Heart Disease Prediction Analysis

Problem Statement

Cardiovascular disease jeopardizes the health of populations across the world and was considered the leading cause of death in the U.S. in 2020 (CDC). The disease can stem from, and be accelerated by, both genetic and behavioral factors. As observed during the COVID-19 pandemic, governments and other authorities can assist in reducing the public health risk that widespread disease poses both to their respective populations and hospitals. Understanding the risk that heart disease can pose to these populations can help groups tailor solutions to specifically address heart disease, whether through spreading awareness of healthy behavioral changes or focusing medical resources on mitigation and prevention.

Such epidemiological exercises can benefit from the use of machine learning models to analyze whether populations with certain traits may be susceptible to contracting heart disease. While models cannot, and should not, be used to diagnose individual patients in place of medical professionals, these algorithms can help professionals understand how a disease may affect populations containing many patients with different characteristics.

Using a dataset compiled by the UCI Machine Learning Repository and further refined by Kaggle user fedesoriano, I have created two models that classify subjects as having or not having heart disease. Such a model could be used to perform similar classification on subjects outside of the dataset provided, although the model is not intended as a diagnostic tool and should not replace the guidance of a medical professional. By using an Extreme Gradient Boosting method, the model achieved a precision metric of 86.5% and a recall metric of 85.0%. A random forest classifier using the same dataset achieved a precision metric of 86.3% and a recall metric of 89.4%.

Data Wrangling

The raw dataset was formatted as a csv file with 918 rows and 12 columns, one of which contained the target variable. Few cleaning steps taken to prepare the dataset for analysis.

Kaggle user fedesoriano provided the following information on the columns, where the bracketed details reflect the units of the numerical values in the column or additional description of the categorical values:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

To limit the cardinality of the age data, I bucketed ages into 10-year ranges and placed this data in a new “Age_Decade” column. This column had 6 possible values but contained the same amount of data points as the Age column. The Age column was later dropped.

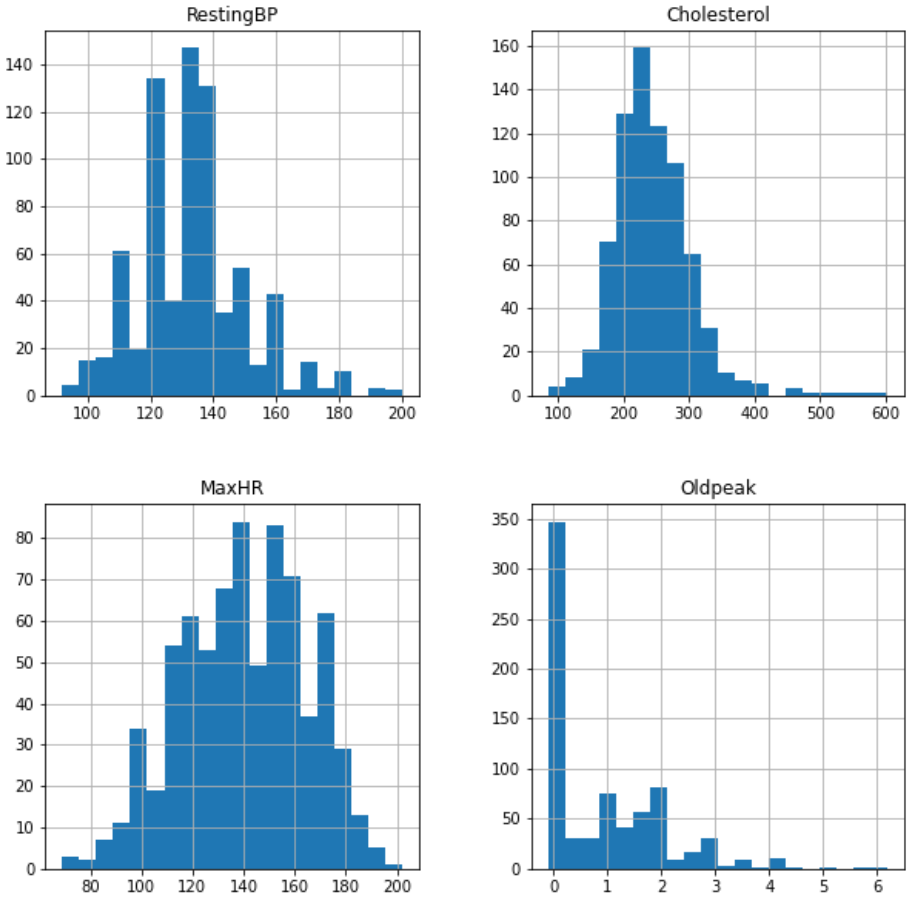
Rows containing a value of 0 for blood pressure, cholesterol, and maximum heart rate measurements were dropped from the dataset prior to modeling. There were no other missing values in the raw dataset.

