Rishitha Boddu and Siddharth Tiwari
Mr. Meyer
CS Seminar: Machine Learning
19 March 2021

Can Machine Learning Algorithms Use Low-Cost Clinical Data to Effectively Predict the Onset of Heart Disease?

Heart disease is an umbrella term for a collection of various heart conditions, some of which can lead to heart attack or heart failure. According to the CDC, heart disease is the leading cause of death in the United States, accounting for 25% of deaths annually (Murphy, SL., 2017). Despite being such a prevalent health issue, it is difficult to predict whether or not an individual has heart disease until they experience the symptoms of a heart attack, heart failure, or arrhythmia. Heart disease is often referred to as "silent" and cannot be diagnosed until these symptoms are identified, which means that clinicians have limited time to discover and treat heart disease. For this reason, doctors encourage their patients to be aware of any risk factors they may have to efficiently treat heart conditions. These include high blood pressure, high cholesterol, diabetes, as well as behavioral factors such as smoking, excessive drinking, unhealthy diet, and poor physical activity ("About Heart Disease"). With many different factors involved, it is difficult to take all of them into consideration and provide forecasting for a patient's risk to develop heart disease.

In this project, we attempt to predict the development of heart disease using 1026 samples from a compilation of data from four medical studies (Janosi, A. *et al*). These samples include thirteen separate demographic and clinical metrics collected from study participants, as well as a column (named "target"), which indicates the presence of heart disease in the

participant - 0 indicates “no heart disease” and 1 indicates “heart disease”. The following chart contains a comprehensive overview of the thirteen metrics in the dataset.

Name	Metric	Metric Summary	Type	Cost
Age	age	Indicates age of the individual based on birth records, easy to acquire	int	$0
Sex	sex	Sex of the individual, female or male	bin	$0
Chest Pain	cp	Chest pain type (typical angina, atypical angina, non-angina, or asymptomatic angina) represented by numbers 1-4 or 0 if inapplicable	cat	$0
Resting Blood Pressure	restbps	Resting blood pressure (mm Hg). Calculated using a sphygmomanometer (consists of stethoscope, arm cuff, pump, and dial)	con	$0
Serum Cholesterol	chol	Serum cholesterol (mg/dl) measured in a blood test called a lipid profile/panel - show levels of high and low-density lipoprotein cholesterol and amount of triglycerides in blood	con	~$25
Fasting Blood Sugar	fbs	Blood test taken after a period of fasting (generally eight hours). Value is either < 120 mg/dl or > 120 mg/dl (>120 indicates diagnosis of diabetes)	bin	~$5
Resting ECG results	restecg	Resting electrocardiography results (normal, ST-T wave abnormality, or left ventricular hypertrophy)	cat	~ $50
Max Heart Rate	thalach	Max heart rate = 220 - age	con	$0
Exercise-Induced Angina	exang	Is angina (chest pain which occurs when the heart does not receive enough oxygen-rich blood) caused by exercise?	bin	~$70
ST Depression	oldpeak	ST depression induced by exercise relative to rest. ST depression is a finding on an electrocardiogram graph	con	~$70
Slope of ST segment	slope	Slope of peak exercise ST segment (upsloping, flat, or downsloping)	cat	~$70
Colored	ca	Number of major vessels colored by fluoroscopy, a	int	~$80

Vessels from Fluoroscopy		test which uses a beam of x-rays to look at parts and movement of parts of the body		
Thalassemia	thal	Thallium stress test result (normal, fixed defect, or reversible defect) - measured through a blood test and analyzed in a lab	cat	~$85

Figure 1. Name column contains the generic name of each metric and the metric column is the name of the column used in the code and in the dataset. Metric summary column provides a brief overview about what tests are needed to acquire the data and what the data means. Type column indicates the data type: binary (bin), integer (int), categorical (cat), or continuous (con). Finally, the cost column shows the cost of acquiring that data whether it be through tests or accessing documentation. Prices all retrieved from (Janosi, A. et al) database's supplemental files.

As seen above, some metrics are easily attainable and others require far more money and resources. While many of these tests have become more accessible to the public, individuals are dissuaded from regular testing and routine checkups due to the costs of some metrics. For this reason, we decided to divide the metrics into two categories. The first category, noted as 'basic parameters', are metrics that are available easily and do not require extensive testing. The second category, noted as 'target parameters', are metrics that require time, money, and extensive testing. In the table above, 'basic parameters' are highlighted in red and 'target parameters' in blue. By separating the metrics into two separate categories, we aim to determine if low-cost metrics, or 'basic parameters', can effectively predict patient outcomes using machine learning algorithms. Our experiment consists of a control group and two experimental groups. We will perform three different classification algorithms (logistic regression, SVM, and k-NearestNeighbors) on each of the three groups and then calculate the confusion matrix, a classification report, and the accuracy score for each of the algorithms. Our features data consists of the thirteen different metrics provided earlier and the labels data is simply the target column.

