Heart Disease – Diagnostic marker in the cohort

Miguel González González

miguel.gonzalez@estudiants.urv.cat

Johanna Ursula Albers

johanna.albers@estudiants.urv.cat

César Merino Fidalgo

cesar.merino@estudiants.urv.cat

Data used:

CardioHealth Risk Assesment Dataset

The CardioHealth Risk Assessment Dataset includes detailed medical and demographic information from patients, such as age, cholesterol levels, blood pressure, and lifestyle factors. This dataset is designed for developing and testing machine learning models to predict the risk of heart disease. It provides a valuable resource for researchers and healthcare professionals aiming to improve diagnostic accuracy and patient outcomes in cardiovascular health.

https://www.kaggle.com/datasets/kapoorprakhar/cardio-health-risk-assessment-dataset

Data available under MIT license with free access and download on Kaggle repository.

Exploratory analysis:

To obtain the different results, some data processing has been needed. The dataset contains a total of 270 rows and 14 different variables. These variables were numerical, categorical and categorical ordinal, with the main feature being 'Heart disease', indicating whether the patient was or was not affected. The rest of the variables were:

- Independent of lifestyle: such as sex and age
- Dependent of lifestyle: such as Cholesterol or Blood Pressure
- Dependent and measured after exercise: a group of variables were taken after completing some exercise to see the response of the body. These include the maximum heart rate or the ST slope, related to the outcome of an ECG.

There were no duplicated rows or missing values, but one of the features, "Chest pain type" didn't have any information about what it was referring to. Therefore, the column has been deleted to only work with interpretable variables.

The coded values of each feature have been properly changed to its corresponding meaning, so the work is done with understandable data without the need of looking up each code.

The distribution of the 'Heart disease' categories is unbalanced, having 120 patients with disease and 150 without it. Since the dataset isn't very large and difference between groups isn't big, we won't try to balance it to have all the information available and work with the whole dataset.

Some of the features such as Cholesterol have extreme values, for example over 500, which might be possible in severe cases, but as there is only one case and we will work with common values, it will be treated as an outlier or error and deleted from the dataset. This will result on more normal and centered distributions in the features with outliers 'Cholesterol', 'ST depression' and 'BP'.

Looking into the difference of values in numerical features by the presence of 'Heart disease', all the differences seem to be significant under alpha=0.1, but the visual differences seem to be significant in 'Max HR', 'ST Depression' and to a lesser extent in age. On the other hand, looking in the distribution of the Heart Disease among the values of the categorical features, there seem to be significant differences in all of them but 'FBS over 120'.

These results gave the idea that the exercise related measures can help to differenciate better between people with and without the disease. To take a deeper look into this fact, we made bivariate plots with the features which seemed to have the biggest difference. This resulted in seeing that plotting 'Max HR' against 'ST Depression' with color being 'heart disease' could help to detect the presence of the disease, as the healthy ones tend to have higher maximum heart rate and lesser 'ST depression'.

Message:

After checking the different plots and statistics from the exploratory analysis, we have come to the next conclusion we want to transmit:

Functional stress-test variables show a much stronger separation between patients with and without heart disease than traditional resting indicators such as cholesterol.

This information is intended for clinicians specialized in cardiovascular health, aiming to assist in the early detection and prevention of heart disease.

Tools:

- Exploratory analysis:

The exploratory analysis was initially than in both Python and R, although the final formal exploratory analysis was performed with Python in a Jupyter Notebook with the following packages and versions:

```
Python implementation: CPython
Python version       : 3.11.8
IPython version      : 8.20.0

pandas    : 2.2.3
numpy     : 1.26.4
seaborn   : 0.13.2
matplotlib: 3.8.0
scipy     : 1.11.4
```

- Explanatory analysis:

The explanatory analysis with the final visualization were created using Microsoft Power BI Desktop, with the following version:

2.137.751.0 (October 2024)