Heart Disease Project Kadiatou KABA

Kadiatou Kaba

12/17/2020

Introduction

Ischemic heart disease (IHD) is a leading cause of death worldwide. Also referred to as coronary artery disease (CAD) and atherosclerotic cardiovascular disease (ACD), it manifests clinically as myocardial infarction and ischemic cardiomyopathy.

Coronary heart disease is often caused by the buildup of plaque, a waxy substance, inside the lining of larger coronary arteries. This buildup can partially or totally block blood flow in the large arteries of the heart. Some types of this condition may be caused by disease or injury affecting how the arteries work in the heart.

Symptoms of coronary heart disease may be different from person to person even if they have the same type of coronary heart disease. However, because many people have no symptoms, they do not know they have coronary heart disease until they have chest pain, a heart attack, or sudden cardiac arrest.

(Source: https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease)

In this report, we tried to learn enough information of this topic to understand the **Heart Disease UCI dataset** and build simple models to predict whether a patient has a disease or not based on features like the heart rate during exercise or the cholesterol levels in the blood.

Dataset

This dataset is hosted on Kaggle (Heart Disease UCI), and it was from UCI Machine Learning Repository. There are records of about 300 patients from Cleveland and the features are described in a following section.

Cleaning the dataset

Thanks to the post of InitPic, this dataset is a bit different from the original one while the description is the same.

Part of these differences is that there were a few null values in the original dataset that have taken some values here:

A few more things to consider:

- data #93, 159, 164, 165 and 252 have ca=4 which is incorrect. In the original Cleveland dataset they are NaNs (so they should be removed)

- data #49 and 282 have thal = 0, also incorrect. They are also NaNs in the original dataset.

Because these are just a few instances, I decided to drop them.

There are also some differences regarding the features of the dataset which are corrected below.

Dataset features

heartDisease Dataset

```
##   age    sex              cp trestbps chol fbs     restecg thalach exang
## 1  63   Male  Typical angina      145  233 Yes Hypertrophy     150    No
## 2  37   Male       No angina      130  250  No      Normal     187    No
## 3  41 Female Atypical angina      130  204  No Hypertrophy     172    No
## 4  56   Male Atypical angina      120  236  No      Normal     178    No
## 5  57 Female    Asymptomatic      120  354  No      Normal     163   Yes
## 6  57   Male    Asymptomatic      140  192  No      Normal     148    No
##   oldpeak      slope ca         thal target
## 1     2.3 Descending  0 Fixed defect     No
## 2     3.5 Descending  0  Normal flow     No
## 3     1.4  Ascending  0  Normal flow     No
## 4     0.8  Ascending  0  Normal flow     No
## 5     0.6  Ascending  0  Normal flow     No
## 6     0.4       Flat  0 Fixed defect     No
```

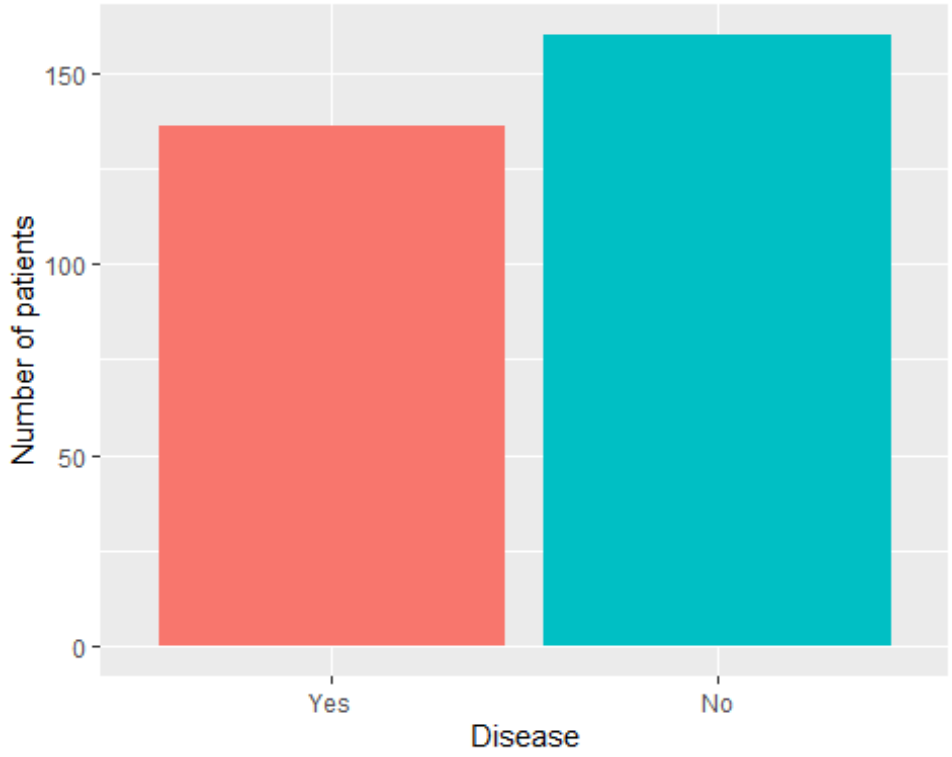
Data vizualisation

Target

Variable *target*: whether the patient has a heart disease or not

Value 0: yes Value 1: no

We can see that the distribution is quite balanced. Thanks to this it wouldn't be a bad idea using accuracy to evaluate how well the models perform.

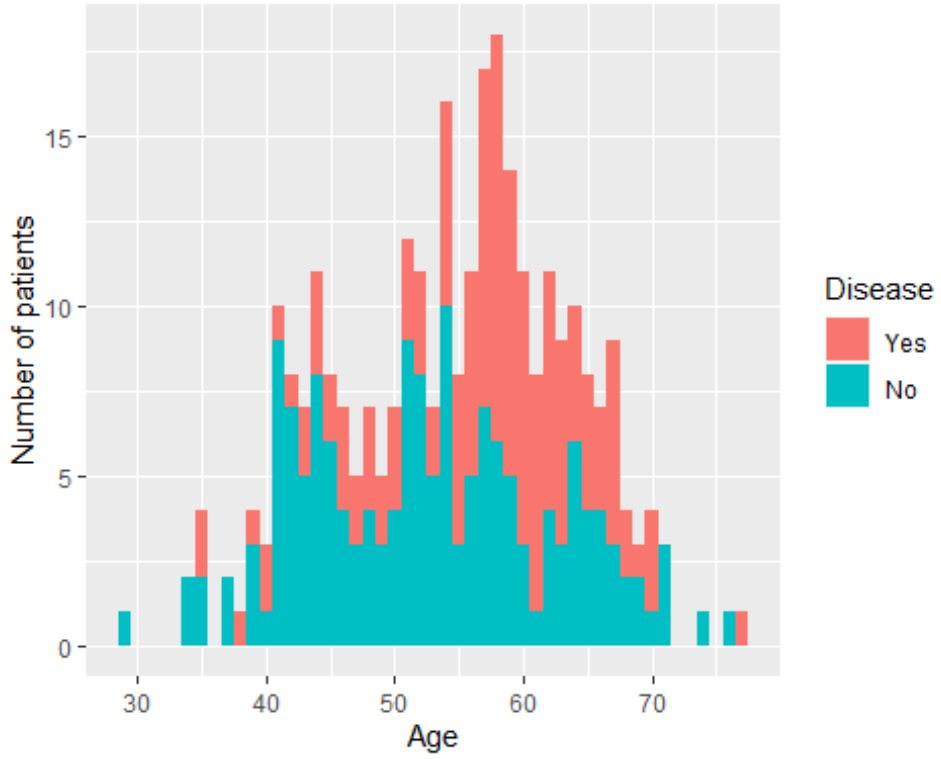



```
##
##       Yes        No
## 0.4594595 0.5405405
```

45.9% of the patients in the dataset have a heart disease.

