UCR | Extension
Professional Studies

PREDICTING PRESENCE OF HEART DISEASES USING MACHINE LEARNING ALGORITHMS

Data Analytics Technologies Final Project

Submitted by
Pedro Corrales De La Hoz
Patricia Martin Ballesteros
March 2020

Index

1. Description of the project

We take patient medical data information from the Cleveland database to analyze it and try to know if studying people medical data we can found out if they are going to have heart disease or not.

One-third of all global deaths are due to heart issues, in USA the half of the deaths are due to heart ailment. Around 17 million people die due to cardiovascular disease every year and in Asia the disease is highly. The lifestyle habits such as eating, smoking or physical inactivity and other medical information (age, sex, cholesterol or diabetes) are considered to be big risky factors for heart disease. There are different types of disease and it is difficult to determine the odds of getting heart disease based on risky factors.

Machine learning techniques are useful to predict the output from existing data. One reason for fatality due to heart disease is due to the fact that the risks are either not identified, or they are identified only at a larger stage. However, machine learning techniques can be useful for overcoming this problem and to predict risk at an early stage. The existing research has used ensemble methods to improve classification accuracy in prediction of heart disease.

The dataset is composed by 76 different attributes but we are using the 13 more important attributes, all experiment published about this problem are using this attributes:

1. Age
2. Sex
3. Chest pain type (4 values)
4. Resting blood pressure
5. Serum cholesterol in mg/dl
6. Fasting blood sugar > 120 mg/dl
7. Resting electrocardiographic results (values 0,1,2)
8. Maximum heart rate achieved
9. Exercise induced angina
10. Oldpeak = ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment
12. Number of major vessels (0-3) colored by fluoroscopy
13. Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
14. Target 1 or 0

Patients in this data has the age from 29 to 79. Male are detonated by a gender value 1 and female gender value 0. We have four types of chest pain type (1-Typical type, 2-Atypical type angina, 3-Non-angina pain and 4- Asymptomatic). We also have attributes for the resting blood pressure(94 to 200), Serum cholesterol (126 to 564), Fasting blood sugar level (1=below 120 mg/dl and 0=above 120 mg/dl), Resting electrocardiographic result (0,1,2), maximum heart rates (71 to 202), the exercise induced angina which is recorded as 1 if there is pain and 0 if there is no pain, oldpeak is the depression induced by exercise (1 to 3),

the slope of the peak exercise (1,2,3), number of major vessels colored by fluoroscopy (0 to 3), Tha; (3=normal; 6=fixes detect: 7=reversible) and Target is class attribute (0 or 1).

2. Review

The heart disease is one of the most complex and life deadliest human diseases in the world, and the rate in the United States it is very high. The European Society of Cardiology (ESC) reported that 26 million adults worldwide were diagnosed with heart disease and 3.6 million were diagnosed every year. Moreover, around 50% of heart disease people suffering from heart disease die within initial 1-2 years, and concerned costs of heart disease management are approximately 3% of health-care financial budget.

The investigation techniques in early stages used to identify heart disease were complicated, and its resulting complexity is one of the major reasons that affect the standard of life. The heart disease diagnosis and treatment are very complex, but it is very necessary to have an accurate diagnosis of the heart disease risk in patients in order to reduce heart issues and improve security of heart.

Here is where machine learning comes because its techniques can significantly benefit the medical field by providing an accurate and quick diagnosis of diseases. Hence, save time for both doctors and patients.

Physicians routinely make treatment decisions using risk scores, which are based on few variables and are typically only moderately accurate for individual patients. Machine learning can use repetition and adjustment to exploit large quantities of data and identify complex patterns that may go unnoticed by humans. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

