Model Selection for Classifying Heart Disease

Oliver Speltz, Dan Tu, Yutaro Sakairi

March 19, 2019

I Abstract

Hundreds of thousands of Americans suffer and pass away from heart disease each year. Due to the lack symptoms, many people don't even know they have heart disease until they are severely ill, or sometimes incurable. Therefore, heart health monitoring is critical to the sustenance of a healthy population. A model that can predict the heart health of patients without expensive diagnosis will be helpful in quickly examining the potential for the occurrence of heart disease. Our experiments show that all three models were able to predict heart disease with 85% accuracy using the given data.

II Introduction

Heart disease is one of the leading causes of death facing Americans today. One in every four deaths in the United States is related to heart health, or about 610,000 deaths per year [1] The most common type of heart disease is Coronary Artery Disease (CAD) in which fatty deposits called *plaque* build up in the coronary artery. The coronary arteries are the arteries which supply blood to the heart. The build up of plaque within these arteries slows the supply of blood to the heart and can result in *angina*, or chest pain, which is the main symptom of CAD [2]. If the flow of blood to the heart is sufficiently decreased the sufferer may experience a heart attack.

There are many factors that put one at risk for heart disease. Among the top risk factors are smoking, poor diet, physical inactivity and excessive alcohol use. Some other conditions also put you more at risk of developing heart issues, such as diabetes and high blood pressure [1].

However, many heart disease don't entail obvious symptoms. Many patients do not realize they have heart disease until they experience complications such as a heart attack or sudden cardiac arrest [9]. The monitoring of heart health is then of large importance for maintenance of a healthy population. Many metrics can be collected to assess heart health, some key ones would be blood cholesterol levels, resting blood pressure and resting heart rate. Patients can also go through *stress tests* where their heart rate and blood pressure will be measured while under going physical exercise like walking on a treadmill. Another important diagnostic tool is the Electrocardiogram (ECG) which measures the electrical signals of the pacemaker cells which cause your heart to beat on time. Figure 1 is a sample of the output of a ECG. Different aspects of a ECG reading can give information about the health of your heart. For example, the length of the ST segment can be indicative of blockage of the coronary arteries [3]. How different pieces of the ECG reading, like the ST segment, respond to exercise can also give information about heart health. However, ECG does not provide all of the details. Further diagnosis by combinations of X-ray images, CT scan, MRI scan and angiography, a special technique that allows for detailed imaging of blood vessels, are often employed to confirm the occurrence of heart disease, which can be costly and time consuming.

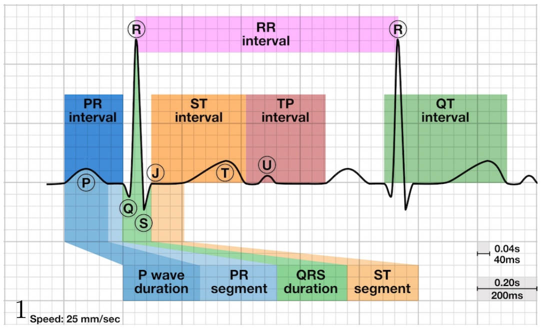


Figure 1: The output of a ECG test [3]

The purpose of our project is to find a model that best predicts the presence of heart disease in a patient given some data related to heart health. Most of the data required to run the model can be obtained through ECG and simple stress test. The model can then be applied to new patients who are uncertain whether they have heart disease or not, and be an indicator of whether they are at risk of heart disease and further diagnosis is recommended.

III Methods

The dataset to be utilized was provided by UC Irvine and was found on Kaggle.com. The data was collected by Dr. Robert Detranot at Cleveland Clinic Foundation in Ohio from 303 patients, 138 of whom are suffering from heart disease. It includes 14 categories such as age, sex, blood pressure, length of the ST interval for each person. A part of the original data is shown below,

patient	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	63.0	1.0	3.0	145.0	233.0	1.0	0.0	150.0	0.0	2.3	0.0	0.0	1.0	1.0
2	37.0	1.0	2.0	130.0	250.0	0.0	1.0	187.0	0.0	3.5	0.0	0.0	2.0	1.0
3	41.0	0.0	1.0	130.0	204.0	0.0	0.0	172.0	0.0	1.4	2.0	0.0	2.0	1.0
4	56.0	1.0	1.0	120.0	236.0	0.0	1.0	178.0	0.0	0.8	2.0	0.0	2.0	1.0
5	57.0	0.0	0.0	120.0	354.0	0.0	1.0	163.0	1.0	0.6	2.0	0.0	2.0	1.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The columns are as follows:

`age` The age of the patient, in years

`sex` The biological sex of the patient, '1' for male, '0' for female.

