Heart Disease Prediction

Problem Statement:

Heart disease is the deadliest condition amongst adult humans in the world; it was estimated to be responsible for more than 30% of deaths worldwide in 2015.[1] Despite this, heart disease is largely preventable through healthy diet and exercise, and can be treated by lowering high blood pressures and cholesterol levels. In both cases, identifying the presence of heart disease early or identifying whether the patient is at risk of developing heart disease is crucial to maximizing their chances of survival. To this extent, this project aims to develop a model to predict the presence of heart disease in patients using readily available medical information, such as age, sex, blood pressure, cholesterol level, etc.

Doctors and other medical professionals who interact with heart disease patients regularly may find the results of this project to be of interest, as it would allow them to make a quick assessment of whether a particular patient has or is at risk of heart disease from simple medical information. Based on the results of the prediction, the doctor could then determine whether additional tests such as CT scans would be necessary if the patient seems to be at risk. While doctors are capable of diagnosing heart disease by themselves, being able to compare their diagnosis to that from a machine learning model could provide additional insight. Even if the result of the model prediction is a false positive, it could indicate that active prevention of heart disease onset for that particular patient may be important.

In addition to predicting the presence of heart disease, this project also aims to identify correlations between variables in the data and ultimately determine which features in the patient data are the strongest indicators of heart disease. Such information provides valuable insight that can be used entirely separately from the model to diagnose heart disease.

Data Wrangling:

The data for this project comes from the Heart Disease Data Set from the UCI Machine Learning Repository, which contains information such as patients' age, sex, blood pressure, cholesterol level, and heart rate, in addition to an indicator of whether the patient in question has heart disease on a sliding scale from 0 (no presence) to 4.[2] The data is split into 4 parts, with each file corresponding to patient information from a specific locale (Cleveland, Hungary, Switzerland, and Long Beach). The dataset consists of 14 columns, described below:

- age
- sex (0 = female, 1 = male)
- cp (chest pain type; 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
- trestbps (resting blood pressure in mm Hg)
- chol (serum cholesterol in mg/dl)
- fbs (fasting blood sugar > 120 mg/dl; 0 = false, 1 = true)

- restecg (resting electrocardiographic results; 0 = normal, 1 = ST-T wave abnormality, 2 = probable/definite left ventricular hypertrophy)
- thalach (maximum heart rate)
- exang (exercise induced angina; 0 = false, 1 = true)
- oldpeak (ST depression induced by exercise relative to rest)
- slope (slope of peak exercise ST segment; 1 = upsloping, 2 = flat, 3 = downsloping)
- ca (number of major vessels colored by fluoroscopy; 0-3)
- thal (MPI defects; 3 = normal, 6 = fixed defect, 7 = reversible defect)
- num (heart disease diagnosis; 0 = absence, 1-3 = presence)

The data is in csv format with null values represented by '?' strings. I imported the data for each of the 4 locales directly from the repository using pandas, making sure to specify na_values='?'. I also specified the column names, as they were missing from the files in the repository.

Upon inspecting the data, I noticed that certain columns in some datasets contained a significant amount of missing values. After calculating the percentage of missing values in each column for each dataset, I chose to drop columns that consisted of more than 40% missing values. These columns corresponded to the slope, ca, and thal columns of the Hungary dataset, the fbs, ca, and thal columns of the Switzerland dataset, and the slope, ca, and thal columns of the Long Beach dataset. The Cleveland dataset was mostly complete, with no column containing more than 1.3% missing values. For the categorical columns (sex, cp, fbs, restecg, exang, slope, ca, thal), I assigned missing values their own category with a value of -1. For numeric columns (age, trestbps, chol, thalach, oldpeak), I filled missing values with the column means from their respective dataset. If the data proves skewed upon further analysis, I may choose to return and fill these missing values with column medians instead.

For this project, I am primarily interested in detecting the presence or absence of heart disease and not the severity, so I converted the num column to be values of 0 or 1 instead of 0 through 4. After applying this transformation, certain datasets revealed themselves to be very skewed, with the Switzerland dataset, for example, consisting of 93% patients with heart disease. Building models for individual datasets with such skewed data would likely not work well, so I decided to focus on models for the Cleveland dataset and a combined analysis of all the locales. These datasets exhibit a 46/54 and 55/45 split of patients without heart disease and with heart disease, respectively. For the combined analysis, I performed an inner join on all the datasets, keeping only the columns I had not thrown away in any one of them. These columns were the age, sex, cp, trestbps, chol, restecg, thalach, exang, oldpeak, and num columns.

When inspecting the summary statistics of the combined dataset, I noticed something strange: The minimum values for the trestbps and chol columns were 0. This made no sense, since those columns represented patient blood pressure and cholesterol levels. Examining further revealed that the Switzerland dataset in particular consisted almost entirely of 0 entries for chol. Presumably, these entries were missing values that were filled in as 0 in the files. To correct this, I replaced the 0 values in those columns with the column mean without taking the 0 values into account.

