**Predicting Heart Disease**

Ethan Munroe

5/10/2024

Heart disease is one of the leading causes of death in the United States of America according to the CDC (*NCHS Data Brief*). I found a dataset collected from Cleveland, Hungary, Switzerland, and Virginia, that includes results from a heart disease study. The participants were observed through various tests, and data was collected on the following…

* Age
* Sex
* Chest pain type
* Resting blood pressure
* Serum cholesterol
* Fasting blood sugar
* Resting electrocardiogram results
* Maximum heart rate
* Exercise induced angina
* Peak of ST segment
* The slope of the peak exercise ST segment
* Heart disease status

I am going to explore that data further with predictive modeling to create an algorithm to classify people and catch future potential heart disease based on the variables above.

Before diving straight into the modeling, I wanted to investigate some of the descriptive statistics, and some distributions of the variables themselves.

**Distributions:**

A graph of different sizes and colors

Description automatically generated with medium confidence

Figure A: The distributions for Age and Cholesterol in the data set.

A comparison of a normal distribution

Description automatically generated

Figure B: The distributions for Resting Blood Pressure and Max Heart Rate in the data set.

We can see that each of these variables seem to be normally distributed. However, the one thing that stands out in the cholesterol distribution is how many observations there are at `0 mg/dL', which according to the University of Rochester is extremely bad. They say that anything below 70 mg/dL is concerning and is a sign that you are at high risk for a heart attack, or you have already had a heart attack.

**Descriptive Statistics:**

Proportion of men: **0.763865546218487**Proportion of women: **0.2361344537815126**Mean Age: **53.72016806722689**Mean Cholesterol: **210.36386554621848**Mean Resting Blood Pressure: **132.15378151260504**Mean Max Heart Rate: **139.7327731092437**

The mean for cholesterol is worrisome because as we can see from the distribution, there is a large mass of data points at ‘0 mg/dL’ which negatively bias the mean. However, the mean is still 210.364 ml/dL, which according to John Hopkins Medicine, is “Borderline High” and above what they say is “normal” at 200 ml/dL. I’m going to assume that the people that have ‘0 mg/dL’ cholesterol just did not have their cholesterol measured and I’ll remove those from the data set.

I also want to dive deeper into the data to see which of the variables has a higher impact on people having heart disease. The first step into this is to use clustering to see which of the variables are. The first step in going about this is to scale the variables that go into the clusters.

**Elbow Method:**

**A graph with a line

Description automatically generated**

Figure C: Visualization of inertia against number of clusters (k) to find optimal number of clusters to use in k-means++ clustering.

By using the elbow method, we can see that somewhere around 4 clusters is when the marginal benefit of adding another cluster starts to really flatten out. Now by using these 4 clusters, I can apply the K-means++ algorithm for clustering to see any natural grouping within the data that can be seen through Age, Cholesterol, Resting Blood Pressure, and Max Heart Rate, the target variable (Heart Disease), and old peak = ST.

**Clustering:**

**(I) (II)**

A chart of different colored dots

Description automatically generated A chart with different colored dots

Description automatically generated

**(III) (IV)**

**A diagram of a number of dots

Description automatically generated** **A chart with different colored dots

Description automatically generated**

**(V)**

**A chart of different colored dots

Description automatically generated**

Figure D: (I): Visualization of Clusters with Age against Resting Blood Pressure. (II): Visualization of Clusters with Age against Cholesterol. (III): Visualization of Clusters with Age against Max Heart Rate. (IV): Visualization of Clusters with Cholesterol against Heart Disease. (V): Visualization of Clusters with Cholesterol against Old peak = ST.

From the first three graphs, you can distinguish the four different clusters. From Figure D(I) we can see that there is a cluster of older people with low to moderate blood pressure, a middle aged to older group with high blood pressure, a “younger” group with low to moderate blood pressure, and a middle-aged group with low to moderate blood pressure as well. Figures D(II) and D(III) follow similar patterns. However, because this is multidimensional data, the last two graphs do not show any clear relationship between the variables that are graphed, and we cannot distinguish four clear clusters.

We can also look at the average values of some of the variables within each cluster:

**Cluster 0 (Purple):**

**Mean Age:** 55.919075 **Mean Cholesterol:** 323.676301 **Mean Resting Blood Pressure:** 132.803468 **Mean Max Heart Rate:** 146.809249  
**Proportion of Heart Disease:** 0.508671

**Cluster 1 (Blue):**

**Mean Age:** 45.443548 **Mean Cholesterol:** 226.680108 **Mean Resting Blood Pressure:** 124.854839 **Mean Max Heart Rate:** 161.774194  
**Proportion of Heart Disease:** 0.231183

**Cluster 2 (Green):**

**Mean Age:** 59.844828 **Mean Cholesterol:** 250.931034 **Mean Resting Blood Pressure:** 159.114943 **Mean Max Heart Rate:** 133.413793  
**Proportion of Heart Disease:** 0.660920