`cp` Chest pain. Categorical 0 - 3 indicating the severity of the chest pain the patient experiences. 0 being none.

`trestbps` The patients resting blood pressure in mmHg, continuous.

`chol` The patients blood cholesterol levels, in mg/dl

`fbs` The patients fasting blood sugar. '1' if the patient has fasting blood sugar over 120 mg/dl, '0' if not. Fasting blood sugar is the measure of blood sugar levels after an overnight fast. Having high fasting blood sugar levels (over 120 mg/dl) is indicative of Diabetes [5].

`restecg` The patients resting ECG results. '0' if normal, '1' if there is abnormality in the ST segment, '2' if there is evidence of hypertrophy (enlargement) of the left ventricle (the main pumping chamber of the heart). Ventricular hypertrophy is evidence that the heart is working harder than it should be [6].

`thalach` The maximum heart rate achieved by the patient during an exercise test.

`exang` Whether or not the patient experiences exercise induced agina. '1' if yes, '0' if not.

`oldpeak` A measure of the length of the ST interval of an ECG taken during exercise relative to resting.

`slope` In a graph of the length of the ST segment vs exercise intensity, the slope. '0' indicates positive slope, '1' no slope, '2' negative slope.

`ca` The number of major blood vessels to the heart colored by fluoroscopy. Takes on the values 0-3.

`thal` Indicates whether the patient suffers from thalassemia. Thalassemia is a inherited blood disorder where red blood cells do not produce enough functioning hemoglobin. Thalassemia comes in a variety of forms, some of which can aggravate heart disease [7]. The different categories of this variable represent the different severities of thalassemia.

`target` Whether or not the patient suffers from heart disease. '1' if yes, '0' if no. This is the variable we will be attempting to predict.

Since our data contains many categorical variables, we need a formal way to encode their values for continuous models. Because the relative values of a categorical variable are hard to determine, we used a technique called OneHot encoding to transform our data. In OneHot encoding, all categorical variables are turned into a collection of binary variables. For example, the column `cp`, would be turned into four binary columns, `cp0`, `cp1`, `cp2` and `cp3`, all would have ones only on patients that had the corresponding value in the original column. Now the relative values of categorical variables is unambiguous and we can go about selecting a model.

We decided to explore a few models implemented in the python module `sklearn`. We chose to examine a range of classifiers from simple Logistic Regression to a more complex Support Vector Machine. The full set of models we considered initially is K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Decision Trees and Logistic Regression (LogR). We followed the procedures of nested cross validation while performing our model selection to avoid over fitting our model to the data. In nested cross validation, a portion of data is withheld from the very beginning, this will be the *testing data*. The remaining data can be split up into many random *training data* and *validation data* sets. This split data is used to initially pick and adjust our models. This allows us to avoid a model selection bias by isolating the testing data until the very end.

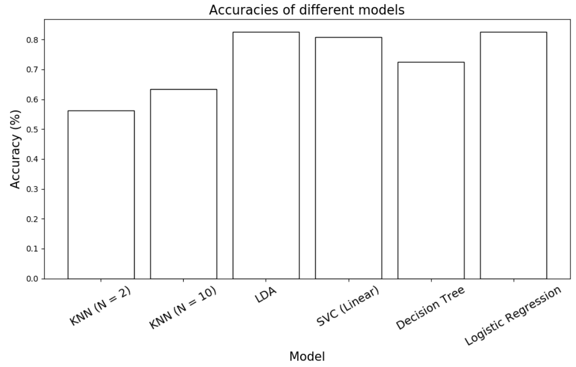


Figure 2: The performance of the initial models on the validation sets.

