Heart disease analysis

Lorenzo Vázquez

2023-02-22

Introduction

This dataset contains detailed information on the risk factors for cardiovascular disease. It includes information on age, gender, height, weight, blood pressure values, cholesterol levels, glucose levels, smoking habits and alcohol consumption of over 70 thousand individuals. Additionally it outlines if the person is active or not and if he or she has any cardiovascular diseases.

Source: https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas

In the dataset are:

- Numeric values for age, height, weight, systolic blood pressure (ap_hi) and diastolic blood pressure (ap_lo).
- Binary values for gender, alcohol consumption (alco), smoking habits (smoke), active person (active), cardiovascular diseases (cardio). In the gender variable it's not defined which value correspond to which gender. In the rest of the cases 0 = No and 1 = Yes.
- Levels in the cholesterol and glucose (gluc) variables. In this cases I will consider that 1 = normal, 2 = above normal, 3 = well above normal.

Hypothesis to be tested

- The lifestyle of the people impact on the risk of having cardiovascular diseases.
- There isn't a correlation between gender and having cardiovascular diseases
- High levels of glucose and cholesterol contributes to developing cardiovascular diseases.
- Older people have more risk to have cardiovascular diseases.
- People with higher body mass index have more risks to develop cardiovascular diseases.
- People with higher blood pressure have more risks to develop cardiovascular diseases.

Analysis

Loading libraries and dataset

```r
heart_data <- read.csv("~/archivos data analytics/Project 1 - Heart disease/heart_data.csv")

library(tidyverse)
library(skimr)
library(ggpmisc)
library(cowplot)
library(caret)
```

Preview of the dataset

```r
head(heart_data)
```

```
##   index id   age gender height weight ap_hi ap_lo cholesterol gluc smoke alco
## 1     0  0 18393      2    168     62   110    80           1    1     0    0
## 2     1  1 20228      1    156     85   140    90           3    1     0    0
## 3     2  2 18857      1    165     64   130    70           3    1     0    0
## 4     3  3 17623      2    169     82   150   100           1    1     0    0
## 5     4  4 17474      1    156     56   100    60           1    1     0    0
## 6     5  8 21914      1    151     67   120    80           2    2     0    0
##   active cardio
## 1      1      0
## 2      1      1
## 3      0      1
## 4      1      1
## 5      0      0
## 6      0      0
```

Complete summary of the dataset

```r
skim_without_charts(heart_data)
```

Data summary

Name	heart_data
Number of rows	70000
Number of columns	14
_______________________	
Column type frequency:	
numeric	14
________________________	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
index	0	1	34999.50	20207.40	0	17499.75	34999.5	52499.25	69999
id	0	1	49972.42	28851.30	0	25006.75	50001.5	74889.25	99999
age	0	1	19468.87	2467.25	10798	17664.00	19703.0	21327.00	23713
gender	0	1	1.35	0.48	1	1.00	1.0	2.00	2
height	0	1	164.36	8.21	55	159.00	165.0	170.00	250
weight	0	1	74.21	14.40	10	65.00	72.0	82.00	200
ap_hi	0	1	128.82	154.01	-150	120.00	120.0	140.00	16020
ap_lo	0	1	96.63	188.47	-70	80.00	80.0	90.00	11000
cholesterol	0	1	1.37	0.68	1	1.00	1.0	2.00	3
gluc	0	1	1.23	0.57	1	1.00	1.0	1.00	3
smoke	0	1	0.09	0.28	0	0.00	0.0	0.00	1
alco	0	1	0.05	0.23	0	0.00	0.0	0.00	1
active	0	1	0.80	0.40	0	1.00	1.0	1.00	1
cardio	0	1	0.50	0.50	0	0.00	0.0	1.00	1

Looking for duplicates

```r
length(unique(heart_data$id))
```

```
## [1] 70000
```

There are 14 columns and 70000 rows, 0 NAs, and 0 duplicates in the dataset

It can be seen a lot of inconsistency in the data:

- The ap_hi and ap_lo columns have values that are biologically impossible.
- In the weight and height column there are also weird values.

