CYO project: Heart Disease prediction

Kiko Núñez

2020/01/06

1. Introduction

This project aims at putting into practice the contents learned during the Data Science course making use of the publicly available dataset Heart Disease UCI | Kaggle and then applying some of the tools and methods explained during the course. In addition, the project explores the use of other R libraries beyond those addressed during the course.

1.1 Dataset description

As explained in the Kaggle documentation of this dataset, the Heart Disease UCI dataset contains 76 attributes, but all published experiments refer to using a subset of 14 of them:

1. *age*: age in years.
2. *sex*: (1 = male; 0 = female).
3. *cp*: chest pain type (typical angina, atypical angina, non-angina, or asymptomatic angina).
4. *trestbps*: resting blood pressure (in mm/Hg on admission to the hospital).
5. *chol*: serum cholestoral in mg/dl.
6. *fbs*: Fasting blood sugar (<120 mg/dl or >120 mg/dl) (1 = true; 0 = false).
7. *restecg*: resting electrocardiographic results (normal, ST-T wave abnormality, or left ventricular hypertrophy).
8. *thalach*: Max. heart rate achieved during thalium stress test.
9. *exang*: Exercise induced angina (1 = yes; 0 = no).
10. *oldpeak*: ST depression induced by exercise relative to rest.
11. *slope*: Slope of peak exercise ST segment (0 = upsloping, 1 = flat, or 2 = downsloping).
12. *ca*: number of major vessels (0-3, 4 = NA) colored by flouroscopy.
13. *thal*: Thalium stress test result; 3 = normal; 6 = fixed defect; 7 = reversable defect; 0 = NA.
14. *target*: Heart disease status 1 or 0 (0 = heart disease 1 = asymptomatic).

1.2 Goal of the project

The main goal of this project is to apply the insights gained during the course to predict whether a subject has (or not) a heart disease depending on a set of related variables.

1.3 Key steps

1. Data analysis and visualization. We'll make a comprehensive analysis of the variables distribution in the overall dataset and also per gender and health status.
2. Partition into train and test datasets. Before training the models, we'll leave the 20% of the samples out for validation purposes.

3. Train the models. We'll train 3 different models to compare their accuracy: Random Forest (RF), Decision Trees (DT) and General Linear Model (GLM).
4. Apply trained models to test dataset.
5. Conclusions, limitations and future work.

2. Data Analysis and Visualization

2.1. Initial Analysis

The first step is to load the dataset, transform it into a tidy format, and then explore it:

Table 1: Head of the dataset. Columns 1 to 7.

age	sex	cp	trestbps	chol	fbs	restecg
63	male	typical angina	145	233	>120	hypertrophy
37	male	non-anginal	130	250	<=120	normal
41	female	atypical angina	130	204	<=120	hypertrophy
56	male	atypical angina	120	236	<=120	normal
57	female	asymptomatic angina	120	354	<=120	normal
57	male	asymptomatic angina	140	192	<=120	normal

Table 2: Head of the dataset. Columns 8 to 14.

thalach	exang	oldpeak	slope	ca	thal	target
150	no	2.3	downsloping	0	fixed defect	asymptomatic
187	no	3.5	downsloping	0	normal	asymptomatic
172	no	1.4	upsloping	0	normal	asymptomatic
178	no	0.8	upsloping	0	normal	asymptomatic
163	yes	0.6	upsloping	0	normal	asymptomatic
148	no	0.4	flat	0	fixed defect	asymptomatic

We notice that we have the 14 variables as explained in the introduction section.

2.2 Data visualization

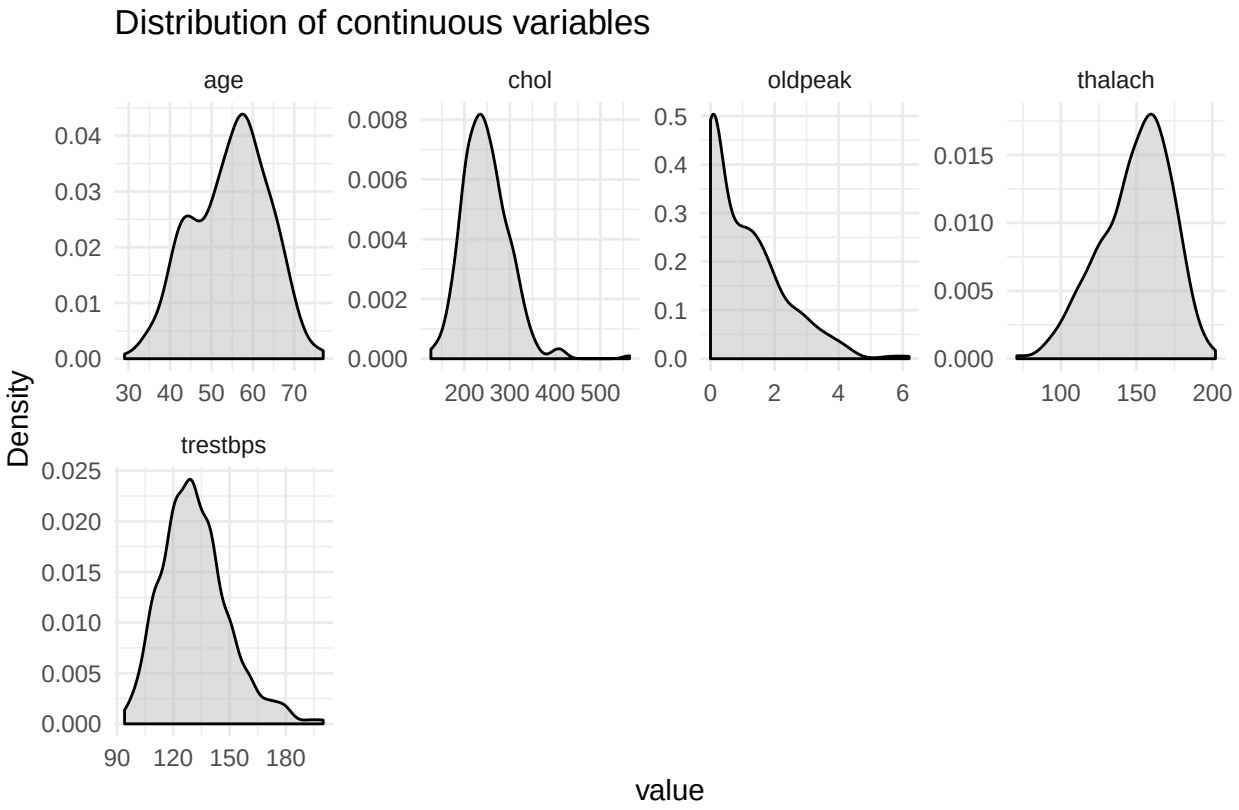
In this section, we'll apply some functions to the entire dataset in order to gain more insights about it through its visualization.

