Heart Disease predicting model

Davide Mori

26/9/2019

HarvardX: PH125.9 - Data Science Professional Certificate: Capstone - "Choose Your Own Project Submission"

You can find this project on Github Website at: https://github.com/davidemori/Cardiac-Disease

Introduction

Of the 56.9 million deaths worldwide in 2016, more than half (54%) were due to the top 10 causes. Ischaemic heart disease and stroke are the world's biggest killers, accounting for a combined 15.2 million deaths in 2016. These diseases have remained the leading causes of death globally in the last 15 years(World Health Organization, 2019). The aim of this project is to analyze the causes and try to find a model to predict, on the basis of some parameters, the occurence of an heart disease.

Many datasets about Medical Sciences, and in particular about heart diseases, are free downloadable on the "WEB". For the following project we have used a subset of the "Cleveland Heart Disease Dataset" from the UCI Archives, that can be found at the following link: http://archive.ics.uci.edu/ml/datasets/heart+disease .

On this dataset we'll try to perform some exploratory analyses, and test a total of six models (5 + 1 ensemble model) to find the best fit for prediction of the heart disease in an adult population.

Methods

cleaning and process the Dataset

The first goal of our cleaning and tidying process is to download the dataset and define the names of the columns. Infact, despite the dataset is provided with a partially tidy format, the names of the variables are missing form the principal file, and hence we have to extrapolate them from another document. At the end of the process we'll have the following variables:

- Age: age in years
- Sex: (1 = male; 0 = female)
- Chest.Pain: chest pain type – Value 1: typical angina – Value 2: atypical angina – Value 3: non-anginal pain Value 4: Asymptomatic
- BP.Rest: resting blood pressure (in mm Hg on admission to the hospital)
- Chol.Liv: serum cholestoral in mg/dl
- Fast.Blood.Sugar: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- ECG.Rest: resting electrocardiographic results: Value 0: normal
  Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- HR.Max: maximum heart rate achieved
- Angina.post.Exercise: exercise induced angina (1 = yes; 0 = no)
- OldPeak.ST: ST depression induced by exercise relative to rest
- Slope.ST: the slope of the peak exercise ST segment
- Vessels.Fluo:number of major vessels (0-3) colored by flourosopy

- Defect:3 = normal; 6 = fixed defect; 7 = reversable defect
- Disease: disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

```
#Load necessary libraries
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
if(!require(foreign)) install.packages("foreign")
```

```
## Loading required package: foreign
```

```
if(!require(GGally))install.packages("GGally")
```

```
## Loading required package: GGally
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
#Downloading and checking dataset
data<-read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cle

#adding column names
colnames(data) <- c("Age", "Sex", "Chest.Pain", "BP.Rest", "Chol.Liv", "Fast.Blood.Sugar", "ECG.Rest",
                    "Angina.post.Exercise", "OldPeak.ST", "Slope.ST", "Vessels.Fluo", "Defect", "Disease"
```

After the first data adjustment we have to check the structure of oure datas and check for any NAs (missing values):

```
#Check for changes after renaming columns, and check for NAs
str(data)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ Age                 : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ Sex                 : num  1 1 1 1 0 1 0 0 1 1 ...
##  $ Chest.Pain          : num  1 4 4 3 2 2 4 4 4 4 ...
##  $ BP.Rest             : num  145 160 120 130 130 120 140 120 130 140 ...
##  $ Chol.Liv            : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ Fast.Blood.Sugar    : num  1 0 0 0 0 0 0 0 0 1 ...
##  $ ECG.Rest            : num  2 2 2 0 2 0 2 0 2 2 ...
##  $ HR.Max              : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ Angina.post.Exercise: num  0 1 1 0 0 0 0 1 0 1 ...
##  $ OldPeak.ST          : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ Slope.ST            : num  3 2 2 3 1 1 3 1 2 3 ...
##  $ Vessels.Fluo        : Factor w/ 5 levels "?","0.0","1.0",..: 2 5 4 2 2 2 4 2 3 2 ...
##  $ Defect              : Factor w/ 4 levels "?","3.0","6.0",..: 3 2 4 2 2 2 2 2 4 4 ...
##  $ Disease             : int  0 2 1 0 0 0 3 0 2 1 ...
```

```r
#Check for NAs
apply(is.na(data), 2,which)
```

```
## integer(0)
```

Despite after check there wasn't any NAs, we have seen that some variables was coded by the question mark "?", so few entries are missed. For thi reason we procede to remove them all by filtering dataset.

```r
data <- data %>% filter_all(all_vars(.!="?"))
```

