Heart Disease Prediction

Andy Chen, Himanshi Gupta, and Joseph Park

COGS 109: Modeling and Data Analysis

Dr. Eran Mukamel

June 14, 2023

1. Abstract

Heart disease remains the leading cause of mortality in the United States, responsible for a staggering number of deaths each year. According to the Centers for Disease Control and Prevention (CDC, 2022), approximately 695,000 individuals succumbed to this condition in 2021 alone, accounting for 1 in every 5 deaths. The impact of heart disease on public health is profound, and there is an urgent need for effective strategies to improve early detection and intervention. Timely identification of individuals at risk of heart disease holds immense potential to significantly improve patient outcomes and reduce mortality rates.

By harnessing the power of predictive modeling and utilizing a diverse multisource dataset, we aim to empower healthcare professionals with a valuable tool for early detection and intervention in heart disease, ultimately making a significant impact on public health.

2. Introduction

Traditionally, heart disease diagnosis has relied on clinical symptoms, medical history, and invasive procedures. However, these approaches may not provide the necessary accuracy and efficiency required for early detection. In recent years, predictive modeling techniques have emerged as a promising avenue for developing accurate prediction models that can aid healthcare

professionals in identifying high-risk individuals. By leveraging advanced data analytics and machine learning algorithms, these models can analyze complex patterns and relationships within large datasets to predict the likelihood of heart disease development.

We will investigate whether age, cholesterol level, resting heart rate, blood pressure, and exercise-induced angina will serve as key predictors of heart disease. Previous studies have consistently shown significant associations between these factors and the development of heart disease. Advancing age, elevated cholesterol levels, high blood pressure, abnormal resting heart rates, and the presence of exercise-induced angina have all been linked to an increased risk of heart disease. By testing this hypothesis, we aim to validate the importance of these factors in predicting heart disease.

Another hypothesis that we will investigate is whether the inclusion of a larger amount of data will lead to improved prediction accuracy compared to using a smaller dataset. This hypothesis is based on the assumption that a larger dataset provides a more comprehensive representation of the underlying population and captures a wider range of variations and patterns related to heart disease.

3. Dataset and Features

"Heart Failure Prediction Dataset" sourced from Kaggle is a comprehensive and diverse dataset consisting of 918 observations and 11 key features. It is a compilation of information from various sources, including Long Beach, Cleveland, Switzerland, Germany, and Hungary and encompasses a rich array of clinical, demographic, and lifestyle attributes that have been associated with heart disease.

A multisource dataset allows us to capture potential variations in heart disease characteristics across different populations, enabling a more comprehensive understanding of the disease. By including data from various countries, we can explore potential disparities in risk factors, incidence rates, and clinical manifestations, providing insights into the generalizability of our prediction model.

The dataset includes the patient's age, sex, and chest pain type, which are important demographic and clinical factors. Additionally, it contains measurements such as resting blood pressure, serum cholesterol levels, and fasting blood sugar, which are commonly associated with heart disease risk. The resting electrocardiogram results indicate any abnormalities in the heart's electrical activity, while the maximum heart rate achieved during exercise provides insights into cardiac function. The presence or absence of exercise-induced angina, as well as the ST segment characteristics (measured in depression and slope), further contribute to the dataset's predictive power. Finally, the dataset includes the output class, which indicates the presence or absence of heart disease.

Moreover, the dataset's substantial size, consisting of 918 observations, provides a robust foundation for analysis and model development. A larger sample size enhances the statistical power of our study, allowing for more accurate and reliable predictions. Additionally, the dataset being sourced from Kaggle implies that it has undergone some level of preprocessing and quality control, ensuring data reliability and reducing potential biases.

4. Methods

4.1 Logistic Regression

We deployed a logistic regression model as a baseline for our predictive models due to its simplicity and interpretability. To ensure compatibility with the logistic regression model, categorical variables in the dataset were converted into numerical format using one-hot encoding. After one-hot encoding, the dataset underwent a standardization process to bring all features to a similar range using the StandardScaler class from scikit-learn which subtracts the mean from each data point and divides by the standard deviation of the samples. In the case of missing values, we imputed them with reasonable values using the SimpleImputer class from scikit-learn, using the default parameter of filling in missing values with the mean, before training the logistic regression model on the data.

To evaluate the performance of the logistic regression model, the dataset was split into training and testing sets, with a test size of 20 percent. Scikit-learn's train_test_split function to split the dataset into four separate arrays. The training set was then used to train the logistic regression model and used 5-fold cross validation to test the accuracy of the model. Afterwards, we implemented the best subset selection method to identify the optimal subset of features for predicting heart disease.

4.2 SVM

The target variable, "HeartDisease," was separated from the features and stored as a separate variable. Categorical variables were converted into numerical format using one-hot encoding to ensure compatibility with the Support Vector Machine (SVM) classifier. The dataset was then split into training and test sets, with a test size of 20% and a random state of 42. The SVM classifier was chosen initialized with default parameters. The training set, consisting of the

transformed features (X_train) and the corresponding target variable (y_train), was used to train the SVM model.

