Quinn Meier
Springboard - DS Career Track
Capstone 2

Final Report
Heart Disease Capstone

Problem Statement

According to the World Health Organization, Cardiovascular Disease is the number one cause of death worldwide. As with many medical conditions, early detection and treatment of heart disease can dramatically improve patient outcomes. A model that can take a small sample of basic health information from a patient, namely factors like age, resting blood pressure, cholesterol levels, type of chest pain, etc., and rapidly assess their likelihood of having heart disease could be a powerful tool for strained healthcare systems around the world.

The data utilized in training these models comes from the famous 1988 Cleveland Clinic Heart Disease dataset hosted on the University of California, Irvine Machine Learning Repository. The dataset consists of 76 health attributes of 303 anonymous patients with and without heart disease, with some missing values.

Several supervised machine learning classification models were used to address this problem, which will be explored more deeply in this report.

Data Wrangling

The raw dataset compiled by the Cleveland Clinic contains 303 rows with 76 columns. The version hosted on the UC Irvine data repository pared down the number of parameters to 14, which is the standard for these types of analyses. Of the 303 rows, 4 had missing values. These rows were dropped because they were categorical variables that didn't seem appropriate to impute.

To treat potential outliers, the z-scores of the continuous features were examined. Any value with a z-score > 3 was removed. These outliers came from the cholesterol and resting blood pressure parameters. Applying this method assumes that all of these parameters are normally distributed, which may not be accurate in all instances. The final shape of the dataset was 290 rows x 14 columns.

Exploratory Data Analysis

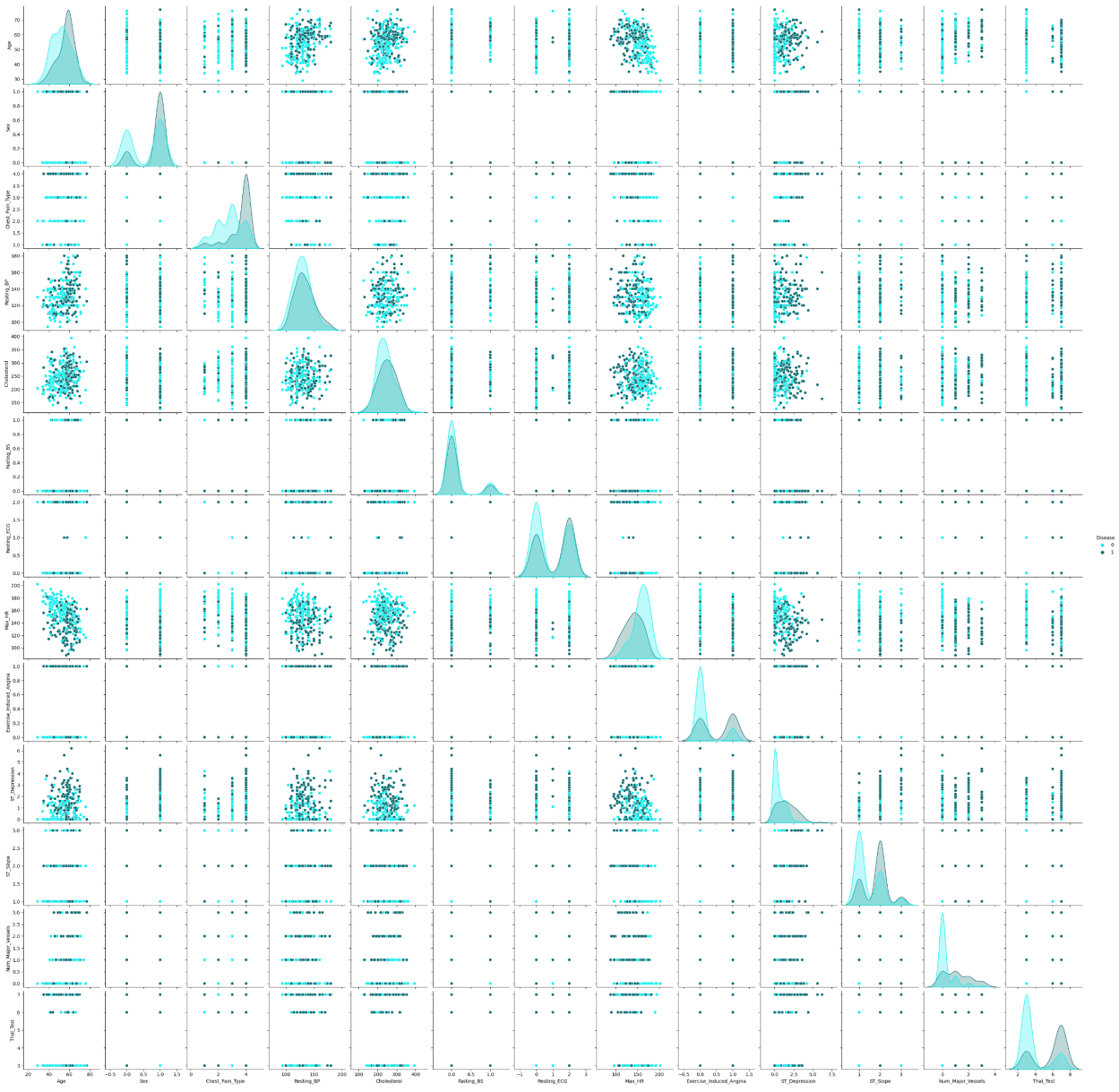
Once the data had been cleansed of outliers and missing values, exploratory data analysis could begin. This included examining the structure and searching for