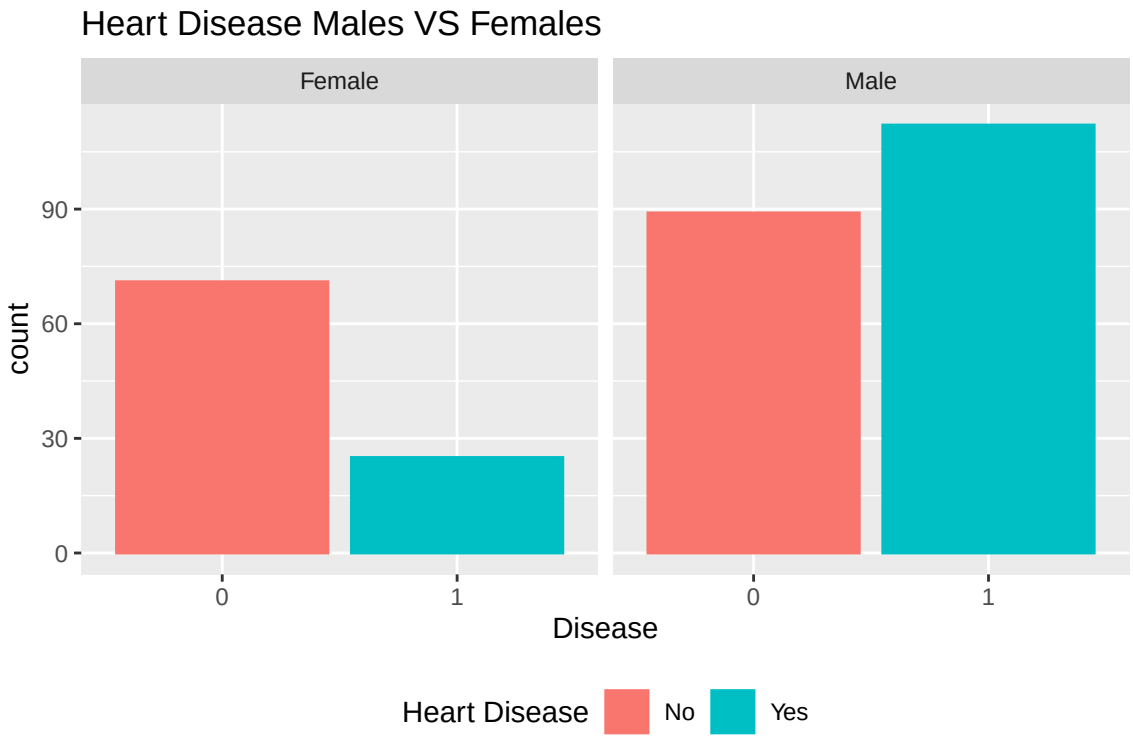
The dataset contains the column `disease`, that is represented by a range of five values 0 to 4, where 0 means "No disease" and the other 4 values are the level of disease progression. Since also a a value of 1 indicates heart disease, we'll convert it to a binary variable where 0 is kept if no disease are present, and 1 otherwise. Additionally, some data loaded as numeric are converted to factor for better analysis.

```r
#Now we have removed 6 entries, but due to the pregress "?", despite his numeric nature, "Vessels.Flu"
#was loaded as Factor. We have to convert it in a numerical variable and remove two values in accord wi

data$Vessels.Fluo<-as.numeric(data$Vessels.Fluo) -2

#For the same reason ("?") "Defect" variable present four levels instead three. We have to correct it,
#other categorical/binary variables that was mistaken read as numeric from the file.

data$Defect<-factor(data$Defect)
data$Sex<-factor(data$Sex)
data$Chest.Pain<-factor(data$Chest.Pain)
data$Fast.Blood.Sugar<-as.factor(data$Fast.Blood.Sugar)
data$ECG.Rest<-as.factor(data$ECG.Rest)
data$Angina.post.Exercise<-as.factor(data$Angina.post.Exercise)
data$Slope.ST<-as.factor(data$Slope.ST)

#For further analyses we also convert Disease variable in binary s reason, where 0 means "No disease",

data<-data%>%mutate(Disease=as.factor(ifelse(Disease==0,0,1)))
```

