LINEAR CLASSIFICATION PREDICTION OF HEART DISEASE

Anjali Ramesh, Saroop Samra, Harmeena Sandhu, Ashna Sood, Urmi Suresh

University of California, San Diego, Department of Cognitive Science

Background

Healthcare is a vital point of research in order to best help patients with certain conditions. Heart disease specifically often has no symptoms, and yet if untreated can be a large contributor to more severe health conditions such as a stroke or heart attack **[2]**.

According to the CDC, 1 in 3 U.S. adults are not even aware of their high blood pressure which is often a leading contributor to heart disease means their blood pressure is going untreated **[3]**.

Due to heart disease presenting itself almost invisibly, we wanted to further understand what can best predict if a patient has heart disease (or is more prone to get this disease), which is useful information that could help diagnose someone and potentially provide early preventiative care.

Research Question

Using the provided dataset, which attribute(s) and classification algorithm can be used to best predict if a patient has heart disease?

Dataset

To answer this question, we are looking at the UCI Statlog (Heart) data set **[1]**.

This dataset consists of 14 attributes with 270 samples. The attributes are the following:

- age
- sex
- chest pain type
- blood pressure
- cholesterol
- blood sugar
- resting ECG
- heart rate
- angina,
- ST depression
- peak exercise ST segment,
- blood vessels
- thal
- **Label:** heart disease (1 or 2)

While performing our analysis, we used one hot encoding to turn the nominal variables into usable values.

Methods

We created **four different linear classification models** with different combinations of variables , which attempt to narrow down the best predictors of heart disease.

Model 1: all continuous and binary variables, excluding the nominal variables

Classifier = Perceptron

Variables = age, gender, cholesterol, blood pressure, blood sugar, heart_rate, angina, ST depression, peak_exercise ST segment, blood vessels

Accuracy = 68.5%
Precision = 67.7%
Recall = 83.3%
F1= 74.6%

Model 2: all variables, using one-hot encoding

Classifier = Logisitic Regression

Variables = Model 1 Variables +
One Hot Encoding chest pain,
One Hot Encoding electrocardiographic
One Hot Encoding thal type

Accuracy = 77.8%
Precision = 65.4%
Recall = 85.0%
F1 = 73.9%

Model 3: Four Highest scoring variables

Classifier = Ridge

Variables = heart rate, angina, ST depression, peak_exercise ST segment, blood vessels, One Hot Encoding chest pain, One Hot Encoding electrocardiographic, One Hot Encoding thal type

Accuracy = 88.9%
Precision = 90.3%
Recall = 90.3%
F1 = 90.3%

Model 4: Two Highest scoring variables

Classifier = Ridge

Variables = One Hot Encoding chest pain, One Hot Encoding thal type

Accuracy = 74.1%
Precision = 71.8%
Recall = 90.3%
F1 = 80.0%

Results

Of the four models of classification we created Model 3, using a ridge classifier, performed the best based on the F1 score, followed by Model 4, Model 1, and Model 2.

Model 3 has an accuracy of 88.9%, meaning this is the chance we get a correct prediction (total number of correct predictions divided by the total number of predictions made for a dataset). However, we also need to take into account the imbalance in the dataset which may have more or less heart disease patients. The precision for Model 3 is 90.3% and represents the number of positive heart disease predictions that actually belong to the positive heart disease class. The recall for Model 3 is also 90.3% and represents the number of positive heart disease predictions made out of all heart disease examples in the dataset. Finally, the F1 score is 90.3% which takes into account the concerns of precision and recall in one number.

From the F1 scores we can see that reduced attributes in Model 3 actually outperforms Model 2 which had 5 more attributes. By adding more attributes, the test score did not improve and as well as a better test score in Model 3 it is a more robust and less susceptible to overfitting.

Discussion

Answering our research question, heart rate, angina, the ST measures, chest pain, electrocardiographic and thal type with a ridge classifier are the best attributes and model respectively to predict if a patient has heart disease.

Ultimately, developing a test/train split and then measuring the F1 score on the 4 models provided gave us confidence on selecting the variables and best classification algorithm yielding over a F1 score of 90%.

Future Plans:
- Implement a decision stump to further validate our models
- Comparing the results in other similar heart disease datasets
- Finding a larger dataset to allow us to be more certain of our results

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.
[2] https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure
[3] https://www.cdc.gov/bloodpressure/5_surprising_facts.htm