Heart Failure Detection

Chiara Esposito 1716809

March 2022

1 Introduction

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. 4 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.

When we talk about heart diseases, we can have multiple conditions where heart is not working the way it should be. What aggravates this situation is that most of these diseases are being diagnosed at later stages at which it is very difficult to control. So, 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 for doctors, especially cardiologists.

The idea is to create an interactive dashboard that allows to see the different elements that may contribute to a heart failure, in order to detect and prevent possible diseases. It will be possible to monitor different elements, such as cholesterol, blood pressure, chest pain, maximum heart rate, etc... according to age and sex of the different people.

In particular, firstly we will see in the **Section 2** the dataset we used, and how we adapted it for our purposes. Then, in **Section 3** we will analyze some **related works** and studies already done on this subject. In **Section 4** we will analyze how we **implemented** our dashboard. And finally, in **Section 5**, we will conclude the research defining the main advantages of Visual Analytics.

2 Dataset

The dataset used in this project is from Kaggle [1] and describes heart failure prediction. It includes a csv file, **heart.csv**, which stores information about patients and which contains 11 features that can be used to predict a possible heart disease. The 11 features for patients are:

- Age: represented in years;

[1] `https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction`

- Sex: distinguished between M: Male and F: Female;
- ChestPainType: there are four types TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic;
- RestingBP: representing the resting blood pressure in mm Hg;
- Cholesterol: represented in mm/dl;
- FastingBS: fasting blood sugar having 1: if FastingBS $>$ 120 mg/dl, 0: otherwise;
- RestingECG: resting electrocardiogram results which could be Normal: Normal, ST: having ST-T wave abnormality, LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria;
- MaxHR: the maximum heart rate achieved, which is a numeric value;
- ExerciseAngina: exercise-induced angina, that could simply be Y: Yes or N: No;
- Oldpeak: = ST, which is the numeric value measured in depression;
- STSlope: the slope of the peak exercise ST segment. This could be of three types Up: upsloping, Flat: flat, Down: downsloping;

Finally, there is a 12th feature which corresponds to the output class: HeartDisease, which is 1 if there is an heart disease or 0 if not.
This dataset was created by combining different datasets already available independently but not combined before. In particular, Cleveland (303 observations), Hungarian (294 observations), Switzerland (123 observations), Long Beach VA (200 observations), Stalog (Heart) Data Set (270 observations), with a final dataset of 918 observations, removing duplicates.

3 Related works

We analyzed two different papers: 1. V. Gupta, V. Aggarwal, S. Gupta, N. Sharma, K. Sharma and N. Sharma, "Visualization and Prediction of Heart Diseases Using Data Science Framework," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1199-1202, doi: 10.1109/ICESC51422.2021.9532790.

2. Rami Lehtinen, Harri Sievänen, Jari Viik, Väinö Turjanmaa, Kari Niemelä, Jaakko Malmivuo, "Accurate detection of coronary artery disease by integrated analysis of the ST-segment depression/heart rate patterns during the exercise and recovery phases of the exercise electrocardiography test", The American Journal of Cardiology, Volume 78, Issue 9, 1996, Pages 1002-1006, ISSN 0002-9149, https://doi.org/10.1016/S0002-9149(96)00524-3.

From the first paper we took the approach, divided into different phases, in particular: A. Data Collection B. Data Cleaning, in which were removed all the noisy data records from the dataset, such as records which do not have identification available, NA data items and other parameters which are not involved in the prediction of heart diseases. In particular, regarding this point it has to be said that the dataset was already well cleaned, what we did was to check if there were noisy data and remove the aspects we found not so relevant for our purposes. C. Data Explore D. Parameter Selection In particular, for this last point we referred to the second paper and some other related information (from the website in the **Section ??**) in order to understand which parameters were more relevant and which not.