Age

Variable *age*: Patient age in years. In the data we can see, as expected, that age is a risk factor. In other words, the higher the age, the more likely that the patient has a heart disease.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   48.00   56.00   54.52   61.00   77.00
```

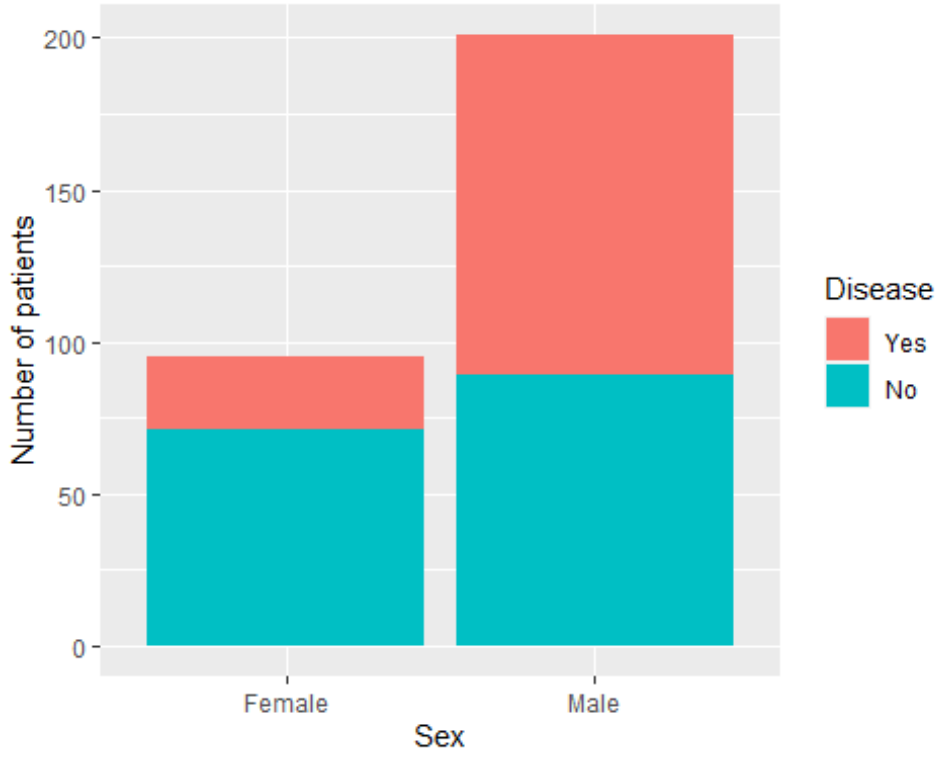
The average age of the patients is **54.5**.

Sex

Variable *sex*: Patient's sex

Value 0: female Value 1: male

There are approximately half the observation of women than men. We can also see that sex is a risk factor, like some of the references indicate, men are more likely to have a heart disease than women.




```
##
##    Female      Male
## 0.3209459 0.6790541
```

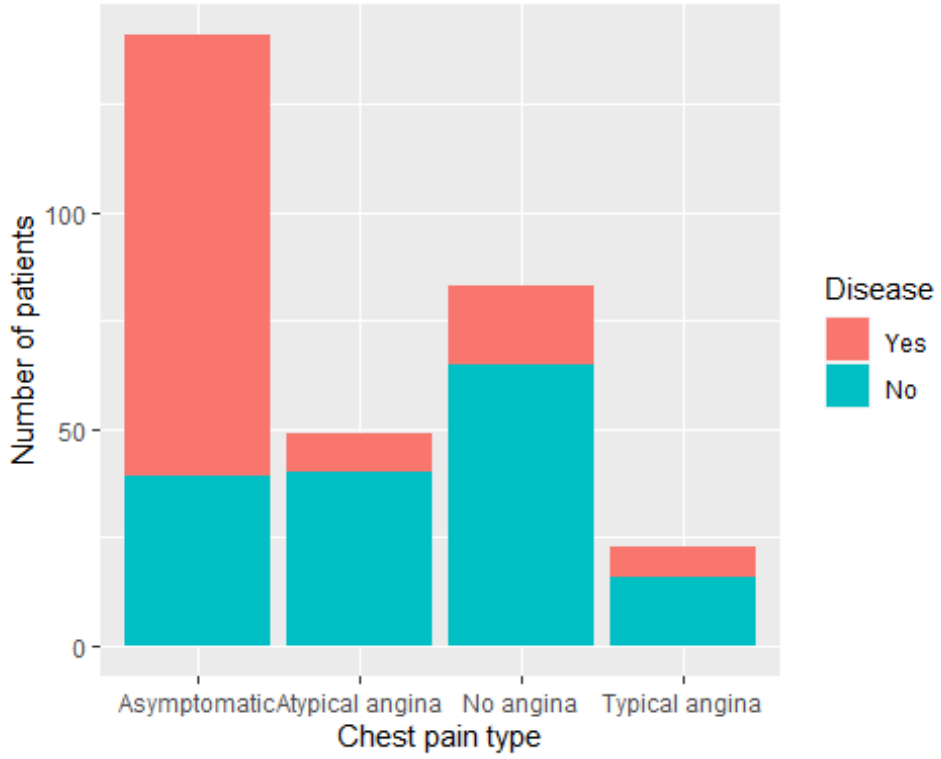
68% of the patients are male.

CP

Variable *cp*: Chest pain type

Value 0: asymptomatic Value 1: atypical angina Value 2: pain without relation to angina Value 3: typical angina

The description of the data doesn't provide information about how this classification of pain was made. But we can see that it is very difficult to tell whether a patient has a heart disease just looking at the symptoms of the patients.



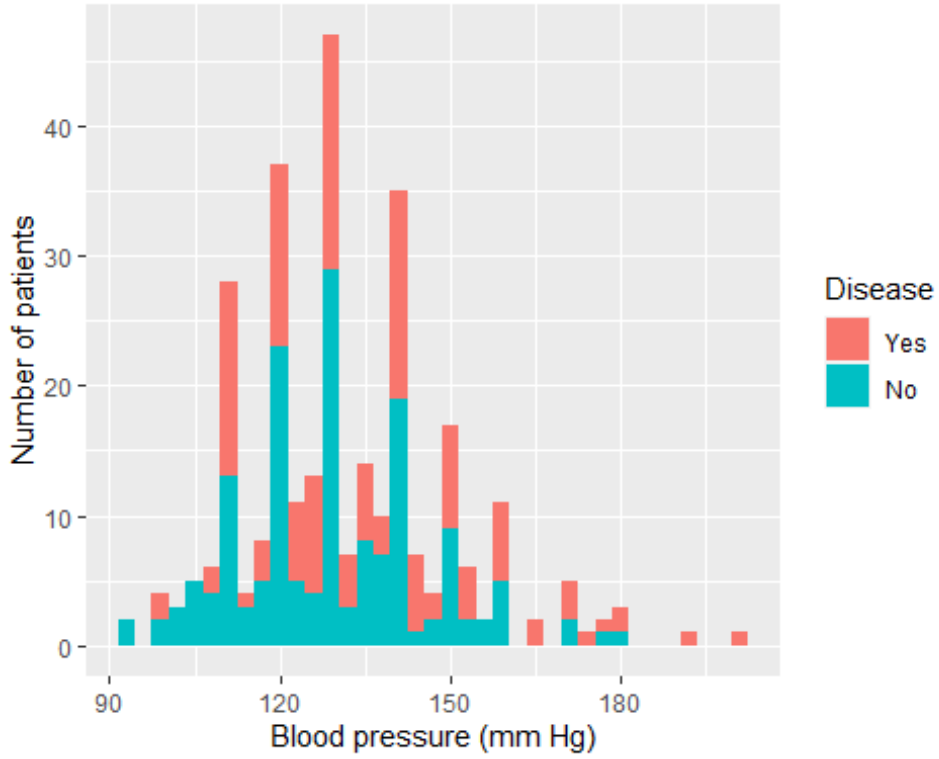