The control group has a features group consisting of all the original dataset with the thirteen different metrics. This acts as the control because the original data is unaltered. We split this dataset into test and train groups to train our three classification algorithms. In the first

experimental group, we filter out all expensive metrics, or target parameters, from the features data. Now, we use these low-cost metrics, or basic parameters, as our features and use this new dataset to train the classification algorithms. Our final experimental group is obtained through a multi-step process. First, we filter out all the target parameters, similar to the way we established the second experimental group's features dataset. Once this step is completed, we individually trained and tested three classification or regression models (SVM, Random Forest, and KNN) on every single target parameter. For each target parameter, we looked at the accuracy scores of the three models and substituted the values of the original dataset with these new predicted values from the model with the highest accuracy. Figure 2 displays the results we received after looking at either the accuracy score (for classification algorithms) or the root mean square error values (for regression algorithms) for all eight target parameters. For example, we saw that the random forest classification algorithm produced the highest accuracy for the fasting blood sugar (fbs) metric and so the values of the original dataset would be replaced with the ones produced by the random forest model. Once this process was completed for all eight target parameters, we were left with a fully reconstructed dataset with the same dimensions as the original. Finally, we used this reconstructed dataset in our classification models to predict the onset of heart disease as we did with the control and first experimental group.

Target Paramater Regression/Classification Results

Root Mean Square Errors for Target Parameter Regressions.

Target Parameter	SVM Regression	Random Forest Regression	KNN Regression
chol	49.08	49.08	37.70
oldpeak	1.04	1.04	0.95
slope	0.55	0.55	0.51

Accuracy scores for Target Parameter Classifications.

Target Parameter	SVM Classification	Random Forest Classification	KNN Classification
fbs	0.844	0.985	0.846
restecg	0.602	0.946	0.631
exang	0.817	0.971	0.760
ca	0.649	0.961	0.659
thal	0.771	0.971	0.751

Figure 2. A table displaying the accuracy scores of each regression/classification model on the eight different target parameters. Three target parameters were used to train regression models and five were used for classification models.

Final Classification Results

Accuracy scores for all classifications. ¶

Classifier	Accuracy Score (Including Target Params)	Accuracy Score (Excluding Target Params)	Accuracy Score (Including *Predicted* Target Params)
Logistic Regression	0.873	0.80	0.854
SVM	0.939	0.812	0.9
k-NearestNeighbors	0.851	0.819	0.849

Figure 3. This chart shows our accuracy scores for each of the test groups with the three different classification models we used to predict the target (presence or absence of heart disease) field.

From our results, it is evident that the SVM classifier produced the most accurate results for the control group and the second experimental group and are the second most accurate for the first experimental group. As for the different groups, we found that the control group was the most accurate as we initially predicted. However, we found that with the second experimental group where predicted the target parameters, the accuracy only decreases an insignificant amount. When compared to the first experimental group in which there are only five metrics being used, the accuracy values of the second experimental group overall was much closer to the accuracy values of the initial dataset. For example, when we trained the k-NearestNeighbors classification model using the original dataset, we got an accuracy of 0.851. The accuracy of the

same classifier with the reconstructed dataset only went down by 0.002. These findings lead us to believe that similar predictions can be achieved through our method of dual-step classifications instead of having the general public spend hundreds of dollars to monitor their risk of heart disease.

Integrating equity into public health services is an important initiative that spoke to our team as many Americans are unable to access affordable and quality healthcare. If machine learning allows all individuals to track their risk levels at an affordable price, we should work to make this technology more widespread. We hope our project would be able to assist individuals in determining prognostic indicators of disease, allowing them to take preemptive measures accordingly.

Works Cited

About Heart Disease. (2021, January 13). Retrieved March 20, 2021, from https://www.cdc.gov/heartdisease/about.htm

Cholesterol Test. (2019, July 12). Retrieved March 20, 2021, from https://www.mayoclinic.org/tests-procedures/cholesterol-test/about/pac-20384601#:~:text=A%20complete%20cholesterol%20test%20%E2%80%94%20also,and%20triglycerides%20in%20your%20blood

Fluoroscopy. (2016, February 01). Retrieved March 23, 2021, from https://www.peacehealth.org/sacred-heart-riverbend/services/imaging/Pages/fluoroscopy

Heart Disease. (2021, January 19). Retrieved March 20, 2021, from https://www.cdc.gov/heartdisease/index.htm#:~:text=Heart%20disease%20is%20the%20leading,can%20lead%20to%20heart%20attack.

Janosi, A., Detrano, R., Pfisterer, M., & Steinbrunn, W. (n.d.). UCI machine Learning REPOSITORY: Heart disease data set. Retrieved March 23, 2021, from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Murphy SL, Xu J, Kochanek KD, Arias E. Mortality in the United States, 2017. NCHS data brief, no 328. Hyattsville, MD: National Center for Health Statistics; 2018.

Target Heart Rate and Estimated Maximum Heart Rate. (2020, October 14). Retrieved March 20, 2021, from https://www.cdc.gov/physicalactivity/basics/measuring/heartrate.htm#:~:text=You%20can%20estimate%20your%20maximum,beats%20per%20minute%20(bpm)

What is Blood Pressure and How is it Measured? (2019, May 23). Retrieved March 20, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK279251/