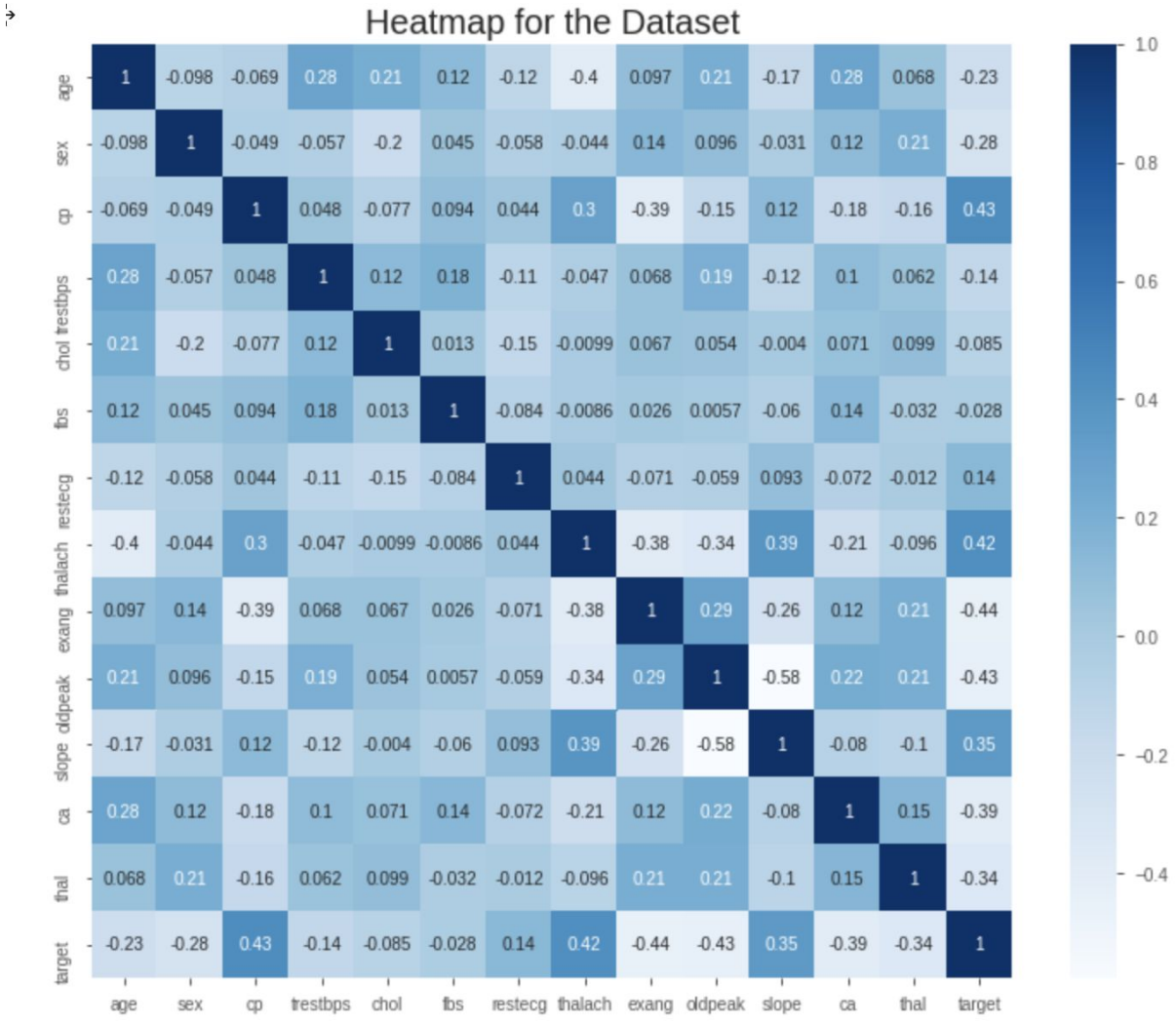
There have been various research projects that have used ML to predict heart diseases - many of them posted on Kaggle, where we found which models work better and give higher accuracy. The most popular algorithms are Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Naive Bayes, Decision Tree and Random Forest.
After checking the result of the accuracy of each algorithm in the different posted notebooks we realized that Decision Tree gives significantly the worst prediction with an accuracy around 70%.
By far, SVM seems to be the best predictor by average, but Random Forest and Naive Bayes share the best accuracy with an 88.52%. Besides, Logistic Regression and KNN have also good results with accuracy values between 79% and 86.89%.
Previous research have also shown that Naive Bayes was found to be the best algorithm due to its simplicity and accuracy. Moreover, Sex, age, smoking, hypertension, and diabetes seem to be the major risk factors for heart disease.

Given the above and if the model has some predictive ability, we should see these factors standing out as the most important.

3. Data Exploration and Visualization

To start understanding the data, we got an overview of all the different attributes. Thanks to the library pandas_profiling we got the mean, maximum, minimum, median, and other statistical measures of the data as well as Frequency Plots to show its distribution through the data.

Then we decided to see the correlation matrix of the features and try to analyse it.

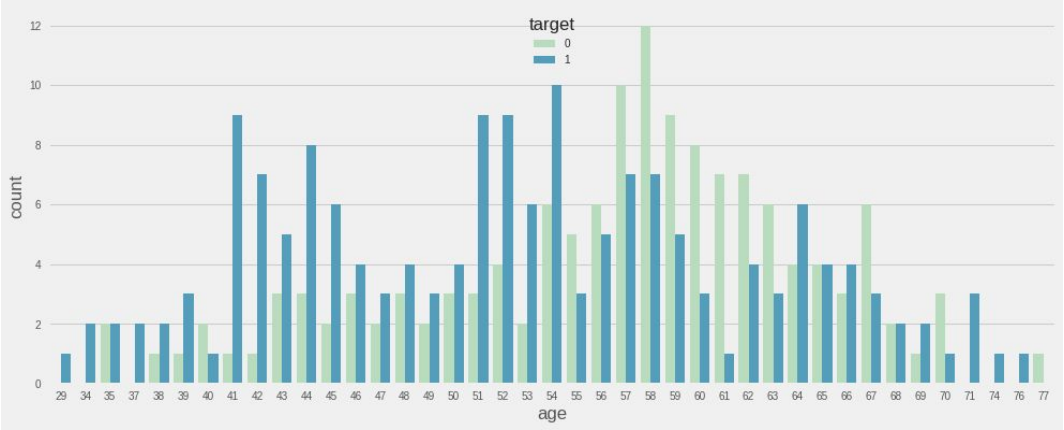


The heat map shows the correlations between the different attributes of the dataset. We can see that almost all of the features given in the dataset are very less correlated with each other. This implies we must include all of the features, as we can only remove those features where the correlation of two or more attributes are really high. We can highlight some of them like:

- Target and chest pain type (cp) are mildly positively correlated (0.43).
- Target and maximum heart rate achieved (thalach) are also mildly positively correlated (0.42).
- Target and exercise induced angina (exang) are gently negatively correlated (-0.44).

- Target and ST depression induced by exercise relative to rest (oldpeak) are also mildly negatively correlated (-0.43).

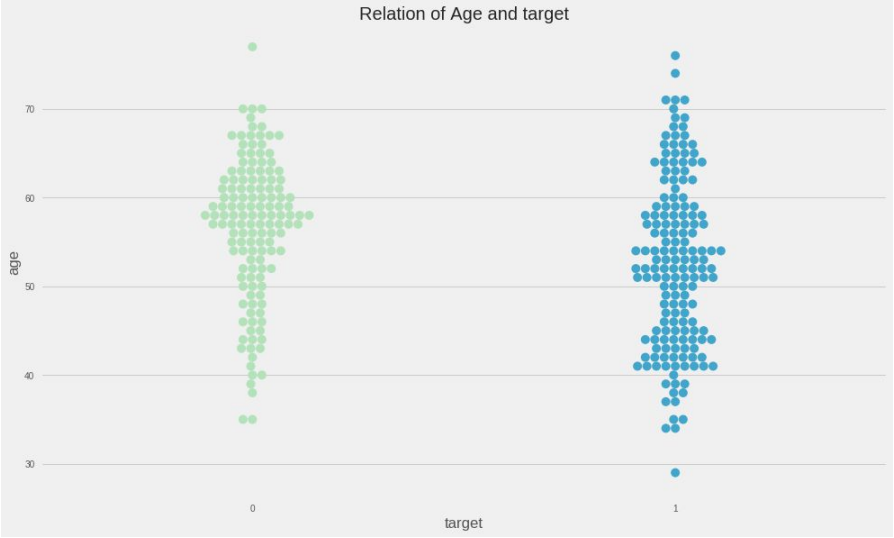
Age



As we can appreciate in the first graph, the age variable distribution is approximately normal. Also, we see that the highest number of people that have from heart diseases are in the age group of 55-65 years. The patients in the age group 20-30 are very less likely to suffer from heart diseases.

Since the number of people in the age group 65-80 has a very low population, hence distribution is also less. we might have to opt for other plots to investigate further and get some more intuitive results.

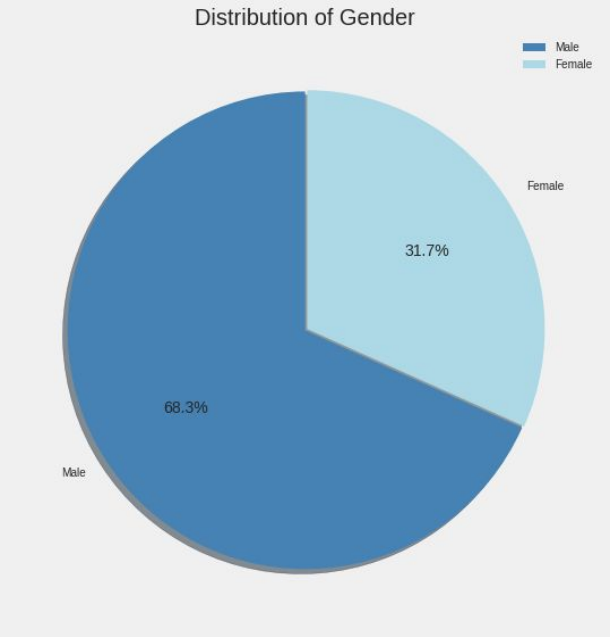
Age and Target



From the above Swarm plot between the age of patients and target,we are not able to find any defined pattern, so age is not a very good attribute to determine the heart disease of a patient. Since a patient of heart diseases range from 30-70, it is not important that all of the people lying in that same age group are bound to suffer from the heart diseases.