In these measured variables there may be errors when recording them. So when analyzing those variables, i will remove the outliers applying the IQR method (1). In that range, the wrong data will be deleted, and the analyzed data will be large enough to be representative of the entire population.

Data processing

I've made some processing in the data:

1. Turn age from days to years.
2. Remove outliers from the age variable since there are only four values corresponding to the age of 30, the rest are in the range of 39-65 years old.
3. Change 0, 1 code in smoke, alco, active, and cardio. E.g "Smoker", "No smoker".
4. Change cholesterol, gluc and gender columns to factor.
5. Calculate body mass index (BMI). Formula = weight(kg)/height(m)^2). And classificate those bmi values according to the World Health Organization classification (2):
   - <18.5 underweight (uw)
   - 18.5-24.9 normal weight (normal)
   - 25.0 - 29.9 pre-obesity (pre-ob)
   - 30.0 - 34.9 obesity class 1 (ob 1)
   - 35.0 - 39.9 obesity class 2 (ob 2)
   - ≥ 40 obesity class 3 (ob 3)
6. Calculate mean arterial pressure (MAP). Formula = (ap_hi+ap_lo*2)/3. And classificate those map values according to the American Hearth Association (2020) (3):
   - <90 Normal (normal)
   - 90 to 91.99 Elevated blood pressure (high-bp)
   - 92 to 95.99 Hypertension stage 1 (hyp1)
   - ≥ 96 Hypertension stage 2 (hyp2)
7. Selecting desired columns

And this are the first ten rows of the dataset I will work with:

```r
head(heart_data)
```

```
##   Age_years Gender      Bmi Bmi_class       Map Map_class Gluc Cholesterol
## 1        50      2 21.96712    normal  90.00000   high-bp    1           1
## 2        55      1 34.92768      ob 2 106.66667      hyp2    1           3
## 3        52      1 23.50781    normal  90.00000   high-bp    1           3
## 4        48      2 28.71048    pre-ob 116.66667      hyp2    1           1
## 5        48      1 23.01118    normal  73.33333    normal    1           1
## 6        60      1 29.38468    pre-ob  93.33333      hyp1    2           2
##       Smoke    Alco    Active     Cardio
## 1 No smoker No alco    Active No disease
## 2 No smoker No alco    Active    Disease
## 3 No smoker No alco No active    Disease
## 4 No smoker No alco    Active    Disease
## 5 No smoker No alco No active No disease
## 6 No smoker No alco No active No disease
```

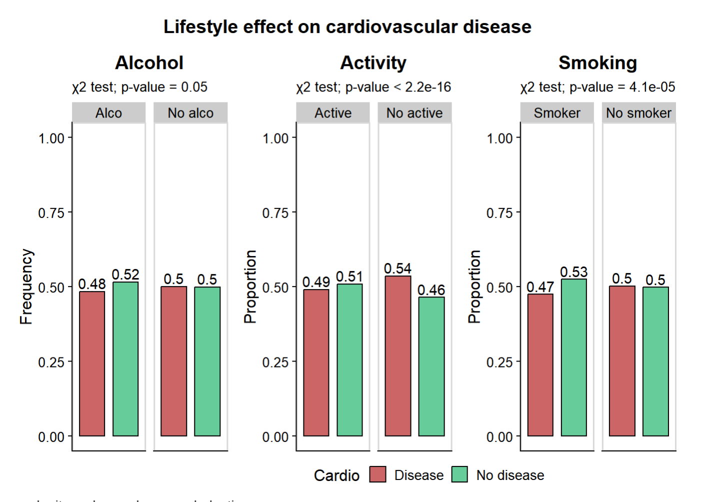
So it will be analyzed the risk factors for developing cardiovascular diseases in adult people from 39 to 65 years old.

Analysis

Lifestyle analisis: smoking, drinking and activity.

First I'm gonna analyze the relationship of lifestyles (smoking, drinking alcohol and being active) and the development of cardiovascular diseases.

- In each case a Chisquare test was made to determine if there are statistical differences between conditions (e.g smoking or not smoking).
- In all cases H0 = no differences between conditions; H1 = differences between conditions.