Visualizing the data for the combined dataset revealed the presence of a considerable number of outliers, particularly for the trestbps and chol columns. Inspecting these values revealed them to be within reason, as I had found reports of more extreme values online. For now, I decided to keep the data as it was, but if the outliers prove to be a problem later on, I may choose to apply a log transformation to those columns. The notebook that I used to perform this data wrangling can be found as data_wrangling.ipynb, and the cleaned datasets were saved as csv files into this repository.

Exploratory Data Analysis:

Visualization of the data revealed the presence of a number of interesting trends and correlations. Most notably, amongst the patients in the dataset, men were nearly 2.5 times as likely to have heart disease than women. A 2 sample t test confirmed the significance of this difference with a p-value of 4.2e-20. This result is not surprising, as it agrees with medical studies that have found men are more at risk of heart disease than women. This result suggests that sex is an important variable in the prediction of heart disease.

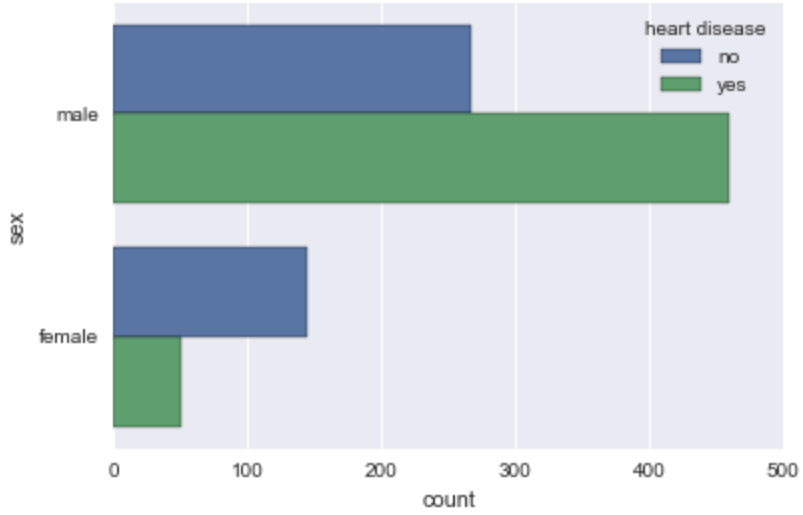

Figure 1. Count plot of male and female patients with and without heart disease.

Another result I noticed was that older patients seemed to be at higher risk of heart disease than younger patients. Overlaying histograms of patient ages for patients with and without heart disease revealed that the age distribution of patients with heart disease was shifted towards the right compared to that of patients without heart disease. A 2 sample t test comparing the mean age of patients with and without heart disease confirmed that there was a significant difference in the mean age between the two groups, with a p-value of 2.1e-17. This finding suggests that older patients are at higher risk of heart disease than younger ones, and that age is an important factor in the prediction of heart disease.

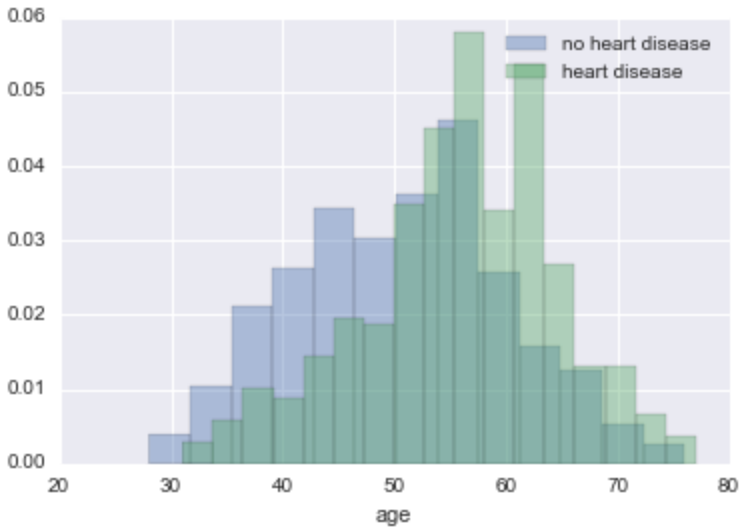

Figure 2. Age distribution of patients with and without heart disease.

One strange trend I found while exploring the data was that patients without any type of pain (asymptomatic) seemed much more likely to have heart disease than patients with chest pain or other types of pain. A 2 sample t test comparing the proportion of asymptomatic patients with heart disease to the proportion of patients with other types of pain with heart disease revealed that the difference was significant, with a p-value of 2.2e-49. This result was surprising, as chest pain is typically a common symptom of cardiovascular disease. One possible explanation for this result is that perhaps many patients admitted to hospitals for heart disease are in critical condition and thus unable to answer questions about chest pain, leading to medical professionals to simply document them as being ‘asymptomatic’. Verifying this would require contacting the people who collected this data, which could be difficult considering how old this study is.