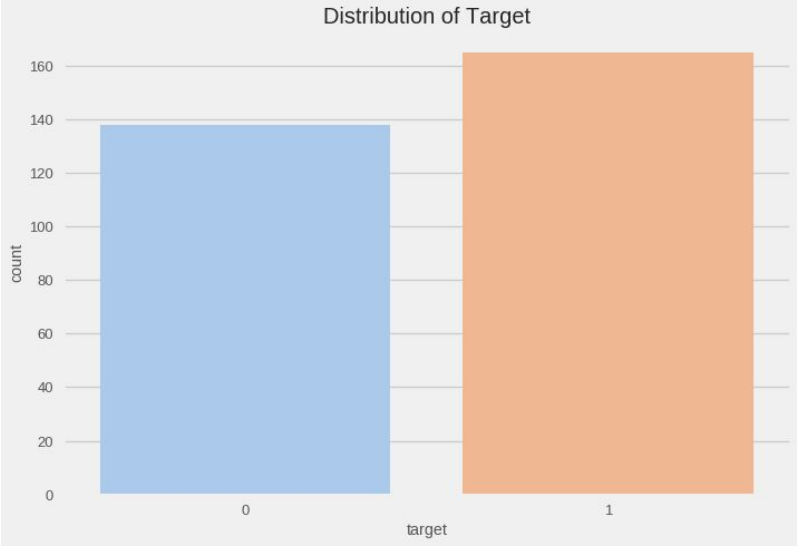
We can see that the people suffering from heart disease (target = 1) and people who are not suffering it (target = 0) have similar ages.

Gender



The above pie chart shows us the distribution of gender in the dataset. By looking at the plot, we can expect that males are two times more likely to suffer from heart diseases in comparison to females - 68% of the patients are men whereas only 32% are women. More number of men took participation in heart disease check ups. But we will continue exploring age to see if which gender really has more heart diseases.

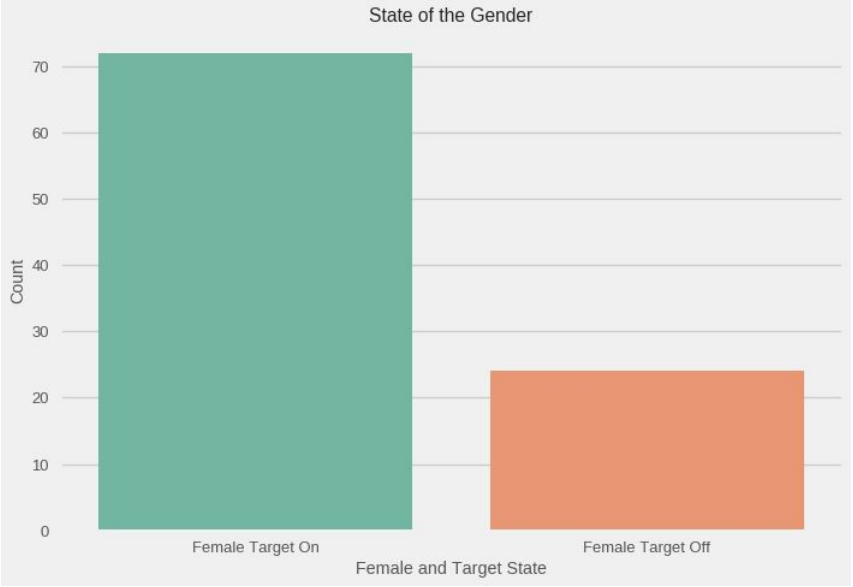
Target



Let's look at the target, the dataset is quite balanced with almost equal number of positive and negative classes. It is important to mention that the positive class says that the patient is suffering from the disease and the negative class says that the patient is not suffering from the disease.

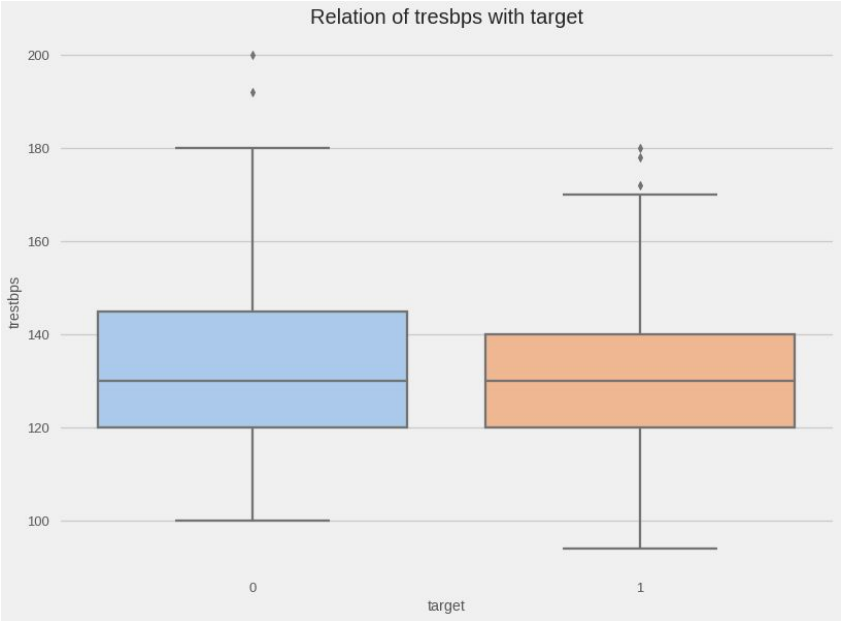
Gender and Target

In the graphic above we can see the relation between males and the target. The number of males who hasn't have heart disease is less than the males who have heart disease. However this difference is not that big.