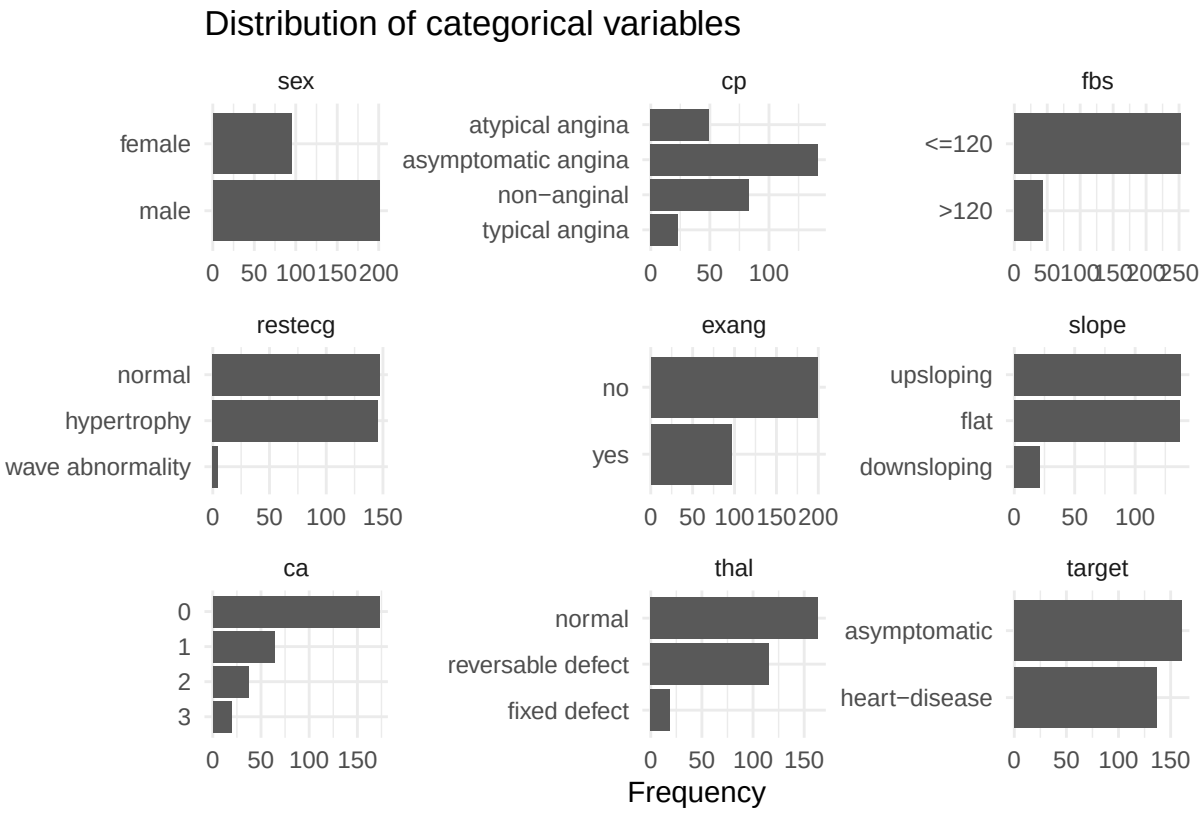
2.2.1 Visualize the data summary and distribution for each variable. A dataset summary can be provided by the `summary()` function:

```
##       age            sex                       cp         trestbps
##  Min.   :29.00   female: 95   asymptomatic angina:141   Min.   : 94.0
##  1st Qu.:48.00   male  :201   atypical angina    : 49   1st Qu.:120.0
##  Median :56.00                non-anginal        : 83   Median :130.0
##  Mean   :54.52                typical angina     : 23   Mean   :131.6
##  3rd Qu.:61.00                                          3rd Qu.:140.0
##  Max.   :77.00                                          Max.   :200.0
##       chol          fbs        restecg             thalach      exang
```

```
##  Min.   :126.0   <=120:253   Length:296         Min.   : 71.0   no :199
##  1st Qu.:211.0   >120 : 43   Class :character   1st Qu.:133.0   yes: 97
##  Median :242.5               Mode  :character   Median :152.5
##  Mean   :247.2                                  Mean   :149.6
##  3rd Qu.:275.2                                  3rd Qu.:166.0
##  Max.   :564.0                                  Max.   :202.0
##     oldpeak              slope     ca                     thal
##  Min.   :0.000   downsloping: 21   0:173   fixed defect     : 18
##  1st Qu.:0.000   flat       :137   1: 65   normal           :163
##  Median :0.800   upsloping  :138   2: 38   reversable defect:115
##  Mean   :1.059                     3: 20
##  3rd Qu.:1.650
##  Max.   :6.200
##            target
##  asymptomatic :160
##  heart-disease:136
##
##
##
##
```

Now we'll use some functions from the `DataExplorer` library to visualize the distribution of the continuous and categorical variables:

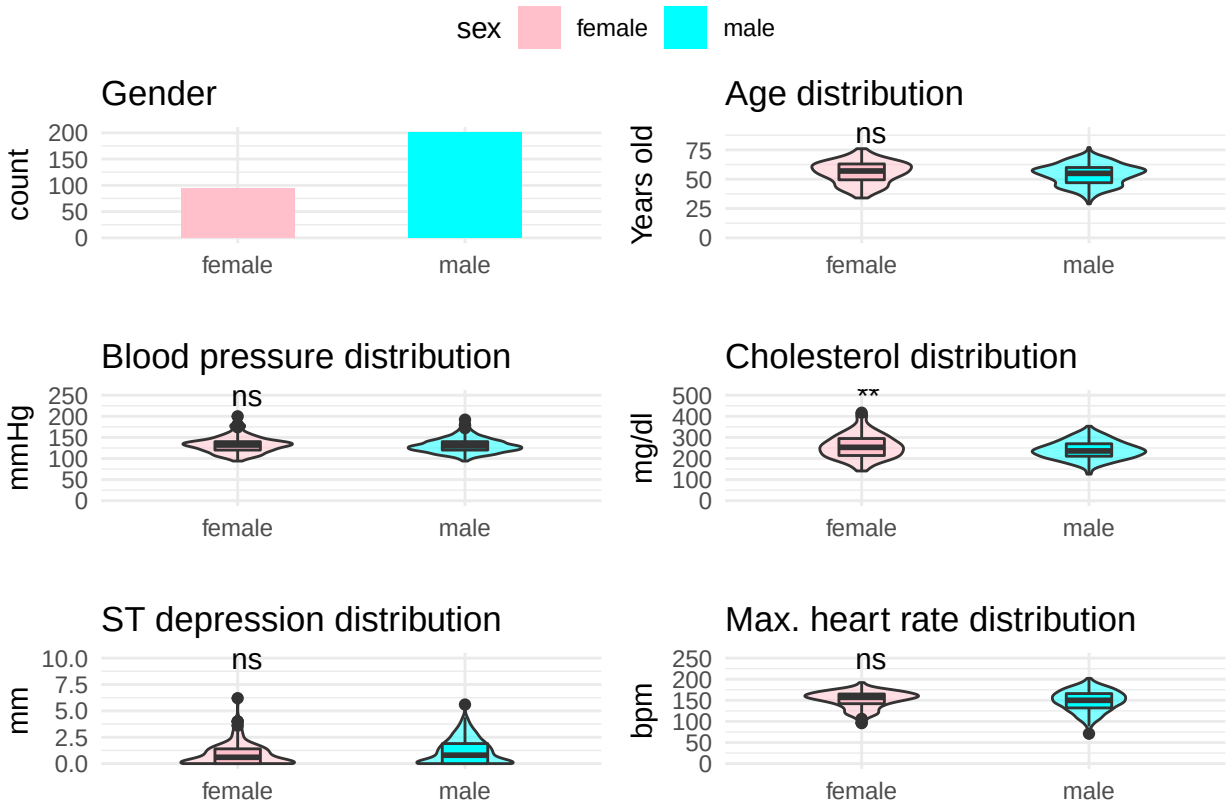




Here we can extract the following information:

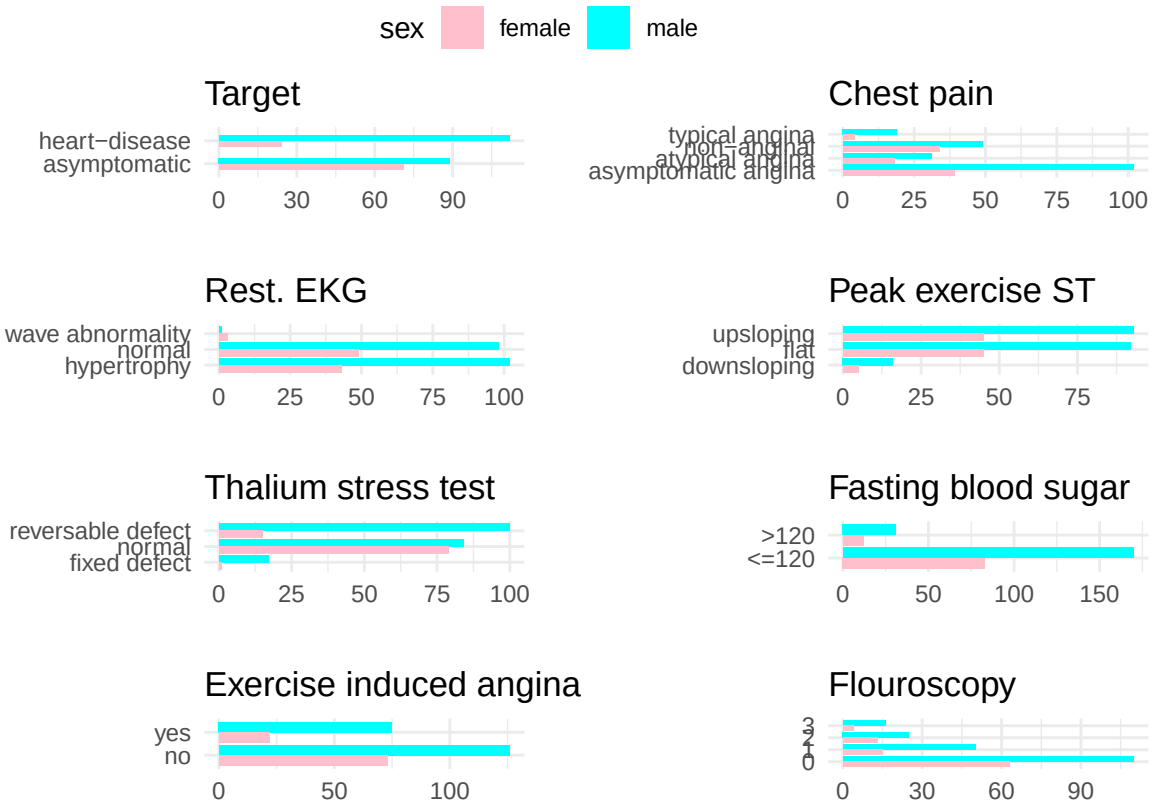
- Age distribution is slightly skewed towards the odest values, being the mean about 55 years old..
- Cholesterol levels are normally distributed around 250, with some outliers in the upper levels.
- Most of ST depression tests are between 2 and 4 mm.
- Maximum heart rate distribution is slightly skewed towards the upper side, being the mean around 160 bpm.
- Resting blood pressure is slgihtly skewed towards the lower end around 130 mm/Hg.
- Regarding gender, there are two times more males than females in the dataset.
- Most of the cases reported asymptomatic angina when asked about chest pain.
- Fasting blood sugar was mostly rated under 120 mg/dl.
- Resting EKG results were almost equally distributed between normal and hypertorphy findings.
- Exercise induced angina was found in 1/3 of the cases included in the dataset.
- Slope of peak exercise ST segment was reported mostly on upsloping and flat in an equal basis.
- Number of major vessels colored in the fluoroscopy was found to be 0 for most of the subjects.
- Thalium stress test resulted in normal or reversable defect for most of the cases.
- The dataset is a bit biased towards asymptomatics subjects compared to heart disease subjects.

2.2.2 Continuous variables distribution per gender. In this section, we are going to analyze more in depth the continuous variables distributions per gender:



With a first visual inspection, in these plots we notice that the datset has two times more males than females and that females have a slightly higher maximum heart beat rate than males. The other variables are more or less equally distributed in both groups.

2.2.3 Categorical variables distribution per gender. In this section, we'll perform a similar analysis on the categorical variables.

