Will Jaekle
Data1030
10/12/21

Heart Disease Preprocessing Midterm Report

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.

Exploratory Data Analysis

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*.

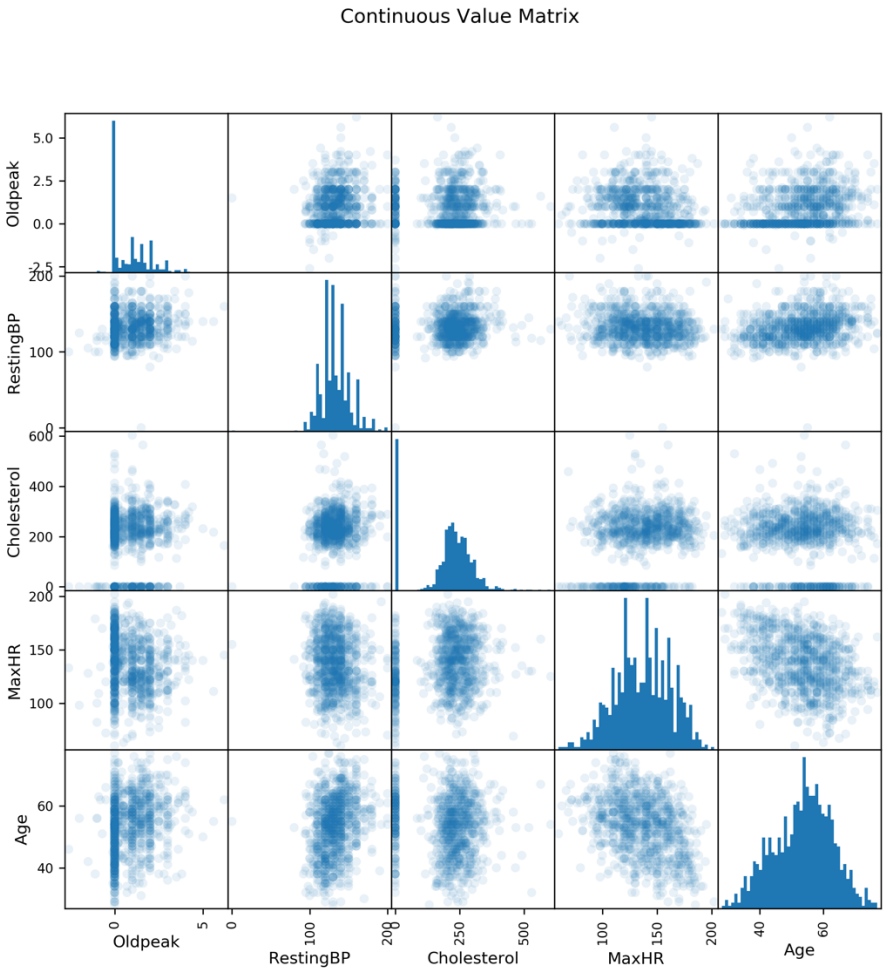

Figure 1. A scatter matrix displaying the continuous features comparatively.

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*.

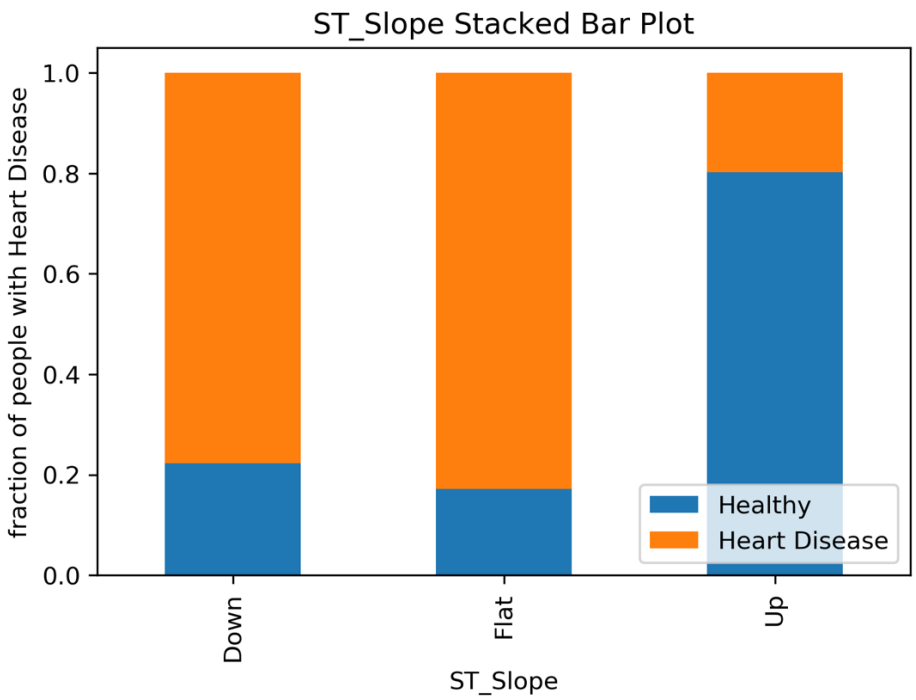

Figure 2. A stacked bar plot displaying the relationship between ST slope and heart disease

