Capstone

Andre

15 may 2019

Introduction

For this project for the Data Science Capstone from EDX we will perform a EDA and create a machine learning model using a choosen dataset.

The data choosed is the Heart Disease Data Set from the UCI Machine Learning Repository. We will use 4 dataset:

1. Cleveland Clinic Foundation (cleveland.data)
2. Hungarian Institute of Cardiology, Budapest (hungarian.data)
3. V.A. Medical Center, Long Beach, CA (long-beach-va.data)
4. University Hospital, Zurich, Switzerland (switzerland.data)

In this dataset we the information about pacients that went to 4 hospitals and had several test and were diagnosed with or without heart desease.

Executive summary

In this project we will:

- load the data
- perform some cleaning and transformation
- analysis the data
- prepare the data
- perform a machine learnig model
- evaluate the model

Loading the data

```
download.file(paste('http://archive.ics.uci.edu/ml/machine-learning-databases/',
            'heart-disease/processed.cleveland.data', sep = ''), 'processed.cleveland.data')
data.1 <- read.csv('processed.cleveland.data',
           col.names = c('age', 'sex', 'cp', 'trestbps', 'chol', 'fbs',
                 'restecg', 'thalach', 'exang', 'oldpeak', 'slope',
                 'ca', 'thal', 'target'))

download.file(paste('http://archive.ics.uci.edu/ml/machine-learning-databases/',
            'heart-disease/reprocessed.hungarian.data', sep = ''), 'processed.hungarian.data')
data.2 <- read.csv('processed.cleveland.data',
           col.names = c('age', 'sex', 'cp', 'trestbps', 'chol', 'fbs',
                 'restecg', 'thalach', 'exang', 'oldpeak', 'slope',
                 'ca', 'thal', 'target'))

download.file(paste('http://archive.ics.uci.edu/ml/machine-learning-databases/',
            'heart-disease/processed.switzerland.data', sep = ''), 'processed.switzerland.data')
data.3 <- read.csv('processed.cleveland.data',
```

```
          col.names = c('age', 'sex', 'cp', 'trestbps', 'chol', 'fbs',
                'restecg', 'thalach', 'exang', 'oldpeak', 'slope',
                'ca', 'thal', 'target'))

download.file(paste('http://archive.ics.uci.edu/ml/machine-learning-databases/',
           'heart-disease/processed.va.data', sep = ''), 'processed.va.data')
data.4 <- read.csv('processed.cleveland.data',
          col.names = c('age', 'sex', 'cp', 'trestbps', 'chol', 'fbs',
                'restecg', 'thalach', 'exang', 'oldpeak', 'slope',
                'ca', 'thal', 'target'))

data <- rbind(data.1, data.2, data.3, data.4) %>%
    mutate(target = as_factor(ifelse(target == 0, 0, 1)),
          sex = as_factor(sex),
          cp = as_factor(cp),
          exang = as_factor(exang),
          slope = as_factor(slope),
          ca = as_factor(ca),
          thal = as_factor(thal))

levels(data$target) <- c('not_desease', 'desease')
levels(data$sex) <- c('female', 'male')
levels(data$cp) <- c('typical_angina', 'atypical_angina', 'non-anginal_pain',
            'asymptomatic')
levels(data$exang) <- c('no', 'yes')
levels(data$slope) <- c('upsloping', 'flat', 'downsloping')
levels(data$thal) <- c('?', 'normal', 'fixed_defect', 'reversable_defect')
```

Exploratory data analysis

In this dataset we have 302 observations (one for each patient) and 14 variables.

```
head(data)
```

```
##   age    sex               cp trestbps chol fbs restecg thalach exang
## 1  67   male     asymptomatic      160  286   0       2     108   yes
## 2  67   male     asymptomatic      120  229   0       2     129   yes
## 3  37   male non-anginal_pain      130  250   0       0     187    no
## 4  41 female  atypical_angina      130  204   0       2     172    no
## 5  56   male  atypical_angina      120  236   0       0     178    no
## 6  62 female     asymptomatic      140  268   0       2     160    no
##   oldpeak       slope  ca             thal      target
## 1     1.5        flat 3.0           normal     desease
## 2     2.6        flat 2.0 reversable_defect     desease
## 3     3.5 downsloping 0.0           normal not_desease
## 4     1.4   upsloping 0.0           normal not_desease
## 5     0.8   upsloping 0.0           normal not_desease
## 6     3.6 downsloping 2.0           normal     desease
```

```
str(data)
```

```
## 'data.frame':    1208 obs. of  14 variables:
##  $ age     : num  67 67 37 41 56 62 57 63 53 57 ...
##  $ sex     : Factor w/ 2 levels "female","male": 2 2 2 1 2 1 1 2 2 2 ...
```