**Cluster 3 (Yellow):**

**Mean Age:** 57.732441 **Mean Cholesterol:** 221.906355 **Mean Resting Blood Pressure:** 126.591973 **Mean Max Heart Rate:** 122.100334  
**Proportion of Heart Disease:** 0.628763

Figure E: Shows typical values of these four variables within each cluster, and the proportion of people within each cluster that he

From the table above, we can see how the clustering algorithm splits up the data and at what values it selected. For Cluster 0 specifically, we can see how high the mean cholesterol is within each cluster. It is also important to note from the proportion of heart disease within each cluster, that is seems like the majority of people within the study have heart disease, which could bias a classification towards having heart disease.

To get a better understanding of which of these variables weigh more into whether people have heart disease or not, I can run a logistic regression to get the coefficients on the log odds function to see which variables are weighed more. This is possible due to the target function being binary, so the coefficients on the log odds will represent how much each variable effect the probability.

To start this process, I created two different logistic regression models, and used k-fold cross validation to calculate the accuracies of the models. The first model only uses `Age`, `Sex`, `Resting Blood Pressure`, `Cholesterol`, `Fasting Blood Sugar`, `Max Heart Rate`, `Exercise`, and `Old peak` as variables. The second model uses every variable available in the data set to predict heart disease.

**Coefficients for Logistic Regression 1**

**Age**: 0.1693385

**Sex**: 0.5191619

**Resting Blood Pressure**: 0.0362069

**Cholesterol**: -0.2802334

**Fasting Blood Sugar**: 0.3674143

**Max Heart Rate**: -0.5180450

**Exercise Angina**: 0.7589122

**Old peak**: 0.7359082

**Accuracy of Model 1: 0.80336134453**

**Coefficients for Logistic Regression 2**

**Age**: 0.1640762

**Sex**: 0.6545488

**Resting Blood Pressure**: 0.127478

**Cholesterol**: -0.3513098

**Fasting Blood Sugar**: 0.3632633

**Max Heart Rate**: -0.1745938

**Exercise Angina**: 0.4149149

**Old peak**: 0.4428267

**Accuracy of Model 2: 0.8462184873**

Figure F: Results from Logistic Regressions and Cross Validation

As you would expect, the model we give more data is more accurate. However, there are some notable differences and similarities that are worth discussing. First, we will start with the variables that stay very similar in both models, like `Age`, and `Fasting Blood Sugar`. At least in our sample your age accounts for around 16 or 17 percent of the probability you have heart disease, and around 36 percent for your fasting blood sugar levels. (That is for the scaled variables in the model). Contrarily, all the other variable’s coefficients vary dramatically in each model. I will use Model 2 to discuss my results from here on out due to the improved accuracy. Like I mentioned before, the coefficients on each variable in the log odd function give us their “weights” to the overall probability of having heart disease. Therefore, we can see the leading variables like `Sex`, `Exercise Angina`, and `Old peak`.

I then used Python’s DecisionTreeClassifier from the sklearn.tree library, which I was able to tune hyperparameters for the decision tree to find the optimal depth and `ccp\_alpha` for the model. The following represents the results from the cross validation to find the optimal hyperparameters.

Hyperparameter Tuning:

Optimal Depth: **17**

Optimal\_CCP: **0.0**

Best mean accuracy: **0.9042**

Figure G: Results from the hyperparameter tuning for the decision tree.

As you can see from the cross validation, I allowed it to test all the way up to a depth of 20, and it shows that with a depth of 17 and a `ccp\_alpha` of 0, the tree will have the greatest accuracy.

**Final Decision Tree**

A screenshot of a computer screen

Description automatically generated

Figure H: Shows the final decision tree with depth: 17, and ccp\_alpha: 0.

I wanted to include the picture of the complex decision tree model, not so we can map any certain case out to predict whether they have heart disease or not, but rather to show the complexity. To acknowledge how both the improved accuracy and complexity of the decision tree shows the complex relationships within the data that the logistic regression did not catch.

This new knowledge of how these different variables can interact with each other to cause heart disease creates an entirely new problem. Even though through modeling using this dataset we can create a model that would be around 90% correct out of sample, there are still many unknowns when it comes to predicting heart disease. For example, how are variables like `Age` and `Cholesterol` interacting?

**Bibliography:**

[www.cdc.gov/nchs/data/databriefs/db456.pdf. Accessed 04 May 2024](http://www.cdc.gov/nchs/data/databriefs/db456.pdf.%20Accessed%2004%20May%202024).

Mexwell. “Heart Disease Dataset.” *Kaggle*, 8 Apr. 2024, www.kaggle.com/datasets/mexwell/heart-disease-dataset. Accessed 05 May 2024.