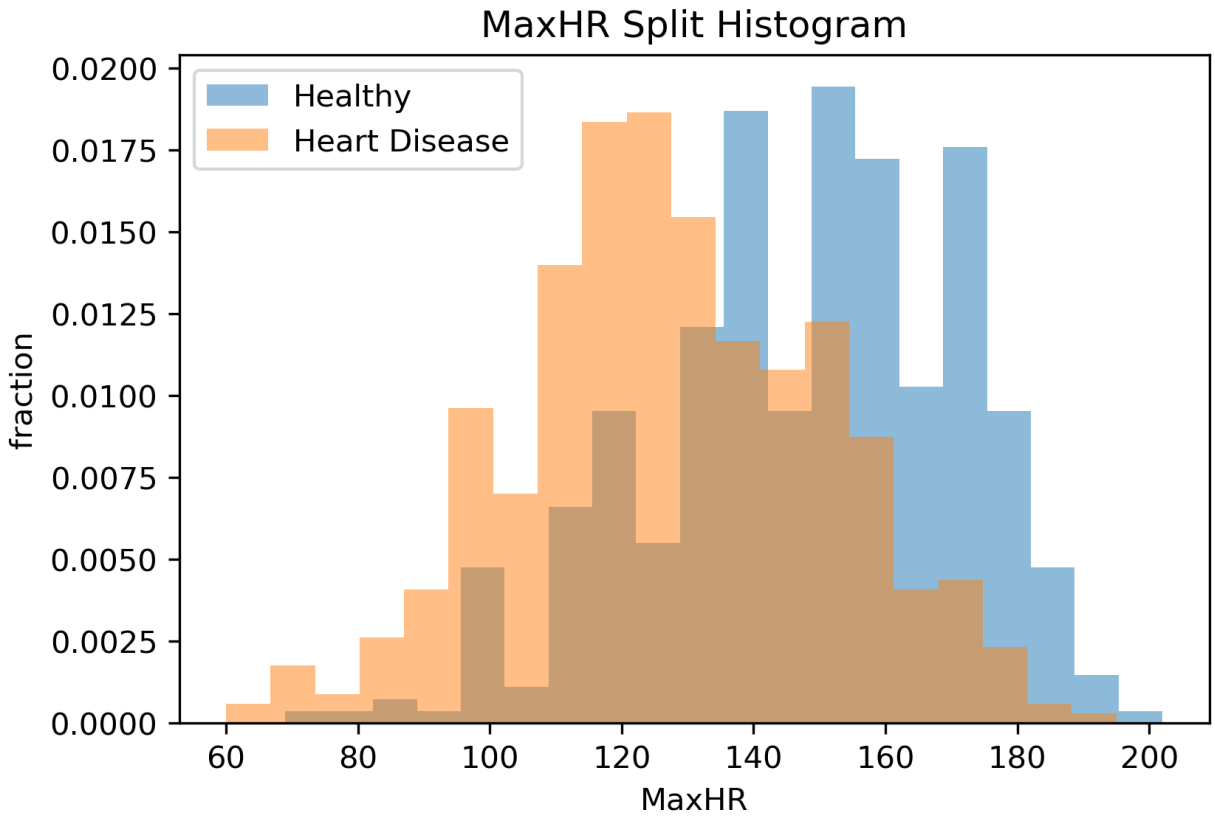

Fig 3. A split Max Heart Rate histogram showing the frequency of MaxHR compared to the presence of heart disease.

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.

Preprocessing

The dataset is comprised of eleven features, five of which are continuous and the other six being categorical. None of the categorical components are ordinal, so we can encode using onehot encoding for every categorical feature. These include Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, and ST_Slope. Since there are no missing values assigning them their own category is unnecessary. Of the continuous data, *Age, MaxHR, Cholesterol,* and *restingBP* are good candidates for minmax encoding while the *Oldpeak* and can be encoded with a standard scaler. The features which have minmax applied to them are such that there is reasonable bounds to their values. Normally, *Oldpeak* would be subject to the bounding of [0,6.5] however this is not true for this dataset so I will apply a standard scaling. Before encoding, values which are obviously missing can be assigned NaN to prevent skewing.

The data set as previously discussed is independent and identically distributed. This allows for the use of train_test_split. With 918 data points the starting split will include a training set comprised of 70% of the data and validation and test sets comprised of 15% of the data each. The splits will be entirely random. As there is no group structure to the data nor is there a time dependence. After preprocessing there are a total of twenty-one features.

References

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

Heart disease related deaths from WHO:
https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

Github Repository:

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