Heart diseases Analysis

# Abstract

Heart disease is the number 1 cause of death according to world health organization (WHO)[1], detecting heart diseases and correctly diagnosing heart diseases in early stages could improve and increase the life expectancy of the population.

In this study we will study the relationship between 13 clinical features, and the presence of heart disease.

we have found that there are no single feature, however we found that 84% accuracy for support vector machine, and the presence of asymptomatic chest pain is the best predictor.

Being able to accurately train a machine to make clinical decisions to help healthcare professionals to make their decisions is an important and could be life saving however, there are not enough data to make an accurate decision without collecting large amounts of accurate data.

# Introduction

According to WHO heart diseases is the leading cause of death worldwide[1], and while the advancement in medicine in treating heart diseases is significant, it still can’t cure it completely, and many cases remain undiagnosed, while still suffering from the underlying symptoms[2]

as the advances in the fields of data science becomes more and more advanced, so does its applications in medicine, and bioinformatics. In this project the aim to apply and evaluate the use of machine learning and data analysis on Heart diseases data set[3].

## Literature review and related works

This dataset was used for extensive ML and analysis methods.

in November 2018 (Ali Bou Nassif et al [4] has trained 3 classifiers, to predict coronary heart disease, where they found out the naïve bayes classifier outperform SVM and KNN with 84% accuracy.

(C.V.SubbulakshmiandS.N.Deepa[5] ) proposed a new algorithm combined the particle swarm algorithm (PSO) with the extreme Learning Machines (ELM) to predict the precense of Heart disease, which 99.78%.

## Problem statement and objective

The dataset is a 13 selected features, where the objective is to use predictive statistics and data mining techniques, to predict the presence of Heart diseases, and the severity, however in this project we are only interested in the presence of heart diseases.

the dataset records features from 303 participants, each feature and their description are interpreted in (Table.1)

Table 1 Features description

|  |  |
| --- | --- |
| Names of columns | Description |
| Age | Years |
| Sex | 1: males, 0: females |
| cp | - Value 1: typical angina  - Value 2: atypical angina  - Value 3: non-anginal pain  - Value 4: asymptomatic |
| trestbps | resting blood pressure (in mm Hg on admission to the hospital) |
| chol | serum cholestoral in mg/dl |
| fbs | (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) |
| restecg | resting electrocardiographic results:  - Value 0: normal  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or  depression of > 0.05 mV)  - Value 2: showing probable or definite left ventricular hypertrophy by Estes'  criteria |
| thalach | maximum heart rate achieved |
| exang | exercise induced angina (1 = yes; 0 = no) |
| oldpeak | ST depression induced by exercise relative to rest |
| slope | the slope of the peak exercise ST segment:  - Value 1: upsloping  - Value 2: flat  - Value 3: downsloping |
| ca | number of major vessels (0-3) colored by flourosopy |
| thal | 3 = normal; 6 = fixed defect; 7 = reversable defect |
| results | num: diagnosis of heart disease (angiographic disease status).  It is integer valued from 0 (no presence) to 4.  Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0) |

This study is concerned in making a predive analysis using xgboosting, SVM and randomForest.

# Experiment setup and results

The first step in any data analysis and data mining project is to prepare the data, and make it easier and clearer for analysis, once we attenuate the noise we can represent the data, and we can extract all the information.

For our study we have to ensure that our data is clear, and doesn’t have any missing values, since our dataset is relatively small, and the number of missing values is also small, we will replace the missing values with the most frequent value the mode of the column, this way we can ensure that it wont significantly mess with our analysis, and will preserve the underlying structure of the data.

The next step in our preparation phase is to check for outliers, where we used the Elliptic Envelope[4], [5], which can look at the distribution and estimate the likelihood of each data point to be within the main distribution, if not it is more likely to be an outlier.

Applying the elliptic envelope to our dataset have detected 30 outliers, most of the outliers are however within are not noisy, and some pattern could be viewed, where the presence of the outlier could predict the outcome of the target variable, where we used the outlier as a new feature for our dataset.

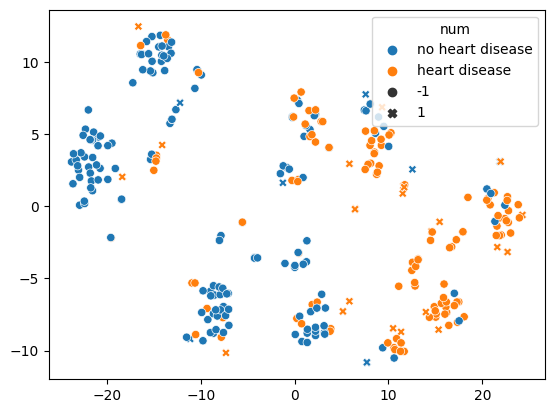


Figure 1 TSNE scatter plot and the location of the outliers, where x marks an outlier and o marks an inlier, orange is the presence of a disease, and blue is the absence of a disease.