Additionally, after the creation of box plots, rows containing numerical data points that were over three standard deviations from the column mean were also removed so as to limit outliers. After this step, the dataset contained 727 rows and 12 columns.

Exploratory Data Analysis

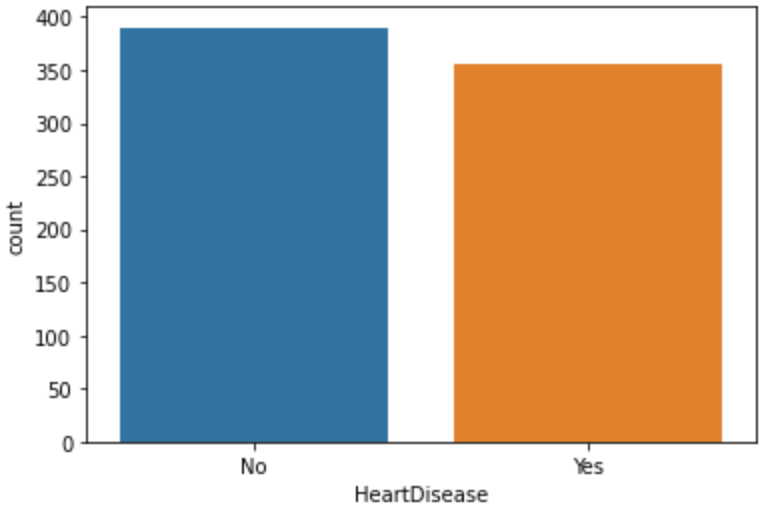
After cleaning the data, I analyzed the data for trends in individual columns as well as relationships between variables.

The below histograms depict the distribution of the numerical data.

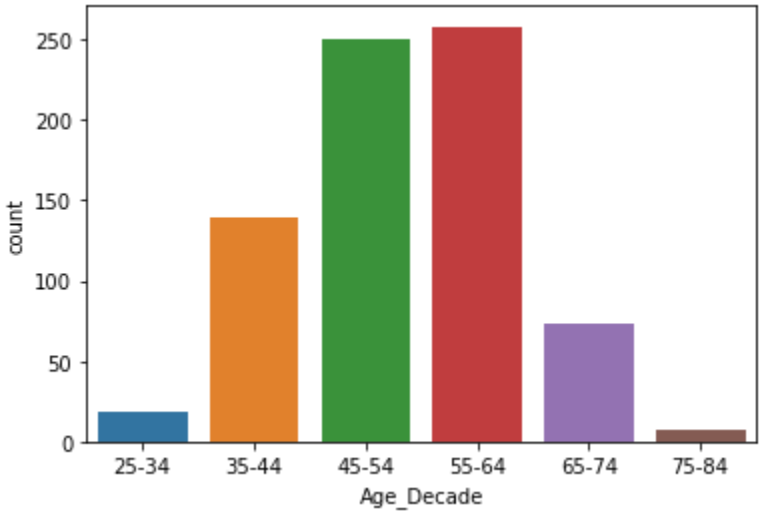


These graphs were generated prior to the removal of outliers, which appear to exist in the RestingBP, Cholesterol, and Oldpeak columns. The Oldpeak column has a more prominent tail than the other graphs; there were only 13 rows in the original dataset that had negative values in this column. Since these rows also had a 0 cholesterol value, they were removed from the final dataset used for analysis. The other graphs have unique characteristics but visually appear to more closely resemble a normal distribution than the Oldpeak graph does. Given the shape of the Oldpeak graph, a log transformation may help address the skew and allow the data to appear more normal. This could be explored in a subsequent exercise.

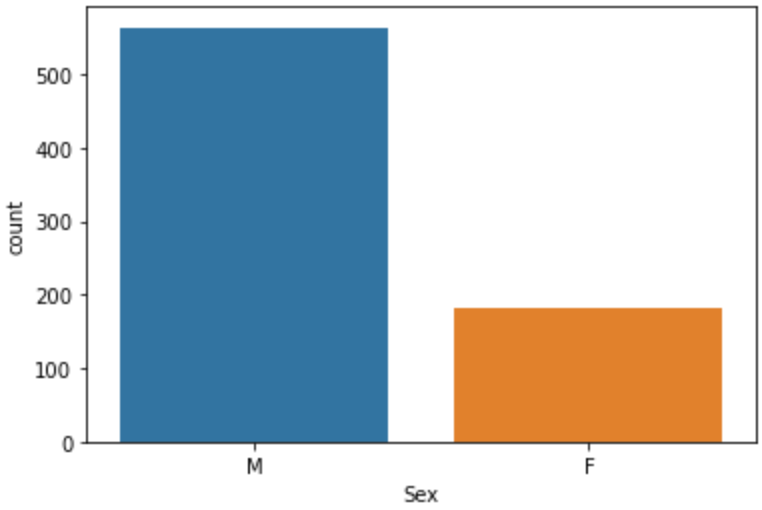
The below graphs depict the distribution of categorical data within individual columns.



