IDENTIFYING PATIENTS WITH HEART DISEASE

December 11, 2019

Students: Alec M. Petrack
Matthew Randles
Andrew Young
Instructor: Dr. Tanvi Banerjee
TA: Fan Yang
Wright State University

Abstract

Heart disease is the leading cause of death in the United States with more than 600,000 deaths each year [8]. Physicians, ideally, want to form a prognosis before their patient's health deteriorates and suffers from a heart attack or stroke. Unfortunately, current methods used to diagnose heart disease are invasive and expensive. In this paper, we investigate whether machine learning can accurately predict patient's with heart disease given non-invasive patient data. An exhaustive search of machine learning models is performed to determine which is most suitable for identifying patients with heart disease. Descriptive statistics are given to provide useful insight along with the justifications of our pre-processing methods. Then, several models are trained to determine which is most suitable for classifying heart disease based upon precision, recall, and F1 score.

I Introduction

Heart disease research is an increasing research topic because more people die from heart disease every year [8]. Unfortunately, current tests to detect heart disease are expensive and potentially dangerous to the patient.

One test that doctors use is cardiac magnetic resonance imaging (MRI), which creates images of the heart and major blood vessels, but the average cost of the MRI is between $1000 and $5000 [15]. Also, in some countries, availability of MRI equipment is limited, and patients may have to wait an extended period of time before an MRI can be taken, which puts the patient at risk during that time [16].

Another standard test is left heart catheterization. In this procedure, a thin tube is passed from the patients wrist, arm, or upper leg into the heart. The catheter is then moved though the aortic value to measure the pressure inside the heart as well as the heart's ability to pump blood. While this method is fairly accurate, it is invasive and runs the risk of complications or death to the patient. Left heart catheterization has a mortality rate of about 0.05%, or about 500 deaths per year in the United States [19].

One newer method of testing for heart disease is the coronary calcium scan. This test uses computed tomography (CT) scans to take pictures of the heart, and then the patient is given a score based on how much calcium is visible from the images. Based on the score, the doctor may recommend additional testing. There are multiple caveats for this test. Not all arteries that have early signs of heart disease have calcium, so it has a chance of missing patients that potentially are still at risk. A patient could also get a high score, but not necessarily be at risk for heart disease, so the patient may be required to have additional tests that are unnecessary. Also, not all health insurance companies pay for this test, so it can be expensive for the patient [17].

Because the current methods of detecting heart disease are expensive and potentially dangerous, it is important to gain more insight to the causes of heart disease so more efficient and safer options can be discovered. This research experiment attempts to understand more about heart disease by implementing multiple machine learning algorithms that explore patient biometrics collected from the Cleveland database [5]. In the next section, the specific biometric features are described. This is followed by section III, where a statistical analysis of features is implemented. Based on the feature analysis, data pre-processing methods are explored in section IV. The data is then fed through multiple machine learning models described in section V. The results are then reported in section VI, with some discussion points in section VII. General conclusions about the

research and results are given in section VIII.

II Descriptions of Features

The data used in the research was recorded by the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation [13]. The original dataset contains 76 features (biometrics) across 303 patients, but the data was reduced to 14 features by the owners to remove biometrics that were incomplete or did not have sufficient records to adequately process. This reduced dataset is the one used in this study.

Categorical Features		
Feature	Categories	Description
Sex	2	Gender
CP	4	Chest Pain
FBS	2	Fasting Blood Sugar > 120 mg
Restecg	3	Resting Electrocardiogram Results
Exang	2	Exercise Induced Angina
Thal	3	Thalassemia
Target	2	Has Heart Disease?