trends, patterns, and statistical relationships between the dataset's features, both visually and numerically.

An explanation of the variables in the dataset is helpful for context.

- age: patient age in years
- sex: patient sex (1 = male, 0 = female)
- cp: Type of chest pain experienced (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
- trestbps: patient resting blood pressure (mmHg measured at intake into hospital)
- chol: patient cholesterol level (mg/dl)
- fbs: patient fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- restecg: patient resting electrocardiograph measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- thalach: patient maximum heart rate achieved
- exang: exercise-induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)
- slope: the slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
- ca: number of major vessels (0-3)
- thal: thallium tracer test result, coronary stenosis detected when healthy heart cells uptake tracer is called reversible defect, while no uptake due to prior heart attack is fixed defect. (3 = normal; 6 = fixed defect; 7 = reversible defect)
- target: presence of heart disease (0 = not present, 1 = present.)

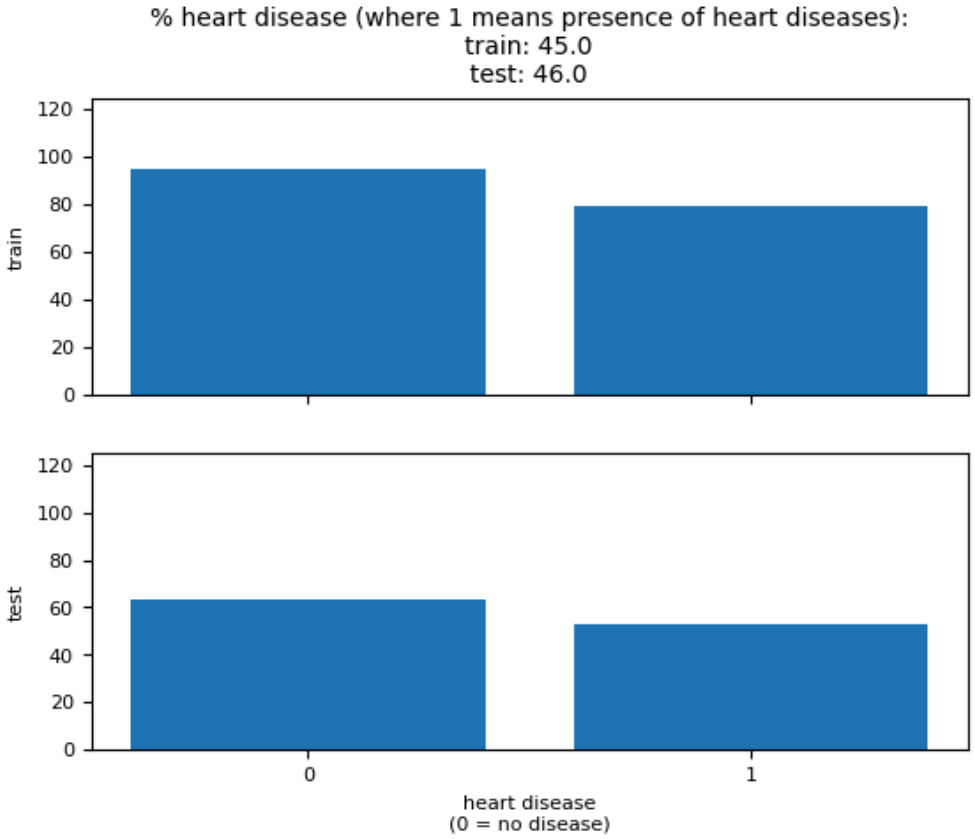
Below is a combined visualization plotting the above variables against each other. This is a portion of the Seaborn library known as a pairplot. The light blue data points represent patients without heart disease, and the grey points represent patients with heart disease. Several trends become apparent, such as patients with heart disease being more likely to be asymptomatic concerning chest pain. This effect is expected because patients without disease experiencing chest pain are perhaps more likely to be enrolled in a study like this. In contrast, patients with disease may be enrolled in the study for other reasons. Similarly, patients with thallium test results of fixed or reversible defect, indicating damage to the heart muscle, are more likely to have heart disease. Additionally, correlation and chi-squared tests were performed on the relevant combinations of variables, leading to some interesting insights.



