Exploratory data analysis for heart failure prediction dataset

Jelena Bozickovic

I Heart disease

Heart diseases are the leading cause of death worldwide, but it is estimated that almost 90% of them could be prevented. Avoiding risk factors and early detection are one of the two main ways to fend off heart diseases. Purpose of this paper is to see how can Machine Learning methods help detect heart diseases.

Dataset was taken from kaggle.com and it contains 918 samples with 11 attributes and an output class – heart disease.

Attributes are shown in the table below.

Age
Sex
Chest Pain Type
RestingBP
Cholesterol
FastingBS
RestingECG
MaxHR
ExerciseAngina
Oldpeak
ST Slope

Table 1: Attributes

Chest pain type are typical angina, atypical angina, non-anginal pain and asymptomatic. RestingBP is blood pressure in rest, Cholesterol represents total/serum cholesterol, fasting blood sugar is split into binary attribute: 1 if blood sugar is above certain threshold, 0 if below. Next is electrocardiogram in rest, which has 3 categories: normal, ST - ST wave abnormality, and LVH - left ventricular hypertrophy. Then there is maximum value of heart rate, exercise angina, oldpeak - numeric measure of ST shift, and ST slope which can be: up, down or flat.

II EDA

Exploratory data analysis gives more insight to the data. It is used through visual tools to gain more information about the dataset. Some of the basic visual representation of the data is used in this research and conclusions from those will be explained in few sentences.

IIa Correlation

Correlation statistically describes two variables, and how they move in accordance to one another. Correlation coefficients are numbers between -1 and 1. If two variables have positive correlation it means that if one variable increases, the other one increases as well. If they have negative correlation, what happens is that if one variable increases the other one decreases. In this dataset, we don’t have very strong correlation coefficients, the biggest positive correlation is 0.4, and biggest negative correlation is -0.4.

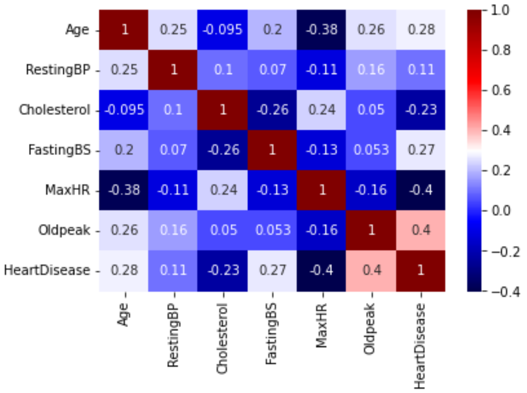

Image 1: Correlation among dataset variables

From the correlation plot we can assume or even prove some of the initial thoughs. Negative correlation is present between maximum heart rate and heart disease. This makes perfect sense, meaning that a person with a heart condition is not able to achive high heart rate, as if he/she is performing highly demanding physical activity.

Negative correlation is also present between cholesterol and heart disease. At first this doesn't seem right. People with higher levels of cholesterol should be in higher risk of developing a heart disease. Given that in dataset cholesterol includes total blood or serum cholesterol, which includes triglycerides, good (HDL) and bad (LDL) cholesterol it is reasonable that negative correlation is present. As a deeper analysis, it could be beneficial to separate 3 different kind of cholesterol and to find out if LDL cholesterol actually has impact on heart conditions.

Positive correlation is present in pretty much all other variables: target variable and oldpeak, the bigger the old peak the higher the risk of heart disease. Same applied to blood sugar, age and blood pressure.

IIb Pairplot

Relationships between attributes are shown in the image below and hue parameter was our output class - heart disease.

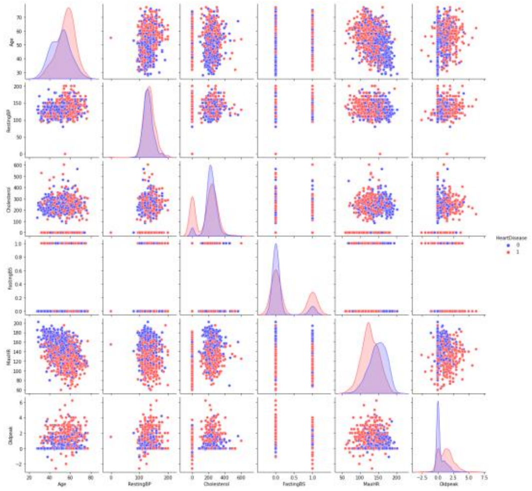

Image 2: Pair plot

On the diagonal of the pair plot we can spot the relationship between each attribute and a target class. For the age attribute it is visible that patients with heart disease tend to be older. For resting blood pressure distributions are pretty similar with values between 100 and 200, with majority around 150. With oldpeak attribute we have the biggest difference between classes. When there is no heart disease, oldpeak is around 0, which makes sense since the lower oldpeak the less chance for heart disease presence. When heart disease is present the values for oldpeak are around 2, but ranging from 0 to 5.