Once the SVM model was trained, it was used to make predictions on the test set (X_test). The predicted values (y_pred) were compared to the true values (y_test) of the target variable to evaluate the model's performance. The accuracy of the model was calculated using the accuracy_score method from the scikit-learn library. The resulting accuracy score represents the proportion of correctly predicted heart disease cases in the test set.

4.3 Random Forest

We wanted a higher accuracy while also using a method that wasn't prone to overfitting by nature. This led to the use of a Random Forest which fit both of those criterias. The other thing that we did differently with this model is the way we pre processed the data. We wanted to see if the addition of data would increase the model accuracy and we did multiple approaches. The first approach was with the dropped columns which gave us a 77% accuracy which wasn't that impressive since this was lower than our baseline logistic regression model.

With our second approach, we got the columns that were dropped due to their non numerical nature and one hot encoded them into binary variables which we then fed into the random forest. With this addition of data, our test accuracy was increased up to 90% which was over a 10% increase in accuracy. We did not have to worry about overfitting because of our model choice but also the parameters that go into it. For the Parameters, we had a 80-20 split for the train to test size. The parameter random state=42 allowed us to set a specific seed which would make the random numbers generated consistent if it ran multiple times.

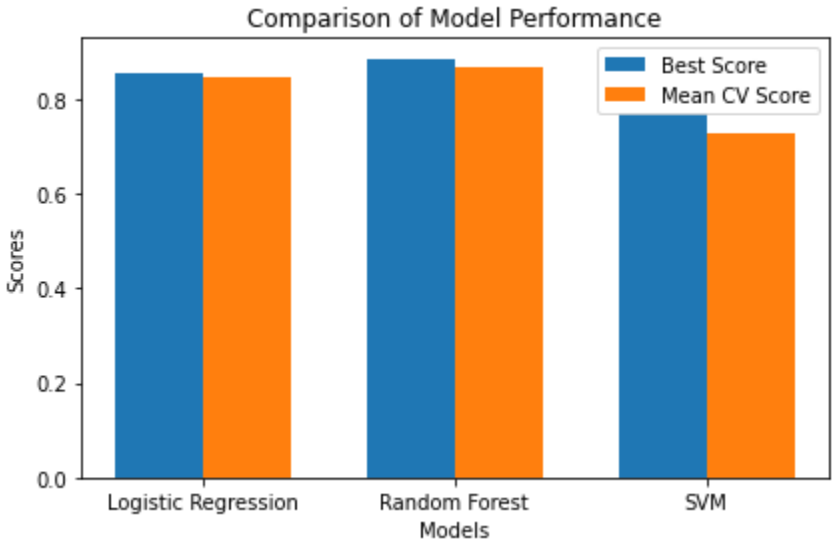
5. Results-model selection

The logistic regression model was able to achieve a mean accuracy of 85.56%, with a best subset selection of age, sex, exercise-induced angina, cholesterol levels, and resting electrocardiogram results.

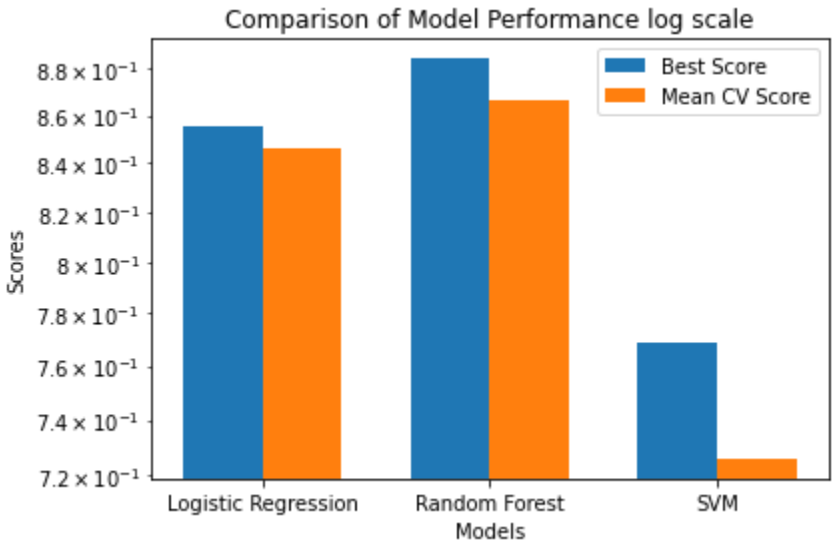
The SVM model achieved an accuracy of 0.6902, indicating that it correctly classified approximately 69.02% of the heart disease cases in the test set. Hence, the Random Forest model performed better than the SVM model in terms of accuracy.

The Random Forest model had the highest score with 88% accuracy which beats both the Logistic regression model and the SVM model. It also has an average accuracy of 0.86 which is just 0.02 off from the best score which is a good sign of no over fitting.

Now, let's look at the graphs of the best scores compared to the average scores for their accuracy.