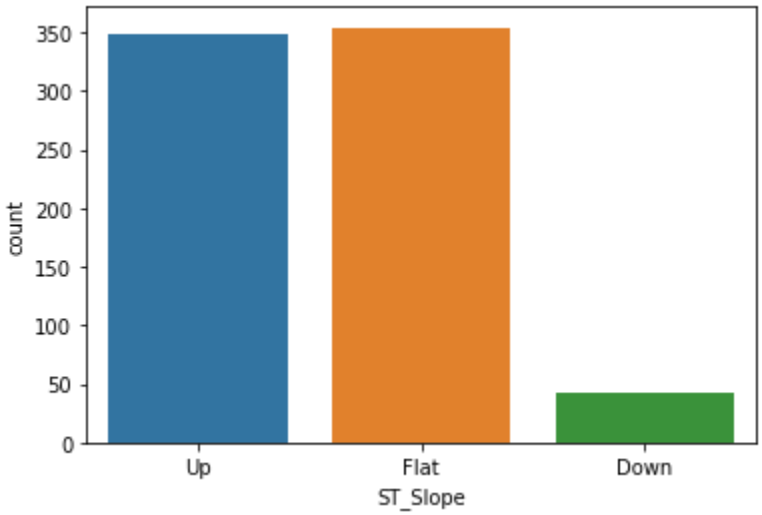
The dataset was nearly evenly split between subjects that were categorized as having heart disease as compared to those categorized as not having heart disease.



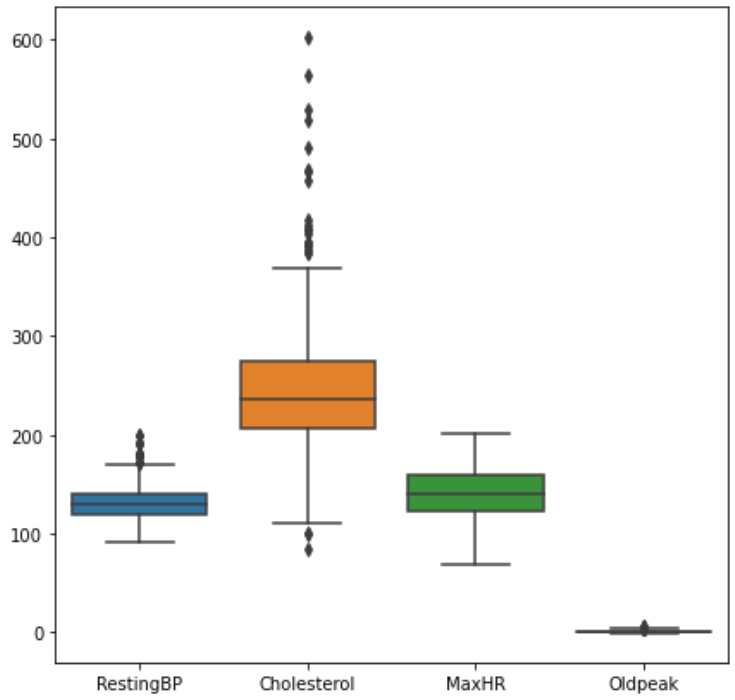
This graph indicates that most of the subjects in the dataset are between the ages of 44 and 65. There are no subjects under the age of 18 included in the dataset.