Figure 1: Categorical Features in Data Set

Numerical Features				
Feature	Mean	STD	Range	Description
Age	54.37	9.10	29-77	Age of patient
Trestbps	131.62	17.54	94-200	Resting blood pressure
Chol	246.26	51.83	126-564	Cholesterol
Thalach	149.65	22.91	71-202	Max Heart Rate Achieved
Oldpeak	1.04	1.16	0-6.2	ST Depression Induced by Exercise Relative to rest
Slope	1.40	0.62	0-2	Slope of the Peak Exercise Relative to Rest
Ca	0.73	1.02	0-4	Number of major blood vessels colored by fluoroscopy

Figure 2: Numerical Features in Data Set

III Statistical Analysis of Features

Our data set consisted of 13 features and 1 target variable with 303 observations. For machine learning, the number of observations with respect to features is very small, so our initial concerns were the models we use would tend to overfit. To help prevent overfitting, first, we performed principal component analysis (PCA) to see how much variance could be explained by each feature. Finally, we computed the correlation between all of the features to see if we could reduce the number of features by combining features into more significant features.

Before training models, we performed PCA see which variables may be the most influential in determining heart disease. Table 1 lists the features in order of most explained variance, to least explained variance. We anticipated cholesterol would play a major role in determining patients with/without heart disease and it makes sense that about 75% of the variance is explained by it. On the other hand, we thought sex would play a bigger

role considering the hormone estrogen naturally helps prevent the build up of calcium in arteries.

Principal Component Analysis		
PC	Feature	Variance (%)
1	chol	74.7564
2	thalach	15.0370
3	trestbps	8.4597
4	age	1.6216
5	oldpeak	0.0384
6	cp	0.0281
7	ca	0.0229
8	thal	0.0100
9	restecg	0.0077
10	slope	0.0059
11	sex	0.0050
12	exang	0.0041
13	fbs	0.0031

Next, we computed the correlation between all features to see if any features could be combined into more significant features. Our objective was to see if the number of features could be reduced, which would help prevent our models from over-fitting the data. The correlation between all of the features are reported in Figure 2.

The tiles represent the correlation between each feature with respect to the color bar on the right. Bright yellow indicates a strong positive correlation between the features, while dark blue represents a strong negative correlation. We considered features with at-least a positive or negative correlation of ± 0.75. None of the feature combinations were strongly correlated enough to justify combining features. Therefore, we didn't combine any features and proceeded with processing the features, so that the models could be trained and perform well.

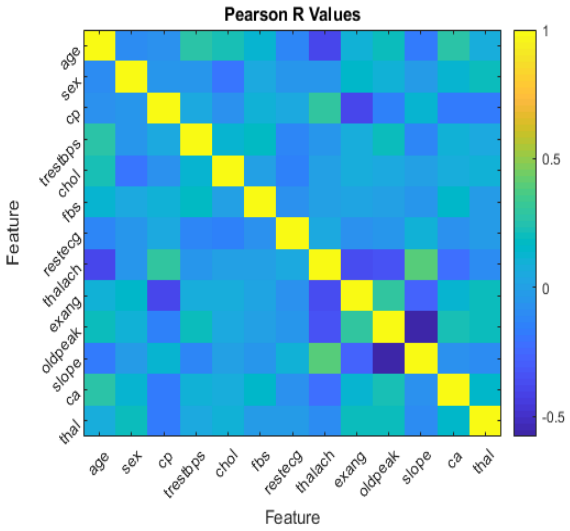

Figure 3: Pearson's Correlation Between All Features

IV Data Pre-Processing

In machine learning, very often, data-sets contain variables with different mean, standard deviation, and ranges of values. This poses a problem during training. If the scales of the variables being trained are not equal, it is likely the cost function associated with the machine learning will not be driven towards a lower value each epoch.

Our data sets variables had varying ranges, means, and standard deviation values as reported in Figures 1 and 2. Before we trained our models, we one hot-encoded the variables, which had more than two categories, and performed z-standardization on all of the numerical features. Our pre-processing methods are summarized in Figure 4.

