What are the Best Medical Markers to Predict Heart Disease?

Hall C, Sara, 20 December 2022

https://github.com/SaraHall22/Data-Science-Assessment

University of Bristol, Langford House, Langford, Somerset, BS40 5DU, UK

Introduction

Heart disease, medically referred to as coronary artery disease (CAD) is the greatest cause of mortality globally.[1] It is not only a disease that can cause significant disability, and be life limiting, but is a huge economic burden to both governments and individuals.

It is therefore of enormous value to be able to accurately predict when a patient may be suffering from heart disease.

The objective of this project was to analyse data collected, in 1988, from a sample set of 1025 patients from four geographical locations. In each case a number of co variates i.e., health markers were measured, and the target variable (y) was a categorical indicator confirming the presence of heart disease.

Objective

The programming language Python was used to visually analyse the dataset and to determine through Machine Learning modelling if there were any variables that can be used within a medical environment to provide a prediction as to whether a patient is likely to have heart disease.

Methods

The original dataset can be downloaded at https://github.com/SaraHall22/Data-Science-Assessment. There is also a README.md file for instruction on loading the Python notebook script file (Heartdata.ipynb) into your preferred interface. The script is fully annotated and can be run in its entirety to see all stages of the analysis.

Initially the data was visually explored, and basic analysis completed; alongside identifying the co-variate health markers as: **age** (in years), **sex** (1=male, 0= female), **cp** | Chest pain type (4 values), **trestbps** (Resting blood pressure in mm

Hg on admission to hospital), **chol** (Serum cholesterol in mg/dl), **fbs** | fasting blood sugar > 120 mg/dl (where 1=true, 0=false), **restecg** (resting electrocardiographic results (0 = hypertrophy 1 = having ST-T wave abnormality, 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), **ca** (Number of major vessels (0-3) coloured by fluoroscopy and **thal** (thalium stress test Result 0 = normal 1 = fixed defect; 2 = reversable defect).

The target variable was identified and separated from the dataset and defined as y. It was a categorical indicator confirming the **presence of heart disease** (0= no disease, 1= disease). The remaining variables were then treated as the co-variate set X.

Multiple cross correlation was explored in the covariate set X and visually demonstrated using a Heatmap. To further interpret the interaction between variables a scatter matrix was plotted. This allowed the identification of “clustering” of patients with disease and without, demonstrated by the separation of colour within the plots. Where clustering was visible for co-variables the relationships were replotted to a larger scale to visually investigate the relationship further.

Nearest Neighbours Analysis (k-NN) was then performed on the two co-variates deemed to show the strongest evident clustering with a known correlation. The model was scored and plotted.

A Pipeline of StandardScaler (to standardise the data in all variables), Principal Component Analysis > (PCA) (to reduce dimensionality and consider all variables and finally> k-NN, for just two principal components (PC1 and PC2) was then run to try and improve the model’s performance. Further performance improvements were then attempted using GridSearchCV, trying a range (1,5) of principal components within the model.

Results

The mean age for the patient sample set was found to be 54 years old and there were 526 patients with confirmed heart disease, 499 without. There was 713 males and 312 female patients. Histograms confirmed there were five continuous variables and eight ca

tegorical.

Multiple Cross Correlation indicated four weak correlations between the covariates a) age and resting blood pressure (0.27), b) cholesterol and age (0.22), maximum heart rate achieved, and ST depression induced by exercise relative to rest (-0.35) and d) age and maximum heart rate achieved (-0.39).

These correlations could all be visualised within a Heatmap and further understood using a scatter matrix.

The Scatter matrix showed that when considering the already identified correlations that there was evident clustering in those patients with heart disease and those without. Where clustering was evident the plots were run individually at a larger scale. The correlation visually showing the greatest clustering was identified in the association between variables oldpeak (ST depression) and thalach (maximum heart rate), highlighting them as predictors of heart disease. However, the variance in values indicated that the data should be standardised. (**Fig. 1**).