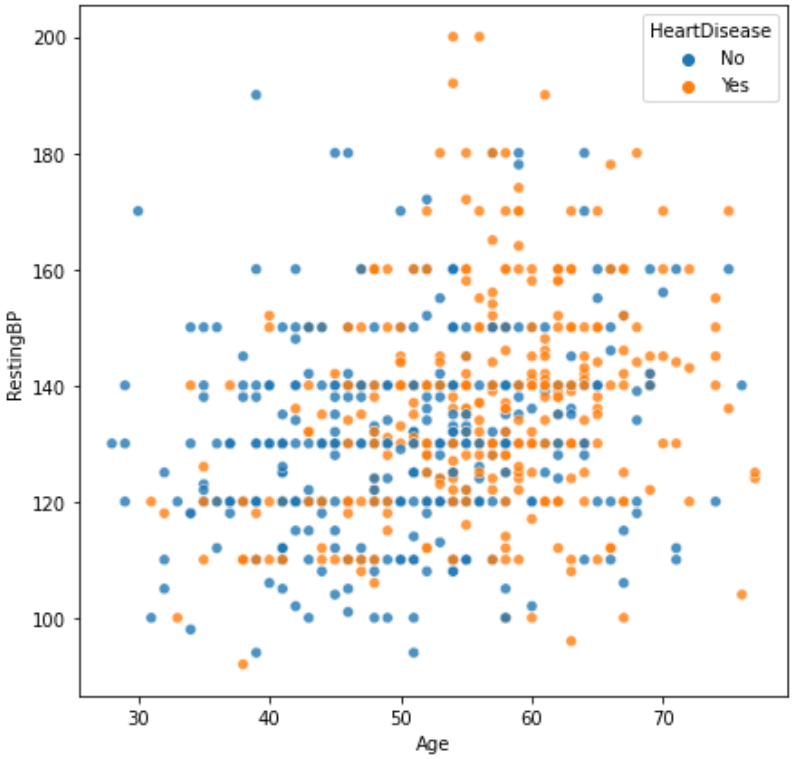
Prior to the removal of outliers, there were over twice as many male subjects as there were female subjects.



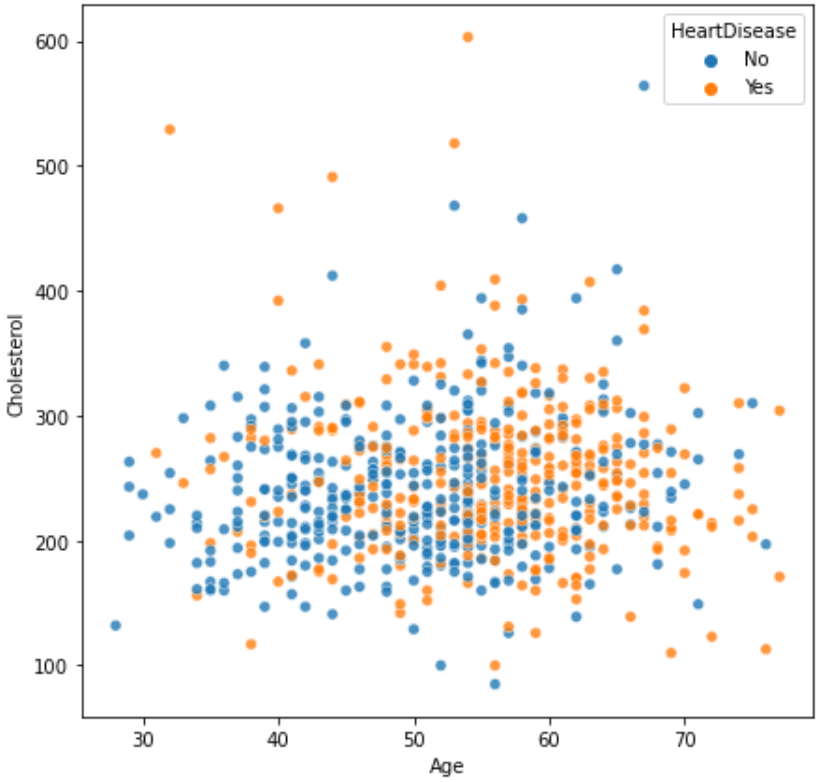
This graph depicts the slope of the ST segment of the subjects' electrocardiograms ("ECG"). The value recorded indicates whether this portion of the visual is recorded to have sloped upward, downward, or flat - the abnormalities in the ST segment can point to underlying disease or negative health conditions . These measurements are related to the Oldpeak metric, which represents the measure of ST segment depression (Domínguez).



This boxplot was created to highlight the presence of outliers in the dataset prior to removal. While the data for individual columns appears on the same scale, one can observe that the cholesterol column in particular seems to contain several outliers.



This graph illustrates the relationship between heart disease detected, resting blood pressure, and age. While there does seem to be a distinct correlation between resting blood pressure and heart disease, the graph seems to indicate that the presence of heart disease increases in older subjects.



This graph illustrates the relationship between heart disease detected, cholesterol, and age. Similarly, there does not appear to be a distinct correlation between cholesterol and heart disease. That said, this graph also seems to indicate that the presence of heart disease increases in older subjects.

Feature Engineering

After performing exploratory data analysis, I moved to perform feature engineering and prepare the dataset for modeling. This entailed creating dummy variables for all columns containing categorical data. Since these columns included data with rather low cardinality, I did not feel the need to restrict the columns that were one-hot encoded.