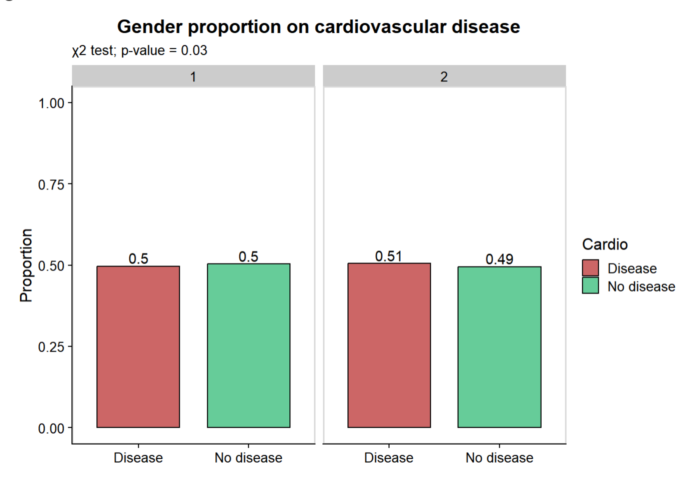


From the previous graphs it can be made some deductions:

- The lifestyle of the adult people between 39 and 65 years, contributes to developing or not a cardiovascular disease.
- Being an active person it's fundamental to minimize the risk of developing a cardiovascular disease.
- In the people who drink alcohol, there aren't significant differences.
- Analyzing the smoking people, statically speaking, there are significant differences (p-value << 0.05), and in the group who smoke there are less people with cardiovascular disease. But this is erroneous and it's related to a small sample size. A lot of bibliography says the opposite (4).
- More detail in the information is needed to be more precisely in the impact of lifestyles in the development of a cardiovascular disease, e.g: the amount of alcohol drinking per day, amount of cigarettes per day, etc.

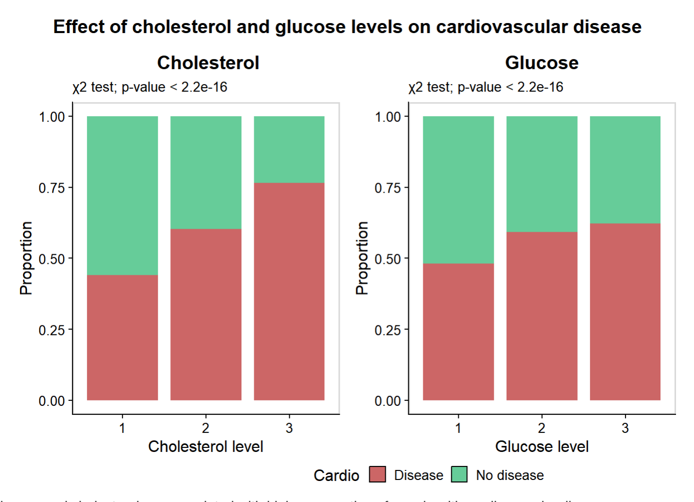
Next I'm gonna analyze if there is a relationship between the personal characteristics of the people (age, gender, cholesterol, gluc, map ,bmi) and cardiovascular diseases

Analyzing gender



It seems that there isn't difference between gender and the develop of cardiovascular disease, the p-value of the chisquare test is near to 0.05 so it cannot reject H0.

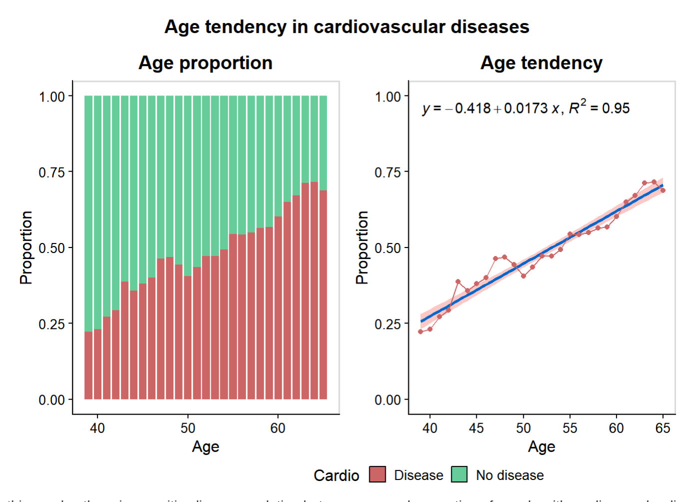
Now analyzing cholesterol and glucose levels



