Predicting Heart Failure
DATA1030 Final Report
Will Jaekle
Brown Data Science Initiative
Project link

Introduction

This report serves to investigate the Heart Disease dataset and prepare its contents for a machine learning algorithm. The data set contains 12 attributes across 918 patients. Through the use of this data we hope to develop an algorithm which can accurately identify those at severe risk of heart disease or a severe cardiac event. The eleven features are common measurements of a patient's heart's health. The target variable, titled *HeartDisease*, is a Boolean which indicates whether the patient has a common cardiac disease. Predicting this trait is a classification problem. Heart disease is the leading cause of death worldwide, accounting for 17.9 million deaths annually according to the World Health Organization. Being able to accurately predict those at risk would help to initiate preventative measures sooner and potentially save lives.

The data is taken from two us states and two European nations, with such a distribution this sample set can be treated as independently, identically distributed. The data presented contains: *Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak and ST_Slope*. The feature *ChestPainType* contains one of four classifications of chest pain: *TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic* which describes the presence of chest pain which can or cannot be described as angina. The feature *FastingBS* contains a boolean which is true if blood sugar is greater than 120 mg/dl while fasting. *MaxHR* refers to the maximum observed heartrate of the patient. *RestingECG* is a feature which categorically describes the electro-cardiogram results of the patient where ST indicates ST wave abnormalities and LVH indicates Left ventricle hypertrophy. *Oldpeak* and *ST_slope* refer to the flat section of the pulse displayed on an electro cardio gram, where ST_slope is a categorical description of the slope and Oldpeak is the value of the ECG during exercise. Lastly, *ExerciseAngina* is a boolean which is true if the patient has exercise induced angina.

This dataset is published on Kaggle.com and has several projects associated with it, all of which attempt to predict the heart disease attribute. The accuracy rate of these models range in the 80th percentile. The best of these results uses logistic regression to obtain a model with 88% accuracy, other models were close behind.

EDA

Initially there are 918 unique rows each of which corresponding to a patient. Of our sample, 508 (55%) patients had some form of heart disease or failure. No null or NAN values were present in the dataset, however there are several values such as cholesterol and blood pressure which patients were listed as having 0 in. Zero in these fields is impossible and thus the data point is considered missing. Additionally, many of the oldpeak values are zero or negative, zero is possible and a healthy result however negative values are considered outside of the range of reasonable values, possible only through abnormal circumstances or measurement error. It is possible that oldpeak values which were not measured were assigned to the zero value as a result.

In general minimal correlation between the features was discovered *See fig 1*.