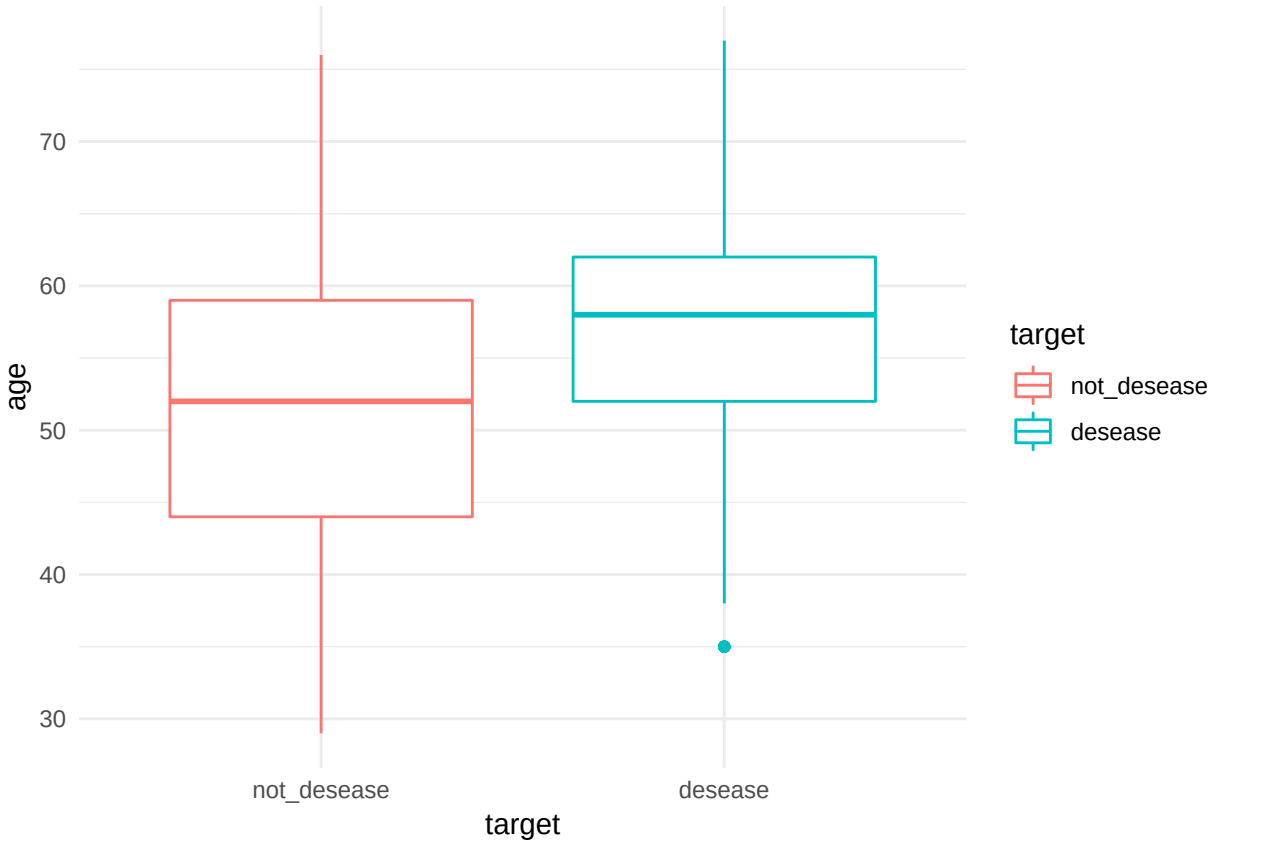
```
##  $ cp      : Factor w/ 4 levels "typical_angina",..: 4 4 3 2 2 4 4 4 4 4 ...
##  $ trestbps: num  160 120 130 130 120 140 120 130 140 140 ...
##  $ chol    : num  286 229 250 204 236 268 354 254 203 192 ...
##  $ fbs     : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : num  2 2 0 2 0 2 0 2 2 0 ...
##  $ thalach : num  108 129 187 172 178 160 163 147 155 148 ...
##  $ exang   : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 1 2 1 ...
##  $ oldpeak : num  1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 0.4 ...
##  $ slope   : Factor w/ 3 levels "upsloping","flat",..: 2 2 3 1 1 3 1 2 3 2 ...
##  $ ca      : Factor w/ 5 levels "?","0.0","1.0",..: 5 4 2 2 2 4 2 3 2 2 ...
##  $ thal    : Factor w/ 4 levels "?","normal","fixed_defect",..: 2 4 2 2 2 2 2 4 4 3 ...
##  $ target  : Factor w/ 2 levels "not_desease",..: 2 2 1 1 1 2 1 2 2 1 ...
```

To start we will analyse eacho of the 13 variables that will be used to explain the model and try to detect visualy patterns and variable that can be removed.

Age

The average age of pacients with heart disease is higher then pacients without a heart disease, this difference is not very big but can be used in the model.

