Sapienza
Università di Roma

Visual Analytics

Heart disease: representation & prediction

Master of Science in Engineering in Computer Science

A. A. 2021/2022

Presented by:

Elisa Berti 1716412
Marian Leonard Mentel 1705340

Abstract - The paper deals with presenting a data visualization regarding the heart diseases, which are the leading cause of heart problems and death for people in the entire world. The term itself refers to several types of heart conditions: the most common is coronary artery disease (CAD), which can lead to a heart attack. You can greatly reduce your risk for them through lifestyle changes and medicine.

This work should try to give doctors a better overview over patients' health; in particular it allows them to predict, with a certain probability, the risk of a heart attack and to prevent it, keeping the situation under observation.

1 Introduction

In order to start the project for the Visual Analytics course, we searched for some datasets on internet and at the end we found a very interesting one on the web site UCI Machine Learning Repository:

https://archive.ics.uci.edu/ml/index.php

Thanks to this, or better thanks to the information that the dataset gives us, we develop some important features for our project such as "visualization" with coordinated graphs or "analytics" having interactive contents.

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5 CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

2 Dataset Information

As we said before, among all the available datasets on the web site we choose the one dealing with the *"Heart Disease" (https://archive.ics.uci.edu/ml/datasets/heart+disease).* [4][5][6][7]

The dataset contains 918 records with 12 attributes, selected by all published experiments over the 76 initial ones.

For every person we have:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

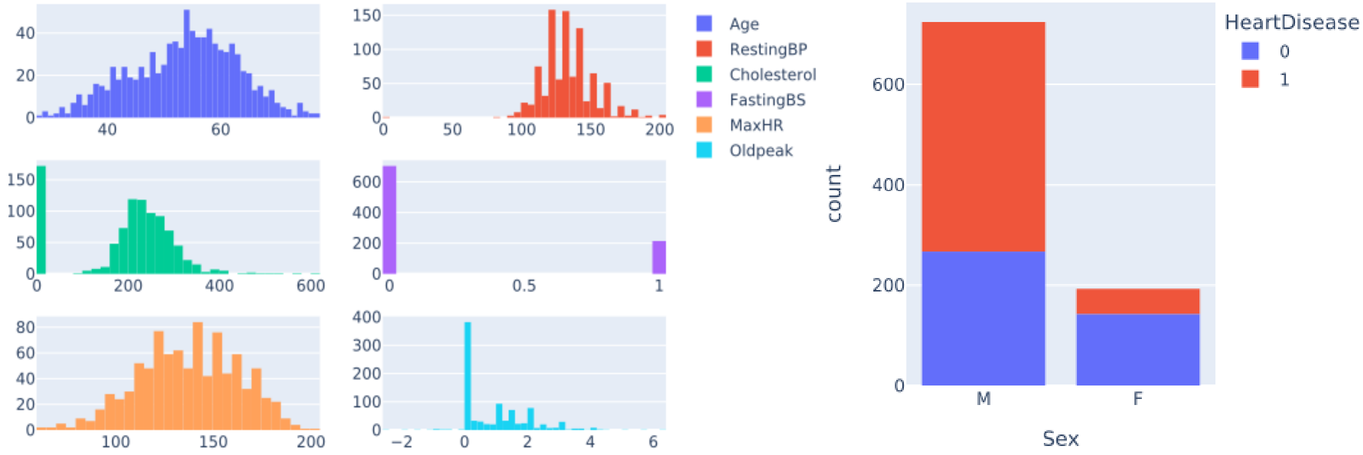


Figure 1. Distribution of numerical attributes (left), Gender and heart disease count (right)

In Figure 1 we can observe different histograms, obtained with the usage of the plotly library in Python; on the left the distribution of all the numerical attributes in the dataset, while on the right there is the count for heart diseases divided between male and female. According to the second one, among 725 males approximately 63,2% suffered from heart disease, while only 25,9% of the 193 females had it. So, the dataset in this case reflects the reality in which men are about twice as likely as women to suffer from heart attack.