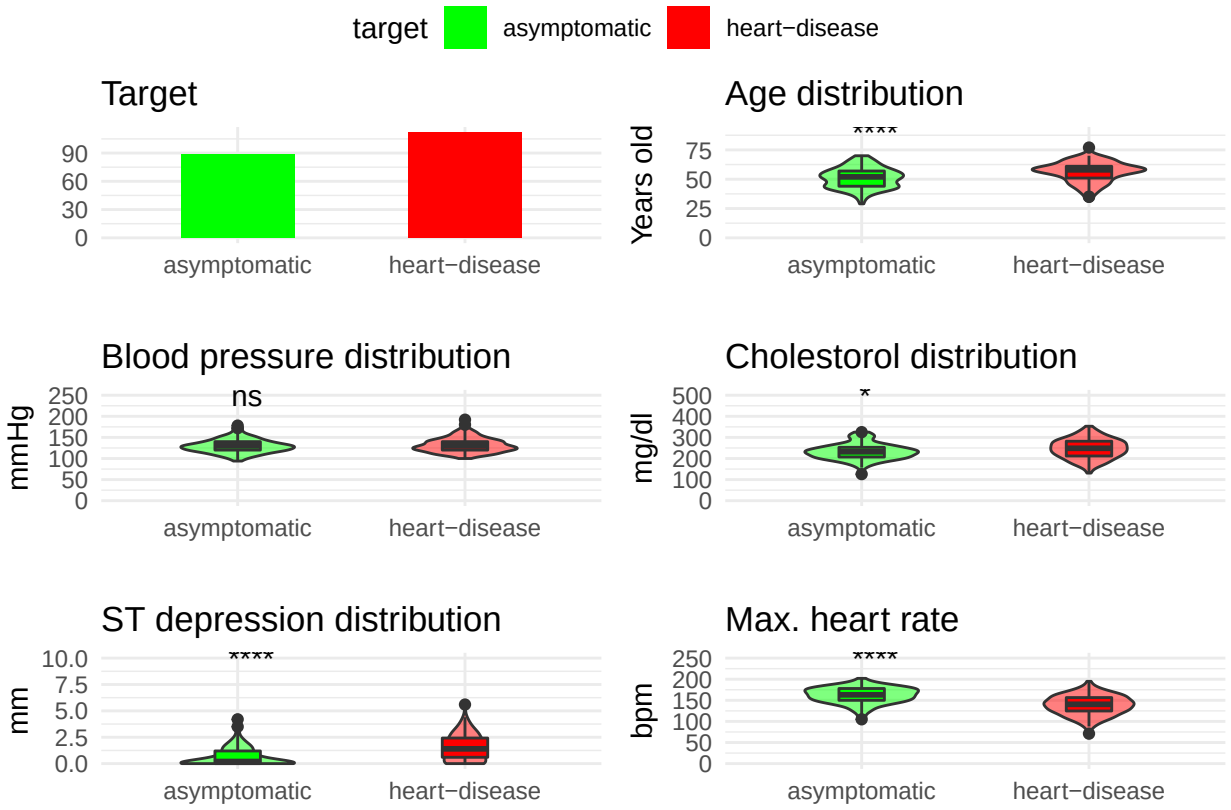


From this set of plots, taking into account that the sample of males is twice the sample of females, we can notice the following differences in the distributions:

- The dataset is biased towards asymptomatic population, with a higher proportion of males belonging to the heart disease group.
- The Thalium stress test resulted in reversable defect in a higher proportion of males.
- Wave abnormalities in the resting EKG were found in a higher proportion of females.
- Males were more likely to suffer exercise induced angina than females.

2.2.4 Continuous variables distribution per gender and disease status. Now let's perform a similar visual analysis but in this case let's put the focus on the disease status instead. We'll also make different plots considering the gender of the patients.

Male subjects:



Male patients with heart disease are significantly older, have higher cholesterol level, higher ST segment depression and lower maximum heart rate response to the Thallium test.

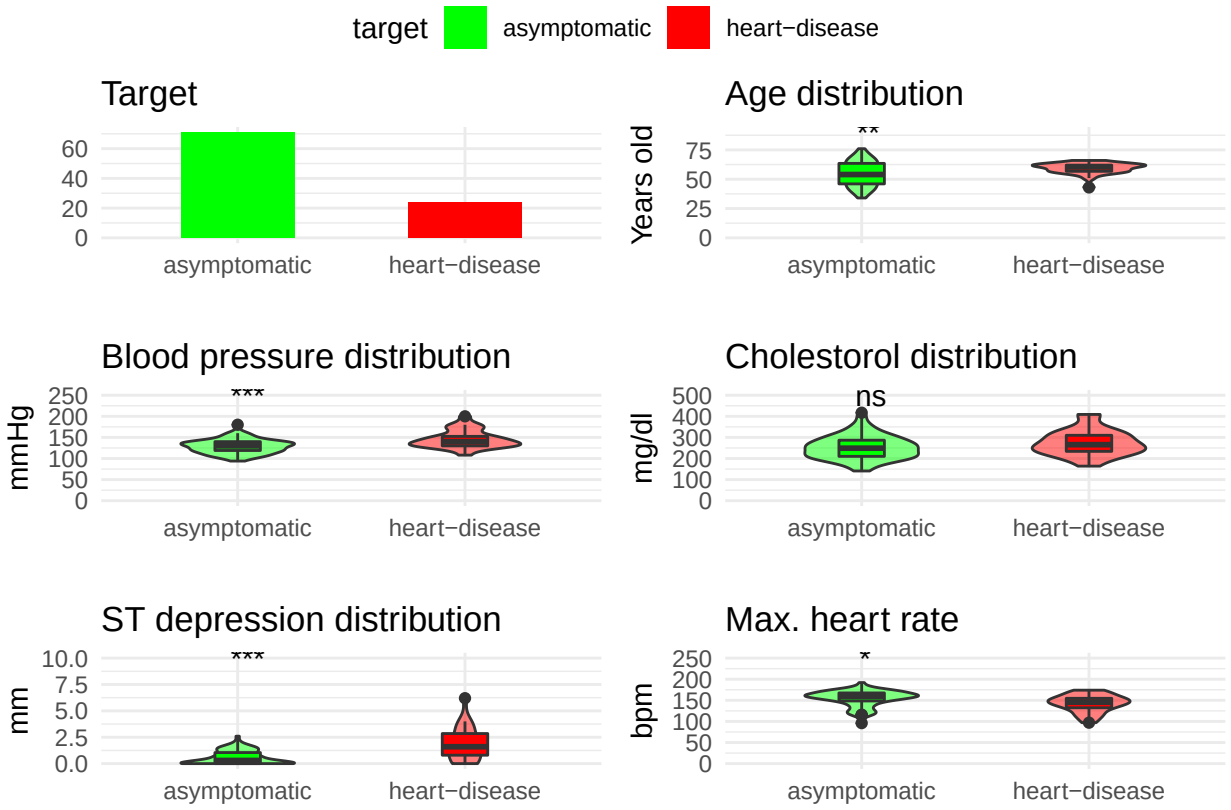
Now let's analyze the categorical variables for males:

From the categorical variables, we could extract the following information:

- The dataset has more heart disease males than asymptomatics.
- Chest pain in males revealed as asymptomatic angina for most of heart disease group.
- Hypertrophy was found in the resting EKG more often in the heart disease group of males.
- ST segment during peak exercise showed to be upsloping in most of asymptomatics males, while being flat for most of the subjects belonging to the heart disease group.
- Thallium stress test resulted in normal for most asymptomatics males while it showed reversable defects in the heart disease group.
- Fasting blood sugar was mostly under 120 for the males belonging to the heart disease group.
- Asymptomatic males were more likely to not to show an exercise induced angina compared to heart disease subjects.
- Fluoroscopy resulted positive for most of the males belonging to the heart disease group.