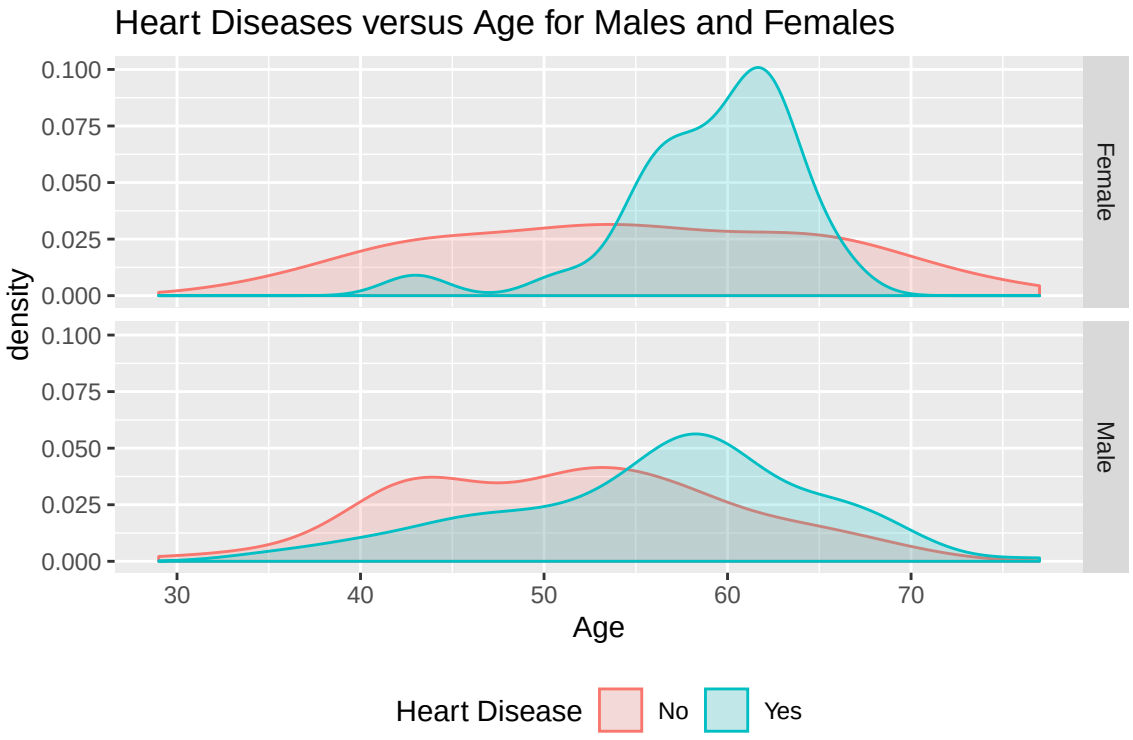
Once we have finally cleaned our dataset, we are ready to starting the explorative analysis.

Data Exploration and testing correlations

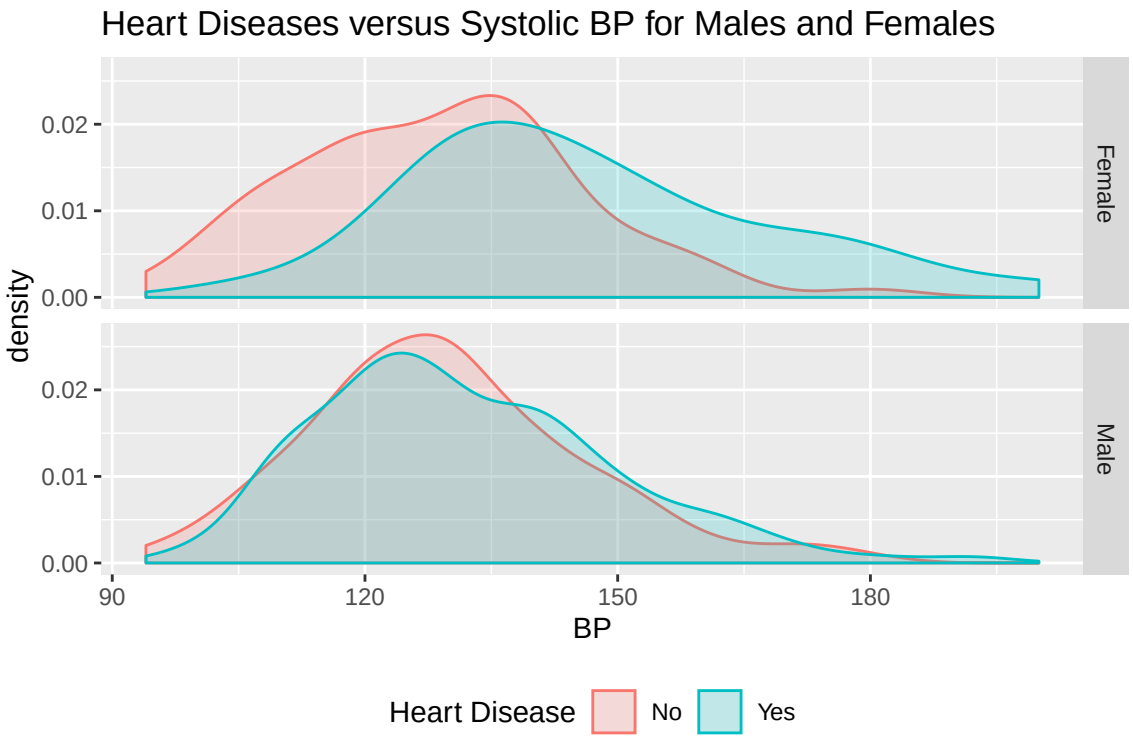
To check our data, will perform a series of density plot and bar plot to better exploring the nature of our independent variables compared to the dependent one (`disease`). Since epidemiologically it is known that the male and the female have a different distribution of heart diseases, we will plot all the variables comparing male and female sex.