Furthermore, we also took the implementation of the first paper as reference. In particular, the first paper uses an implementation of barcharts to check the distribution of all parameters in order to explore the dataset and the interdependence between attributes. We decided to take this implementation but we changed it a little bit as we will see in the **Section 4**. We also saw that the use of an heatmap, developed based on the interdependence of attributes in the dataset. In order to evaluate the results, they have used reciprocal relationship between the attributes. This has been done in order to get to know the association between the parameters or attributes. We decided to take this aspect, but also in this case we modified it a little bit, in ordert to have not the association among all parameters, but just between a few of them, the ones we considered more relevant. We decided not to take the implementation if the Box Plot, but we used the Scatterplot.

The second paper has been chosen since it studies the importance of the slope ST, especially in the ECG. We considered this studies as bases for our purposes and therefore decide to take the ECG as an output for the Scatterplot.

The objective of the study, in fact, was to evaluate whether a novel continuous diagnostic variable, the ST/HR hysteresis (which integrates the ST/HR analysis during both the exercise and post exercise recovery phases of the exercise ECG test) can detect coronary artery disease more accurately than the other factors. The diagnostic performance of the ST/HR hysteresis in terms of coronary artery disease was the best regardless of the partition value selection. It is important to notice that it is impossible to investigate all patients with coronary angiography regardless of the outcome of the preceding ECG. But, having the possibility to have a complessive look of all the factors that may contribute to the heart disease it will be clearer which patients surely need an angiography and which do not.

4 Implementation

4.1 Analytics

Firstly, in order to analyze our dataset and implement the PCA we made our dataset that was qualitative, into a dataset fully quantitative. What we did was to replace for example "Sex" with [0,1] for [Male,Female][2];

```
original__data.replace(to__replace="M",
                       value="0",
                       inplace=True)
original__data.replace(to__replace="F",
                       value="1",
                       inplace=True)
```

"Chest Pain Type" with [0,1,2,3] for [ASY, TA, ATA, NAP]; "Resting ECG" with [0,1,2] for [Normal, ST, LVH; "STslope" with [0,1,2] for [Down, Flat, Up].

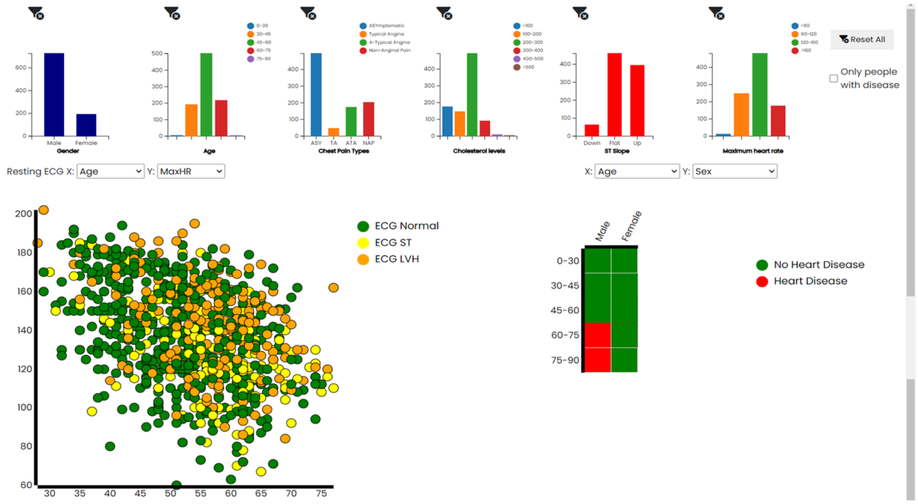
Even though we replaced the qualitative values in the dataset, we will continuously check them and represent them as quantitative in our visualization. We then saved our modified dataset as **heartquantitative.csv**, that is the one we will further use in our study. We also wanted to implement the possibility for the user to check not only all the 918 observations, but only the ones in which the probability for the heart disease is high (Heart Disease = 1). $reduced_data.drop(reduced_data[reduced_data['HeartDisease'] == 0].index, inplace = True)$ The result have been saved as **heartreduced.csv**. We did this thing twice, because we also wanted to have some of the values, such as Age and Cholesterol divided into ranges, so we developed an **heartquantitaive2.csv** file that contains the division in ranges, and **heartreduced2.csv** that is the file with ranges in which we consider only patients with Heart Disease = 1.