Female subjects:

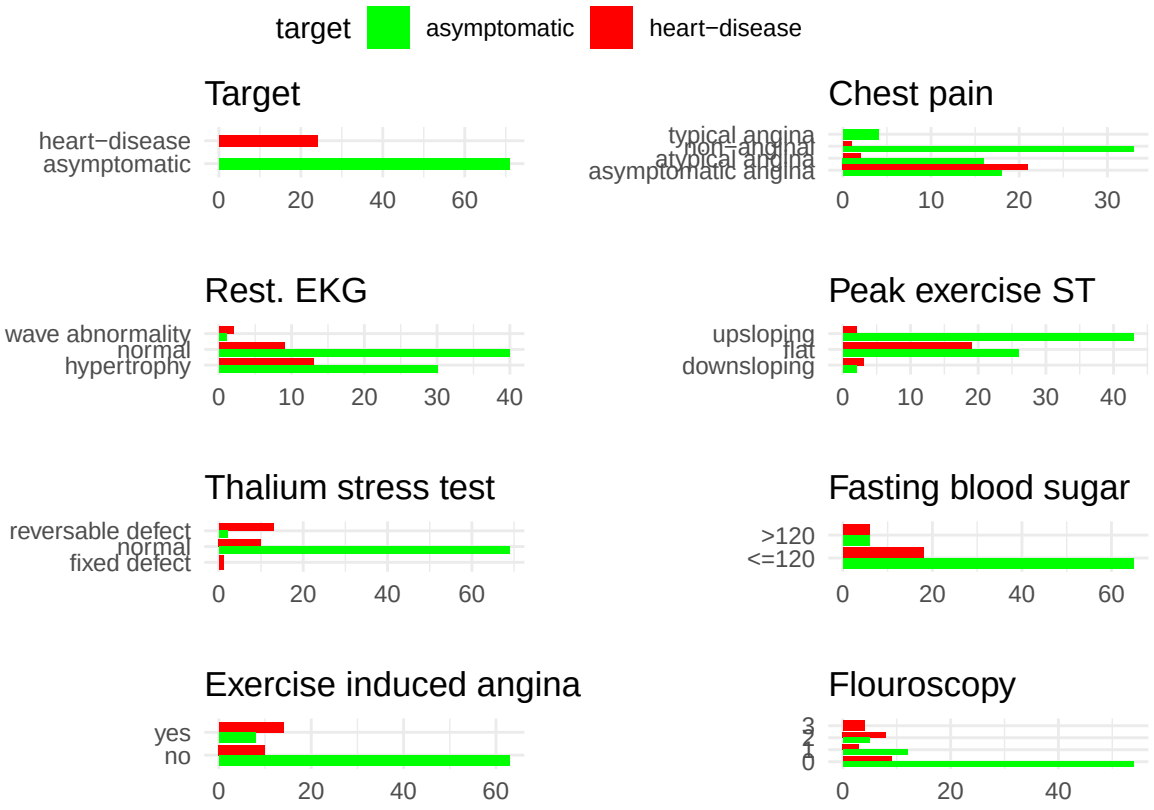
Now let's repeat the same analysis with the females group only.



From the categorical variables in females, we could extract the following information:

- The dataset has less heart disease females than asymptomatics.
- Chest pain revealed as asymptomatic angina for most of females belonging to the heart disease group, similarly to males.
- Hypertrophy was found in the resting EKG for females more often in the heart disease group, while most asymptomatics reported a normal EKG.
- ST segment during peak exercise showed to be upsloping in most of asymptomatics, while being flat for most of the subjects belonging to the heart disease group, similarly to males.
- Thallium stress test resulted in normal for most asymptomatics femlaes while it showed reversable defects in the heart disease group, similarly to males.
- Fasting blood sugar was slightly higher for the heart disease subjects.
- Proportion of asymptomatic and heart disease females reporting fasting blood sugar under 120 is in keeping with the distribution of females.
- Fluoroscopy resulted positive for most of the subjects belonging to the heart disease group, similarly to males.

The following plots show the distributions of the categorical variables for females grouped by target (asymptomatic vs heart disease):



There are less woman with heart disease in this data set. Women with heart disease have a significantly higher resting blood presure contrary to male with heart disease. Similarly to men, women with heart disease have a lower maximum heart rate in response to the thallium test.

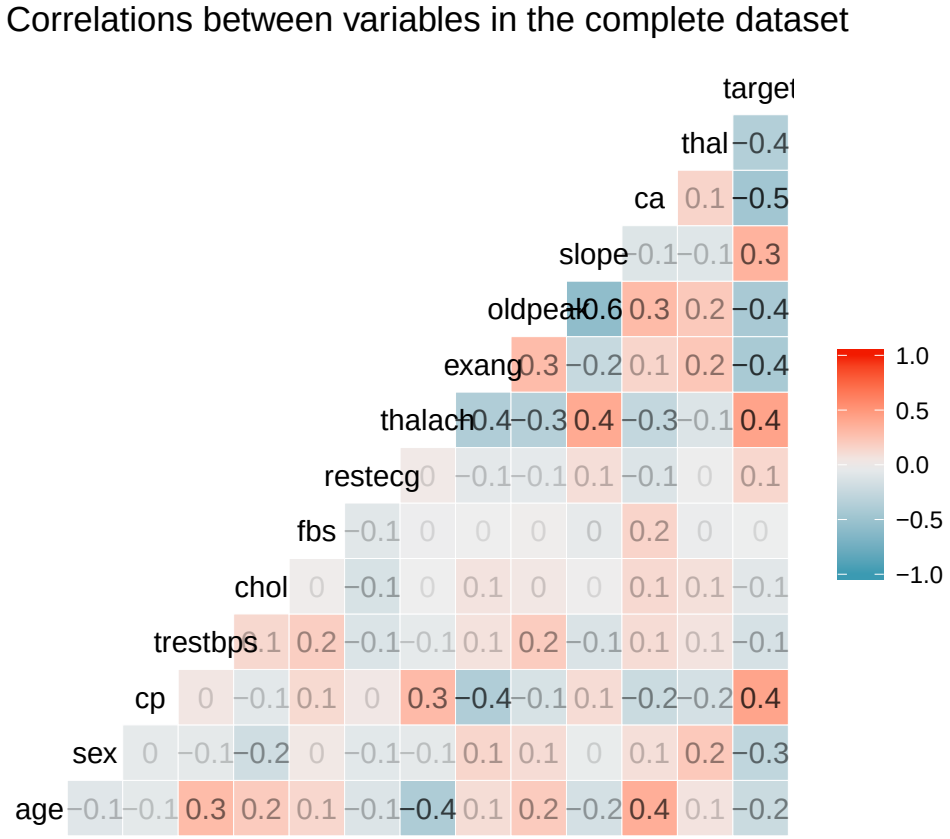
For the continuous variables, we can also inspect their mean and standard deviation grouped by gender and disease status as follows:

Table 3: Mean (SD) of the continous variables grouped by gender and health status.

target	sex	n	age	trestbps	chol	thalach	oldpeak
asymptomatic	female	71	54.58 (10.34)	128.75 (16.65)	257.32 (66.51)	154.58 (18.81)	0.56 (0.65)
asymptomatic	male	89	51.1 (8.63)	129.52 (16.23)	232.46 (37.76)	161.78 (18.72)	0.63 (0.88)
heart-disease	female	24	59.04 (4.96)	146.12 (21.44)	274.96 (60.86)	142.42 (20.26)	1.84 (1.61)
heart-disease	male	112	56.24 (8.36)	131.96 (17.37)	246.43 (45.67)	138.21 (23.23)	1.55 (1.23)

2.3 Correlations

To calculate correlations between the variables in the dataset, we'll make use of the function `ggcorr()` applied to the numerical values in the original dataset after removing the NAs.