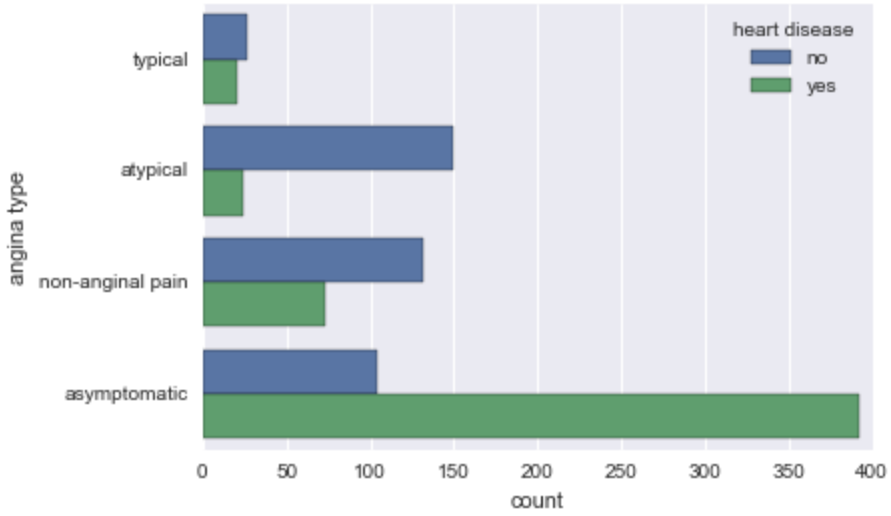

Figure 3. Count plot of heart disease amongst patients with different angina types.

Examining other trends from the dataset, 36% of patients without exercise-induced angina had heart disease, while 84% of patients with exercise-induced angina had heart disease. A 2 sample t test for proportions confirmed the significance of this difference with a p-value of 7.4e-39. The difference in proportions was huge, and suggests that exercise-induced angina could be a crucial predictor of heart disease.

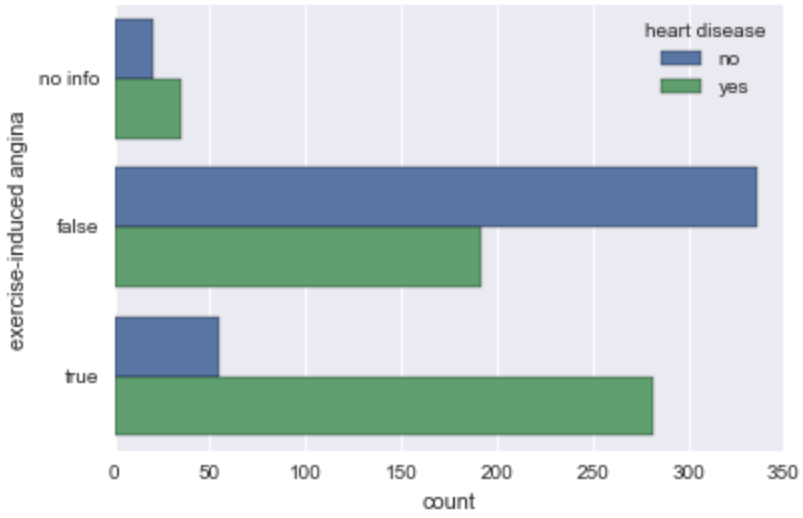

Figure 4. Count plot of heart disease amongst patients with and without exercise-induced angina.

Another surprising result from exploring the data was that patients with higher maximum heart rates seemed to be at lower risk of heart disease, contradicting conventional knowledge that higher heart rates are symptomatic of heart disease. A 2 sample t test comparing the mean heart rate of patients with and without heart disease confirmed that mean heart rate of patients without heart disease was higher than that of patients with heart disease, with a p-value of 6.6e-30. This strange result could be explained if there was a negative correlation between age and heart rate, as I had already determined that older patients were more likely to develop heart disease. In that case, this result could be attributed to the fact that patients with higher maximum heart rates were simply younger and thus healthier.

Figure 5. Heart rate distribution of patients with and without heart disease.

A scatter plot revealed a negative correlation between age and maximum heart rate with a Pearson correlation coefficient of -0.37. The probability of obtaining a p-value as extreme by random chance on the dataset was calculated to be 1.2e-31, meaning that the correlation was significant. This correlation makes sense, as older people are known to have lower heart rates than younger people, and also explains why patients with higher heart rate are seemingly at lower risk of heart disease; it is at least partially due to the fact that they are likely younger.