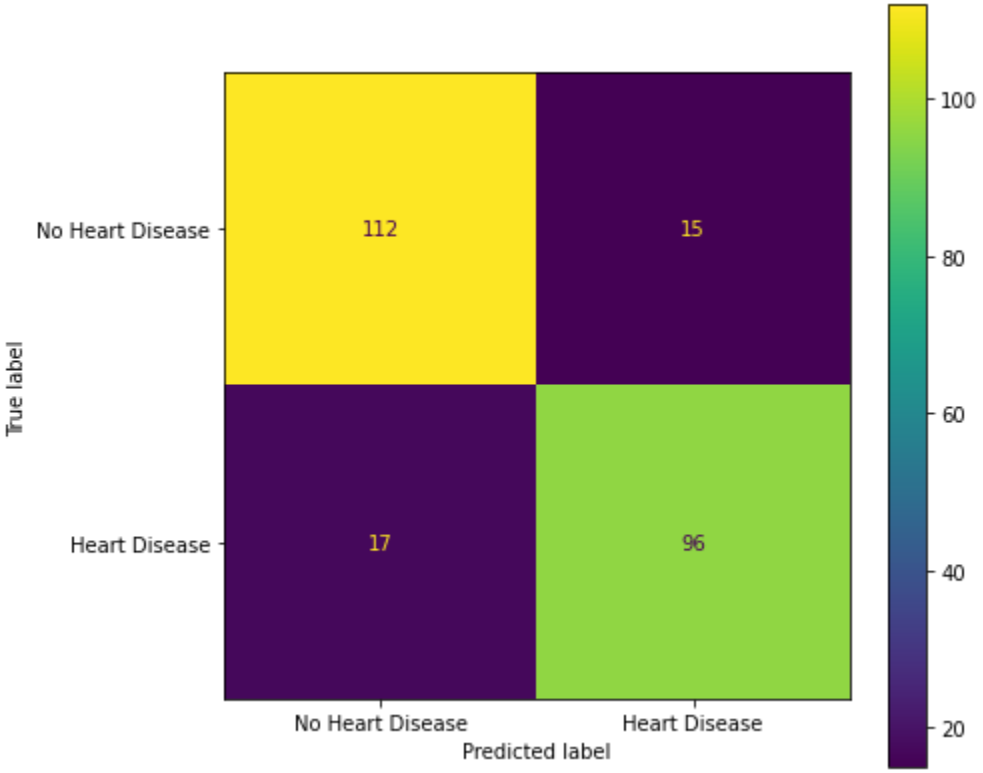
Once outliers were removed, age data was consolidated into buckets, and categorical columns were one-hot encoded, the data was split into training and testing sets. 33% of the data was reserved as the test set. Additionally, the data was split in a stratified manner such that the proportion of subjects having heart disease was nearly the same in both the training and testing data.

Prior to fitting the data to the models, the data was standardized using the StandardScaler class. This step may not have necessarily improved the performance of the model, but it can circumvent any errors in the underlying computation approach stemming from having data with different scales.

Model Selection and Results

Two models were compared: a gradient boosting model and a random forest regression model. 67% of the remaining data was used to train each of these models, while 33% was used as a control group.

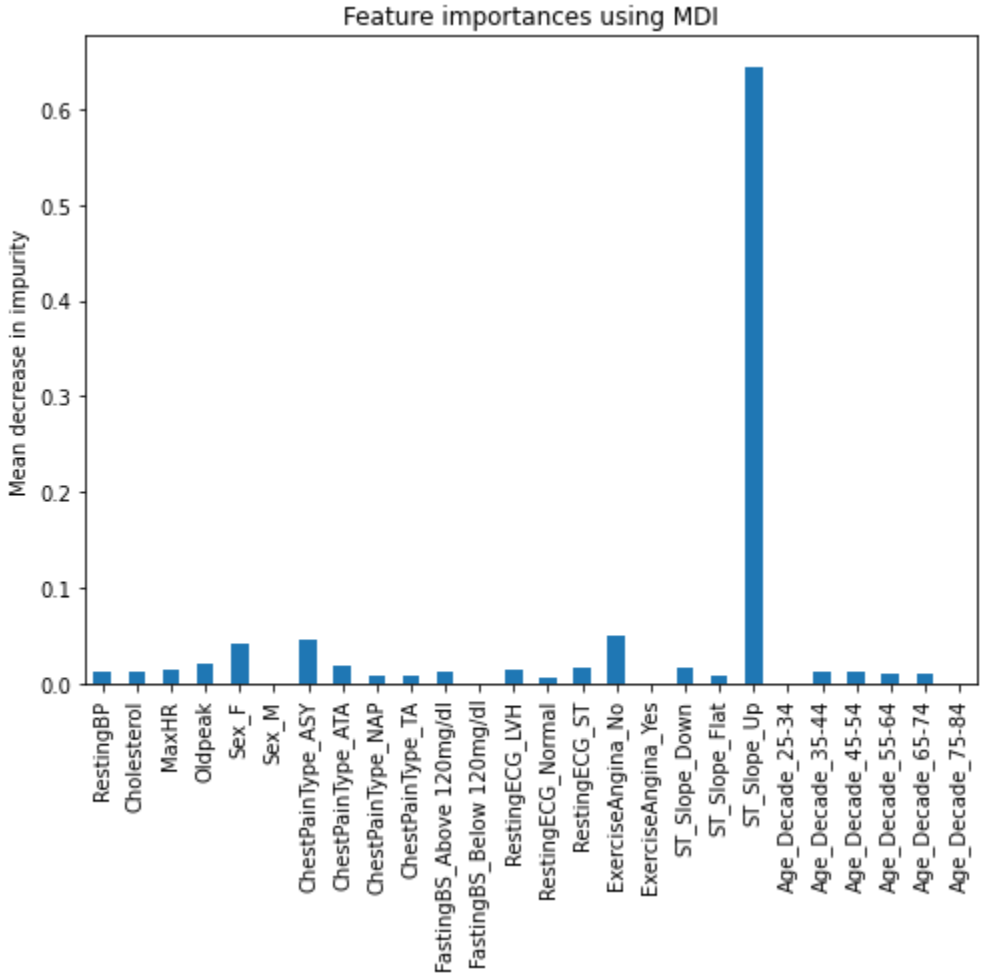
For the gradient boosting model, an implementation called Xtreme Gradient Boost ("XGBoost") was used. This implementation is tailored to controlling overfitting and has been noted to be markedly efficient. After the scaled data was fitted to the XGBoost classifier, I generated a confusion matrix to evaluate the algorithm.