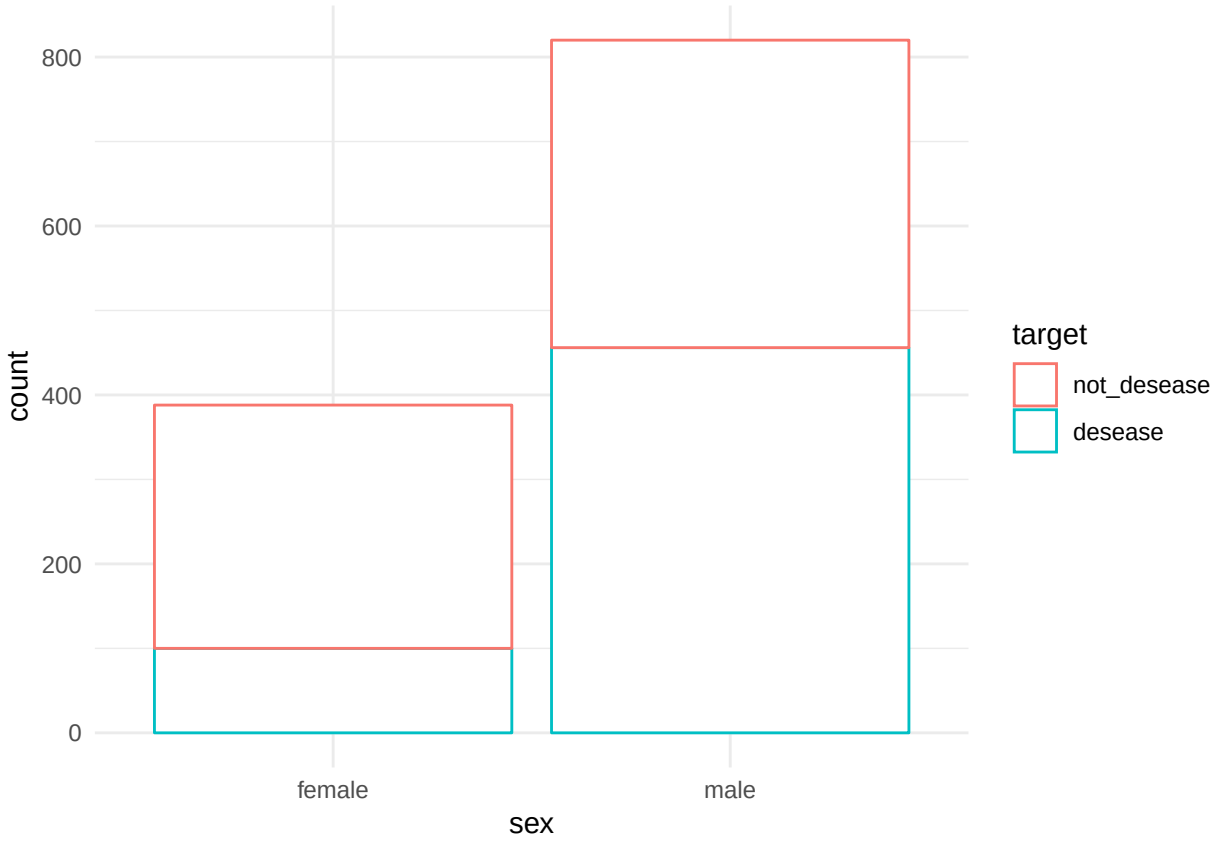
```
data %>% ggplot(aes(y=age, x=target, color=target)) + geom_boxplot() + theme_minimal()
```



Sex

Most of pacients in this dataset are male this could create an gender bias in the model, besides that the percentage of men with heart desease is higher then women.

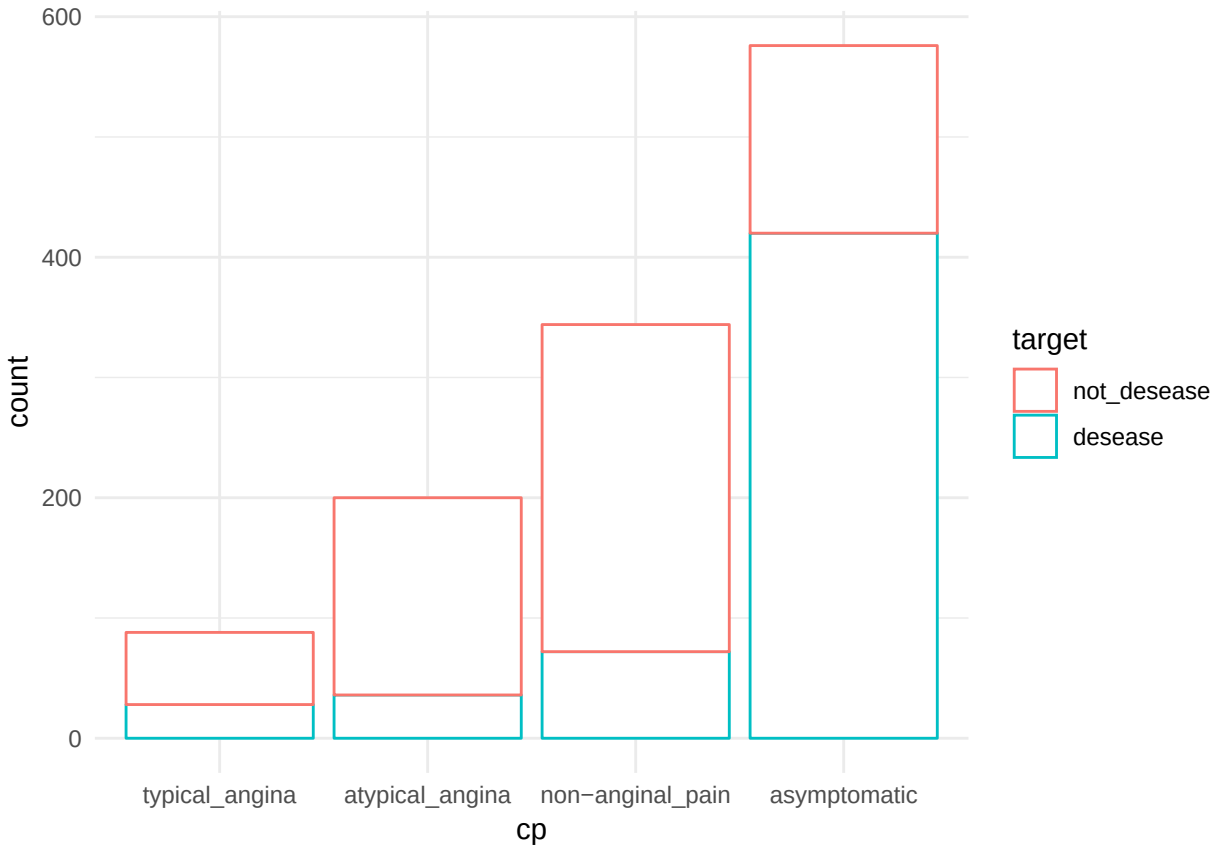
```
data %>% ggplot(aes(x=sex, color=target)) + geom_bar(fill='#FFFFFF') + theme_minimal()
```



Chest pain

Contrary to popular belief, in this dataset most of people with heart desease does not have anginal symptoms.