```
##
##    Asymptomatic Atypical angina       No angina  Typical angina
##       0.4763514       0.1655405       0.2804054       0.0777027
```

47.6% of the patients are asymptomatic and **28%** have no angina.

trestbps

Variable *trestbps*: Resting blood pressure in millimeters of mercury (mm Hg) when the patient was admitted to the hospital.

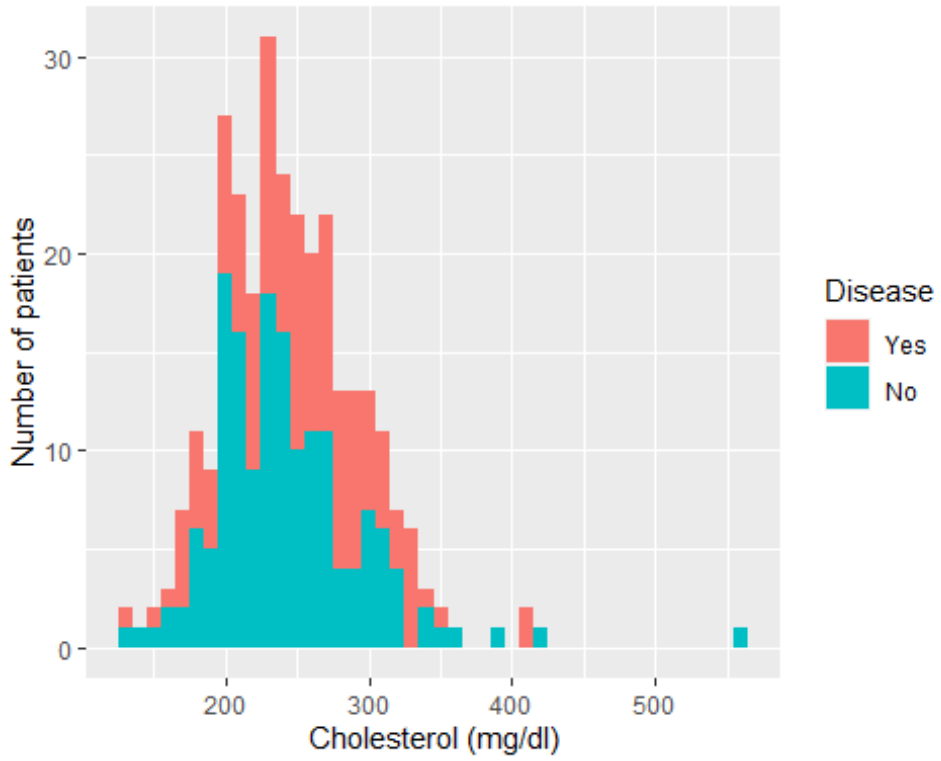
By the different peaks, looks like most people tend to have a normal blood pressure inside certain groups (could be healthy adults, adults that take medication, seniors…). It also looks like very high pressures can indicate that there is a heart disease.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    94.0   120.0   130.0   131.6   140.0   200.0
```

Chol

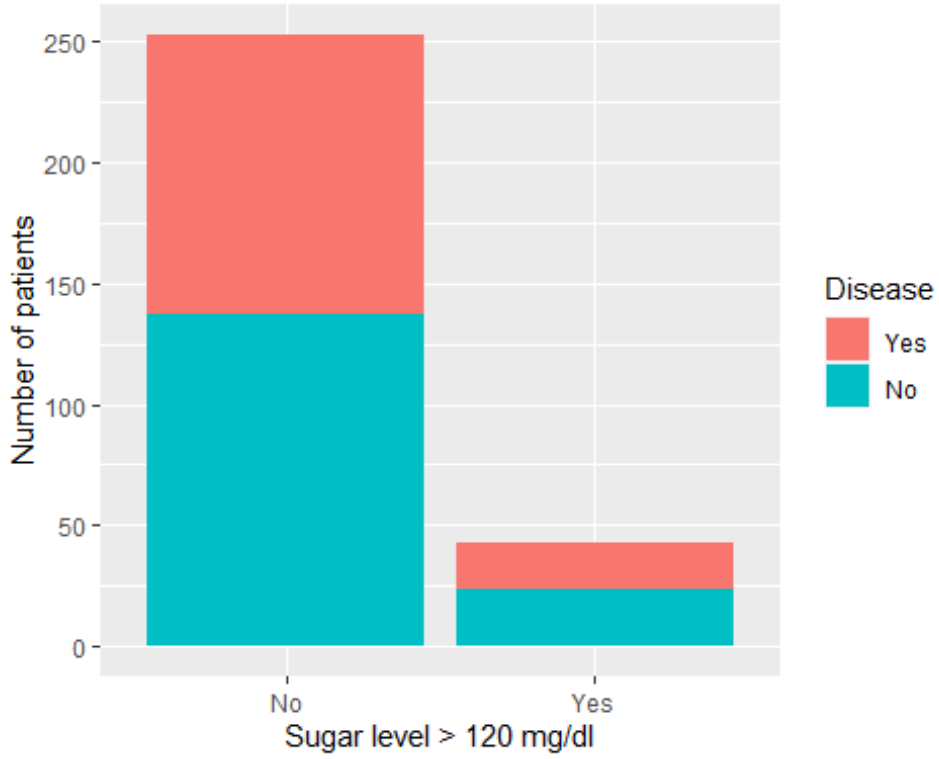
Variable *chol*: Cholesterol level in mg/dl. This is a variable that we can control to prevent the disease. Looks like the majority of the people in the dataset have high levels of cholesterol. It also looks like up to a certain level, the presence of a heart disease is slightly higher on higher cholesterol levels. Though the cases that have the highest levels of cholesterol don't have a heart disease.



Fbs

Variable *fbs*: Whether the level of sugar in the blood is higher than 120 mg/dl or not. This is another variable that we can control. However, by itself it doesn't seem very useful to know if a patient has a heart disease or not. Though we shouldn't drop it right now because it might be useful combined with other variables.