## Exploratory data analysis

In data analysis phase we are trying to understand the data, and figure out how everything is related in the dataset.

First we looked at the variables we have 13 variables, our strategy is to analyze continuous variables, and discover the relationship between it and other categorical and discrete variables, with a focus on the target variable.

At first we looked at the distribution of cholesterol, and resting blood pressure for cases and without heart diseases as in figure 2.



Table 2 cholesterol serum level between males and females

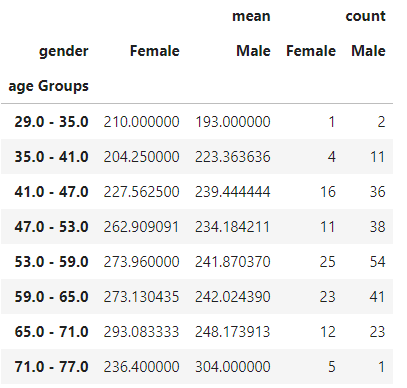
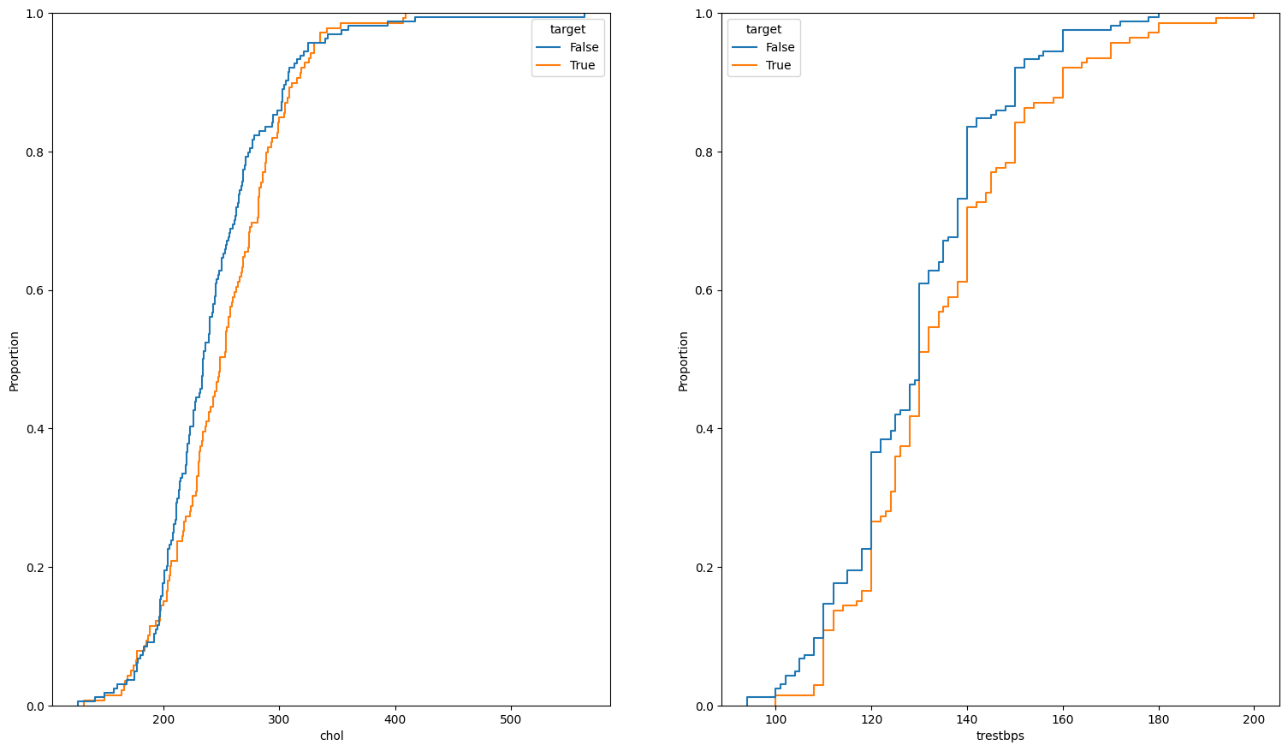


table 3 the distribution of males and females cholesterol levels accross age groups

Figure 2 ecdf plot for cholesterol level left, resting blood pressure right, seperated by the presence of heart disease orange, and absence of heart disease blue.



As shown in figure 2 resting blood pressure have a significant effect on predicting the presence of heart disease.