We then implemented the PCA noticing that there were no correlation among the fetures, meaning that the dataset was already cleaned, as we noticed before. But we analyze some of the features and decided to drop them, such as FastingBP, ExerciseAngina and Oldpeak finding them not so relevant for our purposes. In particular, regarding the Exercise Angina it was only yes or no, so we decided to pass it and focus only on the ST slope and ECG. While for the oldpeak it is said in the description of the dataset that it is related to the ST, so we decided to just take one of the two and in particular the ST Slope, finding it more relevant.

[2] This is an example of the code used. For all other replacement we used the exact same code, only changing values and parameters. For this reason we will not report all examples of code here.

4.2 Visualization

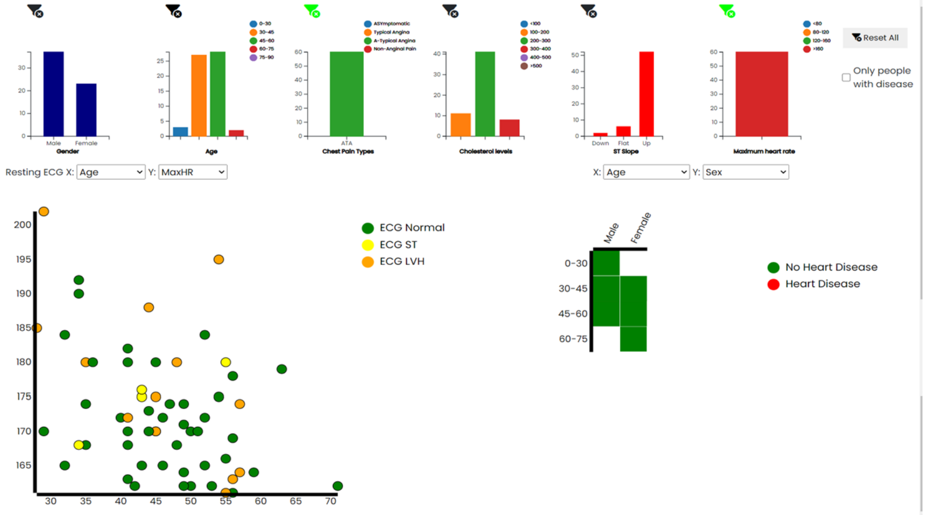
We implemented the visualization in three parts: barcharts, scatterplot and heatmap.



Resting ECG X: Age Y: MaxHR

X: Age Y: Sex

- For the barcharts, we decided to take the most relevant factor, relevant in sense of risk factors, and check how many people present them. This has been done in order to realize how the different factors are related to each other and especially to see if there are some risk factors that combined are more probable to cause an heart failure then others. Furthermore, we used also the single barcharts as filters, in order to allow users to instantly check how each feature influences the others. Once you click on a bar, a filter will be activated and the corresponding icon will change color, to let you know which filter are active at any time. By clicking on the remove filter icon you will remove that filter. You can also decide to reset all filters, in this case there is the reset button.