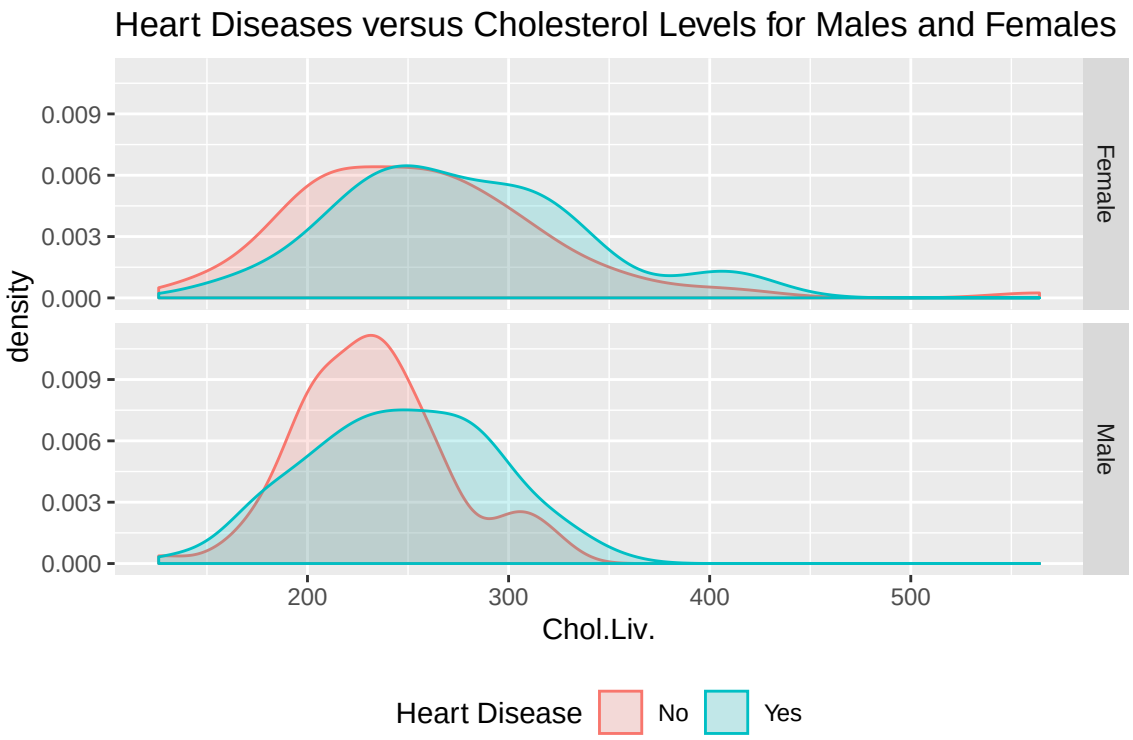
How is possible to see from the first barplot, males have a bigger incidence and a bigger prevalence of heart disease than females.



Not only the sex represent a discriminant variable for the occurrence of heart disease, age too has an important effect on it. Males and females, as is shown in the previous density plot, are both suffer a greater incidence of heart disease between the age of 50 and 70 years old.



If we consider the systolic BP we can observe that there aren't great difference in terms of presence of disease, just in case of female sex, we can note a slightly increasing of incidence with a value of 150 mmHg or greater.



Cholesterol levels are also seeming a non-discriminating variable, although we can observe a little modification in terms of incidence with values greaters of 250.