Value 0: no Value 1: yes



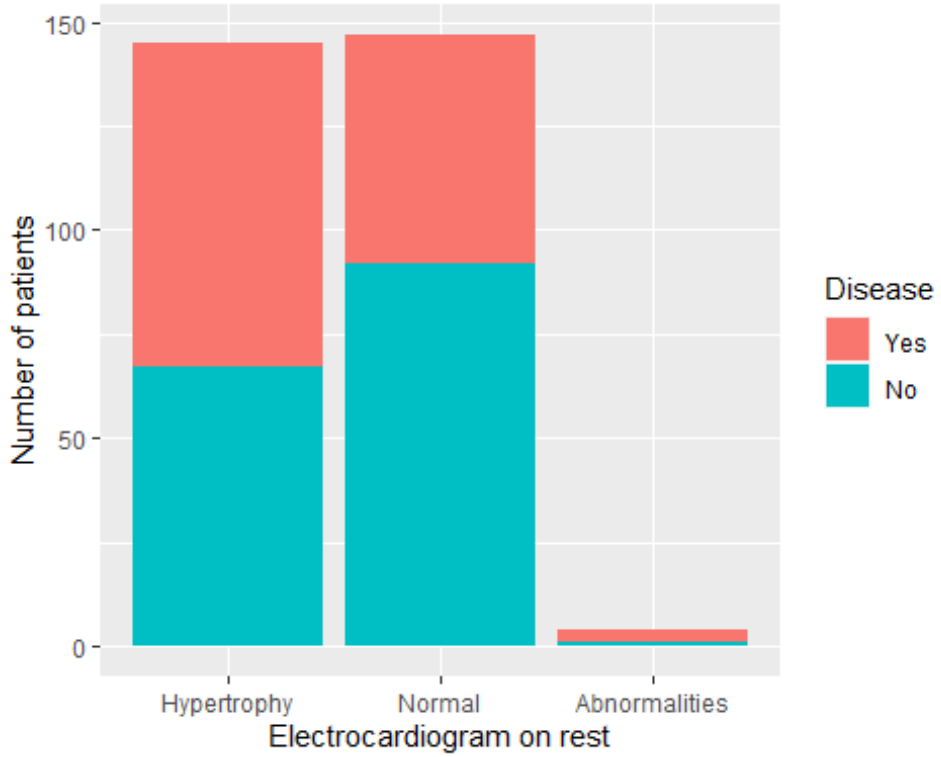
Hereon, variables are related to a nuclear stress test. That is, a stress test where a radioactive dye is also injected to the patient to see the blood flow.

Restecg

Variable *restecg*: Results of the electrocardiogram on rest

Value 0: probable left ventricular hypertrophy Value 1: normal Value 2: abnormalities in the T wave or ST segment When someone has a heart disease the first symptom usually is stable angina (angina during exercise). When angina happens even on rest the disease got worse (usually due to a narrowing of the coronary arteries). This has to be why there are so few patients that show an abnormality on the heart rate on rest, and it is also why seeing this abnormality is very indicative of a presence of a heart disease.

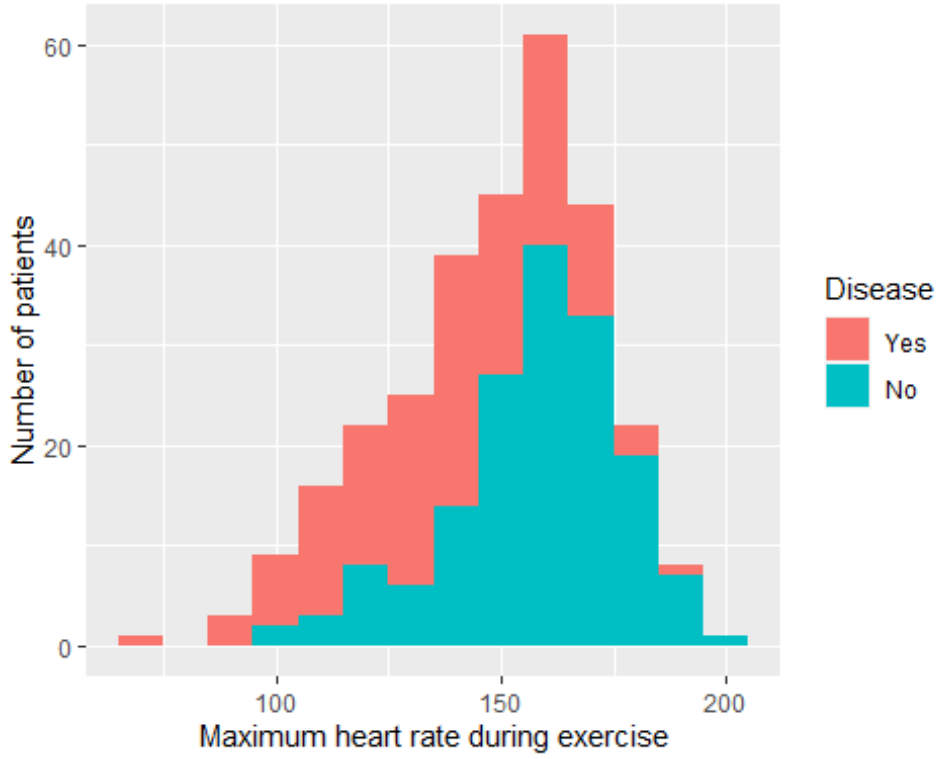
On the other hand, the value 0, probable presence of an hypertrophy, doesn't seem to be very indicative of the presence of a heart disease by itself. It can be because this variable is not very accurate (as noted by the "probable presence").



Thalach

Variable *thalach*: Maximum heart rate during the stress test

At first sight, it may seem weird to see that the higher the heart rate the lower the presence of a heart disease and vice versa. However, it makes sense taking into account that the maximum healthy heart rate depends on the age (220 - age). Thus, higher rates tend to be from younger people.



Eight patients with a heart rate during exercise lower than 100:

```
##     age thalach target
## 137  60      96     No
## 199  62      99    Yes
## 217  62      97    Yes
## 234  64      96    Yes
## 244  57      88    Yes
## 263  53      95    Yes
## 273  67      71    Yes
## 298  59      90    Yes
```

Eighteen patients with a heart rate during exercise higher than 180:

```
##     age thalach target
## 2    37     187     No
## 33   44     188     No
## 45   39     182     No
## 57   48     186     No
## 58   45     185     No
## 63   52     190     No
## 66   35     182     No
## 73   29     202     No
## 74   51     186     No
## 79   52     184     No
## 104  42     194     No
## 122  59     182     No
```