The initial results for the models are given in Figure 2. The SVM, LDA and LogR models performed significantly better than the others, so we decided to investigate them further and tweak hyperparameters that may be associated with each.

LDA is a parameterless classification technique by its nature, so there were no further explorations made with it.

For a SVM, there are a few different structures and parameters to be considered. For one, there is the *kernel* of the SVM. The kernel refers to a transformation of the data into a higher dimensional space before fitting the model. Putting your data into a higher dimensional space can be useful if your labeled data is hard to separate in its natural dimensions but some function of the data separates them.

Additionally, with both SVM and LogR models, there is the regularization parameter. Regularization is a tactic to prevent overfitting of the data. In a SVM, the goal is to find a *decision boundary* between the classes of data, the result is a maximization problem of the margin between the the decision boundary and the data points, while also penalizing for data points that are on the wrong side of the boundary. The regularization is how significant this penalty is. The regularization weight in a LogR is a similar concept.

Once we determined the best kernel to use for SVM and the best regularization parameters for LogR and SVM, we took a closer look at our variables. We wanted to examine them each individually to see if any of them were more important to the success of our model. To do so, we took our best performing model thus far, which was the LogR model, and one by one, excluded one of the variables from the model. The loss of accuracy compared to the model containing all of the variables can be a sort of metric on how important each variable is. Additionally, each variable was used to train a model to predict heart disease on its own, i.e., without any of the other variables. The accuracy of each variable on its own multiplied by the loss in accuracy by excluding that variable gives the importance of each variable.

We tested the importance of each variable on many runs, using as many as 2000 repetitions for each test. Using the importance of each variable, we can try building a reduced model, a model that is trained and tested on a smaller set of predictor variables. For this reduced model, we excluded the variables `fbs`, `chol`, `age` and `restecg`, because they performed the worst.

Now that we have fine tuned our SVM model and our LogR model, we can compare them to the LDA model on the final test set. For the final test, we trained all three models on both the full data set and the reduced data set. The accuracies of all 6 models are then compared.

IV Results

IV.1 Adjusting Hyperparameters

Using cross-validation averaged over 200 trials shows that using linear kernels results in a significantly higher accuracy SVM model as shown in Figure 5a.

After picking the model for SVM, we adjusted the regularization weights (refered to as "penalty" in the figures) for SVC and LogR. The weights ranged from 0.01 to 10.24. By default, each model uses 1.0 as their regularization weight. As shown in Figures 3 and 4, the penalties that resulted in the highest accuracy were 0.64 for LogR, and 1.28 for SVM.

IV.2 Retrying KNN

The best performing datasets on Kaggle included KNN (3, 7, 8 neighbors) and Random Forest classification models with 88.52% accuracy. Because we previously tested KNN with 2 and 10 nearest neighbors, we decided to try analyzing it again using a range of 1 to 10 neighbors. Our results (figure 5b) shows that 7 neighbors was the best choice for this model, but it was only able to predict with around 70% accuracy. The optimal number of neighbors also changes every time we run the cross-validation tests. We decided to stay with the LDA, SVC, and LogR models.