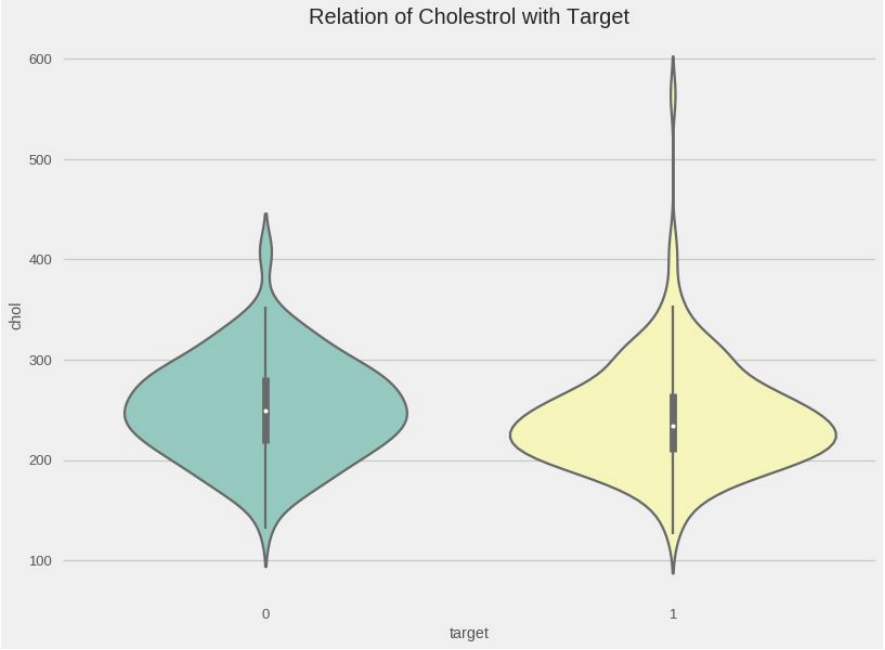
This is interesting: number of women suffering from heart disease are more than men but men population is more than women. The most striking data we can get is the big difference between females with heart disease and females without it.

Resting Blood Pressure and Target



The above bivariate plot between tresbps(the resting blood pressure of a patient), and the target which says that whether the patient is suffering from the heart disease or not. The plot clearly suggests that the patients who are most likely to not suffer from the disease have a slightly greater blood pressure than the patients who have heart diseases.

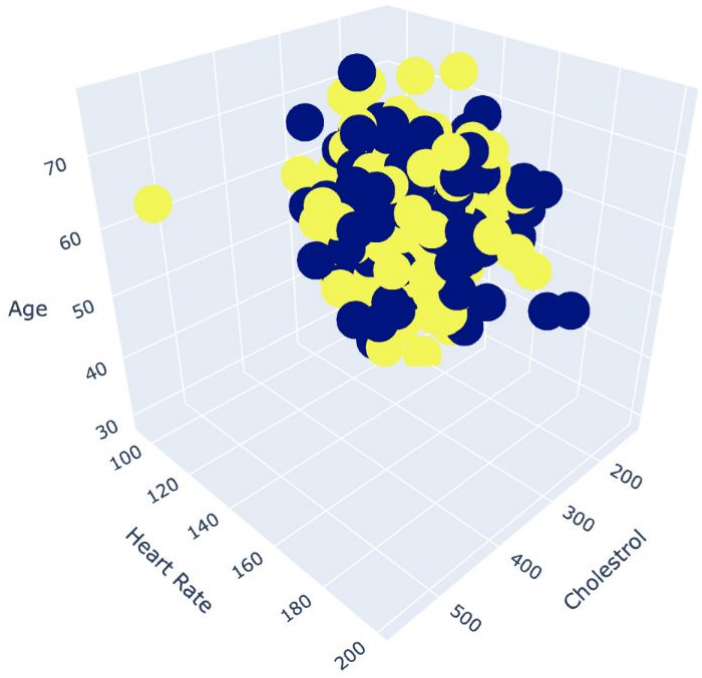
Cholesterol and Target



The above plot between cholesterol levels and target suggests that the Patients likely to suffer from heart diseases are having higher cholesterol levels in comparison to the patients with target 0(likely to not suffer from the heart diseases.