Categorical Features	
Sex	Null
CP	One-hot
FBS	Null
Restecg	One-hot
Exang	Null
Thal	One-hot
Target	Null

Numerical Features	
Age	Z-Score
Trestbps	Z-Score
Chol	Z-Score
Thalach	Z-Score
Oldpeak	Z-Score
Slope	Z-Score
Ca	Z-Score

Figure 4: Pre-processing of Categorical and Numerical Features

V Model Overview

V.I K-Nearest Neighbor

The K-Nearest Neighbor model works on the idea that similar observations will inhabit similar areas in n-dimensional space, where n is the number of features. In the case of this experiment, 'closeness' between points is measured in Euclidean distance. When classifying a point, the algorithm will look at the k nearest points, and use those points (neighbors) to determine the prediction for the class. In this experiment, four different k values were tested: 3, 5, 7, and 9, with 5 giving the best results. K-nearest neighbor was chosen to test because it is presumed that patients with heart disease will exhibit similar bio-metric features, like high cholesterol or high blood pressure, thus inhabiting similar areas of feature space.

V.II Random Forest

A Random Forest model is made of up a large amount of decision trees that operate together as an ensemble. Each decision tree in the structure is responsible for making a classification prediction about the patient, or a vote on whether the patient should be classified as having heart disease. All of the predictions are tallied and the class with the highest total becomes the prediction for that observation. This method is effective because it doesn't rely on one individual prediction, but a collaboration of uncorrelated trees, with the idea that the prediction of the whole ensemble is going to be more accurate than one individual prediction. Another benefit of the Random Forest model is it reduces error caused by individual predictions. If one of the trees in the ensemble learns an incorrect assumption about the data, the other trees (assuming the majority of the trees are correct), will still result in the correct classification. Because of the uncorrelated nature of our features (see **Section III - Statistical Analysis of Features**), it is predicted that the random forest model will be able to accurately determine which features are most important to predicting the target, and as such have high accuracy results. Multiple numbers of trees were tested, with 14 being the optimal choice in this scenario.

V.III Multi-Layer Perceptron

In this experiment, a multi-layer perceptron (MLP) neural network model is tested. The MLP is a four-layer network (one input layer, two hidden layers, and one output layer). The input layer has 21 nodes (in accordance to the 21 features), the first hidden layer

has 8 nodes, the second hidden layer has 3 nodes, and the final output layer has one node. To determine the number of hidden nodes, as well as other parameters including learning rate, regularization, and activation function, an parameter suite was configured with multiple options for each parameter to explore. The neural network is exhaustively tested with each combination of parameters in the suite, and the best resulting model is then reported. The parameter suite is shown below:

```
parameter_space = {
    'hidden_layer_sizes': [(6,), (7,), (8,), (6, 6), (6, 4), (5, 3, 2),
                           (7, 3), (8, 3), (8, 3, 2), (8, 8), (8, 8, 2), (8, 8, 8), (8, 9), (9,)],
    'activation': ['tanh', 'relu'],
    'solver': ['lbfgs', 'sgd', 'adam'],
    'alpha': [0.0001, 0.05, 0.01, 1, 10, 20],
    'learning_rate': ['constant', 'adaptive'],
    'learning_rate_init': [0.001, 0.01, 0.05, 0.1, 0.5]
}
```

Figure 5: Parameter Suite for Multi-Layer Perceptron

The best resulting parameters from this exhaustive test is as follows:

```
clf = MLPClassifier(activation='tanh', alpha=0.05, hidden_layer_sizes=(8, 3), learning_rate='constant',
                    learning_rate_init=0.001, solver='sgd', max_iter=2000)
```

Figure 6: Best Parameters for Multi-Layer Perceptron

The activation function chosen, hyperbolic tangent (tanh), is a re-scaled version of a sigmoid function so the range of the sigmoid is between -1 and 1. There are multiple reasons to chose this activation function over a normal sigmoid. The data that is fed into the neural network is z-score normalized (zero mean and unit variance), so it seems natural to pick an activation function that is also within that scale as to not introduce a systematic bias. Also, because the range of the hyperbolic tangent function is larger than that of a regular sigmoid, the gradients along the curve will be larger, thus resulting in the algorithm converging to a local minima faster (larger increases or decreases in weights).

A multi-layer perceptron model was chosen for the experiment because it is known to perform better with non-linear data, with the ability to 'learn' the most optimal solution though back-propagation techniques if given enough data.