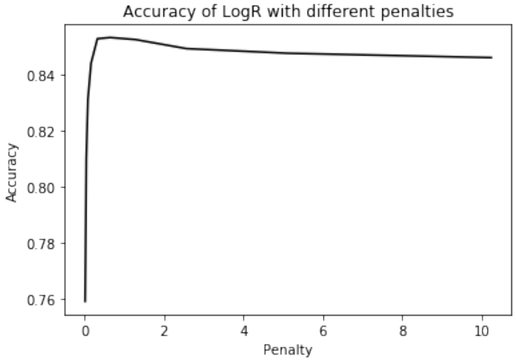


Figure 3: LogR accuracy by penalty

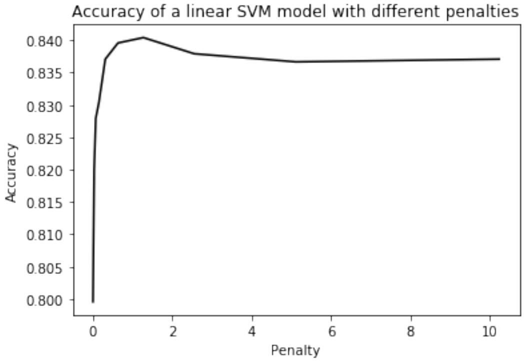
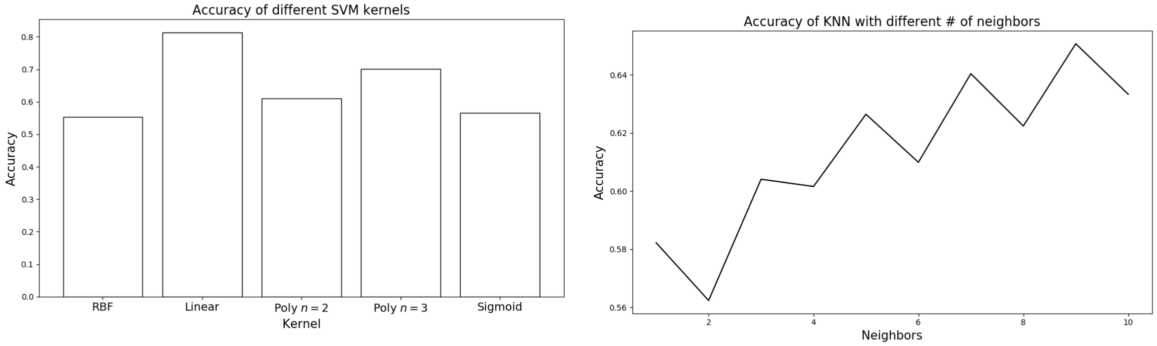


Figure 4: SVM accuracy by penalty



(a) Accuracy of different SVM kernels.

(b) Accuracy of KNN by number of neighbors.

Figure 5: Results of SVM kernel trials and KNN exploration.

IV.3 Variable Importance

When investigating variable importance, we used 500 trials to help avoid the random selection from playing too much of a role. However, many runs confirmed that the variables `chol`, `fbs`, `restecg` and `age` did not contribute well to the model, as noted by their negative values in Figure 6. These variables will be excluded from a reduced model for upcoming tests.

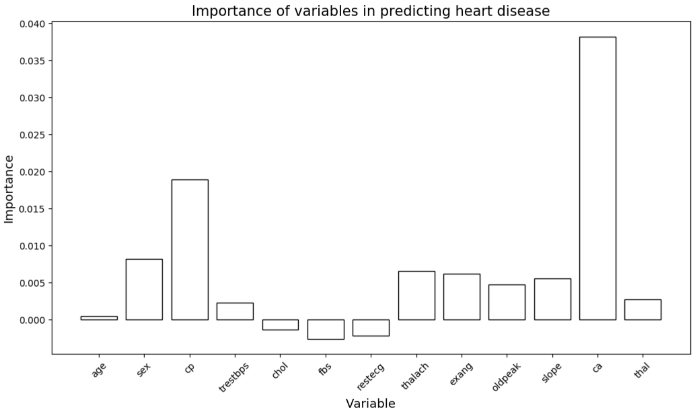


Figure 6: The importance of the variables in predicting heart disease.

IV.4 Test Accuracy

Figure 7 shows the accuracy of each model on the test set withheld at the beginning of the model selection process. All three models, utilizing the reduced data set, were able to accurately predict the test set with 85% accuracy with LDA, LogR and SVM. Removing the variables that decreased our prediction accuracy increased the LDA, SVM and LogR models by 9%, 5%, and 3% respectively. An analysis of the confusion matrix for all three top performing models showed that of the 60 test samples, the models correctly predicted 51 samples and incorrectly predicted 6 false-positives and 3 false-negatives.