2.1 Preprocessing

In order to make the dataset more readable and more comfortable to work with, we decided to modify the original dataset in the following way:

- In *Sex* attribute we replaced male/female with 1/0 respectively
- In *ChestPainType* we replaced [TA: 0, ATA: 1, NAP: 2, ASY: 3]
- In *RestingECG* we changed [Normal: 0, ST: 1, LVH: 2]
- For *ExerciseAngina* attribute we set [Yes: 1, No: 0]
- For *ST_Slope* we used [Up: 2, Flat: 1, Down: 0]

Therefore, we obtained a new dataset which contains only numeric values with a total of 11016 of them to manage.

3 Main Visualization

After observing the dataset we thought about the creation of the visual representation oriented to medics.

More specifically we made an .html file which will display the elaborated data in 3 different ways, such that the medic can observe specific behaviors and study them.

3.1 Scatter plot

This is a type of plot using Cartesian coordinates to display values for typically two variables for a set of data.

Since our dataset contains 12 attributes for each tuple, before displaying it we applied a dimensionality reduction using the PCA algorithm: it produces a (ranked) set of coordinates that allows for an 'optimal' projection, it is a linear transformation and preserves Euclidean distance and does not introduce false positives.

To obtain this we used a Python program that makes use of numpy and sklearn libraries in order to read and parse the original heart.csv file and obtain PCA values, saved into pca.csv.

To make it more user-friendly we made the visual.html file, which thanks to the usage of d3.js, reads pca.csv and displays the scatter plot on a web-browser (preferably Mozilla).

We also decided to introduce an identificator for each entry in the dataset, in order to allow the representations to be interactive with each other. In this way every selection made in one graph is also highlighted in the other one and vice versa.

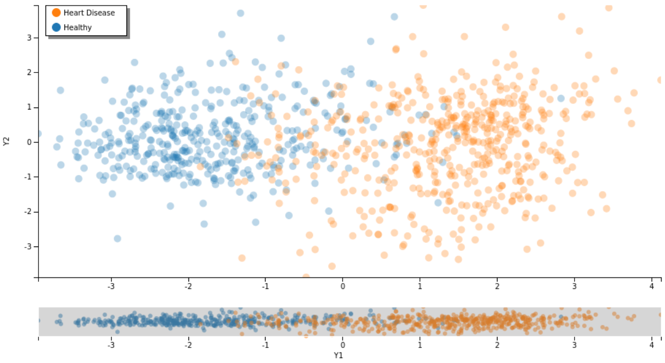


Figure 2. Scatter plot

As we can see in Figure 2, this is the result that we obtained after plotting the data according to the coordinates obtained with PCA and with the help of d3.js functions.
In particular people that suffer from heart disease are represented by orange dots, while the healthy ones are represented by the blue dots.

Here we made an interactive legend, in which by clicking on the specific dot it is possible to highlight all the corresponding ones to observe the details into the table at the bottom of the html page.
Obviously, it is possible to do the same also by selecting the dots of a specific area with the mouse.

To conclude the analysis we can observe that the two categories are quite well clustered and that there is the presence of some outliers, in particular in the central zone of Y1 axis the two classes are quite mixed.

3.2 Parallel coordinates

Parallel coordinates are a common way of visualizing and analyzing high-dimensional datasets.

To show a set of points in an n-dimensional space, a backdrop is drawn consisting of n parallel lines, typically vertical and equally spaced. A point in n-dimensional space is represented as a polyline with vertices on the parallel axes; the position of the vertex on the i-th axis corresponds to the i-th coordinate of the point. The graph is shown in Figure 3.

Similarly to the previous representation it is possible to select specific lines over each axis such that only the chosen ones will remain red, while all the others will go in background and become light gray.
By making a selection over multiple axes their intersection will be displayed.
It is also possible to drag the axis to change its order with respect to the others.

3.3 Table

To make the graphs more readable we decided to add a table containing all the original data of the selected dots/lines.
This table is initially empty, it will be populated only after a selection has been made and it will clear itself after a deselection.

Below an example is shown in Figure 4.