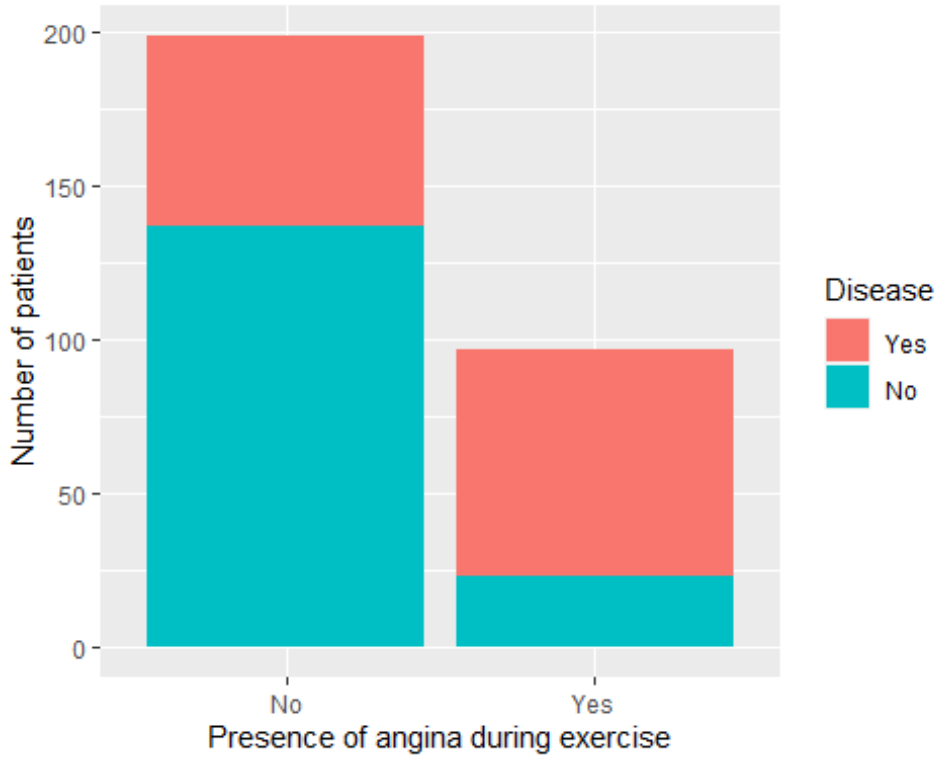
```
## 126   34      192      No
## 142   43      181      No
## 163   41      182      No
## 249   54      195     Yes
## 260   38      182     Yes
## 284   40      181     Yes
```

Exang

Variable *exang*: Whether the patient had angina during exercise

Value 0: no Value 1: yes

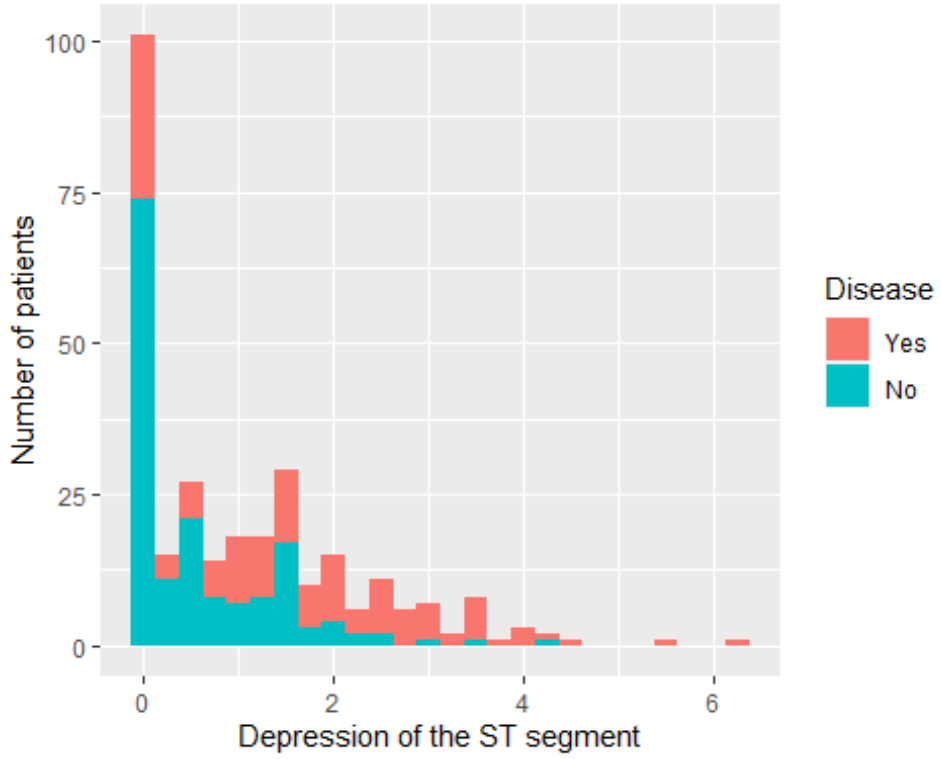
We can see that this feature is a good indicator for the presence of heart disease. However, we can also see that knowing what is angina and what not is not an easy task, it can be confused with other pains or it can be atypical angina.



Oldpeak

Variable *oldpeak*: Decrease of the ST segment during exercise according to the same one on rest.

The ST segment is a part of the electrocardiogram of a heart beat that is usually found at a certain level in a normal heart beat. A significant displacement of this segment can indicate the presence of a heart disease as we can see in the plot.