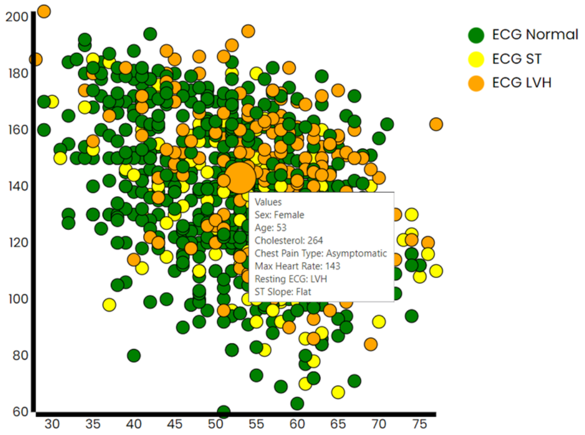
Resting ECG X: Age Y: MaxHR

X: Age Y: Sex

☐ Only people with disease

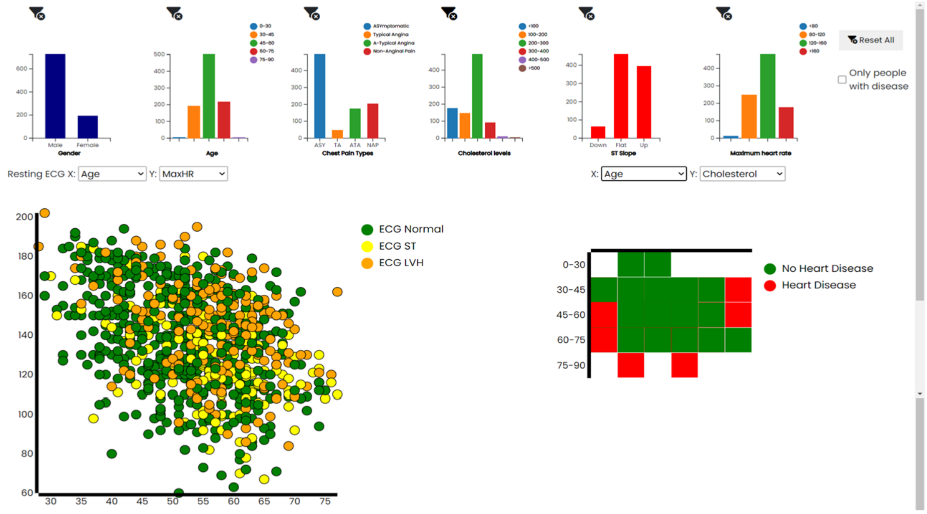
As you can see in the figure, two filters are activated, for Chest Pain Type (in particular ATA, A-Typical Angina) and the Maximum Heart Rate set to > 160. The relative icons above the barchart section are highlithed.

- For the scatterplot, we decided to visualize the distribution of all patients, having the possibility also to focus on a single one and check all its characteristics. We also decided to check as said before the ECG in relation to the different features analyzed by using different colors as output, for the different types od ECG.



Values
Sex: Female
Age: 53
Cholesterol: 264
Chest Pain Type: Asymptomatic
Max Heart Rate: 143
Resting ECG: LVH
ST Slope: Flat

- For the heatmap, as said in the previous section, we wanted to have an association and check the relation not among all parameters, but just the ones we considered relevant for our purposes. In particular in our case the intensity of the heat does not shows how good the attributes are related to each other, but represents if the majority of patients result with an heart disease or not. This has been done in order to have a more accurate visualization and check for patients also grouped for similar characteristics, such as age, type of chest pain and so on, so to immediately have the visualization of which factor are more dangerous.

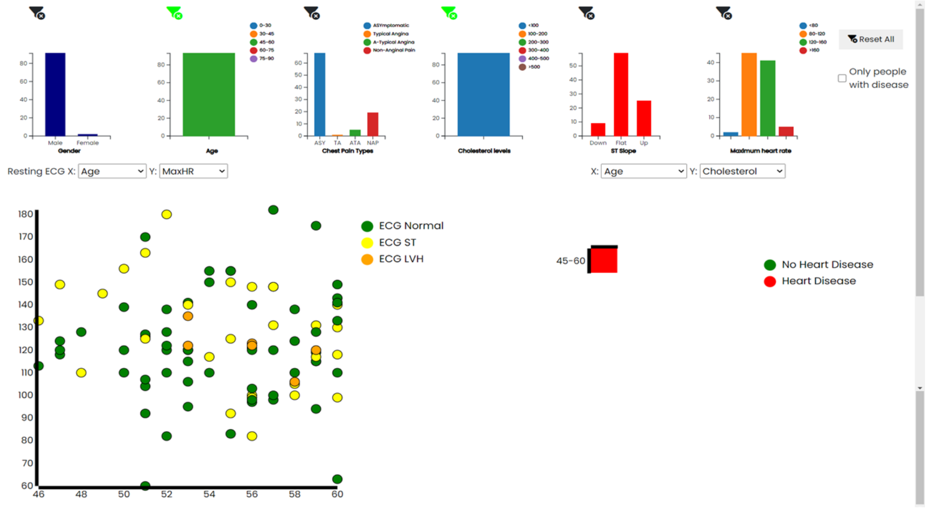


Resting ECG X: Age Y: MaxHR

X: Age Y: Cholesterol

□ Only people with disease

Even in this case it is possible to filter patients. In particular, by clicking on the single cells of the heatmap a filter also on the other visualizations will be activated.