Figure 3 shows that heart disease is only prevalent in patientsolder than 35 years old, and the median seems to have a positive trend for cases diagnosed with heart disease, however there is no clear pattern for cases without heart disease, also there are more cases where the resting blood pressure is significantly higher than the median of the dataset.

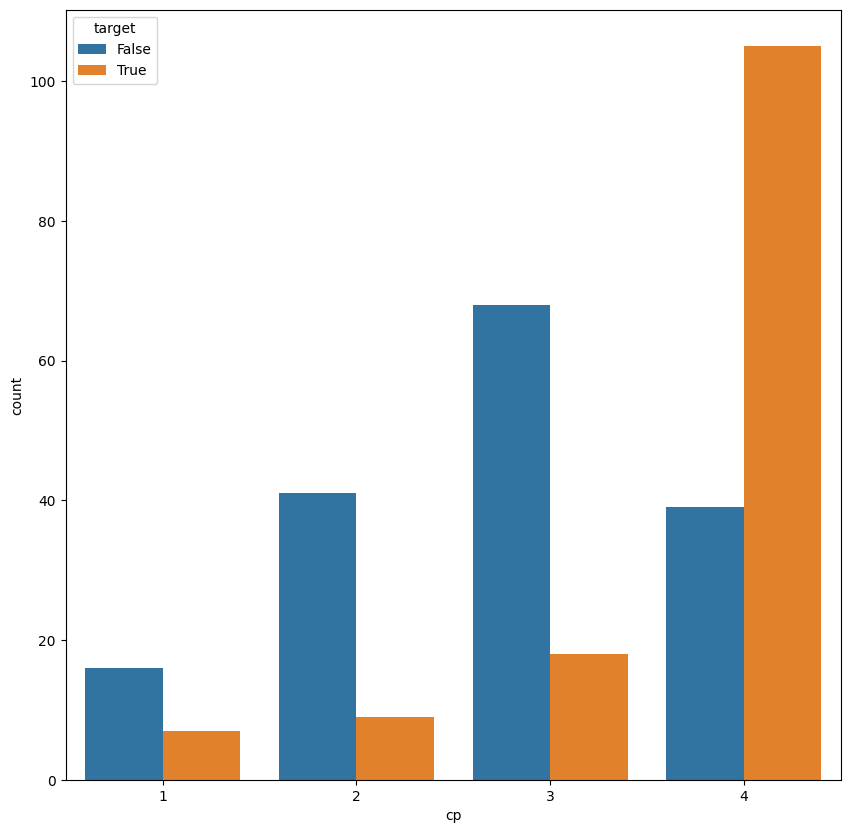


Figure 4 countplot for chest pain and the the presence of heart disease orange, and absence blue

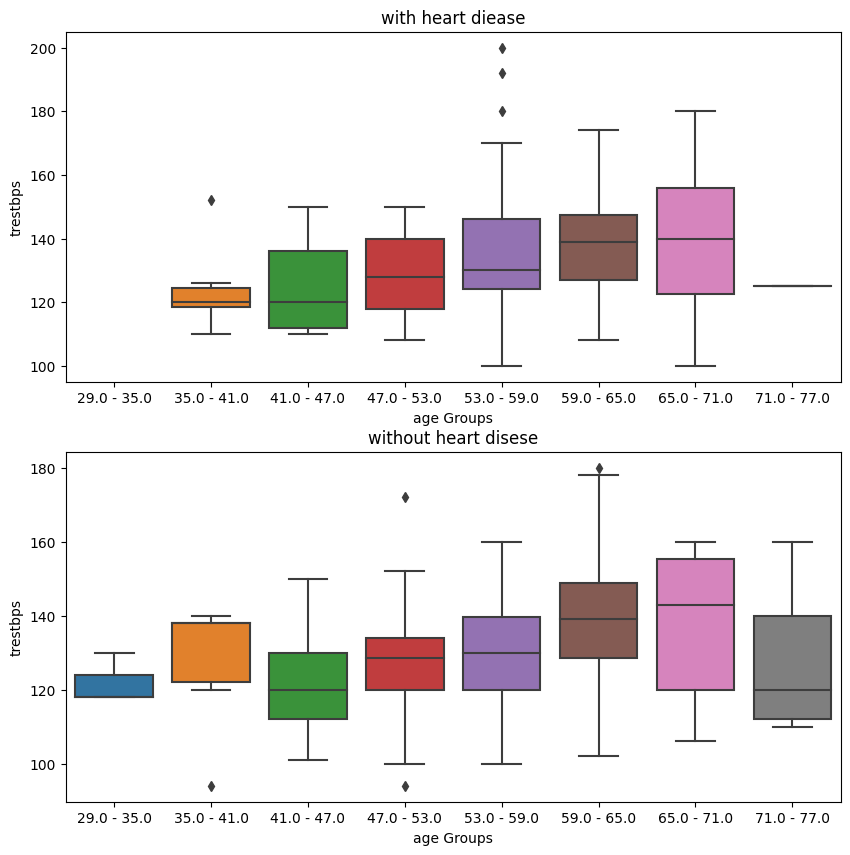


Figure 3 resting blood pressure accross age groups, seperated by the presence og heart disease.