3. Split into train and test datasets

We will train the models with 80% of samples, leaving out the 20% for test purposes only:

```
## [1] "Number of rows in the train dataset: 236"
```

```
## [1] "Number of rows in the test dataset: 60"
```

4. Building classification models

In this section we'll compare accuracy and Kappa values for 3 different methods: Random Forest (RF), Decision Trees (DT) and General Linear Model (GLM) with a 5-fold cross-validation within the train dataset. Besides, we'll classify the variables importance for each one of the models.

4.1 Random Forest

```
fit_rf <- train(target~.,
             data = train,
             method = "rf",
             trControl = trainControl(method = "cv",
                                      number = 5,
                                      p = 0.8))
```

4.2 Decision Trees

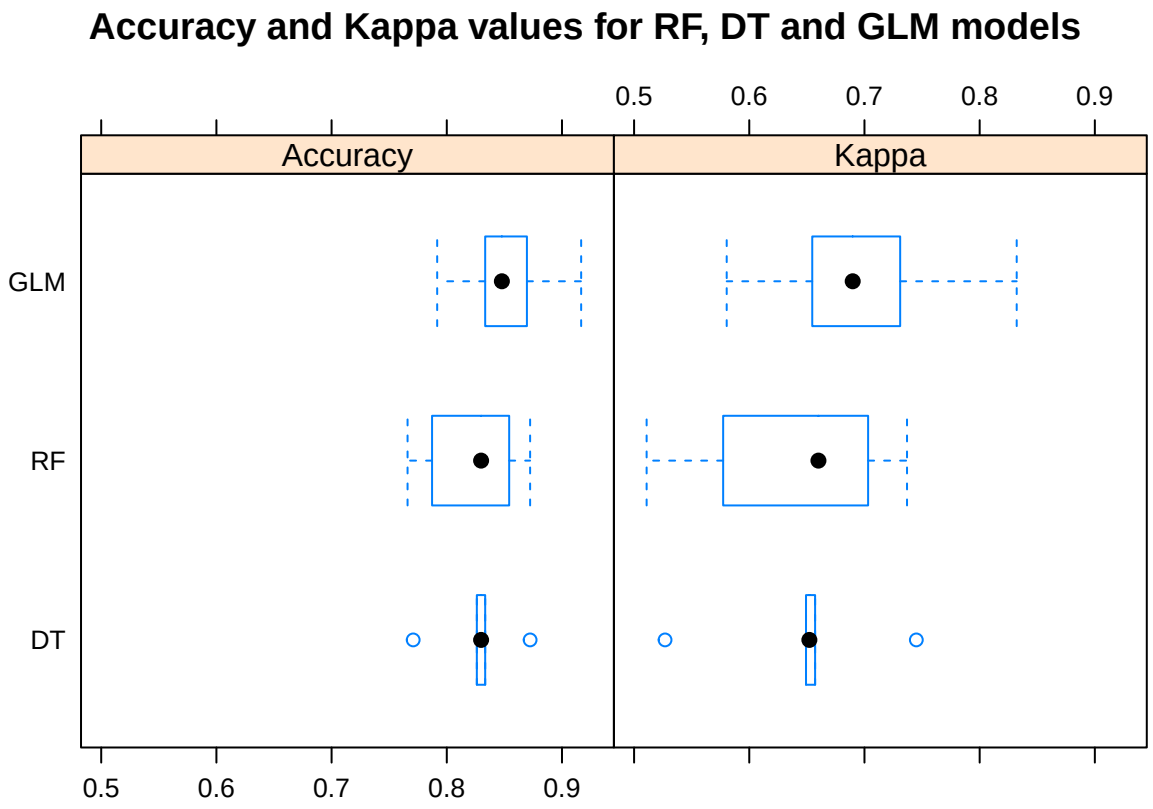
```
fit_rpart <- train(target~.,
             data = train,
             method = "rpart",
             trControl = trainControl(method = "cv",
                                      number = 5,
                                      p = 0.8))
```

4.3 General Linear Model

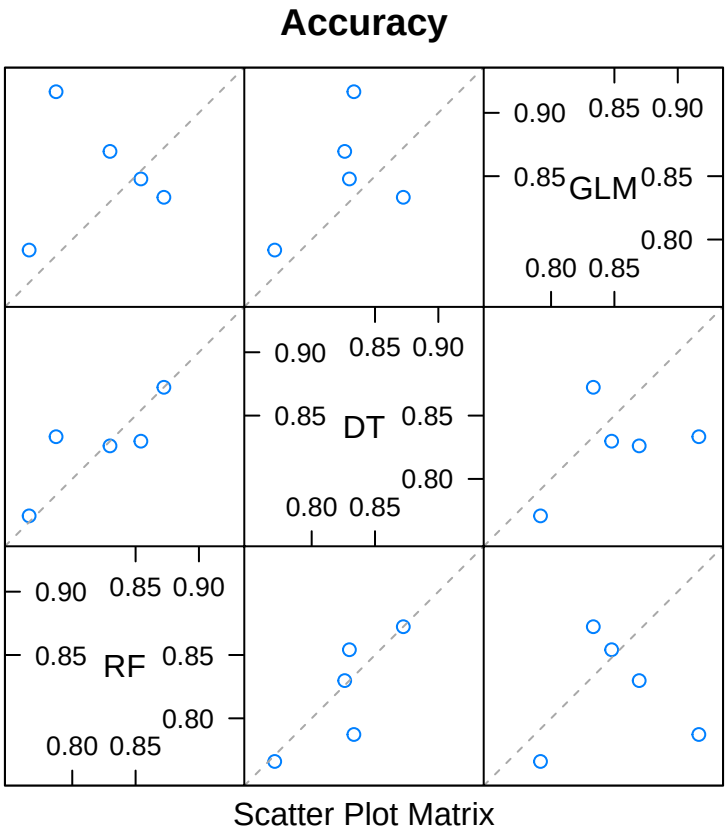
```
fit_glm <- train(target~.,
             data = train,
             method = "glmnet",
                 preProc = c("scale","BoxCox"),
             trControl = trainControl(method = "cv",
                                      number = 5,
                                      p = 0.8))
```

4.4 Evaluating the models

Let's take advantage of the `resample()` function to visualize in a boxplot fashion the results yielded by the 3 models trained:



We can also make use of the `splom()` function to visualize, in a scatter plot matrix, each one of the accuracies yielded during the training process (remember that it was done through a cross-validation with 5 folds)



5. Applying the models

Now let's apply the 3 trained models to the test dataset to try to predict the target variable.