The confusion matrix indicates that 15 subjects in the testing dataset that the model predicted to have heart disease did not actually have heart disease, representing false positives. Conversely, 17 subjects who were predicted not to have heart disease did in fact have heart disease, representing false negatives. These values yield a precision metric (the ratio of true positives to the sum of true positives and false positives) of 86.5% and a recall metric (the ratio of true positives to the sum of true positives and false negatives) of 85.0%. While having fewer false positives would be beneficial, the false negative results could have more severely negative

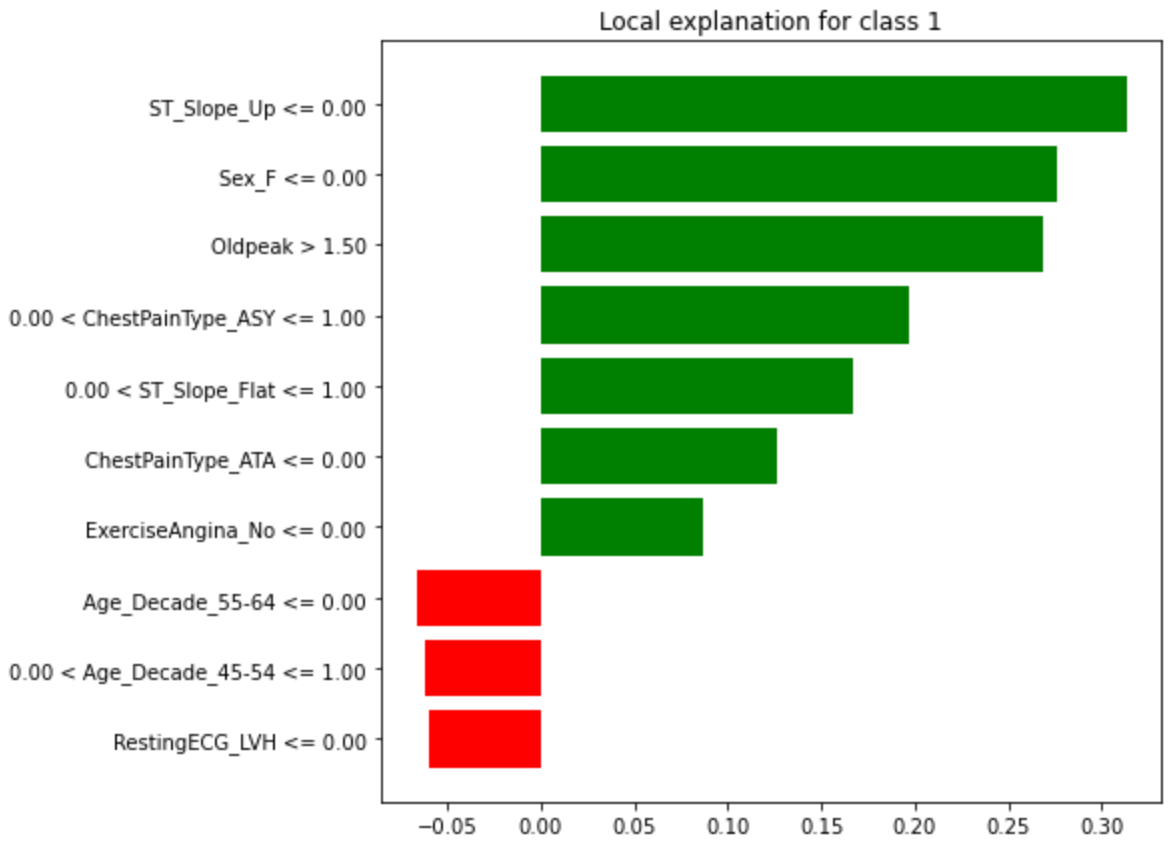
implications, as this could evidence the model underestimating the presence of heart disease in a given population. As such, the recall statistic should be designated as the evaluation metric of focus for this exercise.