Preprocessing and Training

Several additional preparatory steps were necessary before the data could be fed into a machine learning model. Namely, categorical features are split into what are known as ‘dummy variables,’ and the data are divided into training and test sets. Converting categorical variables into dummy variables split up over multiple features allows the treatment of qualitative features in the analysis. The features which were split up for this analysis were 'Sex,' 'Chest_Pain_Type,' 'Resting_ECG,' 'ST_Slope,' 'Thal_Test,' and 'Num_Major_Vessels.' ‘Num_Major_Vessels’ and ‘ST_Slope’ were

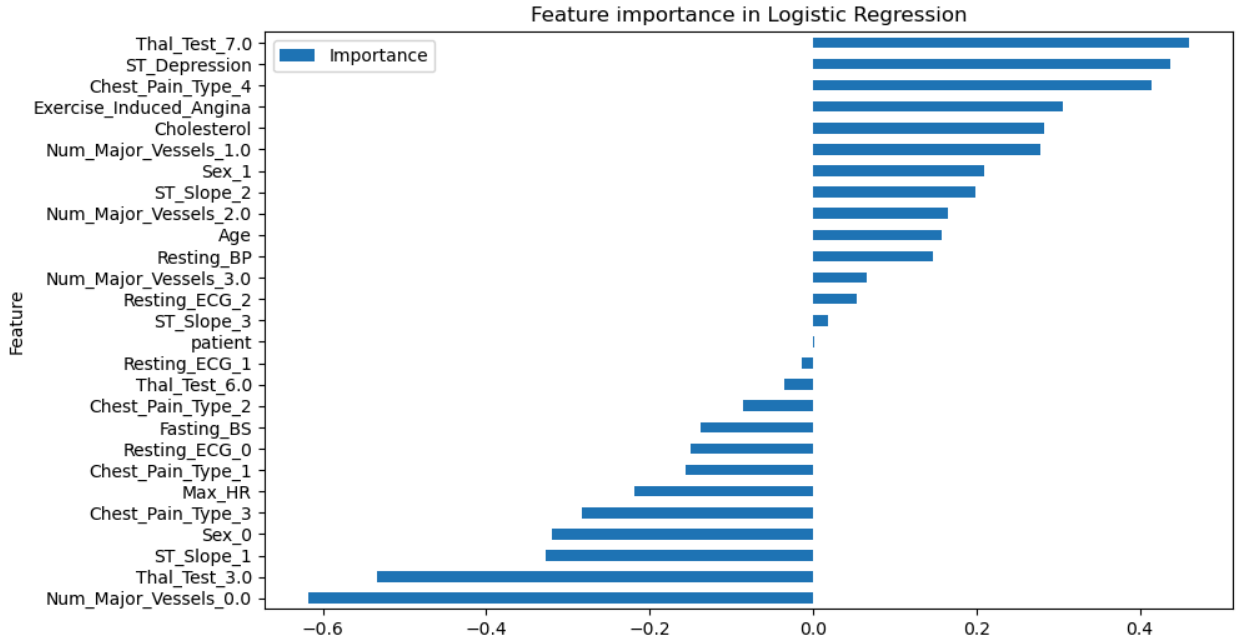
considered categorical variables because they have discrete integer values. A ratio of 60% training data to 40% test data was chosen to segregate the training and test sets. This split was selected to prevent overfitting, which occurred with higher proportions of training data. The data was also stratified along the target variable to ensure similar ratios of healthy and sick patients in the training and test sets. This stratification can be seen in the figure titled '% heart disease,' with 45% of the training set and 46% of the test set consisting of patients with heart disease.