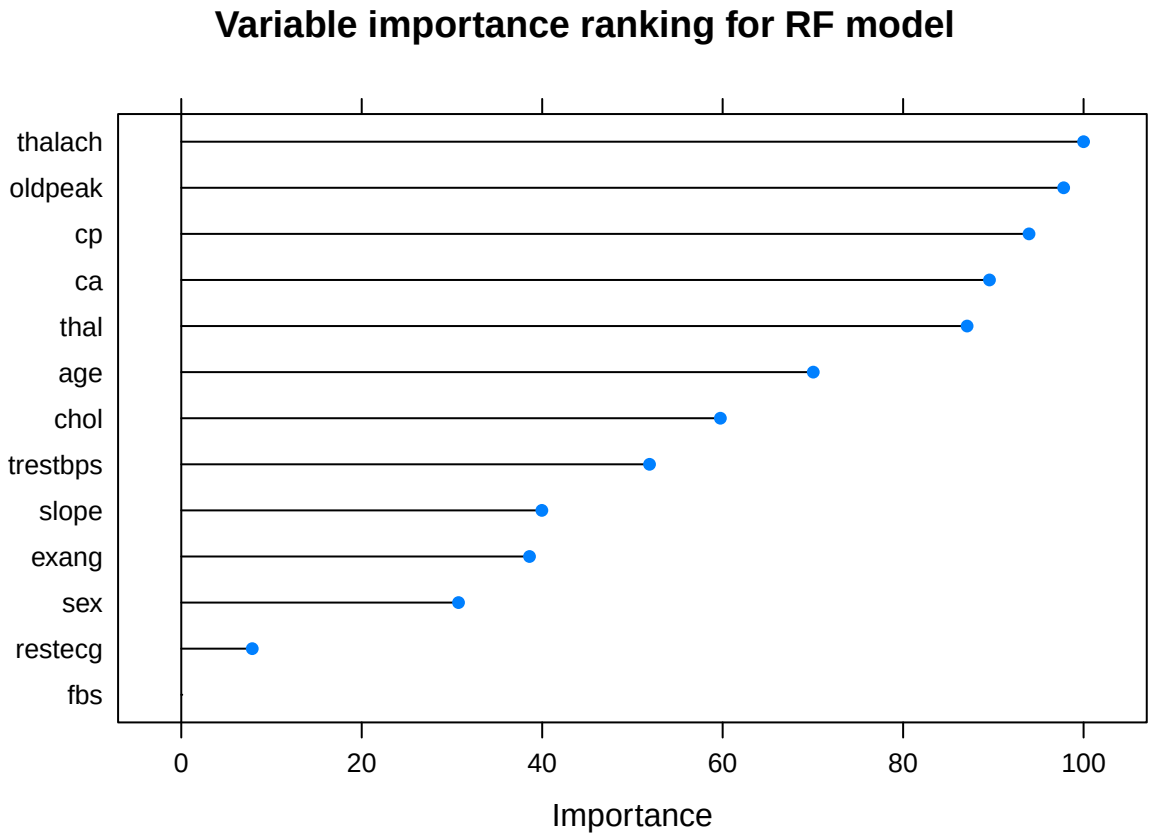
5.1 Random Forest

Accuracy of the Random Forest model is as follows:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 22  3
##          1  6 29
##
##                Accuracy : 0.85
##                  95% CI : (0.7343, 0.929)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 2.293e-07
##
##                   Kappa : 0.6966
```

```
##
##  Mcnemar's Test P-Value : 0.505
##
##             Sensitivity : 0.7857
##             Specificity : 0.9062
##          Pos Pred Value : 0.8800
##          Neg Pred Value : 0.8286
##              Prevalence : 0.4667
##          Detection Rate : 0.3667
##    Detection Prevalence : 0.4167
##       Balanced Accuracy : 0.8460
##
##        'Positive' Class : 0
##
```

We can see which are the most important features for this model by plotting the results of the `varImp()` function:

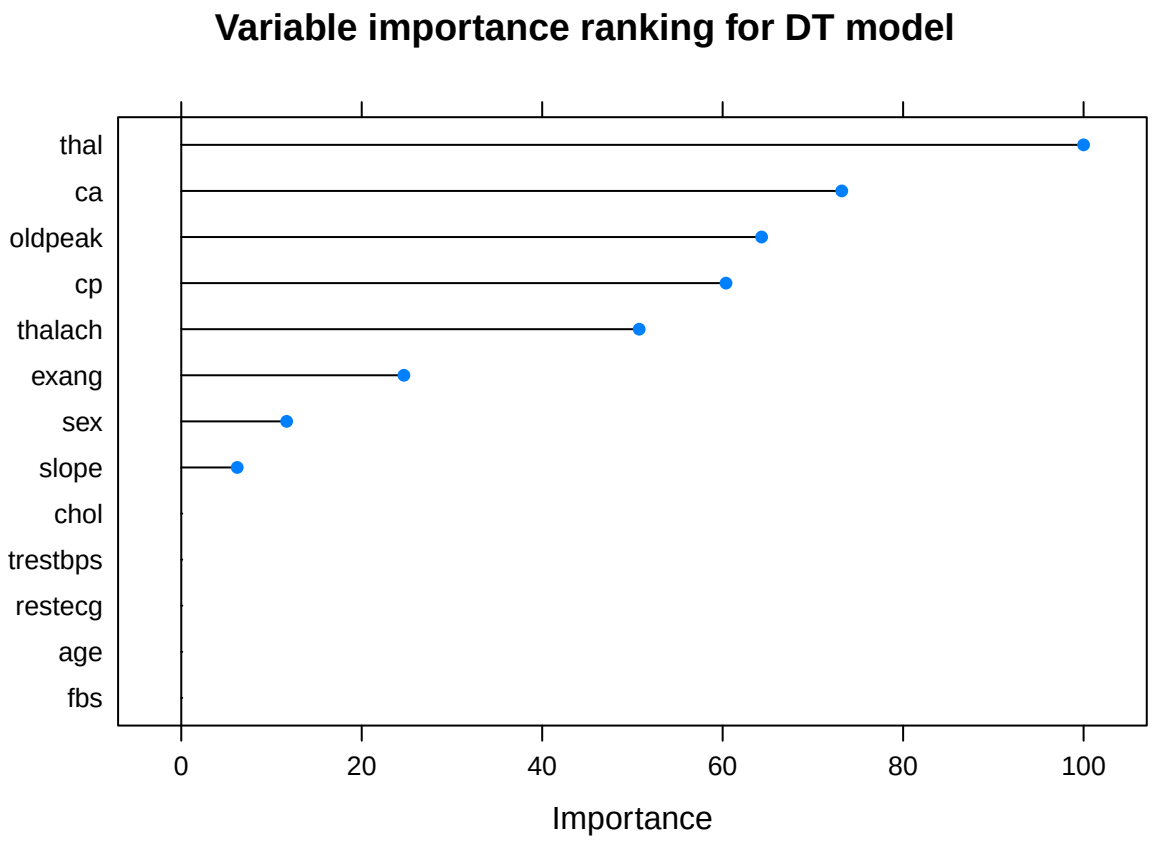


5.2 Decision Tree

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 19  4
##          1  9 28
```

```
##
##                Accuracy : 0.7833
##                  95% CI : (0.658, 0.8793)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 5.405e-05
##
##                   Kappa : 0.5598
##
##  Mcnemar's Test P-Value : 0.2673
##
##             Sensitivity : 0.6786
##             Specificity : 0.8750
##          Pos Pred Value : 0.8261
##          Neg Pred Value : 0.7568
##              Prevalence : 0.4667
##          Detection Rate : 0.3167
##    Detection Prevalence : 0.3833
##       Balanced Accuracy : 0.7768
##
##        'Positive' Class : 0
##
```