After generating the confusion matrix, I used two approaches to visualize feature importance. The first approach assigns scores based on mean decrease in impurity.



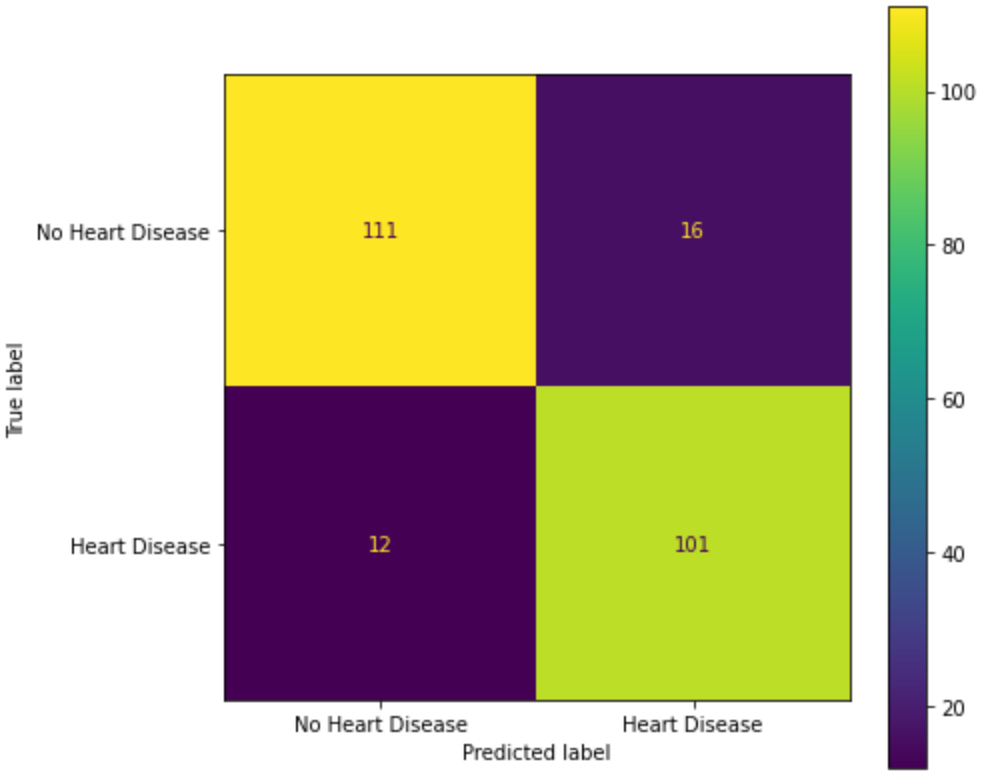
Here, the one-hot encoded “ST_Slope_Up” feature is highlighted as having a high score for classification. This feature has a value of 1 for subjects with an ST segment observed with an increasing slope and a value of 0 for flat or decreasing slopes. The remaining features have much lower scores, although this analysis could be less relevant for the numerical columns, as the approach may encounter problems with features having high cardinality.

The second approach employs local interpretable model-agnostic explanations (“LIME”).