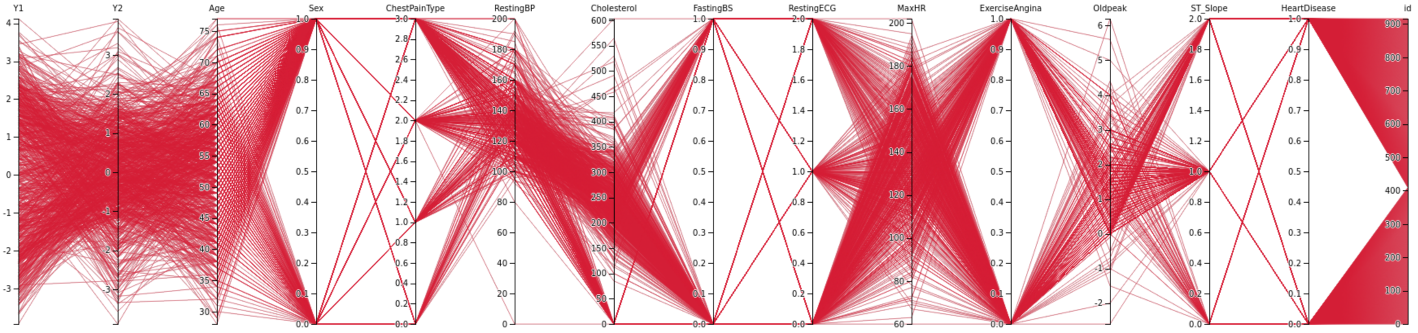
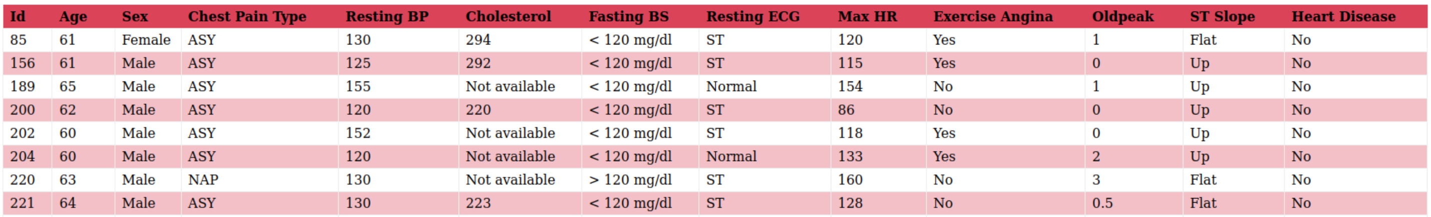


Figure 3. Parallel coordinates graph



Id	Age	Sex	Chest Pain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG	Max HR	Exercise Angina	Oldpeak	ST Slope	Heart Disease
85	61	Female	ASY	130	294	< 120 mg/dl	ST	120	Yes	1	Flat	No
156	61	Male	ASY	125	292	< 120 mg/dl	ST	115	Yes	0	Up	No
189	65	Male	ASY	155	Not available	< 120 mg/dl	Normal	154	No	1	Up	No
200	62	Male	ASY	120	220	< 120 mg/dl	ST	86	No	0	Up	No
202	60	Male	ASY	152	Not available	< 120 mg/dl	ST	118	Yes	0	Up	No
204	60	Male	ASY	120	Not available	< 120 mg/dl	Normal	133	Yes	2	Up	No
220	63	Male	NAP	130	Not available	> 120 mg/dl	ST	160	No	3	Flat	No
221	64	Male	ASY	130	223	< 120 mg/dl	ST	128	No	0.5	Flat	No

Figure 4. Table fulfilled after selection

4 Analytics

The reasoning part of this project is based on performing an estimation of the probability to suffer from heart disease by introducing patients' values inside the correspondent fields on an html page we have created.

The latter one is a process executed on demand interacting with a Python program, that is based on a machine learning algorithm, and with the help of Flask it is capable of communicating with the html page that runs locally.

In order to choose the machine learning classifier that better fits our dataset we made a comparison between the most popular ones and the results are listed below:

1. Dummy classifier with an accuracy score of 0.59
2. Logistic regression with an accuracy score of 0.88 for both with or without Scaler
3. Linear discriminant with an accuracy score of 0.86 for both with or without Scaler
4. SVC obtaining an accuracy score 0.72 and of 0.88 with Scaler
5. K-neighbors with an accuracy of 0.71 and of 0.88 with Scaler
6. Ada boost obtaining an accuracy of 0.86
7. Gradient boosting with an accuracy of 0.87
8. Random forest with an accuracy of 0.88
9. Extra trees classifier with an accuracy score of 0.88
10. XGBoost obtaining an accuracy of 0.82
11. LightGBM with an accuracy of 0.87
12. CatBoost classifier obtaining an accuracy of 0.88 and of 0.90 using tuned parameters

Based on these results we decided to use the last one, that is CatBoost with some tuned parameters being the best one in our case, as we can see also in Figure 5.