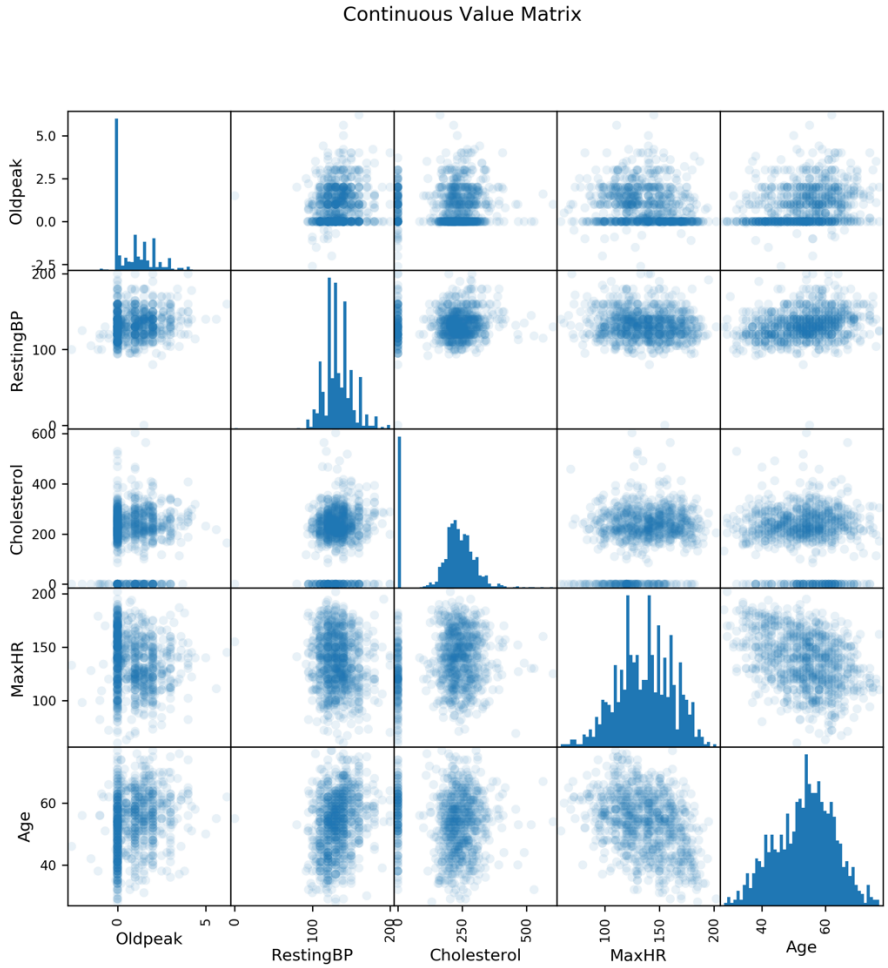

Figure 1. A scatter matrix displaying the continuous features comparatively.

This is unsurprising due to the commonality of the metrics and their frequent use as independent indicators. This lack of correlation will not prevent a sufficient model.

We can also begin to compare our features to the target variable. There are several important relationships to be aware of, the *ST_slope* is an excellent indicator of heart disease within a patient *see fig 2.* We can see that those with a downward sloping ST and those with a level ST are extremely likely to have heart disease in one form or another, so this indicator will likely inform our model to a greater extent than others.

Another interesting feature is the MaxHR feature, where a bimodal distribution appears and those with lower maximum heart rates are more likely to have heart disease, although maximum heartrate by itself is not sufficient to be conclusive *see fig 3*.

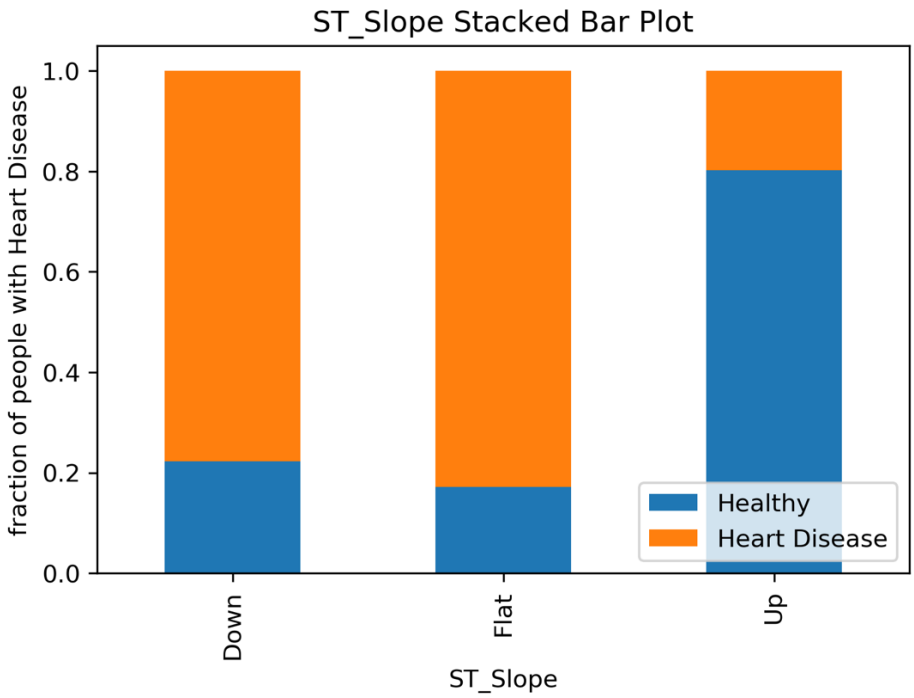

Figure 2. A stacked bar plot displaying the relationship between ST slope and heart disease