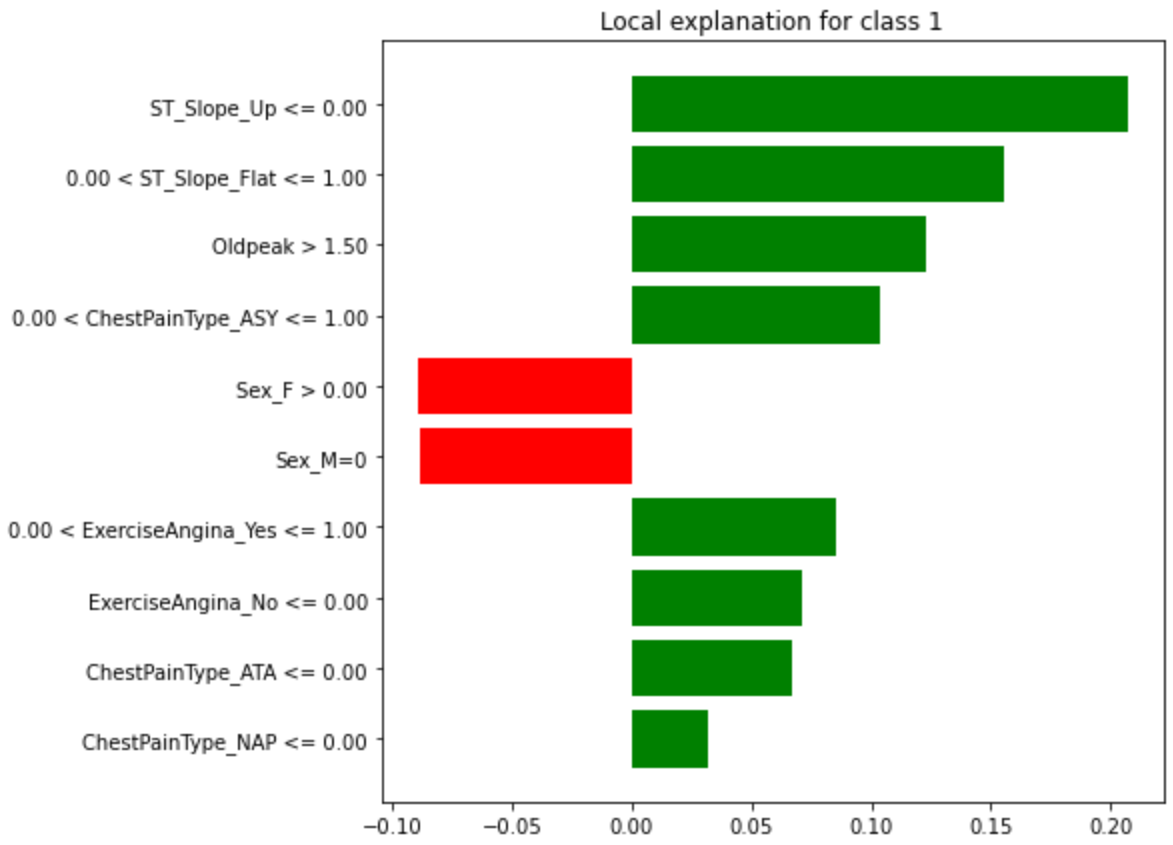
LIME provides a set of weights indicating feature importances for the classification of a single subject in the dataset. For the subject depicted in the above graph, the ST_Slope_Up feature being equal to 0 has a high importance, relative to the other features shown, in classifying this subject as having heart disease.

The second model used in the analysis was a random forest classification model. This model had a specific set of benchmarks chosen as the hyperparameters (max_depth = 10, min_samples_leaf = 3, min_samples_split = 4, and n_estimators = 200). After scaling and fitting the data to the model, I generated the associated confusion matrix.



While there was one more false positive in this analysis as compared to the XGBoost model, there were five fewer false negatives. As such, the precision metric was 86.3%, and the recall metric was 89.4%.

The LIME explanation for a classification using the random forest model is displayed below.



Again, the “ST_Slope_Up” feature is shown to be more important for the classification as compared to the other features.

Takeaways

The two ensemble machine learning algorithms employed for this classification exercise seemed to perform well. Each had precision and recall metrics exceeding 80.0%. That said, we could have improved the performance of the random forest classification model by performing hyperparameter tuning. Through a combination of RandomizedSearchCV and GridSearchCV, we could have a set of hyperparameters to optimize the performance of the random forest model. That said, given the model’s current performance, the additional benefit may be modest.

The feature set for this analysis was somewhat limited, as only 11 independent variables were used for heart disease classification. The compiled dataset also contains observations collected only from 5 data sources, which seem to tie to geographic locations. It would be interesting to see how additional demographic data related to subject race and income as well as more precise subject geographical location could impact the analysis. Beyond the model performance, such information could yield valuable insights as to what specific subpopulations may experience a greater presence of heart disease due to living conditions.

Bibliography

CDC – National Center for Health Statistics – Leading Causes of Death. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. February 3, 2022.

fedesoriano. “Heart Failure Prediction Dataset.” *Kaggle*, 10 Sept. 2021, https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Domínguez, Carlos. “A Detail Description of the Heart Disease Dataset.” *Kaggle*, Kaggle, 20 Apr. 2020, https://www.kaggle.com/carlosdg/a-detail-description-of-the-heart-disease-dataset.