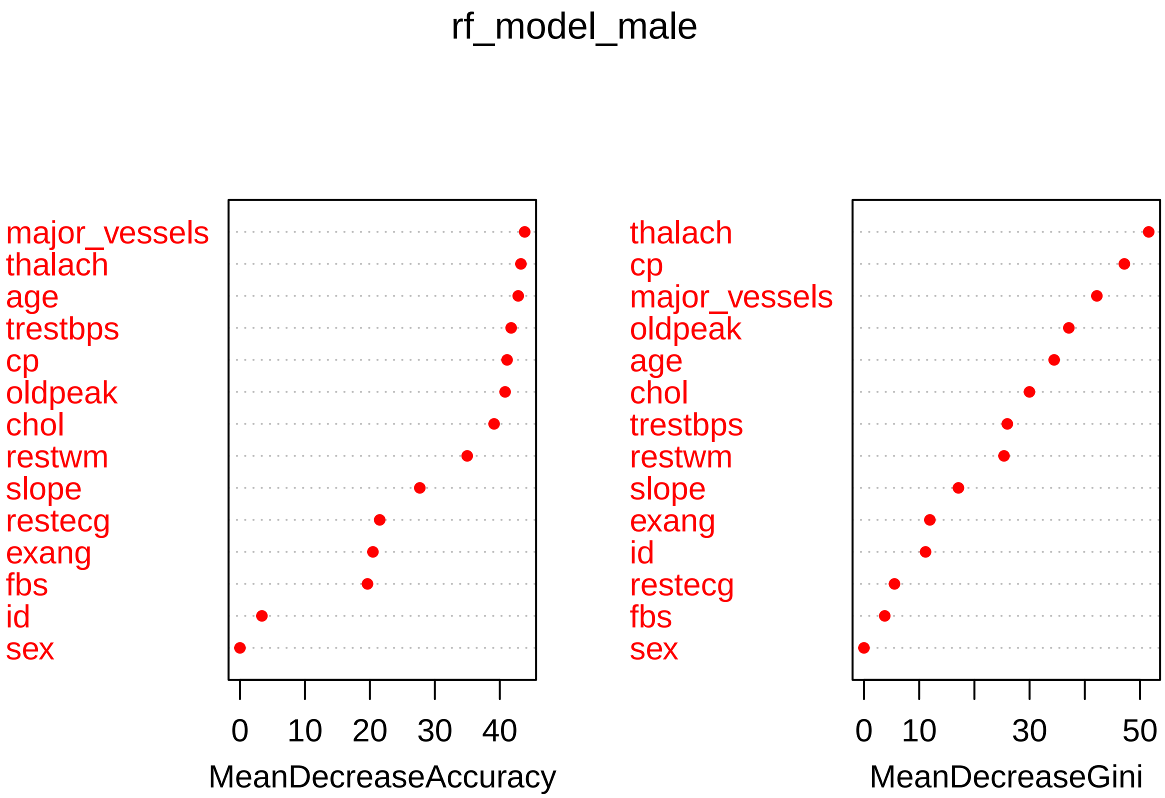
```
set.seed(123)
rf_model_male <- randomForest(target ~ ., data = heart.male, importance = TRUE, ntree = 500)
importance(rf_model_male)
```

```
##                 disease no disease MeanDecreaseAccuracy MeanDecreaseGini
## id             1.606825   3.213858             3.376221        11.154653
## age           33.928469  40.857084            42.838250        34.448119
## sex            0.000000   0.000000             0.000000         0.000000
## cp            38.745483  36.059468            41.121602        47.161819
## trestbps      33.903759  37.024285            41.751265        25.962476
## chol          34.122660  35.756098            39.121768        29.972875
## fbs           16.795603  16.203869            19.636756         3.761825
## restecg       19.083360  19.416802            21.509917         5.531215
## thalach       37.581333  37.899592            43.252118        51.572260
## exang         17.532815  19.401905            20.468760        11.933409
## oldpeak       32.589020  38.902000            40.817055        37.108368
## slope         24.318522  24.566740            27.686364        17.111266
## major_vessels 37.680805  40.115231            43.835239        42.184090
## restwm        31.201322  32.398437            34.984151        25.367047
```