VI Results

Each model was trained using 80% of the observations (243 patients), which was randomized and shuffled beforehand, and 20% of the data was set aside before training for validation (61 patients). After training, the results were recorded by classifying each record in the validation set, then comparing the result to the pre-determined 'target' value for accuracy. Then the data was re-shuffled and a new training and validation set was generated. This process was repeated 5 times (5-fold validation). The performance of each model was then evaluated using multiple metrics that pertain to the percentage or correct and incorrect classifications: accuracy, precision, specificity, sensitivity, F1-score, and AUC (area under the ROC curve). The best results for each model and reported below:

Results from Machine Learning Models						
Model	Accuracy	Precision	Specificity	Sensitivity	F1-Score	AUC
Multi-Variable Logistic Regression	0.88	0.86	0.90	0.76	0.81	0.88
Support Vector Machine	0.88	0.83	0.87	0.76	0.76	0.88
K-Nearest Neighbor	0.87	0.86	0.88	0.86	0.86	0.87
Random Forest	0.89	0.91	0.94	0.80	0.85	0.95
Muli-Layer Perceptron	0.85	0.88	0.88	0.85	0.87	0.95

Figure 7: Accuracy, Presicion, Specificity, Sensitivity, F1-Score, and AUC for Machine Learning models Trained

Overall, the experiment was successful in implementing machine learning algorithms to accurately predict heart disease. It is also noted that there have been multiple machine learning studies using the same data set with comparable, promising results [6][7]. The model that would be selected in a medical environment though is debatable depending on the importance of false positive (FP) and false negative (FN) predictions and the interpretability of features to the predicted output. The random forest model has the best overall accuracy (0.89), as it has the greatest amount of correct predictions, but the k-nearest neighbor model has the best sensitivity rating (0.86 compared to 0.80 from random forest). As sensitivity compares the true positive count (TP) to the false negative count, the higher the sensitivity score, the less likely the model is to incorrectly predict a patient does not have heart disease when a patient could actually be at risk. The random forest model has the highest precision (0.91), which minimizes the number of false positives, so a higher precision would prevent unnecessary additional testing which would occur if the patient was incorrectly diagnosed with heart disease. The multi-layer perceptron model did not perform the best in either precision (0.88) or sensitivity (0.85), but is very close to the best results in both, and has the best F1 score (0.87). For this reason, the multi-layer perceptron might be used if both a high precision and sensitivity were desired. The support vector machine model has a high accuracy (0.88), but fairly low sensitivity (0.76) and F1 score (0.76), so most likely this model would not be used in a medical environment. The logistic regression model did not perform better than the other models, but had a high accuracy (0.88) and precision (0.86), and the model is simple to understand and explain. If interpretability was important, this model may be used as opposed to the other "black-box" techniques.

VII Conclusion and Future Work

The goal of this experiment was to research the causes of heart disease and use the gained knowledge and machine learning techniques to accurately predict the presence of heart disease in patients. Multiple bio-metric features were explored and statistical analysis demonstrated potential to predict heart disease using these features. Multiple machine learning models were then configured and analyzed, and the results were promising. Multiple models demonstrated high accuracy, and the random forest model resulted in precision and specificity ratings of 0.91 and 0.94 respectively. It is debatable which model(s) would be used in a medical environment though, as false positive and false negative rates need to be considered as well as the accuracy.

In future work, more data would be desirable to ensure confidence in the models. The original dataset had many missing features, so more complete data would provide the ability to explore features that weren't included in this experiment. This would be implemented by surveying other hospitals around the United States with high diagnosis accuracy for medical data. Acquiring more data for multiple locations would also help eliminate any inherit geographical biases in the data.

Overall, the experiment reports promising results and concludes research into the predictability of heart disease should be further explored.

VIII References

[1] S. Niculescu and N. Chi Lam, "Geographic Object-Based Image Analysis of Changes in Land Cover in the Coastal Zones of the Red River Delta (Vietnam)," Journal of Environmental Protection, vol. 10, no. 10.4236/jep.2019.103024, pp. 413-430, 2019.