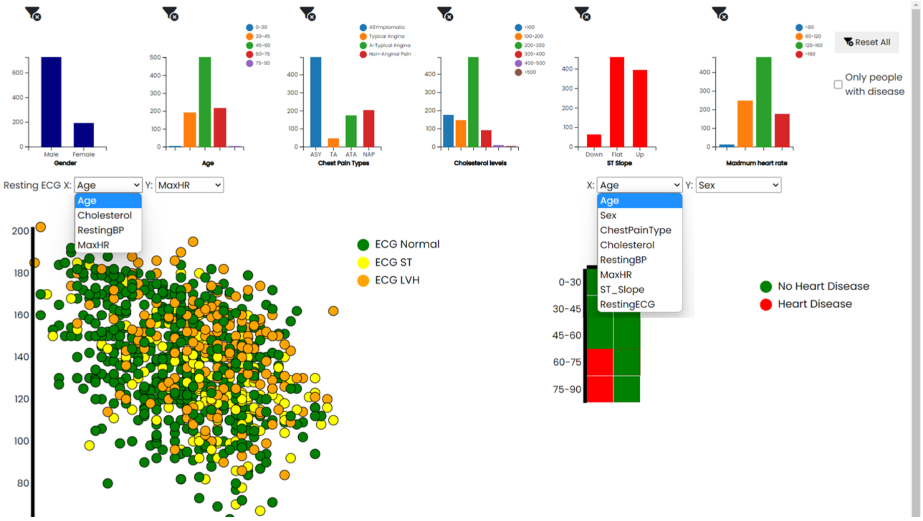
Resting ECG X: Age Y: MaxHR

X: Age Y: Cholesterol

□ Only people with disease

As you can see in the figure, by clicking on the cell having age 45-60 and level of cholesterol < 100, the corresponding filters were activated.

We want to underline the fact that, for both Heatmap and Scatterplot we gave the possibilities to the user to change the axes to see how the two outputs are affected by different components (e.g, how age and level of cholesterol affect the probability to have a heart disease or affect the ECG).



5 Conclusions

The implemented dashboard allows for an interactive, visual analysis for heart failure prediction. The advantages of such analysis are various:

- better risk allocation : you can check which are the most common risk factors that contribute to an heart failure;
- more effective targeted analysis, focused on patients grouped with similar characteristics, such as age, sex or levels of cholesterol;
- improved prevention: in further works, this study could be useful for helping patients in preventing heart failures, by for example addressing them to special medical checkings before synthomps appear.

6 References

V. Gupta, V. Aggarwal, S. Gupta, N. Sharma, K. Sharma and N. Sharma, "Visualization and Prediction of Heart Diseases Using Data Science Framework," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1199-1202, doi: 10.1109/ICESC51422.2021.9532790.

Rami Lehtinen, Harri Sievänen, Jari Viik, Väinö Turjanmaa, Kari Niemelä, Jaakko Malmivuo, “Accurate detection of coronary artery disease by integrated analysis of the ST-segment depression/heart rate patterns during the exercise and recovery phases of the exercise electrocardiography test”, The American

Journal of Cardiology, Volume 78, Issue 9, 1996, Pages 1002-1006, ISSN 0002-9149, https://doi.org/10.1016/S0002-9149(96)00524-3.

https://litfl.com/st-segment-ecg-library/

https://academic.oup.com/eurheartj/article-abstract/3/5/449/424605?redirectedFrom=fulltext