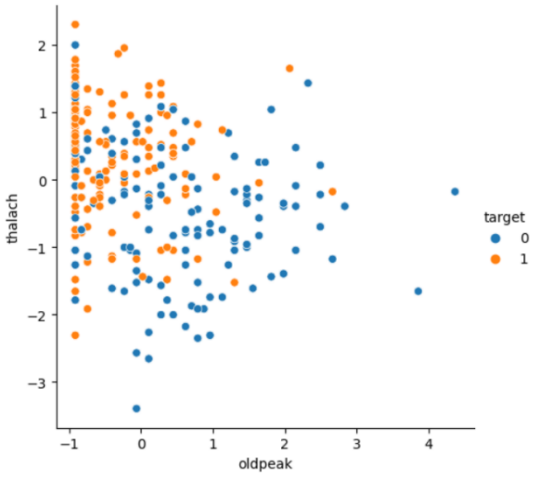

Fig. 1 A plot of the co-variates *oldpeak* and *thalach*, with standardised data. Clustering is evident. Patients with heart disease (orange) appear to have lower ST depression (oldpeak) and higher maximum heart rates (thalach) than those patients with no heart disease (blue).

The supervised technique, K -nearest neighbours algorithm was then run, using the two variables of interest (oldpeak and thalach) and a score of 0.87 was achieved.

The K -NN model was then re-run with all variables included, having first transformed and standardised the data. The model score improved to 0.97 indicating that the model was accounting for all but 3% of the data.

Finally, a pipeline StandardScaler>PCA>K-NN where the PCA was set to two principal components was run. The score remained at 0.97 but it was possible to plot a 2-dimensional decision boundary, for heart disease (**Fig. 3)**

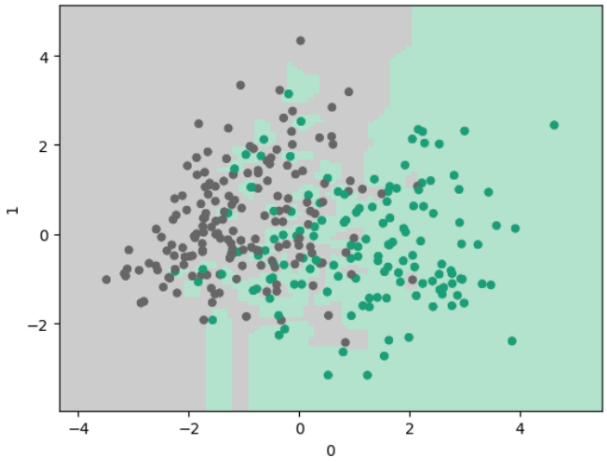

Fig. 3 A plot of the decision boundary for the K-nn algorithm when run at the end of a pipeline where all variables (X) were standardised before running PCA. PC1 = 0 and PC2=1. Grey = Heart disease, Green = No heart disease.

The variance in data, explained by PC1 and PC2 was 33% but the GridSearchCV tool indicated that the optimal number of principal components to use was two and that inclusion of further components would lead to no further improvement in the models score.

Analysis of the PCA loading showed cholesterol within PC1 and thalach (the patients maximum heart rate) within PC2 as the most indicative markers.

Conclusion

A patient's maximum heart rate and cholesterol score were not shown to be correlated within the heatmap or scattermatrix and so can be considered independent variables that are useful variables to be used for the prediction of heart disease.

Further analysis within Python should look to split the data set into male and female patients to consider whether the predictors differ dependent on gender.

The PCA did not improve the models score but in reducing the dimensionality it is possible to create a decision boundary plot that with further development could potentially be used by medical professionals to determine if a patient is likely to have heart disease.

Ref: [1] Ralapanawa U, Sivakanesan R, Epidemiology and the Magnitude of Coronary Artery Disease and Acute Coronary Syndrome: A Narrative Review, Journal of Epidemiology and Global Health, Vol. 11(2), June (2021) pp. 169-177