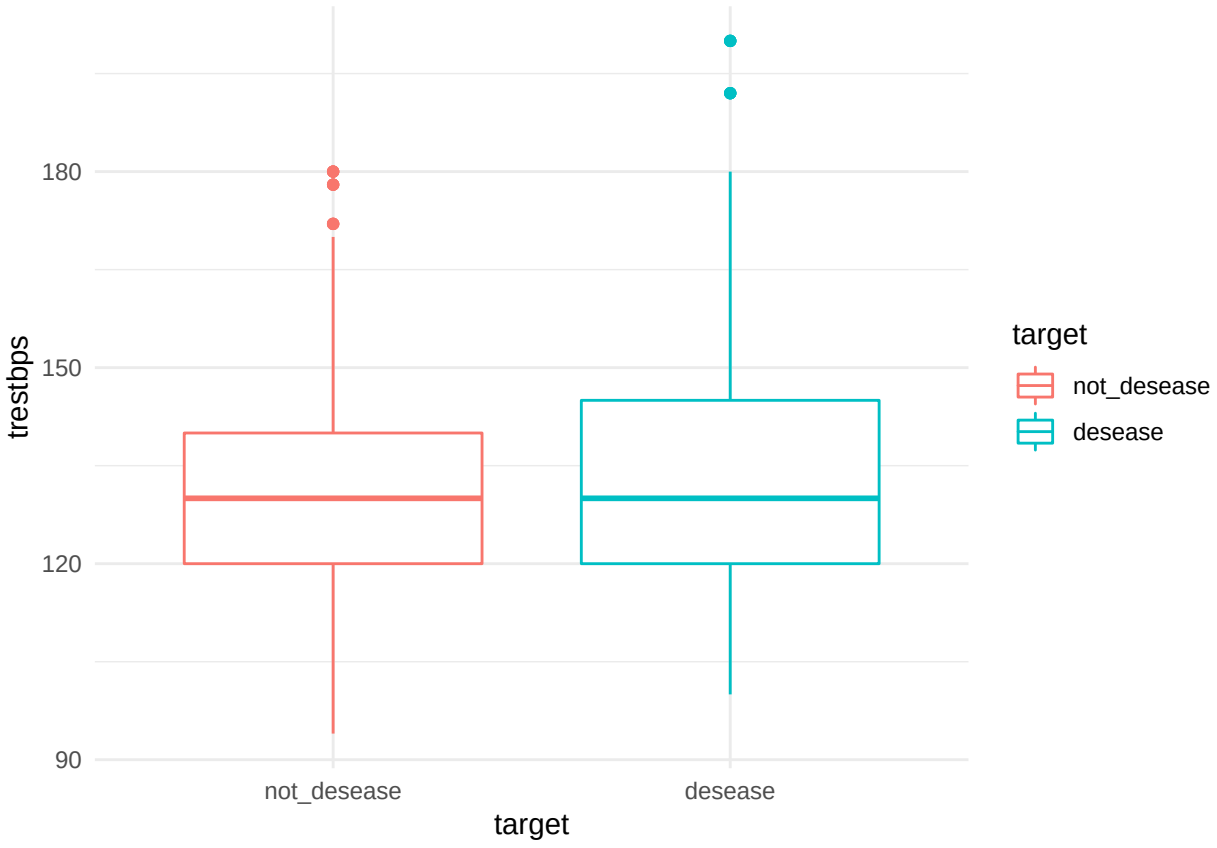
```
data %>% ggplot(aes(x=cp, color=target)) + geom_bar(fill='#FFFFFF') + theme_minimal()
```



Rest Blood Pressure

Both desease and health pacients have almost the same average rest blood pressure, so this variable would not be a good predctior of heart desease.