Even though there are some differences among attributes between two classes, we can see that there aren’t any attributes which are mutually exclusive, meaning that there isn’t a single attribute which could make a clear decision between two output classes, heart disease or not heart disease.

IIc Distribution

Among distribution, only some were analyzed. In image 3 there are distributions of sex among output class. We can see two things from this distribution. First, there are more males in dataset. Second, more males have heart disease than females. To confirm first observation, male and females number was counted and in the dataset there is almost 4 times more male than female.

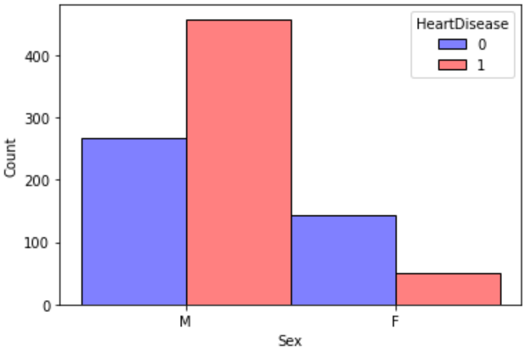

Image 3: Distribution of sex among output class

In image 4, distribution of three types of ECG signals in resting are shown. It is visible that most of the patients have normal ECG signal, and that there are half as much with clear deformities of ECG signal. In the group of patients with normal ECG signal, we have slightly higher number of people with heart disease, which indicates that even a normal ECG is not a clear witness of heart disease. Then in ST and LVH groups, there is higher number of people with heart disease, which makes sense, but there is also quite a lot of people without heart disease, but with irregular ECG signal.

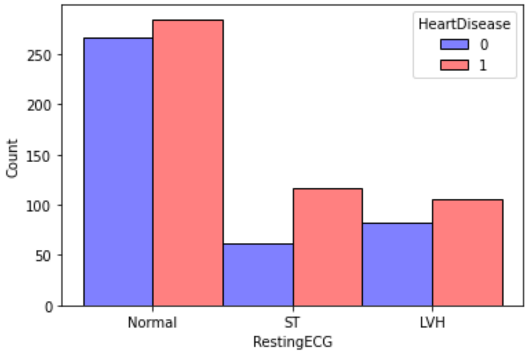

Image 4: Distribution of resting ECG among output class

From image number 5, we can see the distribution of 4 different kinds of chest pain. It is clearly visible that most of the people with heart disease have asymptomatic chest pain.

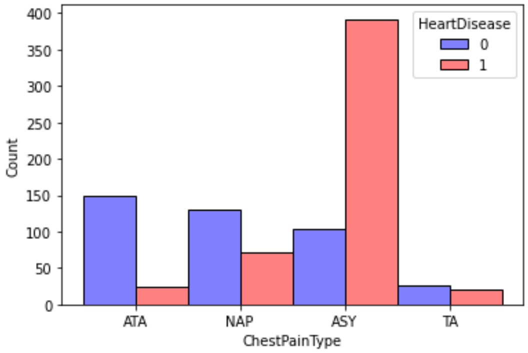

Image 5: Distribution of Chest Pain among output class

In previous image we realized that many of the patients with heart disease have asymptomatic chest pain. In image 6 it is shown that each type of chest pain usually has normal ECG signal. So we might not be able to say that specific irregular ECG signal causes specific chest pain.

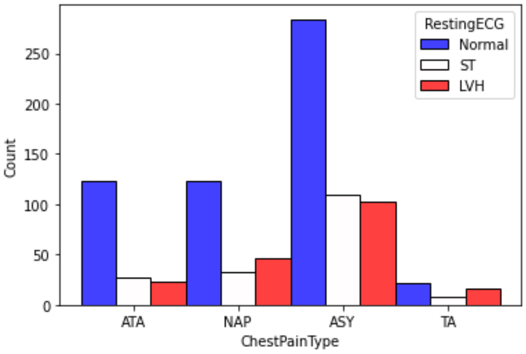

Image 6: Distribution of Chest Pain among resting ECG attribute

What happens with chest pain when different ST Slope is present is shown in image 7. It is visible that most of the chest pain types occur when ST slope is up, except in asymptomatic. Atypical and typical angina are almost never caused by down ST slope.