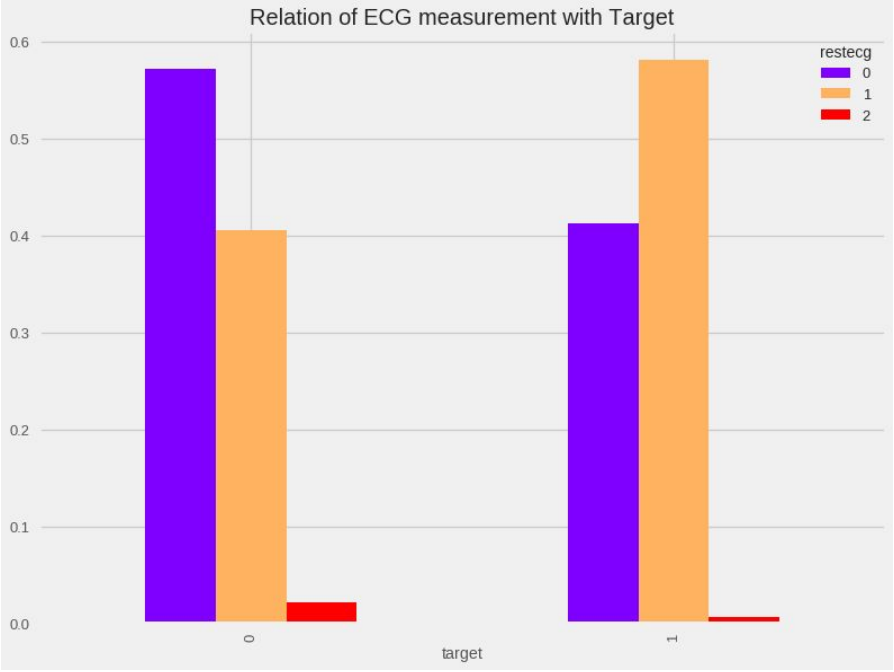
Hence, we can infer from the above plot that the cholesterol levels plays an important role in determining heart diseases. We all must keep our cholesterol levels in control as possible.

Cholesterol, Resting Blood Pressure and Age with Target



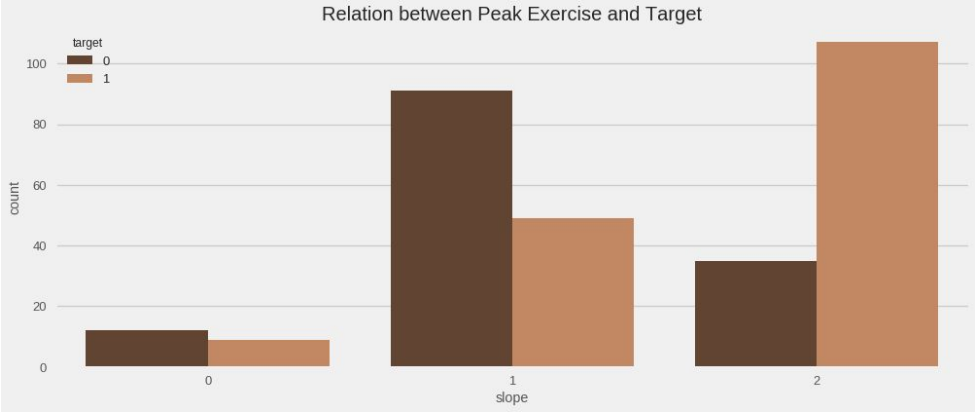
In this graphic we can see all the data in relation to the Cholesterol(x), Heart Rate(y) and Age(z). All of them are moving in the same numbers of cholesterol except a few highlighting one with 564 mg/dl. The normal average of cholesterol is between 125 and 200 mg/dl and several data are more than 200 mg/dl. As for Heart Rate is more dispersed than cholesterol. The normal blood pressure is 120/80 mm/Hg. The are a lot of cases that are below the normal pressure but they are really close to the limit and the rest are above this limit.The population is between 29 and 77 years old.

Resting electrocardiographic measurement and Target



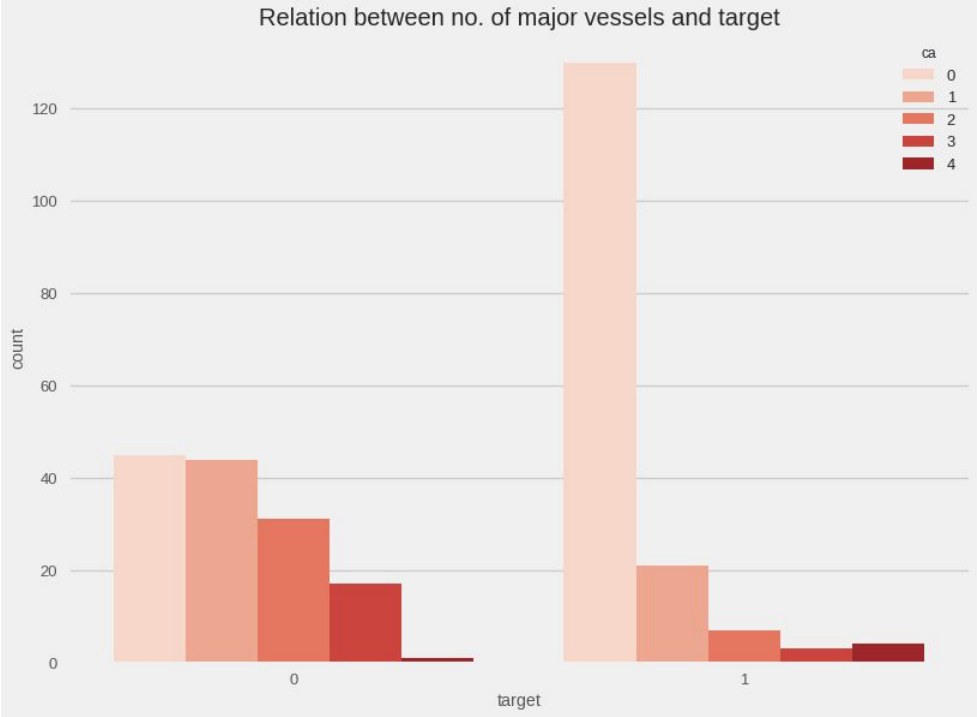
The above plot is column bar chart representing target vs ECG Measurements(ElectroCardioGram), The above plot shows that the more number of patients not likely to suffer from heart diseases are having restscg value 0 whereas more number of people have restecg value 1 in case of more likelihood of suffering from a heart disease.