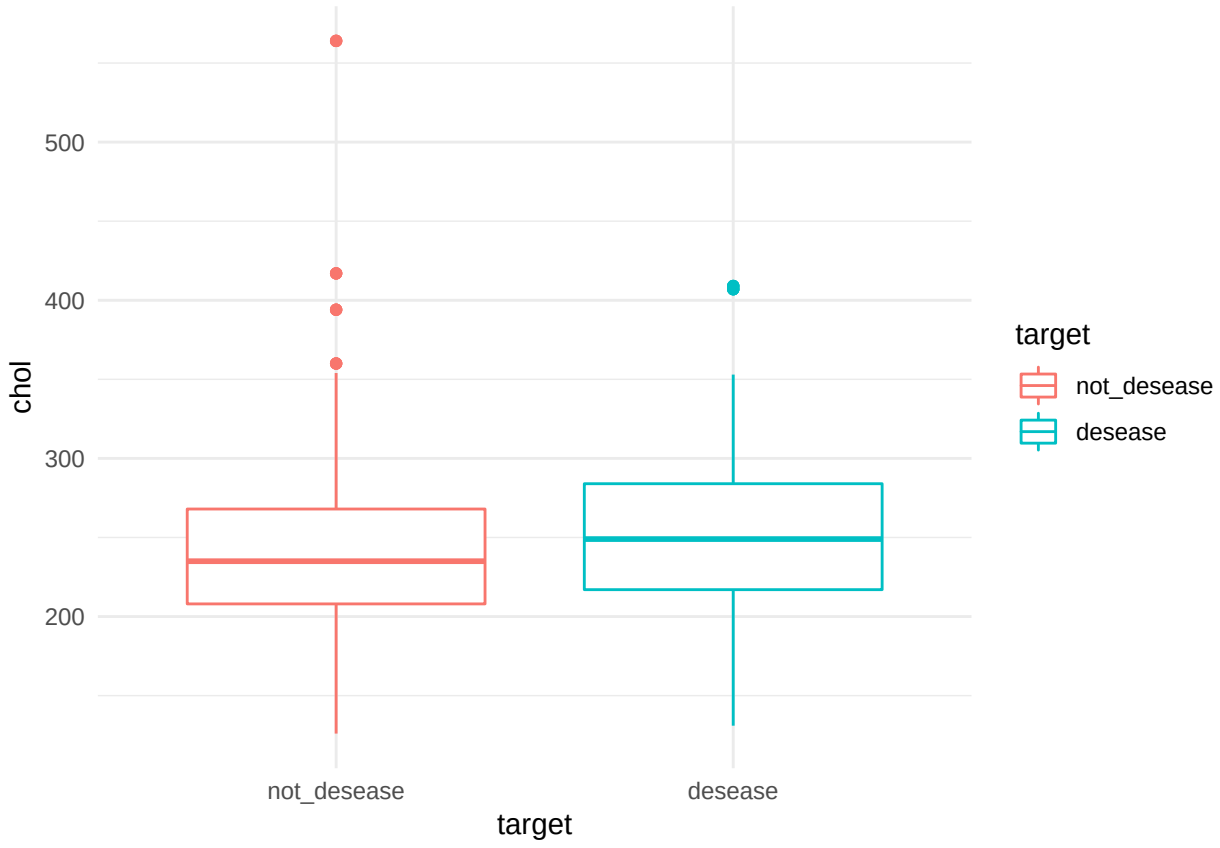
```
data %>% ggplot(aes(y=trestbps, x=target, color=target)) + geom_boxplot() + theme_minimal()
```



Serum cholestoral in mg/dl

As the rest blood pressure the Serum cholestoral probably will not be a good predictor, the values for desease and health pacients are almost the same.

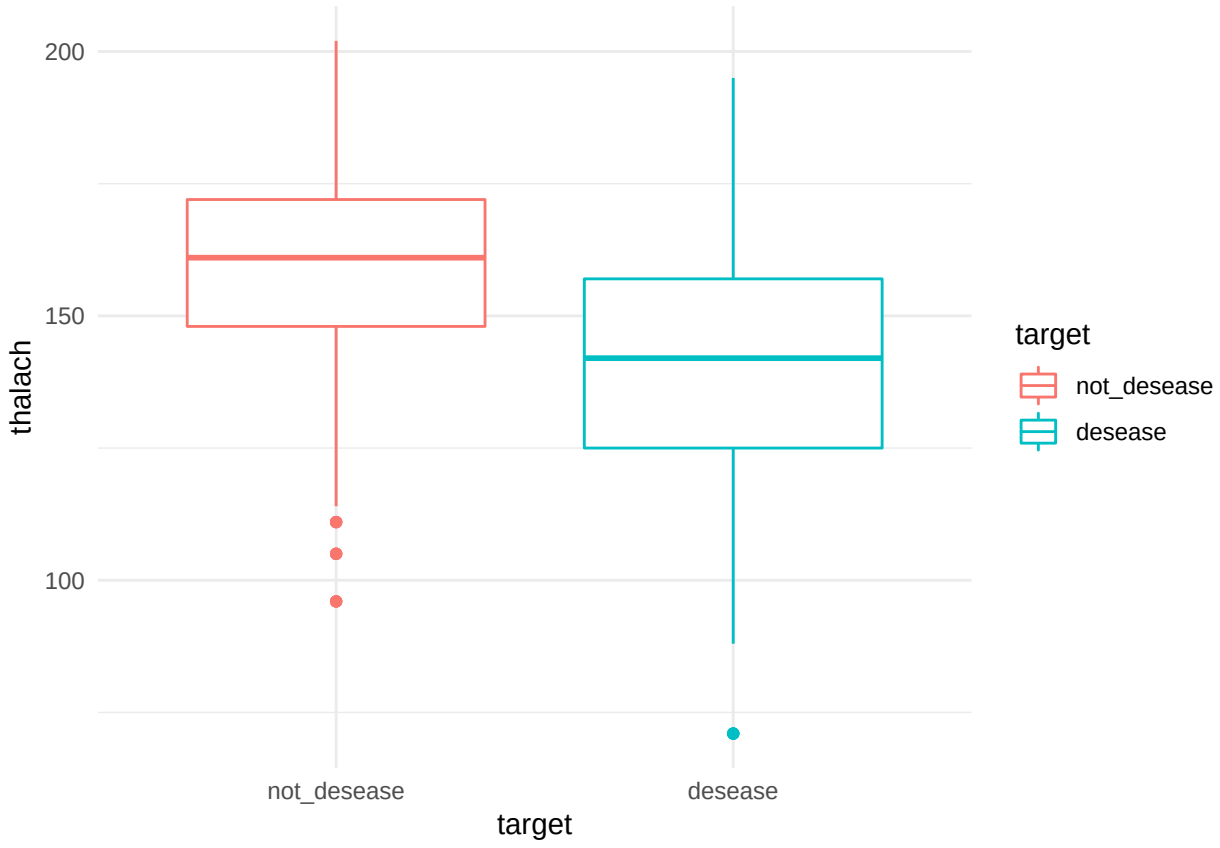
```
data %>% ggplot(aes(y=chol, x=target, color=target)) + geom_boxplot() + theme_minimal()
```



Maximum heart rate achieved

The maximum heart rate achived during the exames is a good predictor, there is a good separation between pacients with and without a heart desease.