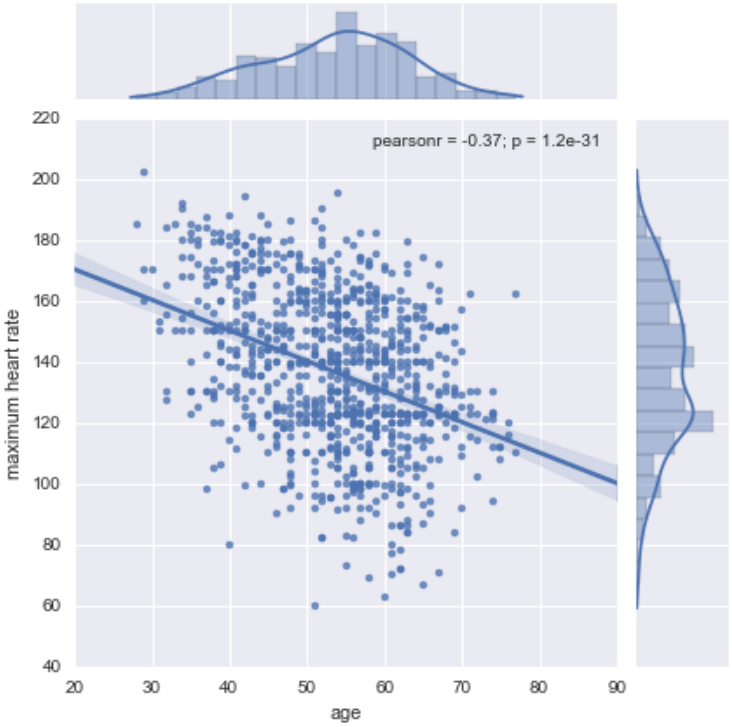

Figure 6. Scatter plot of maximum heart rate vs. age.

Lastly, I decided to test the normality of the distributions of all the numerical variables in the dataset. While normality is not a strict condition for the central limit theorem to apply, making sure distributions are normal can allow for one to leverage the powerful statistics associated with normal distributions. Using the D'Agostino-Pearson omnibus test for normality, I tested the distributions of age, heart rate, blood pressure, and cholesterol with the null hypothesis that the data was drawn from a normal distribution. Using $\alpha = 0.05$, I had to reject the null hypothesis for all the distributions and conclude that none of them were normal. However, this did not invalidate my hypothesis tests, as with a large enough sample size, the tests were still valid. I simply will not be able to use normal statistics on these distributions directly.

Machine Learning Results:

The data was split into 85% training data and 15% test data, and three separate models were developed to predict heart disease: logistic regression, SVM, and random forest. Of the models, SVM achieved the highest accuracy on the test set of 81.2%, followed by logistic regression with 79.7% test accuracy, and random forest with 78.3% test accuracy. The notebook containing the implementation of these models is modelling.ipynb. More detailed analysis of the models follows.

Logistic Regression:

The features of the data were scaled to mean 0 and standard deviation 1 in order to prevent numerically larger features from overshadowing smaller features, and to ensure that the resulting coefficients from the logistic regression model could be interpreted as the strength of the correlation between that feature and heart disease. 10-fold cross validation was performed with various values of the inverse regularization parameter C. The best value of C was found to be 0.01, and the resulting model achieved 79.9% average cross validation accuracy, and 81.0% training accuracy. The testing accuracy was found to be 79.7%. ROC and precision-recall curves for the model were plotted using the test data. The AUC of the model was 0.883.

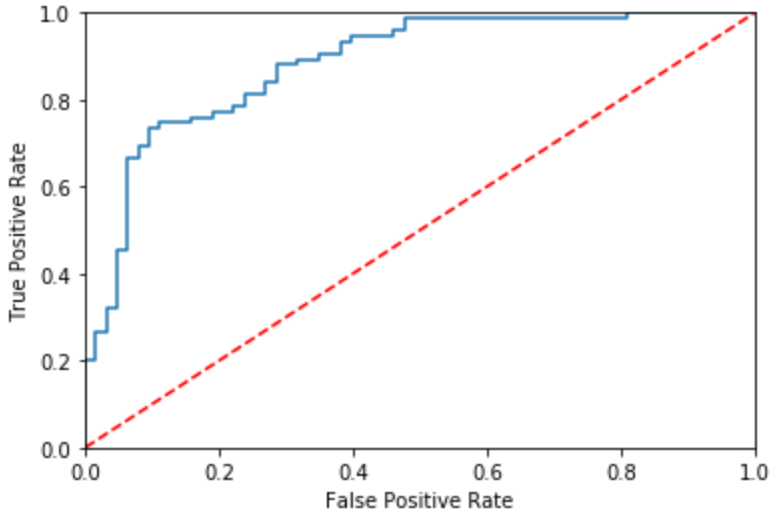

Figure 7. ROC curve for logistic regression model.

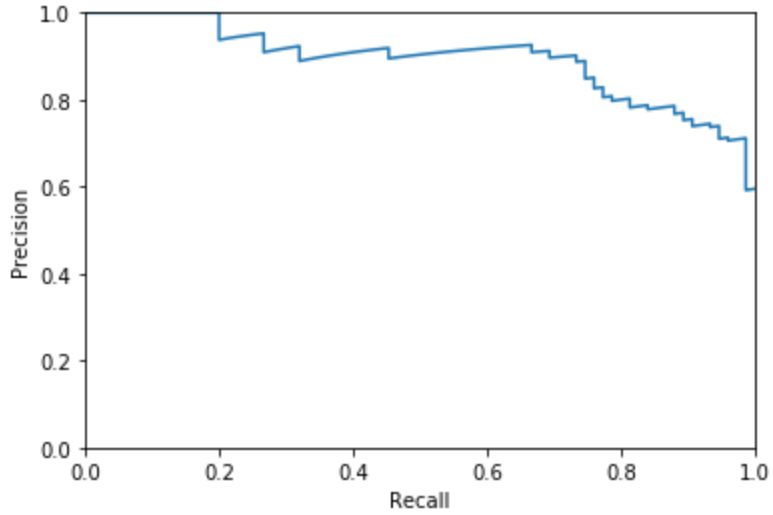

Figure 8. Precision-recall curve for logistic regression model.