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)
```



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, importance = TRUE, ntree = 500)
importance(rf_model_female)
```

```
##                disease no disease MeanDecreaseAccuracy MeanDecreaseGini
## id            1.932703  -3.148355           -0.5202009         2.590325
## age          23.376089  24.724758           28.2364781        12.018921
## sex           0.000000   0.000000            0.0000000         0.000000
## cp           19.702221  20.271296           22.9715941        12.537614
## trestbps     18.409296  18.651176           22.1370717         8.429236
## chol         19.941082  22.898914           25.6314494         8.321454
## fbs           7.611649   9.737944           10.2944996         1.418907
## restecg       14.299858  15.552097           17.3754050          2.982791
## thalach       19.430050  20.969639           24.2566433          8.272732
## exang         18.303884  21.574578           22.6248431          9.167964
## oldpeak       20.677704  21.784792           24.3639560         16.486048
## slope         17.218824  19.973871           21.5111294          8.150365
## major_vessels 19.573486  20.510840           22.7582749         10.407631
## restwm        22.156482  24.907233           26.5256308         22.290165
```

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

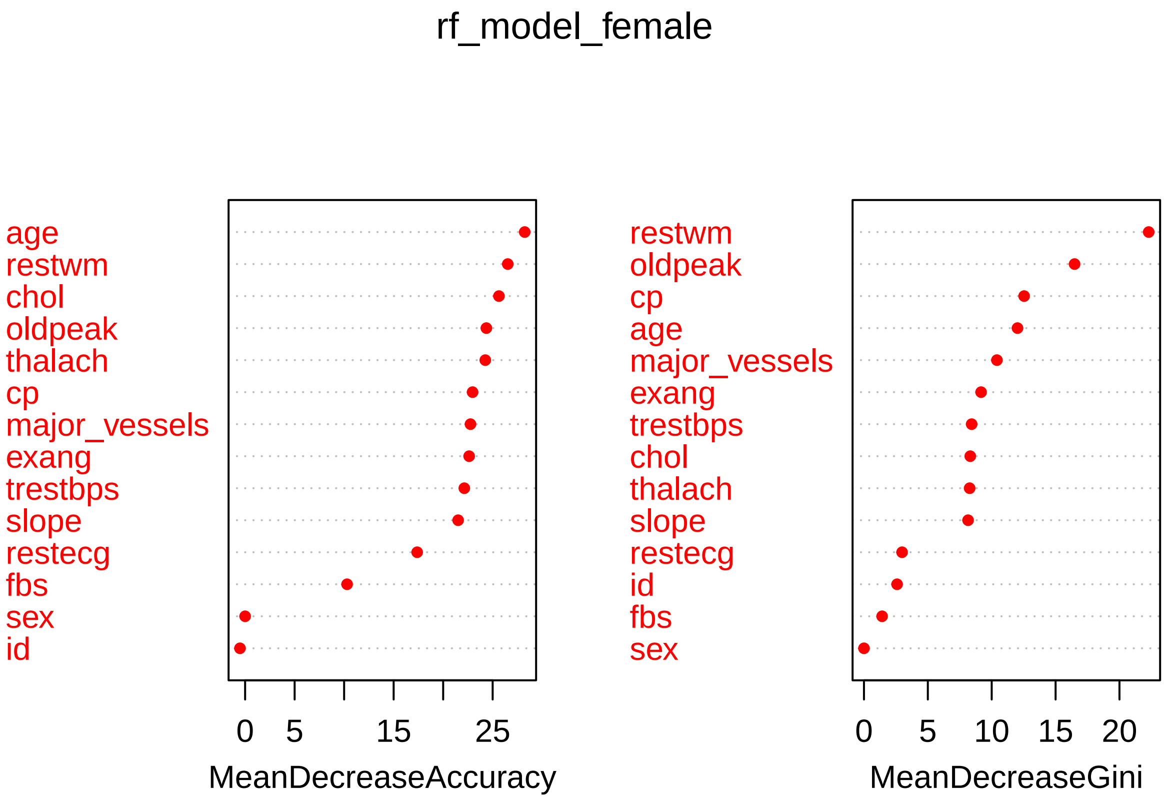
- Age: It exhibits a high correlation with both disease and non-disease and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
- cp (Chest pain type): It exhibits a moderate correlation with disease and non-disease, and has high values for MeanDecreaseAccuracy and MeanDecreaseGini.
- thalach (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)
```



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", "thalach", "exang", "oldpeak", "maj

# Save new dataset as CSV file
write.csv(heart_features_selected, "heart_features_selected.csv", row.names = 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 angina" "typical angina" ...
##  $ thalach      : int  168 155 125 161 106 122 140 145 144 116 ...
```