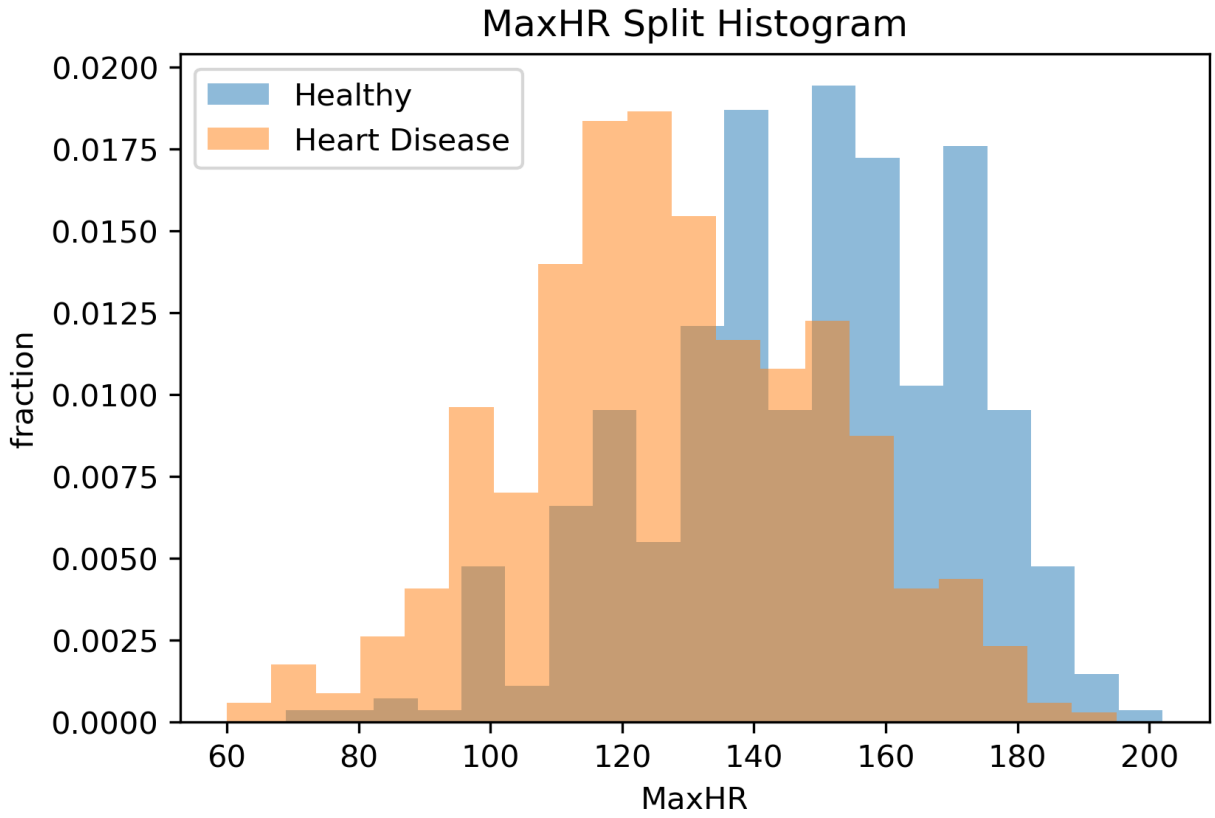

Fig 3. A split Max Heart Rate histogram showing the frequency of MaxHR compared to the presence of heart disease.

Overall, very little work needs to be done on this dataset during preprocessing. With some missing values clearly being written to zero it is unclear what other values were fabricated in the same manner, however there is no reasonable way to address this issue.