Modeling

Several different machine learning classification algorithms were utilized for the modeling phase of this analysis. Each will be explored individually to discuss its respective strengths and weaknesses. Each model underwent grid search cross-validation to optimize its hyperparameters and was evaluated using accuracy, recall, ROC-AUC, and confusion matrices to determine which performed the best. A table of those hyperparameters and the evaluation scores for each model is provided below.

The first model considered was logistic regression. Logistic regression is the standard baseline for almost all binary classification problems and should be the starting point for any modeling of this kind. It is straightforward to implement, interpret, and explain. It works well for a medical application because it provides probabilities for outcomes, which can help make clinical decisions based on risk thresholds. Applying this model assumes that each entry in the dataset is unrelated, that linear relationships exist between the parameters and the target, and that multicollinearity is absent. In this case, all assumptions are satisfied. This fact can be seen from the correlation heat map from the EDA step of the analysis. With that consideration in mind, the model performed reasonably well, as expected, with an accuracy of 85%, a recall of 86%, and an AUC score of 0.92. The model feature weights are seen in the figure below. Positive feature weights indicate contributions to the model leaning towards a positive diagnosis, while negative feature weights work oppositely.



The second model considered was support vector machines (SVM). SVM performs well for datasets with many features. It measures some degree of “closeness” in the hyperspace of the parameters of the dataset. These features of SVM allow it to capture complex relationships between parameters that would otherwise be virtually impossible to tease out. On the other hand, these same factors cause SVM to be less interpretable than other models like logistic regression or decision trees and become computationally prohibitive with large datasets. In this case, the latter was not an issue. SVM also requires careful tuning of its hyperparameters to achieve the best results. In

this case, SVM scored an accuracy of 84%, a recall of 80%, and an ROC-AUC score of 0.92.

The third model considered was decision trees. Decision trees are non-parametric models that recursively partition the data into subsets, making them excellent for handling complex pattern recognition tasks involving variable interactions. This non-parametric nature makes them incredibly robust and versatile but also gives them a strong tendency for overfitting unless steps are taken to control tree size. Decision trees are highly interpretable and can handle classification and regression tasks with numerical and categorical data. In this instance, decision trees were the model with the weakest overall performance, with accuracy and recall scores of 78% and an ROC-AUC score of 0.83.

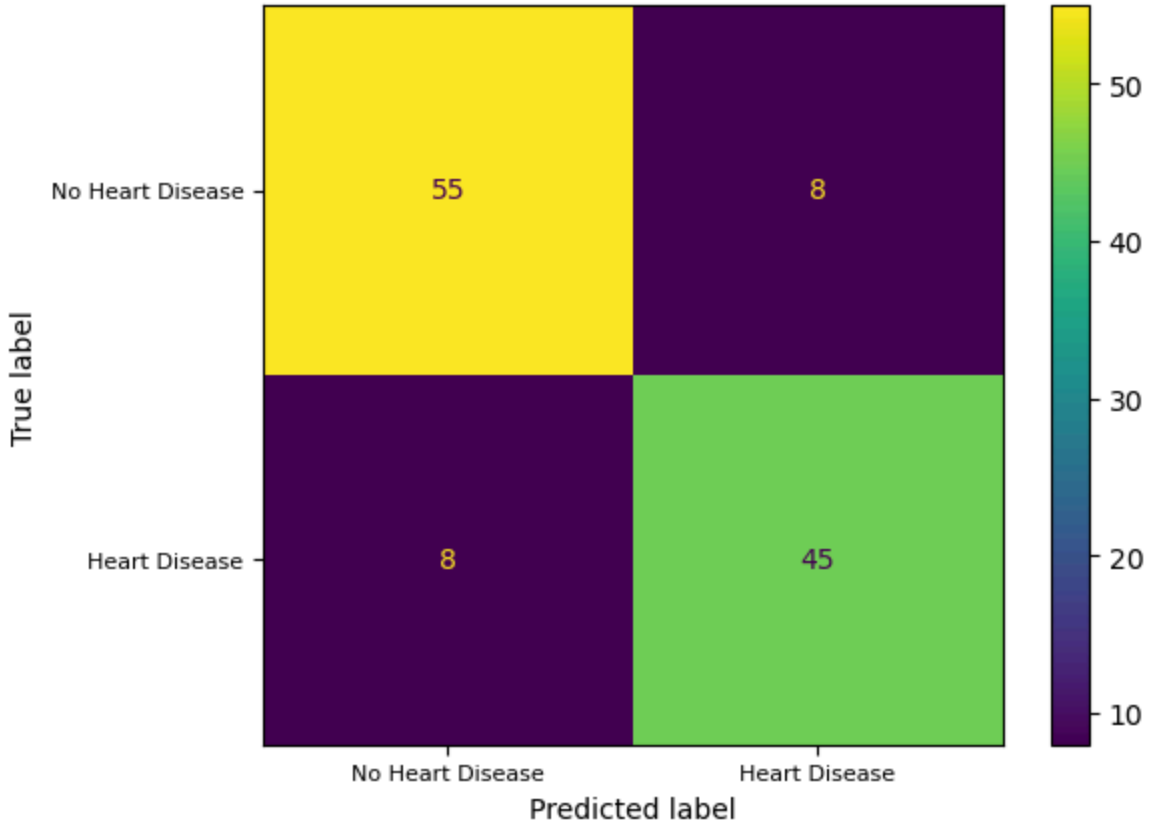
The fourth model considered was random forest. Random forest is an ensemble of decision trees that improves prediction accuracy and controls overfitting by averaging multiple deep decision trees trained on different parts of the training set. RF generally represents a significant improvement in accuracy over a single decision tree at the cost of reduced interpretability. In this instance, RF performed well with accuracy and recall scores of 86% and the best ROC-AUC score of 0.93.