Higher levels of glucose and cholesterol are associated with higher proportion of people with cardiovascular diseases:

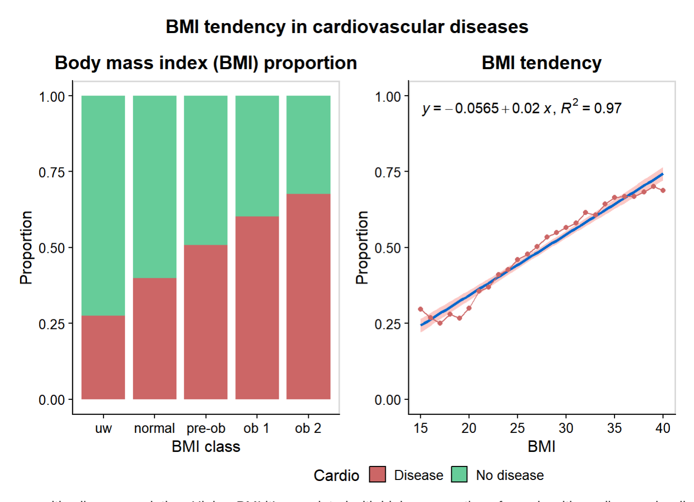
- Having "well above normal" levels of cholesterol it's highly correlated with a high proportion of people with cardiovascular diseases (77%).
- Having "above normal" and "well above normal" levels of glucose have similar proportions of people with cardiovascular diseases (59% and 62% respectively).

Relationship between age and cardiovascular diseases



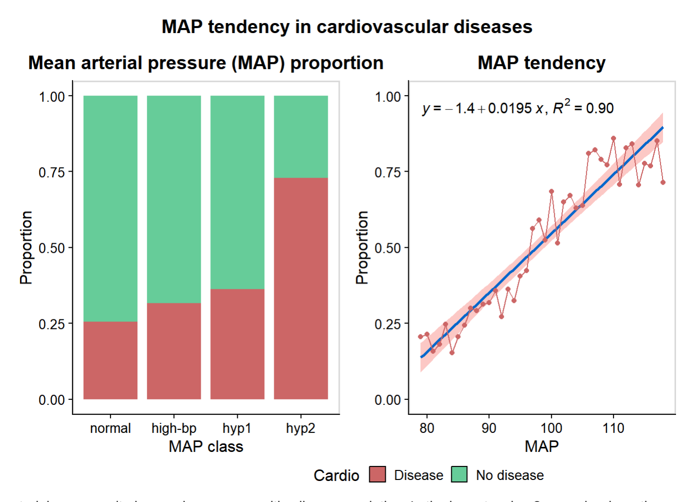
As can be seen in this graphs, there is a positive linear correlation between age and proportion of people with cardiovascular diseases. In older people, the proportion of people with cardiovascular disease is much higher than in the youngest: in the range from 60 to 65 years, the proportion of people with cardiovascular diseases is around 70%, meanwhile in the youngest group from 39 to 45 years old the proportion range of people with cardiovascular diseases is between 22 and 40%.

Now working with the BMI variable



Here it can be seen a positive linear correlation. Higher BMI it's correlated with higher proportion of people with cardiovascular diseases.

Analyzing MAP



Viewing the mean arterial pressure, it also can be seen a positive linear correlation. In the hypertension 2 group is where the proportion of people with cardiovascular diseases is really higher (73%).

So these three measurable variables (age, BMI and MAP) can help to determine the probability of getting a cardiovascular disease.