[2] M. Bazmara, S. Movahed and S. Ramadhani, "KNN Algorithm for Consulting Behavioral Disorders in Children," Journal of Basic and Applied Scientific Research, vol. 3, pp. 325-358, 2013.

[3] "Underlying Cause of Death 1999-2017," CDC, November 2019. [Online]. Available: `https://wonder.cdc.gov/wonder/help/ucd.html`. [Accessed November 2019].

[4] Statista Research Department, "U.S. Physicians - Statistics & Facts," Statista, 28 November 2019. [Online]. Available: `https://www.statista.com/topics/1244/physicians/`. [Accessed November 2019].

[5] ronit, "Heart Disease UCI," Kaggle, 25 June 2018. [Online]. Available: `https://www.kaggle.com/ronitf/heart-disease-uci`. [Accessed November 2019].

[6] R. Harrand, "What Causes Heart Disease? Explaining the Model," Kaggle, 4 March 2019. [Online]. Available: `https://www.kaggle.com/tentotheminus9/what-causes-heart-disease-explaining-the-model`. [Accessed November 2019].

[7] C. Dabakoglu, "Heart Disease - Classifications (Machine Learning)," Kaggle, 7 August 2019. [Online]. Available: `https://www.kaggle.com/cdabakoglu/heart-disease-classifications-machine-learning`. [Accessed November 2019].

[8] Centers for Disease Control and Prevention, "Heart Disease Facts," 28 November 2017. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm. [Accessed November 2019].

[9] Office of Disease Prevention and Health Promotion, "Heart Disease and Stroke," 05 December 2019. [Online]. Available: https://www.healthypeople.gov/2020/topics-objectives/topic/heart-disease-and-stroke. [Accessed November 2019].

[10] K. Bhanot, "Predicting presence of Heart Diseases using Machine Learning," 12 February 2019. [Online]. Available: https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c. [Accessed November 2019].

[11] E. Strickland, "AI Predicts Heart Attacks and Strokes More Accurately Than Standard Doctor's Method," IEEE Spectrum, 01 May 2017. [Online]. Available: https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/ai-predicts-heart-attacks-more-accurately-than-standard-doctor-method. [Accessed November 2019].

[12] S. F. Weng, J. Reps, J. Kal, J. M. Garibaldi and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," 4 April 2017. [Online]. Available: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0174944&type=printable. [Accessed November 2019].

[13] A. Janosi, W. Steinbrunn, M. Pfisterer and R. Detrano, "UCI Machine Learning Repository: Heart Disease Data Set," 01 July 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease. [Accessed November 2019].

[14] J. Beckerman, "Can I Learn if I Have Heart Disease With an MRI?," WebMD Medical Reference, 5 September 2018. [Online]. Available: https://www.webmd.com/heart-disease/diagnosing-mri#1. [Accessed November 2018].

[15] CostHelper, Inc., "How Much Does a Cardiac MRI Cost?," 2019. [Online]. Available: https://health.costhelper.com/heart-mri.html. [Accessed November 2019].

[16] M. Mikulic, "Number of magnetic resonance imaging (MRI) units in selected countries as of 2017," Statista, 9 August 2019. [Online]. Available: https://www.

statista.com/statistics/282401/density-of-magnetic-resonance-imaging-units-by-country/. [Accessed November 2019].

[17] Healthwise Staff, "Coronary Calcium Scan: Should I Have This Test?," Healthwise, 9 April 2019. [Online]. Available: https://www.uwhealth.org/health/topic/decisionpoint/coronary-calcium-scan-should-i-have-this-test/av2072.html. [Accessed November 2019].

[18] e. ekoulier (https://stats.stackexchange.com/users/196038/ekoulier), "Why is tanh almost always better than sigmoid as an activation function?," 26 February 18. [Online]. Available: https://stats.stackexchange.com/q/330565. [Accessed November 2019].

[19] Y. R. Manda and K. M. Baradhi, 13 November 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK531461/. [Accessed November 2019].