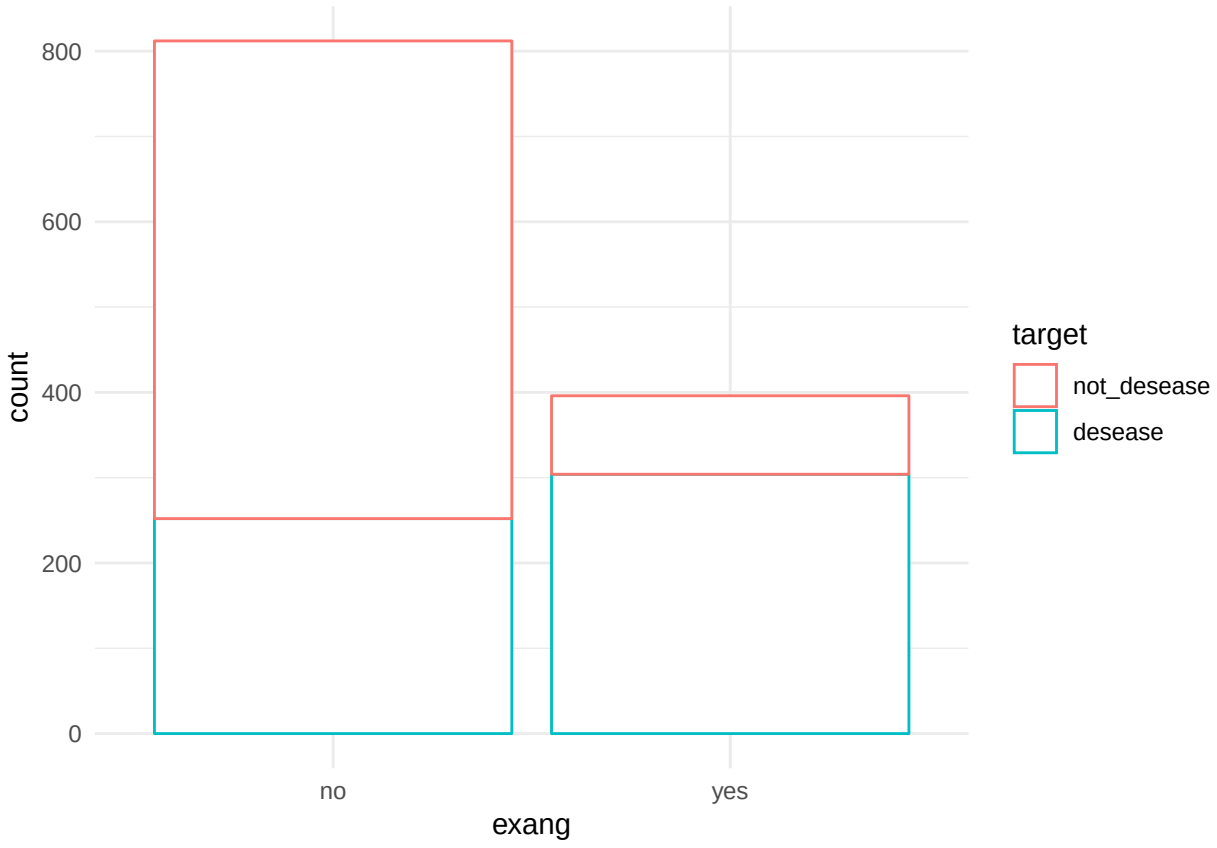
```
data %>% ggplot(aes(y=thalach, x=target, color=target)) + geom_boxplot() + theme_minimal()
```



Exercise induced angina

There is not differente in pacients with a heart desease in the angina induced by exercises, but on pacients without a heart desease this variable is significative.