Logistic regression model

Now I will see if a logistic regression model with that 3 measurable variables can help to predict the risk of getting cardiovascular disease.

```r
#removing the outliers for both bmi, map
no.outlier <- IQRmethod(heart_data, heart_data$Map)
no.outlier <- IQRmethod(no.outlier, no.outlier$Bmi)

#Creating boolean column
no.outlier$Boolean <- no.outlier$Cardio=="Disease"


# Split the data into training and test set
set.seed(123)
training.samples <- no.outlier$Boolean %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- no.outlier[training.samples, ]
test.data <- no.outlier[-training.samples, ]

 #Analyzing the glm model
glmMod <- glm(Boolean ~ Bmi + Map + Age_years,
              data = train.data,
              family = "binomial")

summary(glmMod) #Here it can be seen that all variables are significant when explaining the probability of gettin
g a heart disease.
```

```
##
## Call:
## glm(formula = Boolean ~ Bmi + Map + Age_years, family = "binomial",
##     data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5236  -0.9744  -0.4589   1.0058   2.3956
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -14.106442   0.158937  -88.75   <2e-16 ***
## Bmi           0.039556   0.002386   16.58   <2e-16 ***
## Map           0.102305   0.001389   73.67   <2e-16 ***
## Age_years     0.059249   0.001553   38.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68522  on 49460  degrees of freedom
## Residual deviance: 57938  on 49457  degrees of freedom
## AIC: 57946
##
## Number of Fisher Scoring iterations: 3
```

The model adjusts really well. All variables are significant.

Now testing the model accuracy

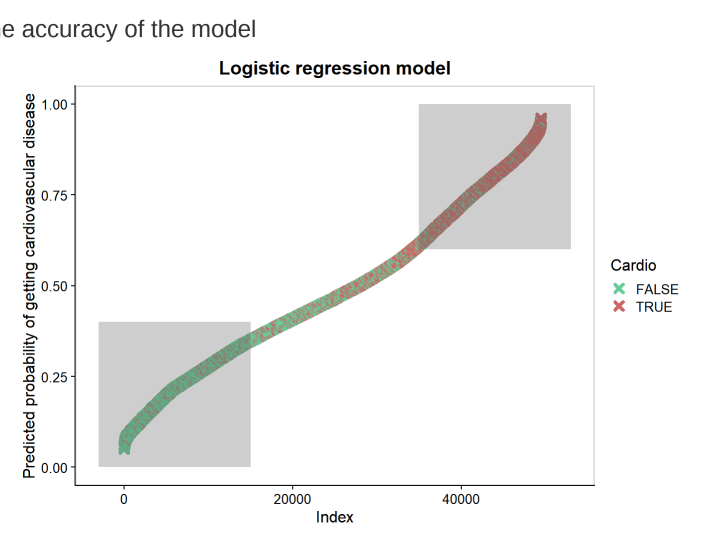
```r
# Make predictions in the test data
probabilities <- glmMod %>% predict(test.data, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "TRUE", "FALSE")

# Model accuracy
mean(predicted.classes == test.data$Boolean)
```

```
## [1] 0.7041411
```

```r
#Model accuracy ~70%
```

Plotting the accuracy of the model



In the lowest gray area, most of the people don't have heart disease, and in the upper gray area most of the people have cardiovascular disease. So here it's represented the 70% of accuracy of the model

Final conclusions

From these dataset it can be concluded the next things for adults between 39 and 65 years old:

- Lifestyle: being an active person reduces the probabilities of having cardiovascular diseases.
- Gender: there isn't difference between genders and risk of having cardiovascular diseases.
- Cholesterol and glucose levels: higher cholesterol and glucose levels are highly correlated with cardiovascular diseases.
- Age: Age and proportion of people with cardiovascular disease are highly correlated. In older people (from 60 to 65 years old) the risk is much higher than in the less-older ones (from 39 to 45 years old).
- BMI: positive correlation between BMI and proportion of people with cardiovascular disease. Obesity class 2 group is the most risky one.
- MAP: poositive correlation between MAP and proportion of people with cardiovascular disease. Hypertension stage 2 group is the most risky one.
- Predictive model: a logistic regression model with the age, BMI and MAP variables was made. This model has a 70% of accuracy when predicting the probability of having cardiovascular diseases.

References

- 1: https://online.stat.psu.edu/stat200/lesson/3/3.2
- 2: https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations
- 3: https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.120.14929
- 4: https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-cardiovascular-disease