Peak Exercise and Target



This plot clearly shows that the patients who are not likely to suffer from any heart diseases are mostly having value 1 means upsloping, whereas very few people suffering from heart diseases have upsloping pattern in exercises. Also, Flat Exercises are mostly seen in the cases of Patients who are more likely to suffer from heart diseases.

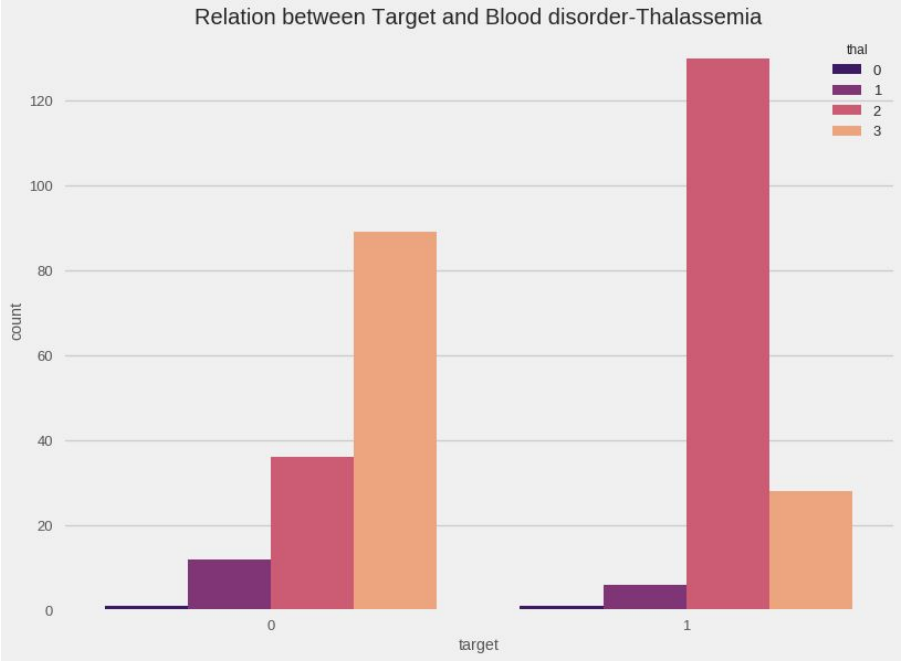
Number of Major Vessels and Target



The above Bivariate plot between Target and Number of Major Vessels, shows that the patients who are more likely to suffer from Heart diseases are having high values of Major Vessels whereas the patients who are very less likely to suffer from any kind of heart diseases have very low values of Major Vessels.

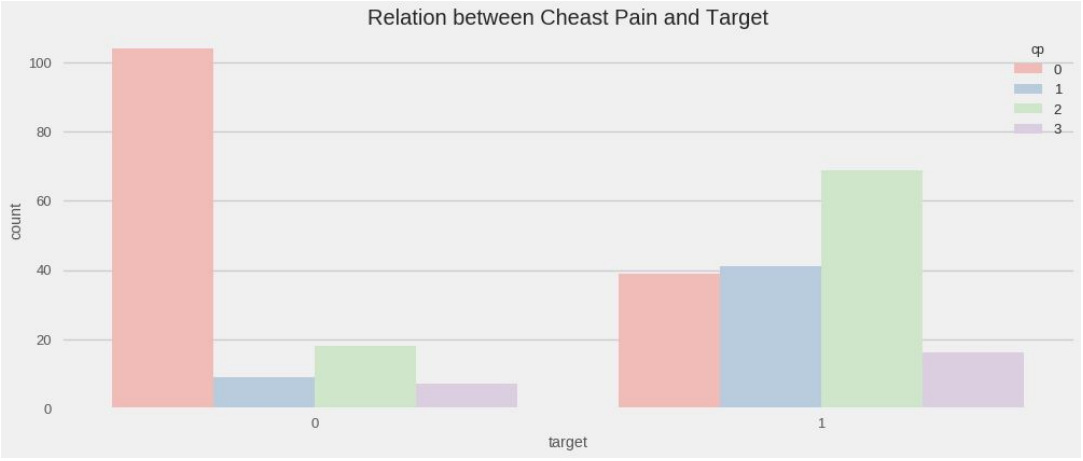
Hence, It is also helpful in determining the heart diseases, the more the number of vessels, the more is the chance of suffering from heart diseases.