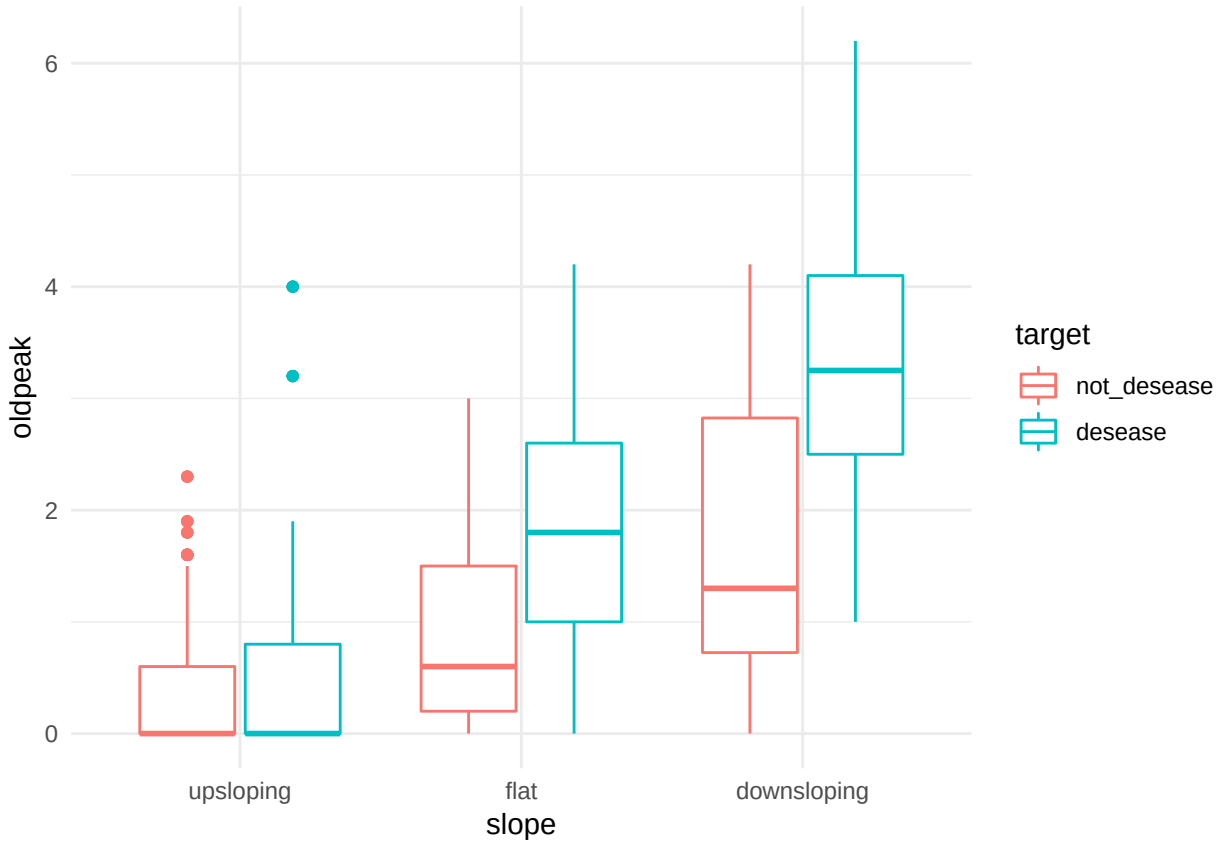
```
data %>% ggplot(aes(x=exang, color=target)) + geom_bar(fill='#FFFFFF') + theme_minimal()
```



ST depression induced by exercise relative to rest vs Slope of the peak exercise

Since both variable are correlate, the ST depression in the eletrocardiogram during the exercise and slope of the exercise, we plot together.
In the up sloping there is no difference between health and desease pacients, but with flat and down sloping the difference increase between health and desease pacients.

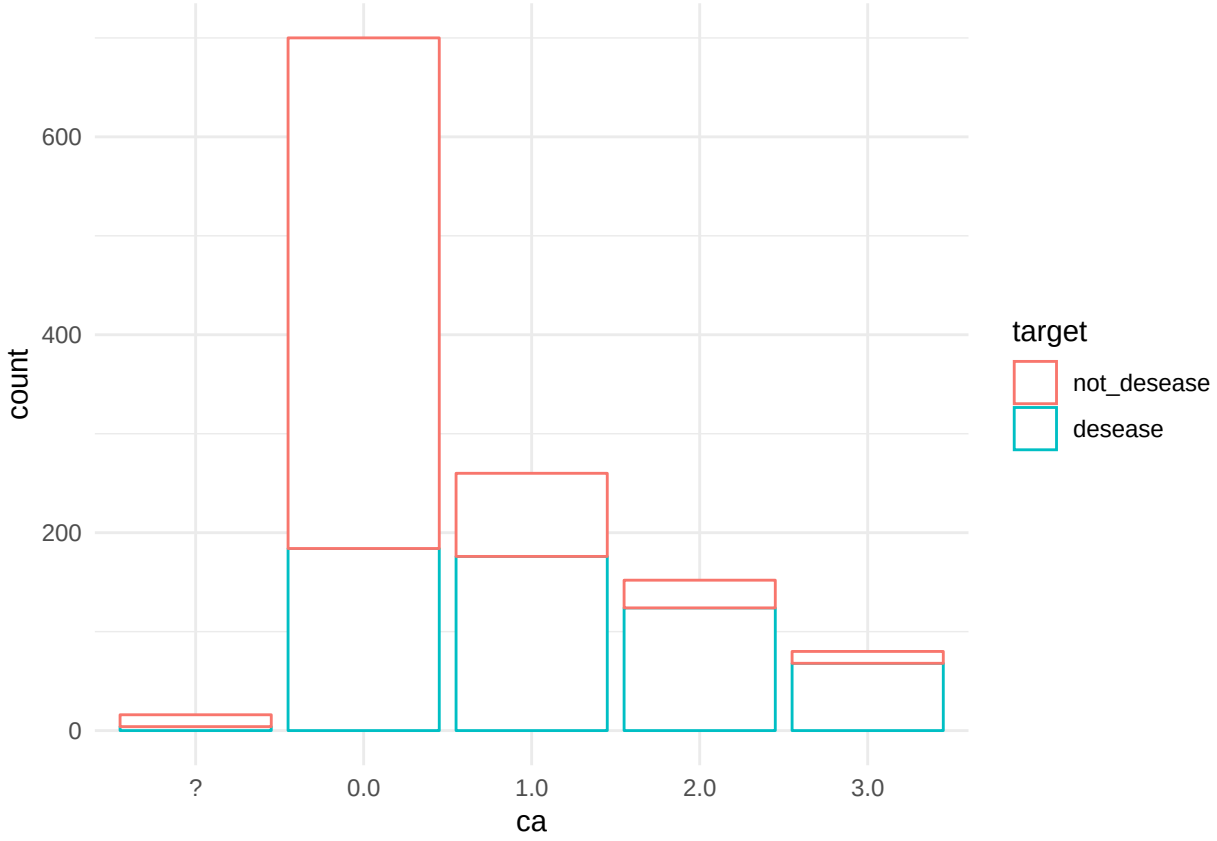
```
data %>% ggplot(aes(x=slope, y=oldpeak, color=target)) + geom_boxplot() + theme_minimal()
```



Number of major vessels colored by flourosopy

Most of health pacientes will have none of the major vessels colored during the flourosopy, pacients with heart desease does not have this pattern.