The confusion matrix for the model revealed that the model performed fairly well, with a precision of 0.85 and recall of 0.76. The F1 score was 0.80. Overall, the model was more prone to false negatives than false positives.

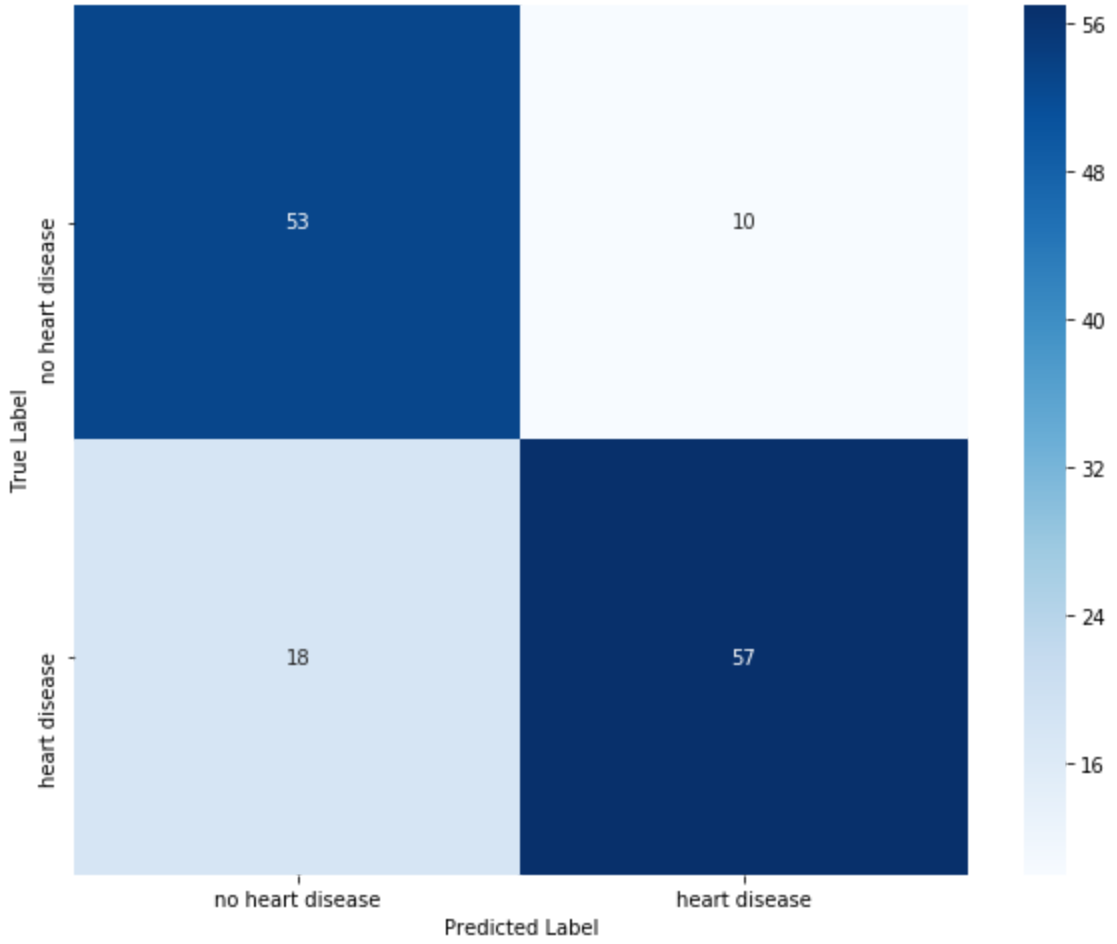

Figure 9. Confusion matrix for logistic regression model.

Investigating the values of the model coefficients for each feature revealed which features were strongly correlated with the presence or absence of heart disease. From Figure 10, it is clear that factors such as older age, being male, having exercise-induced ST depression, and having exercise-induced angina are strongly indicative of heart disease. On the other hand, factors such as higher maximum heart rate and having atypical angina or non-anginal pain are highly correlated with the absence of heart disease. Interestingly, blood pressure did not seem to be strong predictor at all. These correlations and possible reasonings for such relationships were analyzed previously in the exploratory data analysis section of the report.