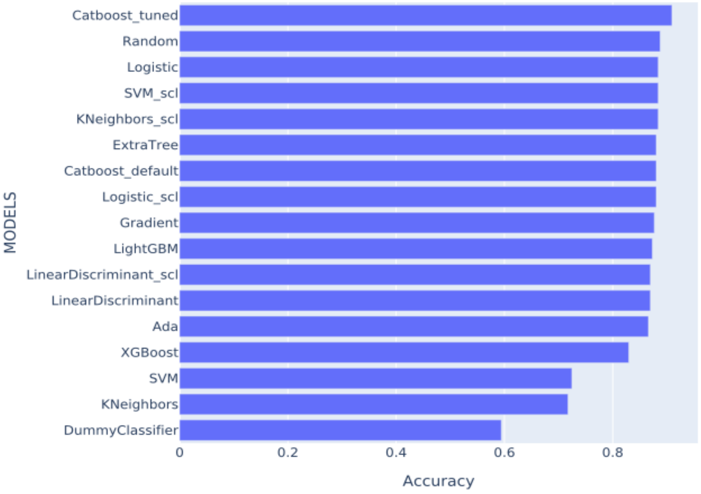


Figure 5. Model comparison

5 Insights

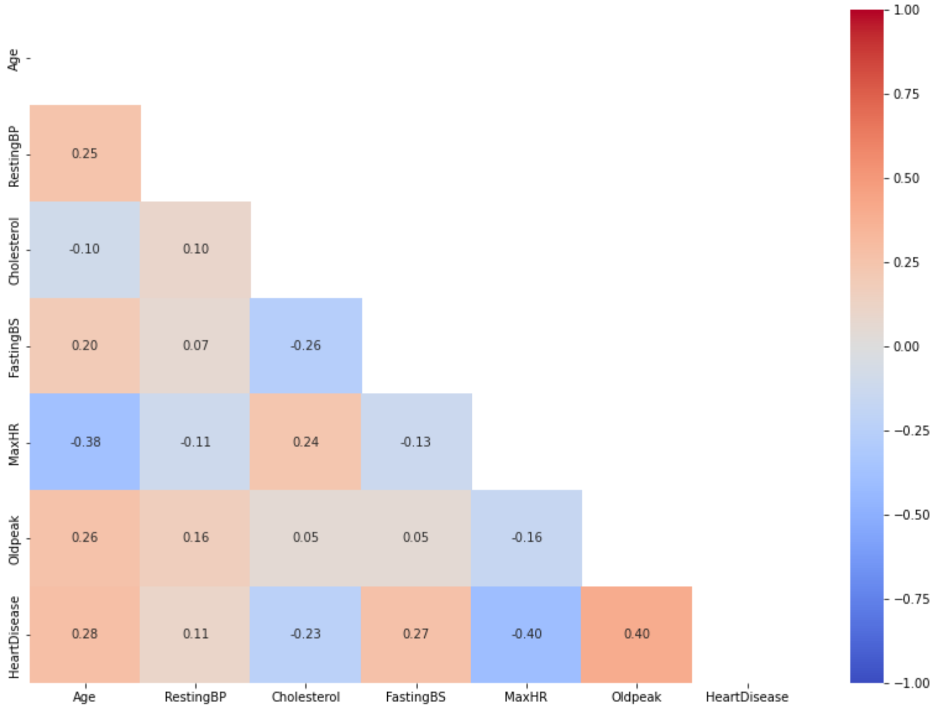


Figure 6. Correlation Matrix

This section gives some intuitions about the capacity of the system to answer the questions it is intended for. Is it possible to perceive some correlations between the values of a specific patient and its probability to suffer from heart disease?

With the help of a Python program, which makes use of the pandas, seaborn, numpy, matplotlib.pyplot and plotly libraries, we have been able to better understand the data by creating different histograms in order to visualize and analyze them.

In Figure 6 we can see the correlation matrix, based only on the numerical features, that shows us a general weak correlation between them and the target variables. More in detail Oldpeak has a positive correlation with heart disease, while maximum heart rate has a negative one with it.