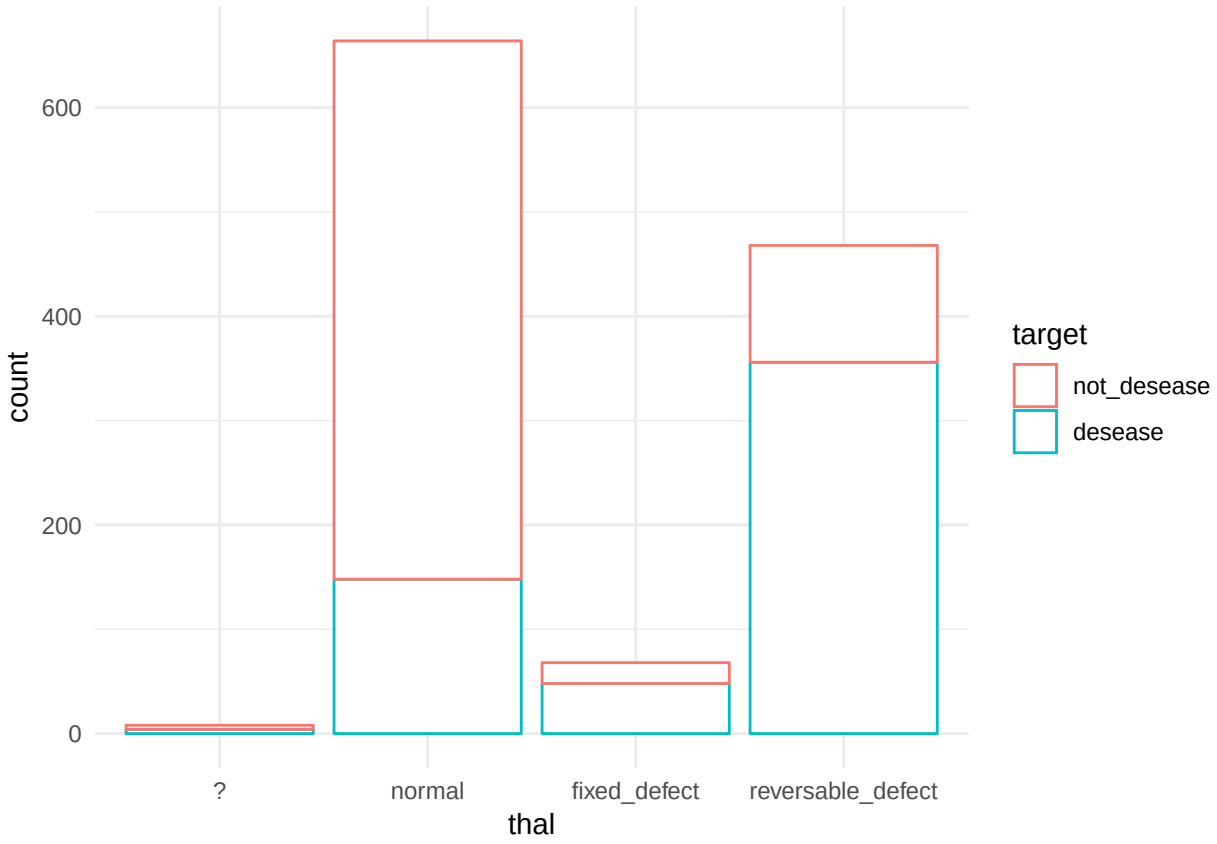
```
data %>% ggplot(aes(x=ca, color=target)) + geom_bar(fill='#FFFFFF') + theme_minimal()
```



Thal

The result of the Thal test show that health pacients have usually normal results and desease pacients have most of time a reversable defect.

```
data %>% ggplot(aes(x=thal, color=target)) + geom_bar(fill='#FFFFFF') + theme_minimal()
```



Data preaparation

Splitting the data

The model will always have a better performance in the training dataset that in the real world, to help to have a better idea of the real world performance we will split the data in 2 datasets, one for training and another for evaluation.

We will use 80% of the dataset to train the model and 20% to evaluate

```
set.seed(123)
train.index <- data$target %>% createDataPartition(p =0.8,list =F)
train.data <- data[train.index,]
test.data <- data[-train.index,]
```

Machine learning model

Random Forest

The random forest is a machine learning model that will test the data against multiple decision tree models tunning the parameters to find a better predictor.

For this model we will use 500 trees,

```
rf.model <- randomForest(target~., data=train.data, ntree=500)
```

With the model we will predict the testing dataset

```
rf.prediction <- predict(rf.model,newdata=test.data)
```

Results

For the testing dataset we have the following results:

```
confusion.matrix <- table(test.data$target, rf.prediction)
print(confusion.matrix)
```

```
##              rf.prediction
##               not_desease desease
##   not_desease         130       0
##   desease               0     111
```

```
print(100*sum(diag(confusion.matrix))/sum(confusion.matrix))
```

```
## [1] 100
```

With a accuracy of 100% we have a good model to predict pacients with heart desease using the 13 variables provided.