The fifth and final model considered was gradient boosting (GB). GBMs are ensemble methods that combine multiple weak learners to reduce a loss function. GB updates weights by computing the negative gradient of the loss function with respect to the predicted output. The loss function is the difference between the actual and the predicted variables. Gradient boosting can achieve some of the best accuracies possible with machine-learning models; however, it tends to overfit if not tuned properly and is less interpretable than other models. In this case, gradient boosting was the second-best model with respect to overall performance, with an accuracy score of 86%, a recall score of 88%, and an ROC-AUC score of 0.92.

models	Logistic Regression	SVM	Decision Tree	Random Forest	Gradient Boosting
hyperparameters					
random_state	123	123	123	123	123
training sample size	60%	60%	60%	60%	60%
solver / kernel / criterion	liblinear	linear	gini	n/a	n/a
C	0.1	0.1	n/a	n/a	n/a
max_depth	n/a	n/a	3	3	2
max_features	n/a	n/a	17	17	17
min_samples_split	n/a	n/a	3	3	3
min_samples_leaf	n/a	n/a	2	1	1
learning_rate	n/a	n/a	n/a	n/a	0.1
subsample	n/a	n/a	n/a	n/a	0.8
n_estimators	n/a	n/a	n/a	2500	20
performance					
accuracy	0.85	0.84	0.78	0.86	0.86
recall	0.86	0.80	0.78	0.86	0.88
ROC	0.92	0.92	0.83	0.93	0.92

Conclusions:

Applying machine learning in a medical context brings about an exciting set of challenges and considerations, such as the interpretability of the decisions made by models and ensuring that as many patients suffering from disease are correctly identified. In this case, logistic regression is the best all-around model. Although LR performed slightly worse in accuracy and recall than random forest and gradient boosting, it makes up for it by being more straightforward to implement and interpret, more generally applicable and scalable, and having the lowest incidence of false negatives, as seen in the confusion matrix below. Having the lowest possible incidence of false negatives is crucial in this application, as it is far worse for patient outcomes to incorrectly ‘clear’ a sick patient than to misdiagnose a healthy one.



Future Research:

This analysis could be extended to a stratified study along different racial and ethnic groups to ensure that the models can perform adequately when these factors are considered and aren't underperforming in their analysis of a particular group. Similar studies could also be performed for many medical conditions other than cardiovascular disease, assisting physicians in quickly diagnosing patients and ensuring they receive the required treatment. Even further, this type of analysis could be combined with other machine learning techniques to predict severe medical emergencies, such as heart attacks and strokes, before they happen, allowing the patient to seek out preventative care. Ultimately, medical applications represent one of the most fascinating and impactful opportunities for machine learning to be used to create a better world for all.