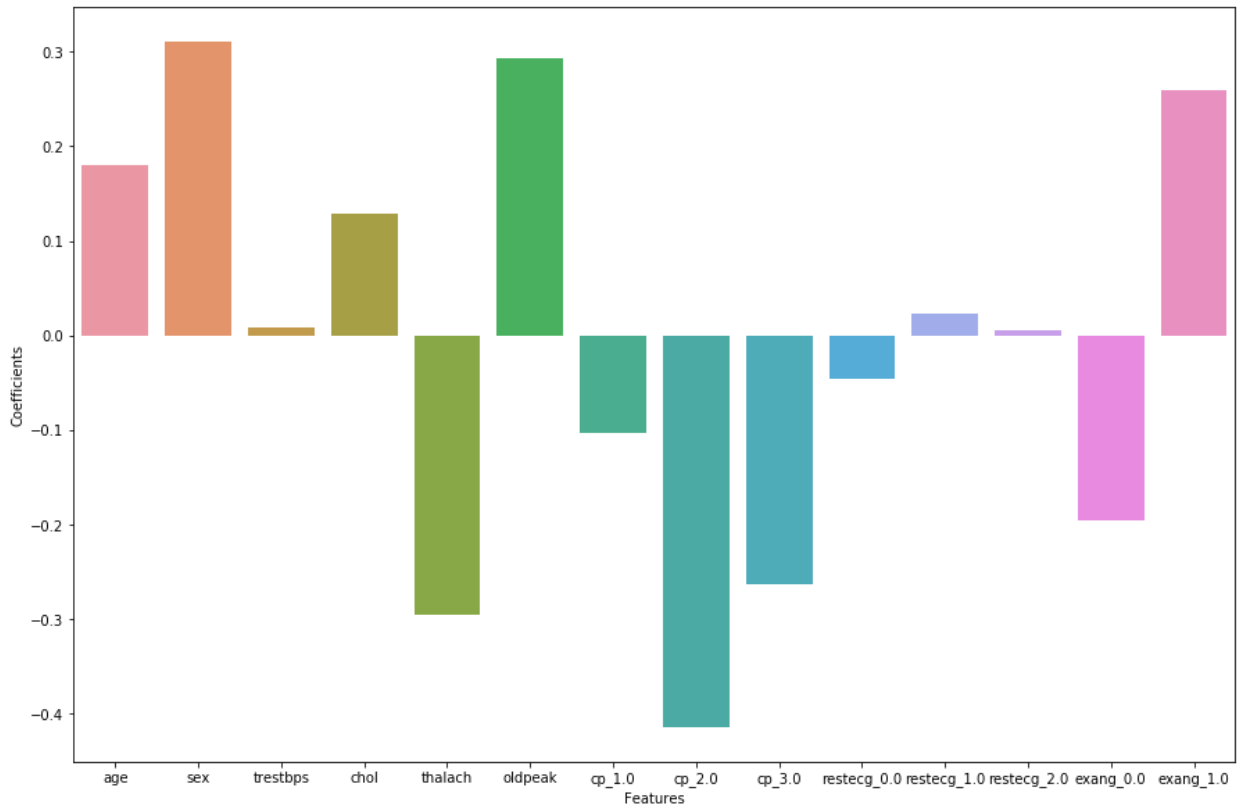

Figure 10. Coefficients from logistic regression model.

SVM:

The features of the data were scaled to mean 0 and standard deviation 1 in order to prevent numerically larger features from overshadowing smaller features. A linear kernel was chosen for the SVM, and 5-fold cross validation was performed with various values of the penalty parameter C. The best value of C was found to be 100, and the resulting model achieved 79.5% average cross validation accuracy, and 80.6% training accuracy. The testing accuracy was found to be 81.2%. ROC and precision-recall curves for the model could not be plotted, as SVM models are unable to predict probabilities for samples. Compared to the logistic regression model, the SVM model performed very similarly on the training and validation data, but seemed to generalize better to the test data.

The confusion matrix for the SVM model confirms that it performs very slightly better than the logistic regression model, with a precision of 0.85 and recall of 0.80. The F1 score was 0.82. The model was still more prone to false negatives than false positives.