Methods

Before splitting the data set, I utilized scikit learns iterative imputation, which it should be noted, is an experimental package. It was used to calculate cholesterol for 172 patients and resting blood pressure for 1 patient. This is a medical scenario and as such imputation needs to be used with a caution, however in this case I believe the other features can provide meaningful insight to the cholesterol value. Cholesterol is a routine test for heart health and is only omitted frequently and intentionally in cases where heart disease is unlikely, such as when treating children, it is never omitted intentionally for patients at risk of heart failure and as such I believe the missing values to be random.

After imputation the data is split into the relatively standard split of 70% train, 15% validation and 15% test. Here paths diverge slightly, as one of the machine learning models, XGBoost, performs without imputed data, so it instead was trained on the original data, split with the same ratios. Overall 4 models were trained, XGBoost, Random forest, Logistic regression, and support vector machine.

The evaluation metric focused on primarily is recall, as this is a medical machine learning algorithms there are serious consequences to false negatives. Specifically when considering heart failure rates, false positives would amount to more intensive monitoring of those at lower risk, the consequences of which are significant but pale in comparison to the dire consequences of telling someone at high risk they do not need to be as cautious. For context the baseline recall score is .525, all of the models outperformed this baseline by at least two standard deviations. Other metrics such as accuracy and recall were considered however the models' scoring mechanic centered around recall. Here is the summary plot of how our models performed *see fig 4*.

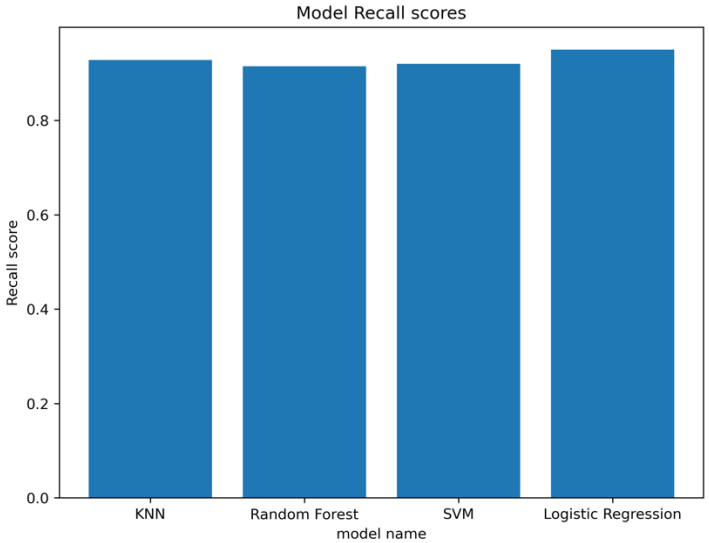

Fig 4. All models recall scores.

It is clear that the trade off of using more complex models is not necessarily worthwhile when training off of recall; logistic regression and KNN managed to stay competitive while also being easier to interpret than say SVM. Grid search was used in all cases, although run time was a serious limiting factor with random forest.

Results

The scores easily exceed the baseline recall of .5, all models exceeded a 90% recall which is above 2 standard deviations away from the baseline recall. Baseline accuracy was 50% as our data set was well balanced. Baseline precision was 55% and baseline f1_score was 52%. The model which was the most predictive was SVM, although not with the highest recalls, it still managed a better accuracy than other models.

The most important feature was easily ST_slope, and this makes sense as it was a metric created by doctors to identify high risk patients, as such it naturally has a nigh correlation with heart failure. The same can be said for old peak, which was not as influential as ST_slope, however once again is a recently fabricated metric specifically designed to help identify heart health.

Outlook

The interpretability of the support vector machine model is not as poor as some other more complex models, but it is not easily understood either. In order to improve the models within this report I would hope to engineer or obtain more features and data likewise. One could collect family history or perhaps feed in a larger swath of tests that might have a less obvious correlation to heart health. I also would like to allow for more detailed parameter fitting through longer times for computation.

References

Word Health Organization. (n.d.). *Cardiovascular diseases*. World Health Organization. Retrieved December 8, 2021, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

Data source:
fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Github Repository:

Included is all of the code, data and figures used and generated during this report.
https://github.com/wjaekle/heart_project.git