As shown in figure 4 asymptomatic chest pain is highly prelevent with cases with heart disease, and might be a good indicator for the presence of heart disease.

# Models results and discussion.

In this study we have experimented with three models to try and predict the presence of heart disease.

Because each model needs a some transformations before it can make an accurate predictions, we have used a pipeline designs in which each pipeline contains all the transformation steps and finally the estimator always as the final step.

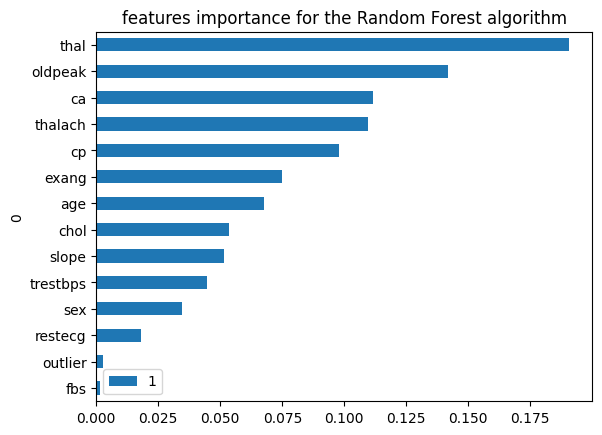
This approach will allow us to optimize all the pipeline parameters and search for the best fitted model.

For optimizing hyperparameters, we used the grid search cv method from scikit-learn[6],

The cv stand for cross validation, in which we train the model on part of the dataset, and holding part of it as a validation set, in each iteration the model is trained and evaluated with different hyperparameters, then a the best fitted model is selected, and can be used for further evaluation and deployment.

In our experiment, we have used, minmax scaler, for all our model, and PCA, with svm, and xgboost classifier, for random forest we used only minmax scaler, however in theory we don’t have to scale the values, for tree based model, because there are no distance calculations between datapoints.

Figure 5 features importance random forest



Svm scored 86% Random Forest and xgboost both scored 83% accuracy with comparable f1-scores. The results from SVM was significantly higher than both RF and xgboost, although we haven’t used PCA with random forest, the removal of this step in svm or xgboost didn’t yield a better result, while adding it to the random forest model lowered the performance of the model.

Figure 5 showed the importance of features, which is related to the mean decrease of impurity in the tree[7], which can show us which feature did the model used the most to be able to detect the outcome.

We can see that chest pain thalassemia, and old peeks in ecu were the top 3 features for the model.

Interestingly enough, our model did not rely on the outlier feature, to make the predictions, we suspect that because the model have already made most of the predictions, the outliers feature become redundant.

None of the model was in fact perfect or near perfection, we suspect that there are not enough data to be able to cover all basis.

# Conclusion

While this study have explained and helped us understand the relationship between heart disease and clinical features, it haven’t been able to accurately predict the presence of heart disease, which might suggest that we need more data, and more complex models to make the predictions.

Machine could significantly help healthcare professionals, in the diagnosis of diseases, however more research needs to be done in order to improve the relationship between machine and doctors, to make the patient life easier.

# References

[1] WHO, “https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death#:~:text=The%20top%20global%20causes%20of,birth%20asphyxia%20and%20birth%20trauma%2C.”

[2] K. E. Di Palo and N. J. Barone, “Hypertension and Heart Failure: Prevention, Targets, and Treatment,” *Heart Fail Clin*, vol. 16, no. 1, pp. 99–106, Jan. 2020, doi: 10.1016/j.hfc.2019.09.001.

[3] R. C. Detrano *et al.*, “International application of a new probability algorithm for the diagnosis of coronary artery disease.,” *Am J Cardiol*, vol. 64 5, pp. 304–10, 1989, [Online]. Available: https://api.semanticscholar.org/CorpusID:23545303

[4] IEEE Computational Intelligence Society and Institute of Electrical and Electronics Engineers, *2020 IEEE Symposium Series on Computational Intelligence (SSCI).*

[5] S. Howard, “The Elliptical Envelope,” 2007.

[6] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830.

[7] Mohammed Saad, “https://medium.com/@eng.mohammed.saad.18/detailed-explanation-of-random-forests-features-importance-bias-8755d26ac3bc.”