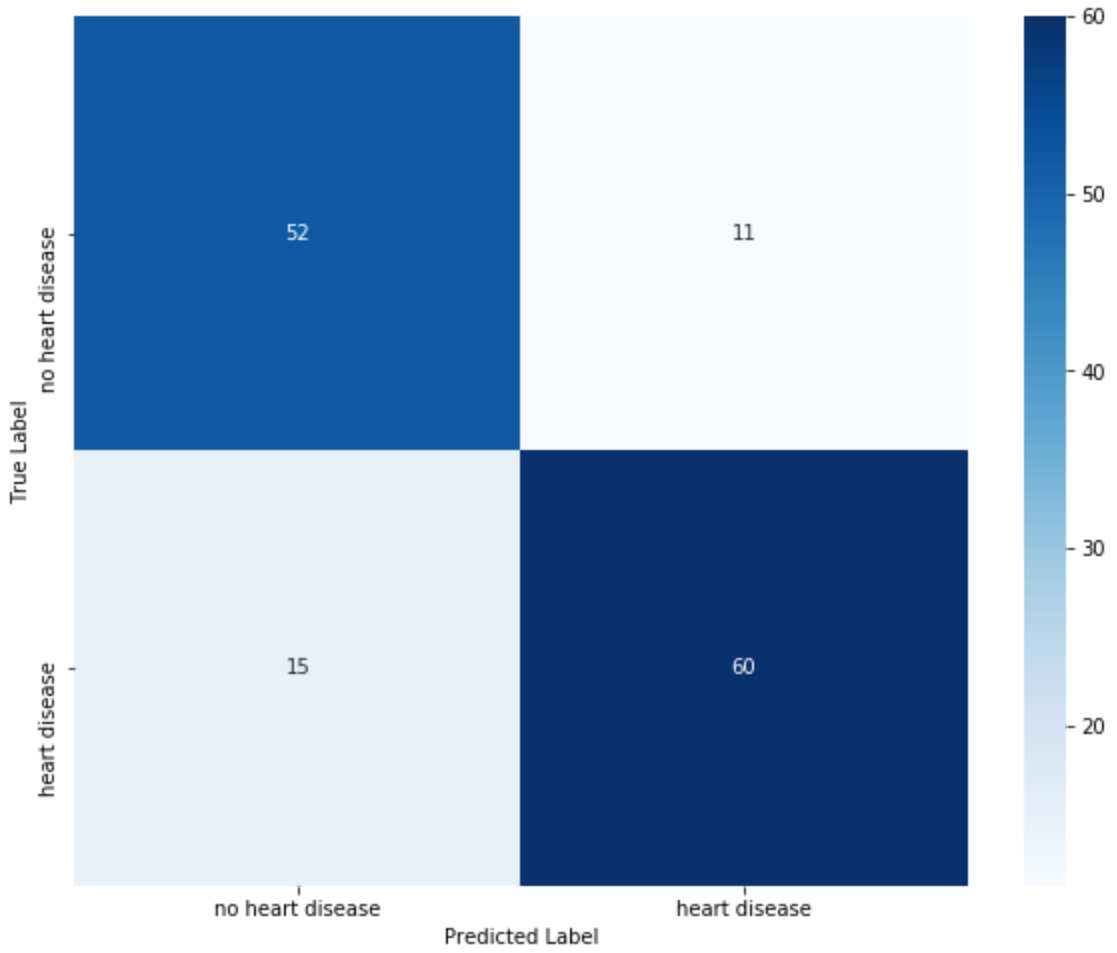

Figure 11. Confusion matrix for SVM model.

Random Forest:

Features were not scaled for the random forest model, as it is unnecessary for decision trees. The split quality was measured using the Gini impurity, and the number of features to consider at each split was chosen to be the square root of the total number of features. 5-fold cross validation was performed with various values of the number of estimators and maximum tree depth. The best number of estimators was found to be 200, and the best maximum tree depth was found to be 5. The resulting model achieved 80.3% average cross validation accuracy, and 85.0% training accuracy. The testing accuracy of the model was 78.3%. ROC and precision-recall curves for the model were plotted using the test data. The AUC of the model was 0.880.

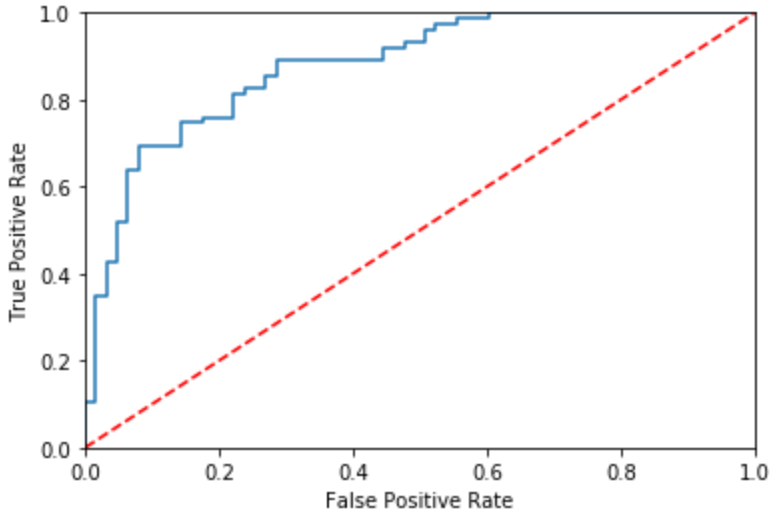

Figure 12. ROC curve for random forest model.

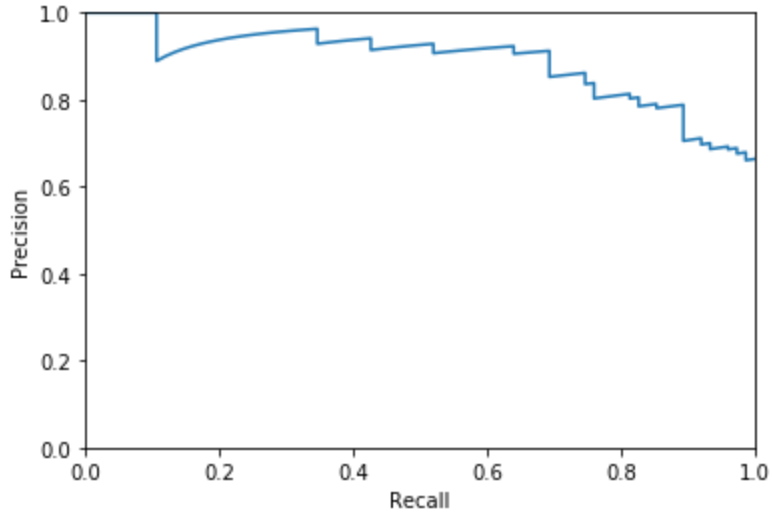

Figure 13. Precision-recall curve for random forest model.