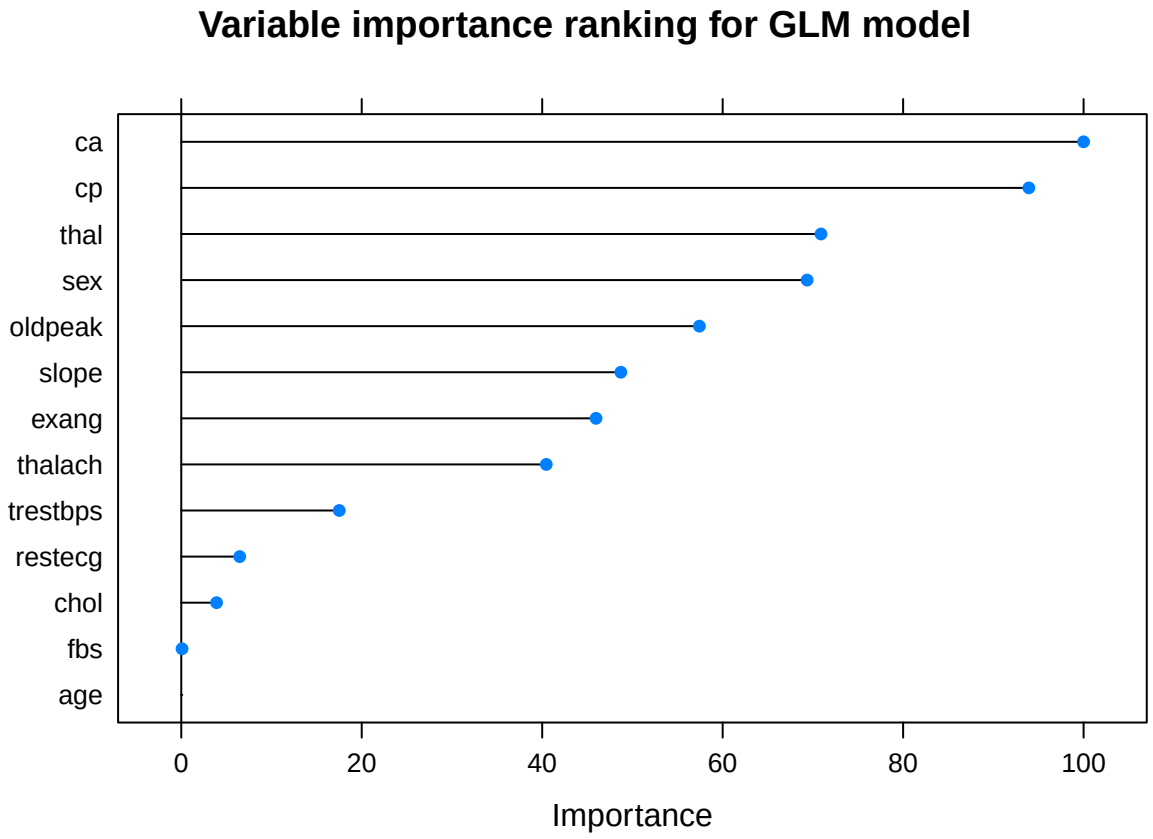
We can see which are the most important features for this model by plotting the results of the `varImp()` function:



5.3 General Linear Model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 22  4
##          1  6 28
##
##                Accuracy : 0.8333
##                  95% CI : (0.7148, 0.9171)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 1.056e-06
##
##                   Kappa : 0.6637
##
##  Mcnemar's Test P-Value : 0.7518
##
##             Sensitivity : 0.7857
##             Specificity : 0.8750
##          Pos Pred Value : 0.8462
##          Neg Pred Value : 0.8235
##              Prevalence : 0.4667
##          Detection Rate : 0.3667
##    Detection Prevalence : 0.4333
##       Balanced Accuracy : 0.8304
##
##        'Positive' Class : 0
##
```

We can see which are the most important features for this model by plotting the results of the `varImp()` function:



The table below summarizes the accuracy and kappa values achieved by each trained model:

Table 4: Summary of the accuracy and Kappa values achieved.

	Accuracy	Kappa
Random Forest	0.8500000	0.6966292
Decision Tree	0.7833333	0.5598194
General Linear Model	0.8333333	0.6636771

6. Conclusions, limitations and future work

6.1 Conclusions

This project aimed at applying the knowledge gained during the Harvard edX Data Science course over a publicly available dataset (Heart Disease UCI dataset, available at Kaggle).

After downloading the dataset, we have performed the following operations over it:

1. Transform the dataset into a tidy format, making use of the `tidyverse` library.
2. Proceed with a comprehensive data visualization process, differentiating continuous from categorical variables, and grouping by the variables of interest, in this case, gender and health status. For this purpose, we've made an extensive use of several libraries, such as `ggplot2`, `knitr`, `DataExplorer`, `gridExtra`, `ggpubr` and `GGally`.

3. Making use of the `caret` library, then we proceeded to generate train and test subsets in order to evaluate the performance of three different machine learning methods when predicting the target variable (health status of a subject).
4. These three machine learning methods (Random Forest, Decision Trees and General Linear Model) were then trained and their performance assessed through a 5-fold cross-validation on the train subset making use of the libraries `e1071`, `rpart` and `glmnet` respectively.
5. Finally, we evaluated the trained models against the test subset in terms of accuracy and Kappa values achieved, and visualized the ranking of the variables importance for each one of the models trained.

In this project, we've successfully applied the techniques learned during the course and some other libraries and functions have been explored.

6.2 Limitations

This work has two main limitations: on one hand, the dataset is not large enough to extract meaningful conclusions about the correlation between the variables and health status of the subjects. However, it poses a good example about how some clinical variables are more determinant than others when trying to predict the health status of a subject.

On the other hand, the dataset is biased, accounting with two times more males than females and a higher population of asymptomatic subjects, which burdens the training capacity of the machine learning models.

6.3 Future work

Future work could address the assessment of other machine learning models, such as Extreme Gradient Boosted Decision Trees (XGBoost), Support Vector Machines (SVM) and/or other clustering methods such as k-Nearest Neighbors (kNN), among others, and compare the accuracy and Kappa values obtained.

Besides, it could be further investigated other performance measurements such as the area under the receiver-operator curve (AUC).