Blood Disorder - Thalassemia and Target



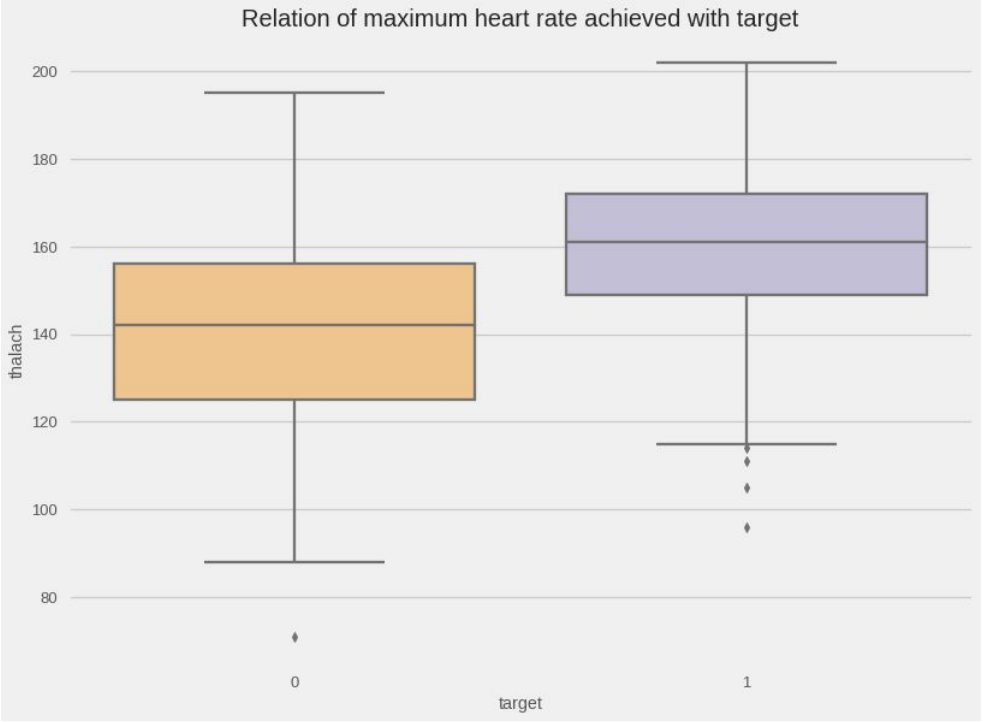
In the above Boxen plot between Target and a Blood disorder called Thalassemia, It can be easily inferred that the patients suffering from heart diseases have low chances of also suffering from thalassemia in comparison to the patients who are less likely to suffer from the heart diseases. Hence, It is also a good feature to classify heart diseases.

Chest Pain and Target



This graphic show us the chest pain of the people studied in this data. Mostly persons without heart disease has typical type of chest pain(red column) and the other three types they are really similar highlighting the non-angina type (green column). Chest pain in persons with heart disease are more distributed, standing out the Non angina type over the others. The Asymptomatic type is the lower case in both situations.

Maximum Heart Rate Achieved and Target



The plot above shows the relation between the maximum heart rate achieved and the target - whether the patient has a heart disease or not. As we can see in the graph, the patients who have heart disease have higher heart rate (between 150 and 170) while those who do not have it are have a rate between 125 and 155.

Fasting Blood Sugar and Target

Fasting blood sugar is used to determine if a patient has diabetes or not. When the sample is higher than 120 mg/dl, then the patient is considered to have prediabetes or diabetes. Looking at the plot above, we see that diabetes does not really affect when having heart disease since there is a greater number of people with heart disease and low fasting blood sugar levels.

Exercise Induced Angina and Target

Looking at the relationship between target and exercise induced angina, we can appreciate that most of the people who have heart disease do not have had exercise induced angina before, which is quite relevant.

References

- https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124
- https://www.heartfoundation.org.au/your-heart/living-with-heart-disease/medical-tests
- https://www.bhf.org.uk/informationsupport/risk-factors
- https://www.kaggle.com/ronitf/heart-disease-uci/kernels
- http://www.ehnheart.org/cvd-statistics.html
- https://www.scirp.org/journal/paperinformation.aspx?paperid=73781
- https://www.hindawi.com/journals/cmmm/2017/8272091/
- https://www.researchgate.net/publication/328031918_Machine_Learning_Classification_Techniques_for_Heart_Disease_Prediction_A_Review
- https://seaborn.pydata.org