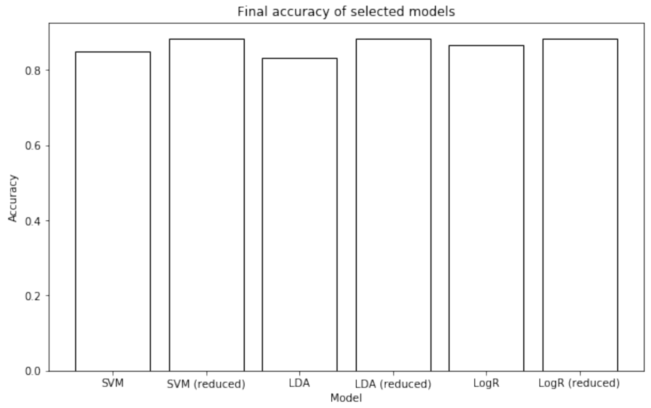


Figure 7: Final accuracy of each model on the testing set.

V Conclusion

The goal of this analysis was to see if machine learning models could be used to accurately predict heart disease given patient data before conducting more expensive diagnoses. The highest ranking model on Kaggle was able to achieve 88.52% accuracy but with a few caveats. Their models only ran one training-testing trial whereas ours was averaged over 200 trials. They also did not analyze or adjust any hyper-parameters for their models.

We found it interesting which variables actually resulted in lower model accuracy. Blood cholesterol levels, for example, intuitively seem like a good metric of heart health considering that heart disease is caused by build-ups of plaque, which contains cholesterol. However, the results of our exploration of the variables showed us that cholesterol generally reduced the accuracy of our model.

Using the given data, we were able to achieve a reasonable level of accuracy. Our SVM, LDA, and LogR models were able to predict unseen data with 85% accuracy after adjusting hyper-parameters. Since the sample size ($n = 303$) is small compared to the population with heart disease, it would be inaccurate to say that our models would generalize well to the public. There are also other factors that may contribute to predicting heart disease that were not included in our data set such as weight or body fat percentage.

To improve our models, we can obtain more data for training. Previous plots of sample size vs accuracy shows that increasing the number of training samples increases both the cross-validation and training accuracy. Following this trend, increasing the sample size should also result in an overall increase in accuracy. We could also test a combination of variables that were removed from the original data set. There may also be models such as Random Forest (88.52% on Kaggle) or neural networks that may provide a higher accuracy.

As of now, these models would not serve as a proper tool to aid in the diagnosis of heart disease. However, this analysis may provide some insight into the use of machine learning to aid health care professionals in their diagnosis and treatment. We hope that doctors may eventually be able to use models such as the ones shown in this study to use as guidelines and improve confidence in their diagnoses.

For future studies, we would like to increase our sample size and perhaps try to use neural networks to provide a more complex and accurate model. We would also like to focus on detecting heart disease in its early stages and the chance that an unaffected patient is likely to get heart disease.

VI Authors' Contributions

Oliver Speltz worked on the introduction, coding for the importance of each variable, and methods. Dan Tu worked on parts of the methods section, results, conclusions, and optimizing hyper-parameters. Yutaro

Sakairi worked on the abstract, a part of the introduction, a part of the methods section and optimizing hyper-parameters.

References

[1] United States Center for Disease Control
Heart Disease Facts
`https://www.cdc.gov/heartdisease/facts.htm`

[2] United States Center for Disease Control
Coronary Artery Disease
`https://www.cdc.gov/heartdisease/coronary_ad.htm`

[3] Life in the Fastlane
The ST Segment
`https://litfl.com/st-segment-ecg-library/`

[4] Kaggle
`https://www.kaggle.com/ronitf/heart-disease-uci`

[5] Mayo Clinic
Diabetes Diagnosis
`https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451`

[6] Mayo Clinic
Left Ventricular Hypertrophy
`https://www.mayoclinic.org/diseases-conditions/left-ventricular-hypertrophy/symptoms-causes/syc-20374314`

[7] Mayo Clinic
Thalassemia
`https://www.mayoclinic.org/diseases-conditions/thalassemia/symptoms-causes/syc-20354995`

[8] WebMD
Heart Disease Treatment
`https://www.webmd.com/heart-disease/guide/understanding-heart-disease-treatment#1`

[9] U.S. Department of Health & Human Services
Ischemic Heart Disease
`https://www.nhlbi.nih.gov/health-topics/ischemic-heart-disease`