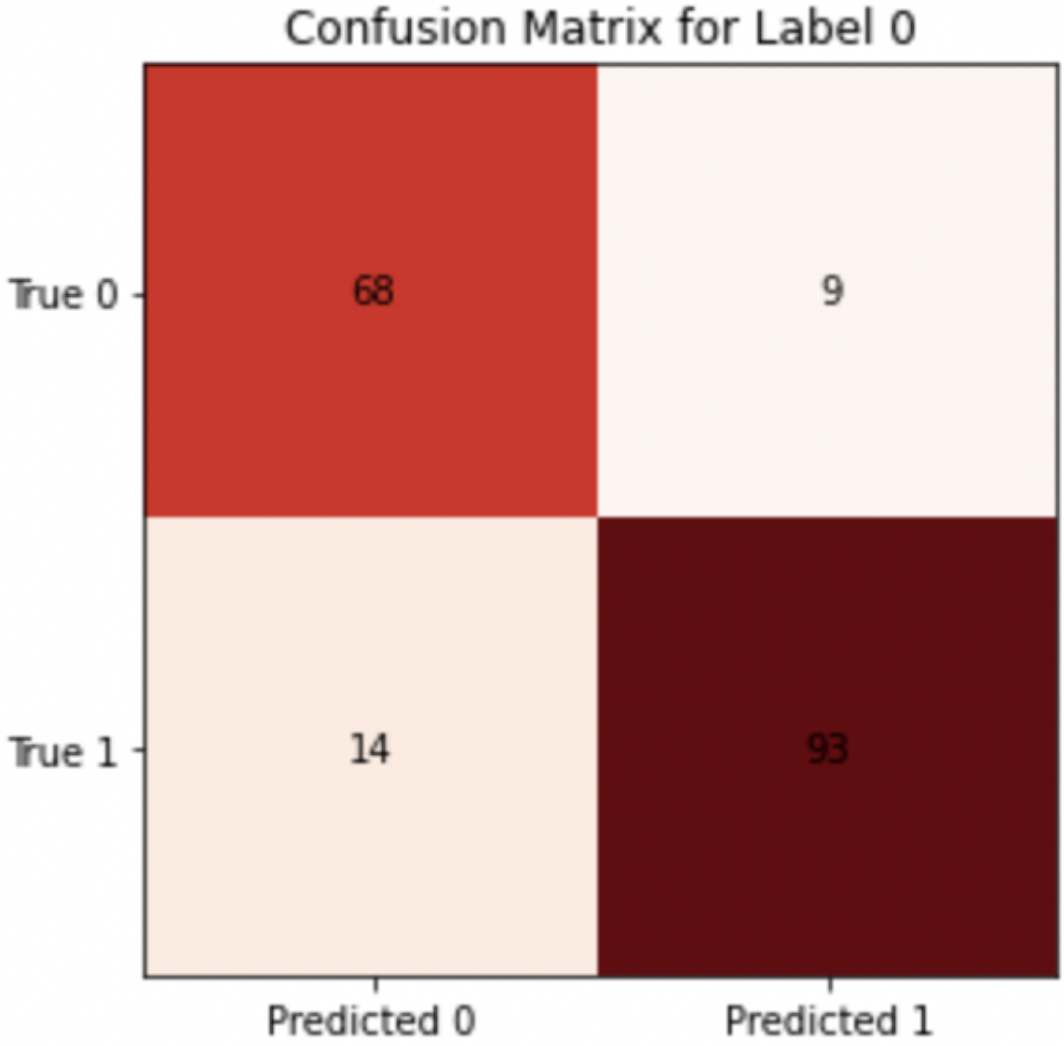
All of the models look good and are well abÍÍove the 70% prediction level but lets log scale it to look into the comparison of the models further.



Here is the log scale comparison of these models and now we can see the difference more clearly. We can see that the random Forest model outperforms the SVM model by over 10% and its mean CV score is higher than the best score of the logistic regression model. With this cross validation, we can see that the accuracy of the random forest model beats the other two in terms of accuracy.

5.1 Model Estimation

The final parameter estimates are a 80-20 split for the train and test data along with a random seed for consistency. We also use a 5 fold cross validation throughout the models. Our final best accuracy score was 0.88 accuracy with a mean of 0.86 with the Random Forest model. We can see the visualized results below with a confusion matrix highlighting true positives, true negatives, false positives, and false negatives.



6. Discussion

This study investigates the use of modeling to predict the risk of heart disease in patients given certain parameters that have been associated with heart disease. We use a logistic regression model as a baseline, random forest, and support vector machine (SVM) to predict heart disease, achieving the highest accuracy score of 88 percent with the random forest model. The results of the logistic regression and random forest models suggest that predictive modeling can be used with relatively high success to identify patients with a risk of heart disease.

Furthermore, our findings are consistent with previous research, which has shown that advancing age, elevated cholesterol levels, high blood pressure, abnormal resting heart rates, and exercise-induced angina are all associated with an increased risk of heart disease. These findings support our hypothesis that these factors are key predictors of heart disease.

7. References

CDC: Link

8. Appendix

GitHub Repository: Link