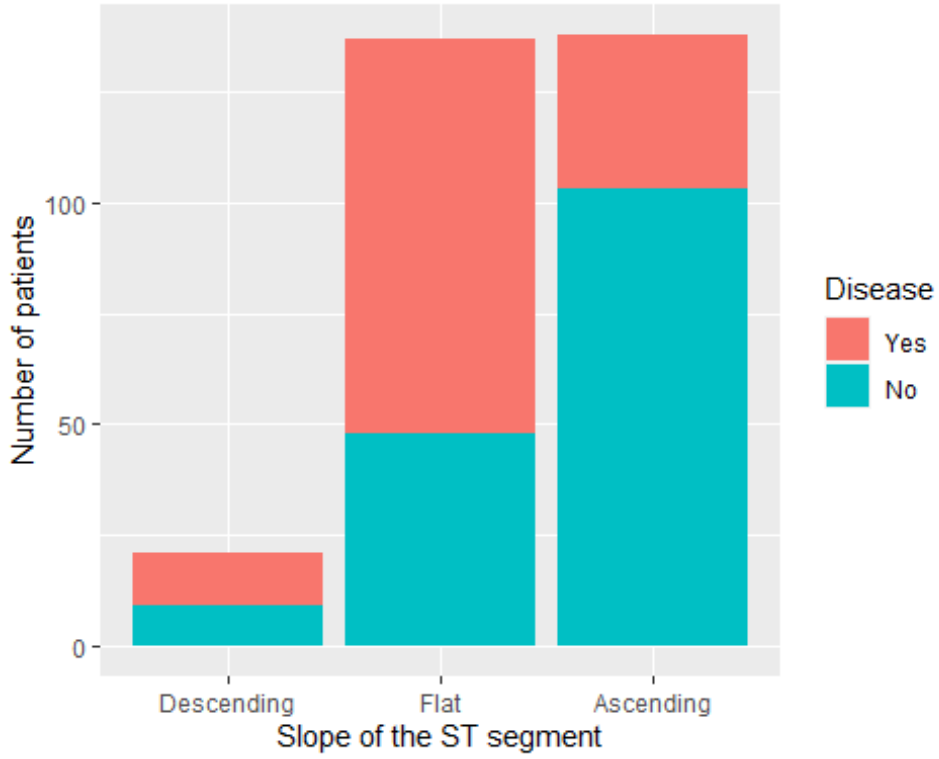


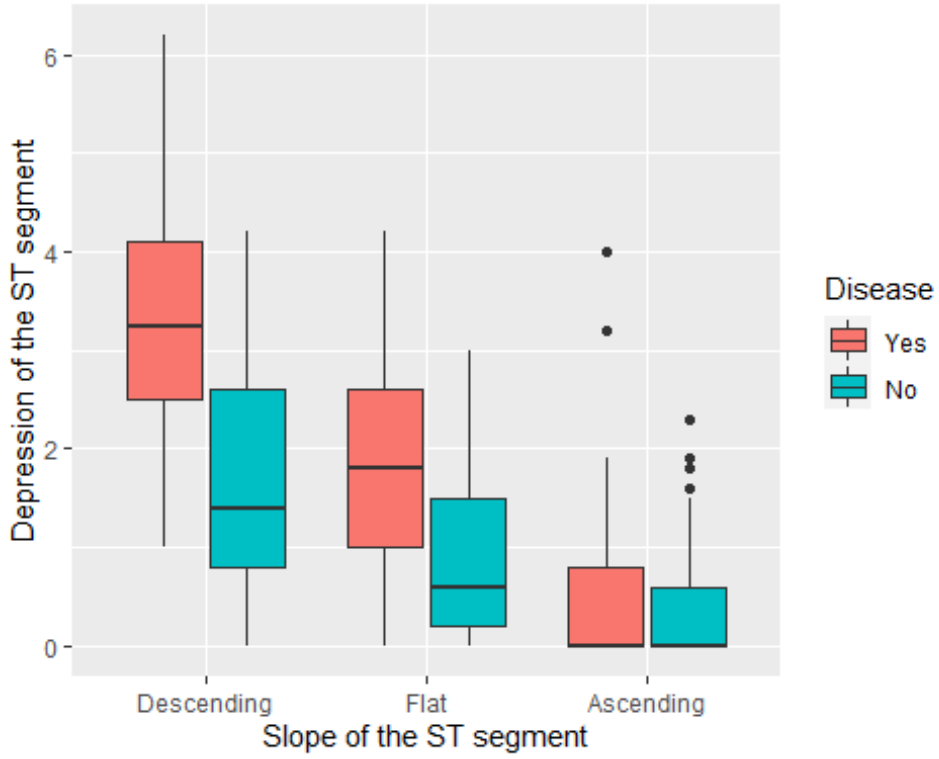
slope

Variable *slope*: Slope of the ST segment during the most demanding part of the exercise

Value 0: descending Value 1: flat Value 2: ascending

In the first graph, we can see that the slope by itself can help determine whether there is a heart disease or not if it is flat or ascending. However, if the slope is descending it doesn't seem to give much information. Because of this, in the second graph a third variable was added and we can notice that, if the slope is descending, the depression of the ST segment can help to determine if the patient has a heart disease.



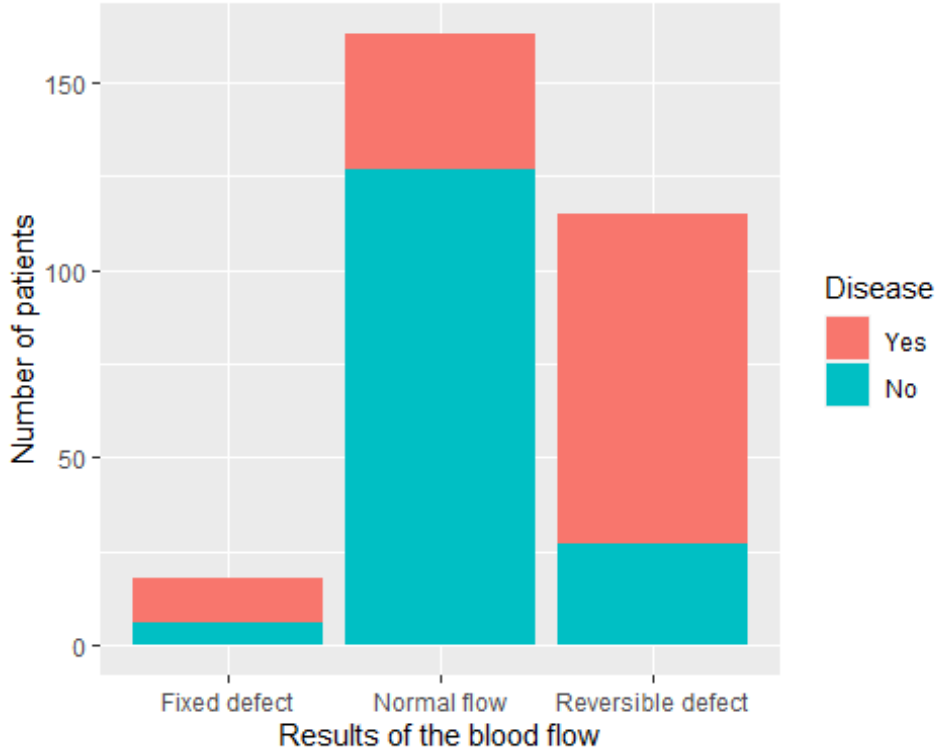


Thal

Variable *thal*: Results of the blood flow observed via the radioactive dye.

Value 0: NULL (dropped from the dataset previously) Value 1: fixed defect (no blood flow in some part of the heart) Value 2: normal blood flow Value 3: reversible defect (a blood flow is observed but it is not normal)

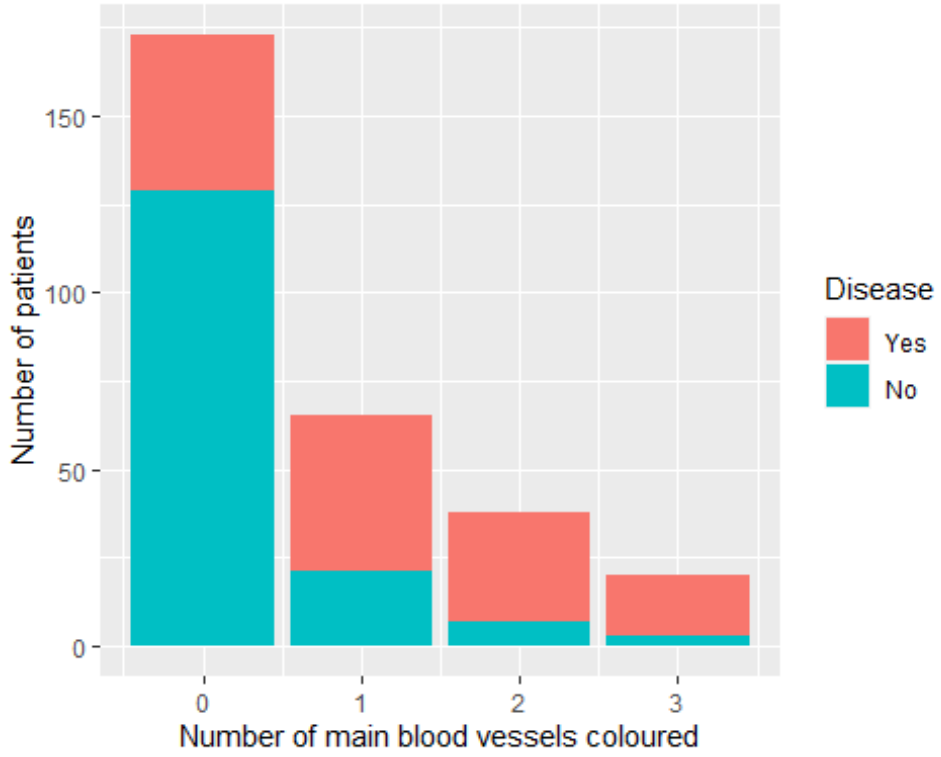
This feature and the next one are obtained through a very invasive process for the patients. But, by themselves, they give a very good indication of the presence of a heart disease or not.



