Machine Learning Classification Final Project

By: David Wilczynski

1. Main Objective

The main objective of this analysis is to use Python and 3 different classification models to predict whether a patient has heart disease or not using various clinical covariates including age, cholesterol, resting blood pressure, and maximum heart rate. The dataset used for this analysis is a publicly available dataset from the UC Irvine Machine Learning Repository entitled Clevland Heart Disease. The model will be focused on prediction and could be utilized in a healthcare setting for early detection and targeted interventions potentially improving patient outcomes and reducing hospitalizations.

1. Brief Description of the Dataset

The dataset is comprised of 303 patients with the target variable being heart disease coded as a binary variable called Condition where 1 indicated heart disease and 0 indicates no presence of heart disease. The dataset then has 13 other variables to assist in prediction, and they are as follows: age, sex, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and Thallium stress test. All variables will be utilized in analysis with the ultimate goal of training a model to be able to predict the presence of heart disease in a patient through the various clinical test results.

There was no missing data within the dataset, and therefore, no major data cleaning step were needed. The dataset was loaded into Google Colab for Python analysis where the data frame head was quickly analyzed to assess the data as seen below.

A screenshot of a white table

AI-generated content may be incorrect.

A quick correlation heatmap was also coded just to quickly examine the data and possible relationships as well.

A screenshot of a graph

AI-generated content may be incorrect.

After the data was examined and no major cleaning steps were deemed necessary, the data was split into the Train and Test sets (80/20) with the variable “Condition” marked as the target variable (y variable). All of the other covariates (X variables) were scaled using Standard Scaler. Now the data is ready for the three models.

1. Models and Evaluation

The three classification models utilized in this analysis were a logistic regression, random forest, and SVM. The logistic regression was run first and produced the following coefficients.

A screenshot of a computer

AI-generated content may be incorrect.

Next, the random forest and SVM were run on the same Training/Test split and the random forest produced the following feature importances.

A graph of blue vertical bars

AI-generated content may be incorrect.

The models were then evaluated on their accuracy, F1, and cross validation scores as seen on the next page.

A number of numbers in a row

AI-generated content may be incorrect.

A number of numbers in a row

AI-generated content may be incorrect.

A number of numbers on a white background

AI-generated content may be incorrect.

A graph showing a graph of a model

AI-generated content may be incorrect.

A close up of numbers

AI-generated content may be incorrect.

1. Recommended Model

Based on the results of the model evaluation step, I would recommend the logistic regression model. It resulted in the highest accuracy, F1, and cross validation score as well as being overall the most interpretable as the coefficients for the model are available for further statistics like determining the odds ratios. The accuracy score was 0.73 and the cross-validation score was 0.798. While the SVM model resulted in almost exactly the same metrics, the logistic regression ultimately has the greatest interpretability for clinical use. The coefficients can be used to assess each variables significance on developing heart disease which could benefit targeted treatments and preventative measures.

The next steps of this analysis would be to add more participants and variables are to achieve a greater reach on the factors that can lead to the development of heart disease. One limitation of this analysis is the small sample size which could be an area for further exploration as the necessary training and test sets further shrink the size of the data. Finally, a ROC curve was coded which displayed all the results on one plot.