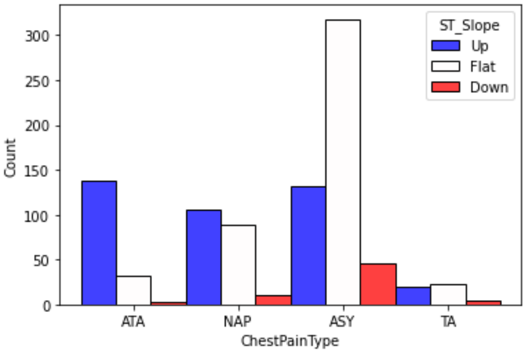

Image 7: Distribution of Chest Pain among ST slope attribute

From image 8 the same conclusion was made as from the correlation. People without heart disease tend to have higher maximum heart rate. As heart rate increases, presence of heart disease is lower, which confirms the negative correlation from the correlation plot. Image shows two normal distributions, one with mean around 120, for people with heart disease, and one with mean around 150 for people without heart disease.

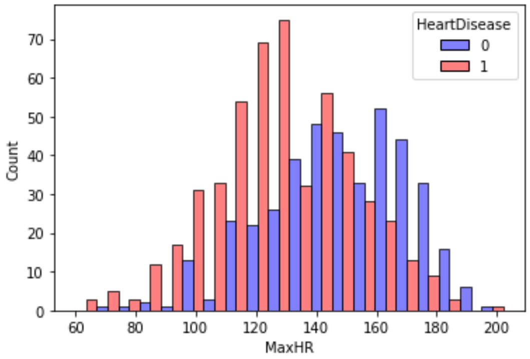

Image 8: Distribution of MaxHR among output class

IIe Boxplot

From box plot images interesting findings are shown in the images 7-9 below.

Image 9 shows box plots of oldpeak. The interquartile range is much wider for people with heart disease then for those without it. The values in class 1(heart disease) range between 0 and 2, and in class 0, between 0 and 1.

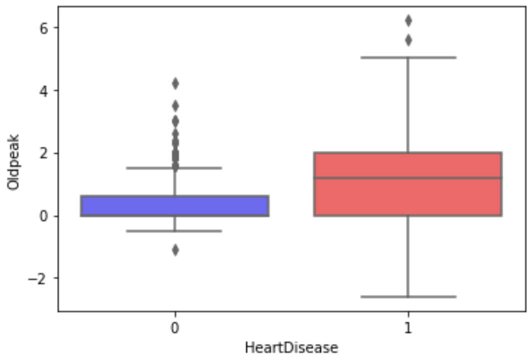

Image 9: Boxplot of oldpeak between output classes

Image 10 shows boxplot of maximum heart rate, and it is again confirmed that higher values of heart rate are achieved within patients without heart disease.

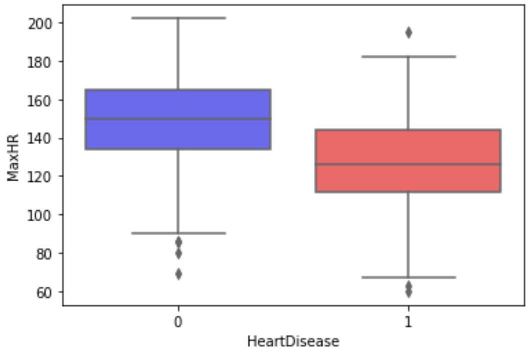

Image 10: Boxplot of MaxHR between output classes

Last, but not the least is shown in image 11. Patients with heart disease have irregular values of total blood cholesterol, and it ranges from 0 to 280, whereas patients without heart disease have their cholesterol levels between constant values 200 and 280, which are slightly higher than referent values, 120-200.

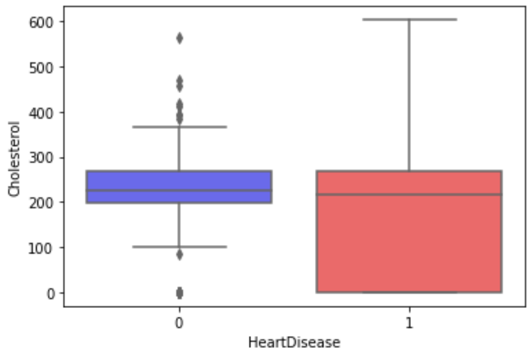

Image 11: Boxplot of cholesterol between output classes

III Conclusion

As a conclusion of this short exploratory data analysis it is stated that none of the attributes make a clear decision between two output classes. This conclusion makes sense realizing the subject that we are dealing with - if we knew one and only clear attribute for heart disease, if wouldn't have been any heart disease in the world, and it wouldn't be one of the leading causes of death worldwide.

However, deeper analysis could be quite beneficial, in two different ways. One way would mean acquiring more attributes meaningful for heart failure risk, and the other way is deeper exploratory analysis.