Ca

Variable *ca*: Number of main blood vessels coloured by the radioactive dye. The number varies between 0 to 4 but the value 4 represents a null value and these have been dropped previously.

This feature refers to the number of narrow blood vessels seen, this is why the higher the value of this feature, the more likely it is to have a heart disease.



Models

For this part, we chose to build four models. Three simple ones: logistic regression, naïve bayes and decision trees. And one a bit more complex: random forest.

Null values where already dropped in a previous cell. Also, categorical variables have already been converted to R factors. Apart from this no more explicit preprocessing was done in this notebook to keep it simple and easy to follow.

To compare the models we first divide the dataset in a training set with **70%** of instances and a test set with the rest of the instances. And this taking into account that the distribution of the target has to be the same in both sets.

The test set mimics data in the real world, it will only be used at the end of the project to get a more robust measure of the models on unseen data.

The training set will be used to evaluate the models via 10 fold cross-validation. For simplicity, we'll leave the hyperparameter selection of the models by default, this means that some random combinations will be chosen and the models will be trained via cross-validation for each combination, keeping the hyperparameter combination that gave the best result.

Logistic regression

Looks like this implementation of logistic regression doesn't have any hyperparameter to tune. However, the results are not bad. A decent accuracy and a Kappa of almost 0.7 is usually considered good.

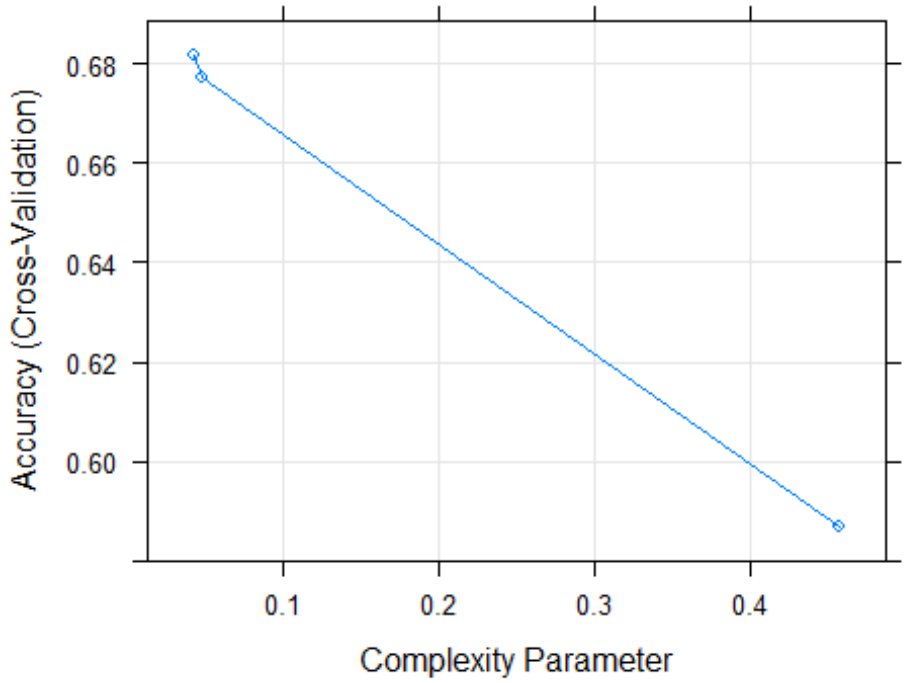
```
## Generalized Linear Model
##
## 208 samples
##  13 predictor
##   2 classes: 'Yes', 'No'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 187, 187, 187, 187, 187, 187, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7971429  0.5919854
```

The accuracy is **0.797**.

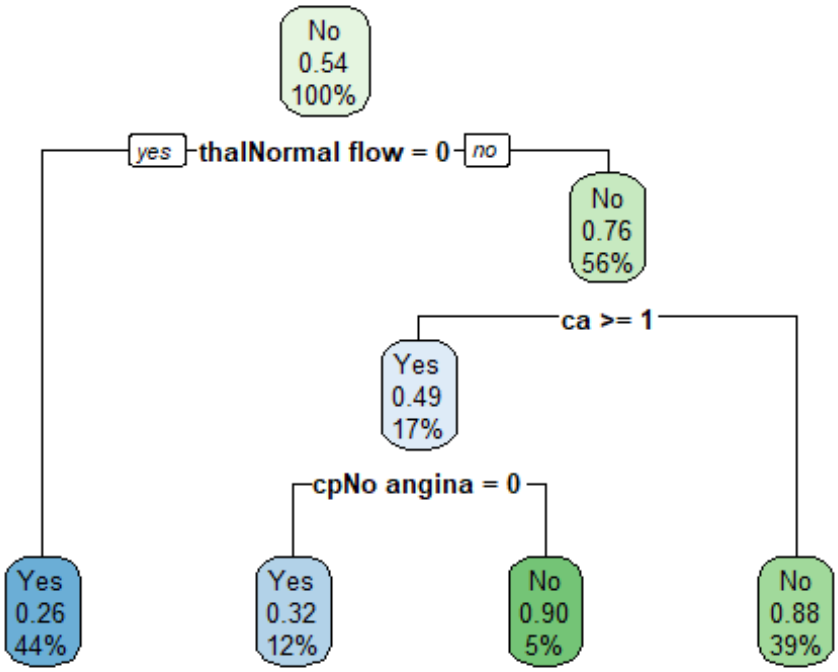
Decision tree

Looks like the tree was build for three different values of a cp hyperparameter. This comes from rpart and it is a hyperparameter that governs the complexity of the model: lower values give more complex (bigger) trees.

```
## CART
##
## 208 samples
##  13 predictor
##   2 classes: 'Yes', 'No'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 187, 187, 187, 187, 187, 187, ...
## Resampling results across tuning parameters:
##
##   cp          Accuracy   Kappa
##   0.04166667  0.6819048  0.3604422
##   0.04687500  0.6771429  0.3511199
##   0.45833333  0.5866667  0.1252208
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.04166667.
```