Compared to the other two models, the random forest model performed slightly worse, and noticeably overfit the training data, which is a common problem with decision trees. The confusion matrix showed that the precision of the model was 0.84 and the recall was 0.75. The F1 score of the model was 0.79.

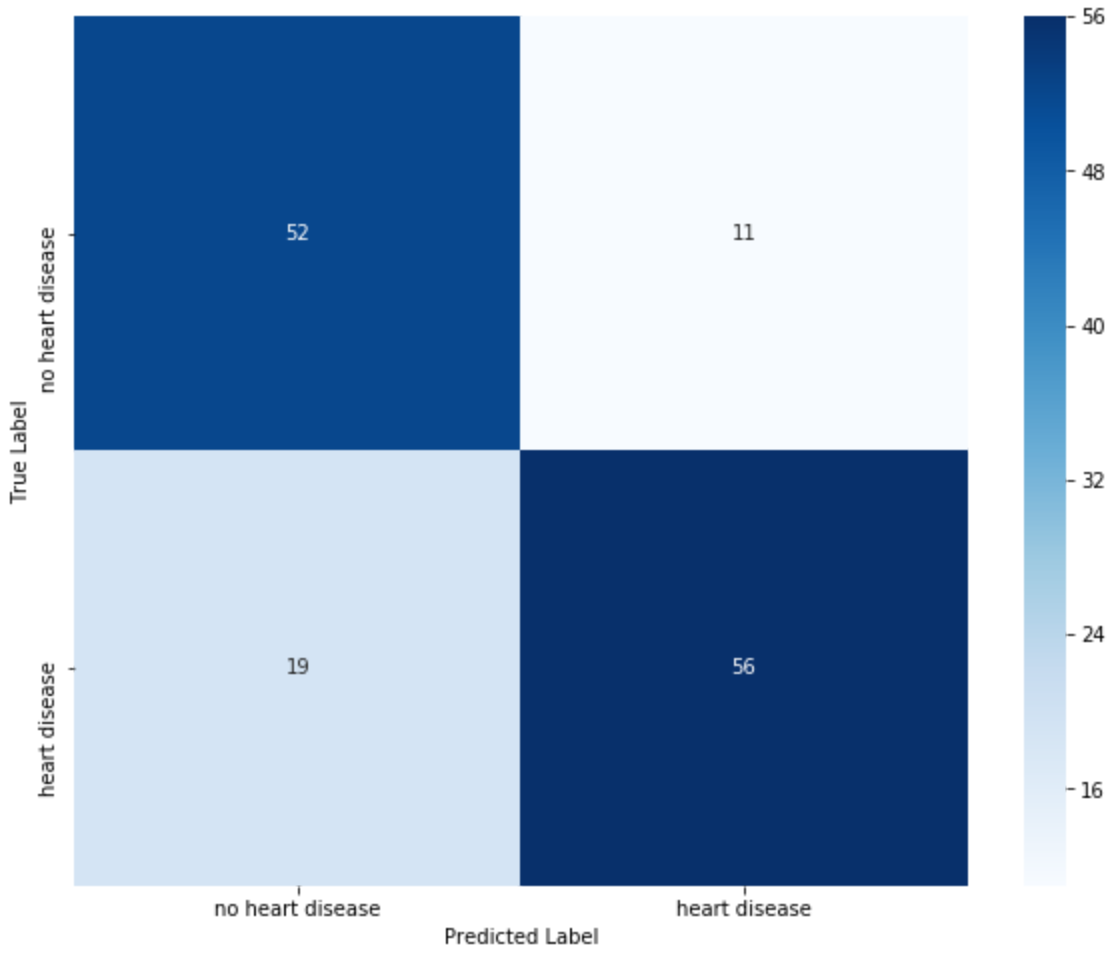

Figure 14. Confusion matrix for random forest model.

Conclusion:

Overall, the SVM model performed the best, achieving 81.2% accuracy on the test data. Analysis of the data as well as the logistic regression model revealed that factors such as age, sex, heart rate, angina, and ST depression were all very important predictors of heart disease. Interestingly, blood pressure did not seem to be a strong indicator of heart disease, and presence of angina seemed to be negatively correlated with the presence of heart disease.

The findings of this report and models created provide a good starting point for the prediction of heart disease from patient data. Further improvements to the model could be achieved by obtaining more robust data sets with more features and/or samples, or using more advanced model such as neural networks. Overall, however, the model performs fairly well, and incorrectly categorized patients could prove to be interesting cases to study.

References:

1. GBD 2015 Mortality and Causes of Death Collaborators (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet (London, England)*, *388*(10053), 1459-1544.

2. Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease Data Set. Retrieved from http://archive.ics.uci.edu/ml/datasets/Heart+Disease