On the other hand observing the categorical values it is possible to claim that:

- Men are almost 2.44 times more likely to have a heart disease than women
- There are clear differences among chest pain types
- Persons with ASY (asymptomatic chest pain) has almost 6 times the probability to have a heart disease than the ones with ATA (atypical angina chest pain)
- Having ST-T wave abnormality is more likely to have a heart disease than others
- Exercise-induced angina with 'Yes' is almost 2.4 times more likely to have a heart disease than exercise-induced angina with 'No'
- ST_Slope Up reduces significantly the probability to have heart disease than the other two segment

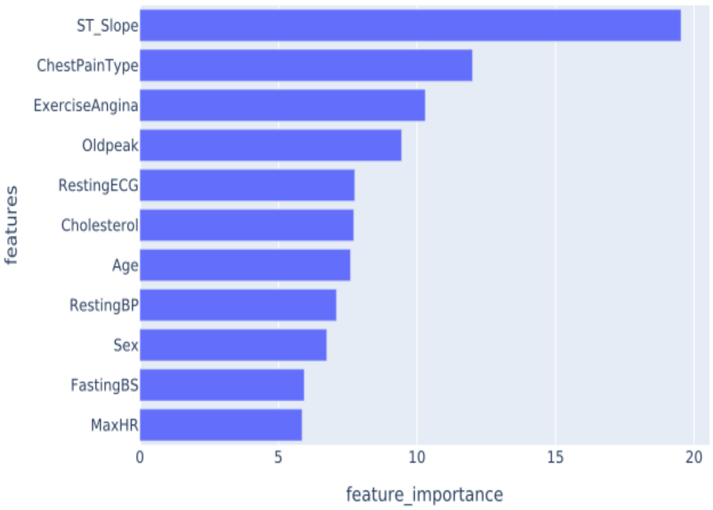


Figure 7. CatBoost feature importance

Finally, by plotting the Catboost feature importance (shown in Figure 6), we can observe which are the attributes that more affect the people's health in a bad way.

6 Related works

By doing research we found some papers related to the dataset we have decided to work with; in this section we will analyze the difference and equalities between them and our one.

The first paper is called *"Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier"* [1], which is similar to ours because of the usage of machine learning, but different for the purpose and the used algorithms. In fact it proposes a hybrid methodology based on the machine learning paradigm integrating the self-regulated learning capability of the particle swarm optimization (PSO) algorithm with the extreme learning machine (ELM) classifier. Moreover this methodology is able to classify also some other UCI medical datasets, namely, Wisconsin Breast Cancer, Pima Indians Diabetes, Heart-Statlog, Hepatitis, and Cleveland Heart Disease, with a very high accuracy but making no predictions based on it.

Another paper we found is *"Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System"* [2] in which a machine learning algorithm is proposed for the implementation of a heart disease prediction, similar to ours, but over two open access heart disease prediction datasets. Another contribution, which differs from our work, is the presentation of a cardiac patient monitoring system using different physiological signal sensors, like heartbeat, temperature, humidity and Arduino microcontrollers.

So even if both of the two papers mentioned above make use of differents machine learning classifiers like we did in our project, the main difference is that we also introduced a visualization part in order to let doctors observe and better study the data with the help of graphs, that are also able to interact among them.

7 Conclusions and future works

This project tried to offer a visual environment of support for doctors in order to give them a better overview on heart failures. More in detail, it gives them the possibility to use the available information of the dataset, combined with machine learning, to obtain a percentage prediction by inserting patients' values and possibly save them from heart diseases.

Taking into account the work developed into the paper [2], a possible future implementation of our project could be the integration of some physiological sensors applied directly on patients, in order to have a continuous monitoring system which will also alert doctors in case of complications.

8 References

[1] Subbulakshmi C. V., Deepa S. N. "Medical dataset classification: a machine learning paradigm integrating particle swarm optimization with extreme learning machine classifier." The Scientific World Journal. 2015

[2] Nashif, S., Raihan, Md. R., Islam, Md. R. and Imam, M.H. (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. World Journal of Engineering and Technology, 6, 854-873.

[3] UCI Machine Learning Repository, "Heart disease", UCI Machine Learning Repository: Heart Disease Data Set

[4] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

[5] University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

[6] University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

[7] V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.