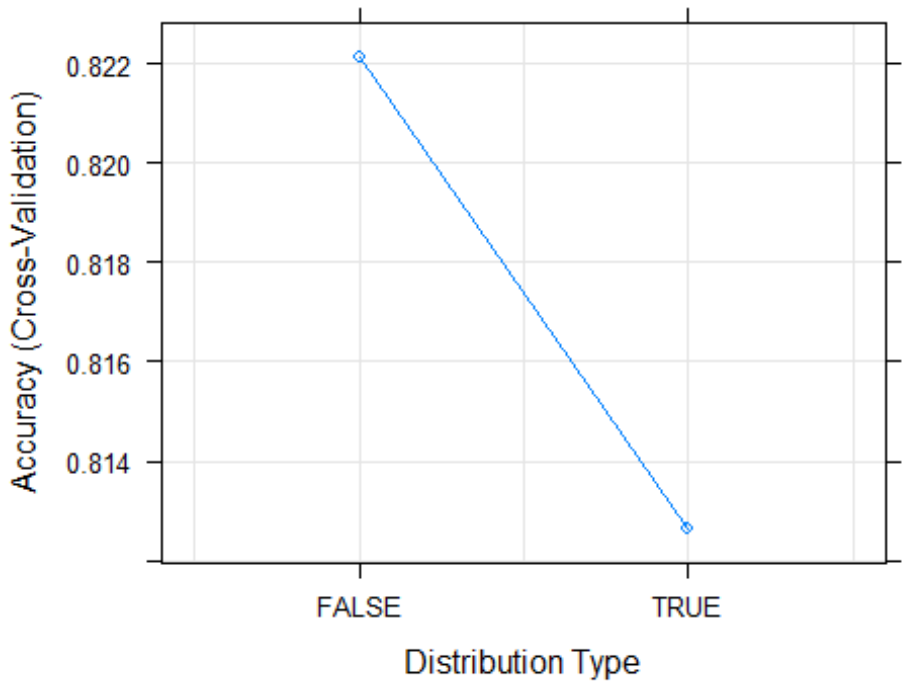


The results are not that great (**0.682**) but the tree is small and easy to interpret as we can see below.



Naive Bayes

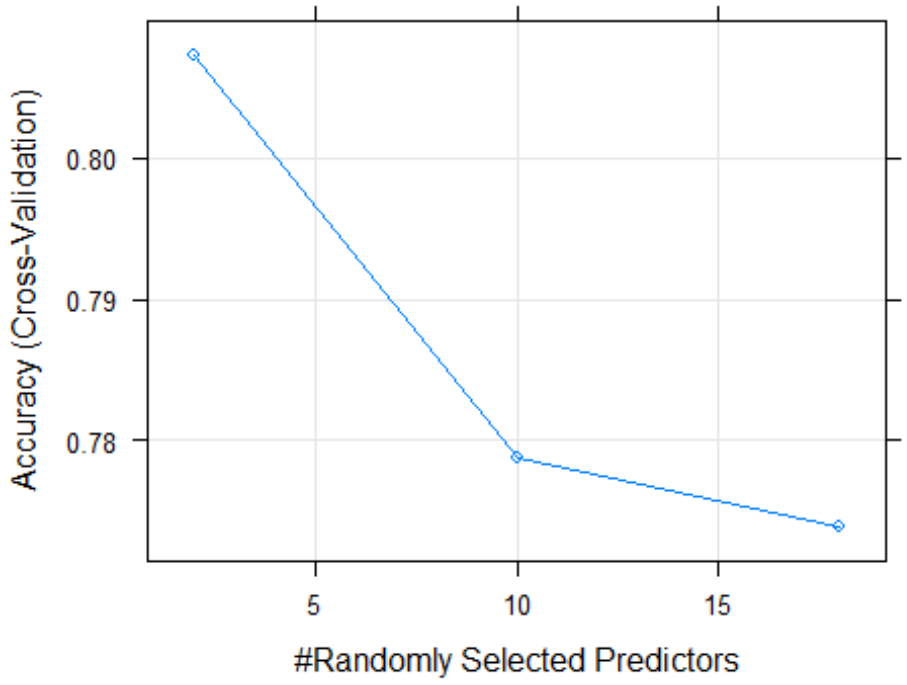
```
## Naive Bayes
##
## 208 samples
##  13 predictor
##   2 classes: 'Yes', 'No'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 187, 187, 187, 187, 187, 187, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa
##   FALSE      0.8221429  0.6403881
##    TRUE      0.8126190  0.6233218
##
## Tuning parameter 'laplace' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were laplace = 0, usekernel = FALSE
##  and adjust = 1.
```



These results are the best for now (**0.822**).

Random Forest

```
## Random Forest
##
## 208 samples
##  13 predictor
##   2 classes: 'Yes', 'No'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 187, 187, 187, 187, 187, 187, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.8073810  0.6108090
##   10    0.7788095  0.5543190
##   18    0.7738095  0.5440596
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```



These results are also good.

Results

Comparison of the trainings

We can see in the following tables a summary of the results of the 10 fold cross-validation.

```
##
## Call:
## summary.resamples(object = results, metric = c("Kappa", "Accuracy"))
##
## Models: LogisticReg, Tree, NaiveBayes, RandomForest
## Number of resamples: 10
##
## Kappa
##                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
NA's
## LogisticReg  0.2079208 0.4533566 0.6458059 0.5919854 0.7829239 0.8108108
0
## Tree         0.0625000 0.3056452 0.3287111 0.3604422 0.4756128 0.7123288
0
## NaiveBayes   0.3287671 0.4743243 0.6457302 0.6403881 0.8055046 0.9041096
0
## RandomForest 0.3287671 0.5248869 0.6019802 0.6108090 0.7096774 0.9041096
0
##
## Accuracy
##                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
NA's
## LogisticReg  0.6000000 0.7261905 0.8285714 0.7971429 0.8928571 0.9047619
0
## Tree         0.5238095 0.6541667 0.6666667 0.6819048 0.7380952 0.8571429
0
## NaiveBayes   0.6666667 0.7357143 0.8297619 0.8221429 0.9047619 0.9523810
0
## RandomForest 0.6666667 0.7619048 0.8047619 0.8073810 0.8571429 0.9523810
0
```

Naive Bayes and random forest were the best models (**0.822** and **0.807** respectively).

On the other hand, the rest of the models have metrics that vary more, this indicates that they might not be reliable to deal with unseen data.

As for final results, we'll see how Naive Bayes and Random forest perform on the test set. And we'll see which one does a good job.

Final results

Naive Bayes

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Yes No
##        Yes  34  5
##        No    6 43
##
##                Accuracy : 0.875
```

```
##                  95% CI : (0.7873, 0.9359)
##     No Information Rate : 0.5455
##     P-Value [Acc > NIR] : 3.53e-11
##
##                   Kappa : 0.7474
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8500
##             Specificity : 0.8958
##          Pos Pred Value : 0.8718
##          Neg Pred Value : 0.8776
##              Prevalence : 0.4545
##          Detection Rate : 0.3864
##    Detection Prevalence : 0.4432
##       Balanced Accuracy : 0.8729
##
##        'Positive' Class : Yes
##
```

The accuracy with Naive Bayes model is **0.875**.

Random Forest

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Yes No
##        Yes  32  6
##        No    8 42
##
##                Accuracy : 0.8409
##                  95% CI : (0.7475, 0.9102)
##     No Information Rate : 0.5455
##     P-Value [Acc > NIR] : 4.384e-09
##
##                   Kappa : 0.6778
##
##  Mcnemar's Test P-Value : 0.7893
##
##             Sensitivity : 0.8000
##             Specificity : 0.8750
##          Pos Pred Value : 0.8421
##          Neg Pred Value : 0.8400
##              Prevalence : 0.4545
##          Detection Rate : 0.3636
##    Detection Prevalence : 0.4318
##       Balanced Accuracy : 0.8375
##
##        'Positive' Class : Yes
##
```

The accuracy with the Random Forest model is **0.841**.

In conclusion, the *Naive Bayes* model is the one with the highest accuracy. We also have a confusion matrix with most observations on the main diagonal.