Predicting the Presence of Heart Disease William Mao

According to the CDC, heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One in every four deaths (before 2020) annually are attributed to heart disease. The goal of this project then, is to build a predictive model based on given features and determine if they can give an accurate diagnosis of heart disease based on those features to treat patients earlier and prevent death.

The data used in this predictive model was scraped from: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/. Features were various health metrics (such as heart rate) taken from patients from the Cleveland Clinic. Data cleanup was required to make the data usable. All health metrics were selected for the model and overall, the dataset contained 14 columns and 296 usable observations. Presence of heart disease in this dataset and model was defined as any patient presenting with >50% artery diameter narrowing.

Examining **Figure 1**, we can see the distribution of cholesterol and maximum heart rate levels for individuals with and without heart disease. The distributions for cholesterol appear similar while there appears to be some difference in max heart rate between the two groups. Indeed, the average and standard deviation cholesterol level of those without heart disease was 243.59±53.92 mg/dl while those with heart disease was 251.85±49.68 mg/dl. The average and standard deviation bpm of patients without disease was 158.63±19.09 bpm while those with heart disease was 139.11±22.71 bpm. Overall, this initial visualization of the data suggests that some features may not be helpful in distinguishing the two categories.

Moving on to **Figure 2**, we used a Principal Component Analysis on our 14 features to explain the cumulative explained variance per number of components. Looking at the unscaled plot, we can see that unscaled, using just 2 features are able to capture 89.6% of the variance. After repeating PCA with scaled data, we can see that reducing features from 14 to 10 still allows us to capture 88.6% of the variance.

Next, we wanted to predict whether a patient with given health metrics presents heart disease or not. To create the model, we created a sklearn pipeline with (1) make column transformer with Polynomial features and OneHotEncoder, (2) StandardScalar, and (3) LogisticRegression. Components 2 and 3 were used as this was a classification problem. After doing a 25%/75% train/test split on the data. Our model achieved an accuracy score of 85.14%, a recall score of 78.79%, and a precision score of 86.67%. As such, the precision score indicates that the model fairly accurate when it predicts that the individual has a heart disease.

Lastly, for **Figure 3**, we have plotted the 10 strongest features with their coefficients. The most important factor for predicting whether someone has heart disease is the patient's resting blood pressure with a coefficient of 0.58. However, the strongest factor in predicting against a patient having heart disease is getting the fewest vessel marked during fluoroscopy with a coefficient of -0.69. This was not unexpected since fewer blood vessels marked during fluoroscopy generally signifies fewer blockages in the blood vessel. Note that this was a categorical feature rather than a continuous feature that placed patients on a rank.

In conclusion, using these features can help indicate the presence of heart disease in a patient. However, it is not infallible and the possibility of a misdiagnosis using this model is very real. Therefore, we recommend that this model be used as a supplemental tool in conjunction with existing medical diagnostic procedures.

Figure 1: Distributions of Features for Individuals with and without Heart Disease

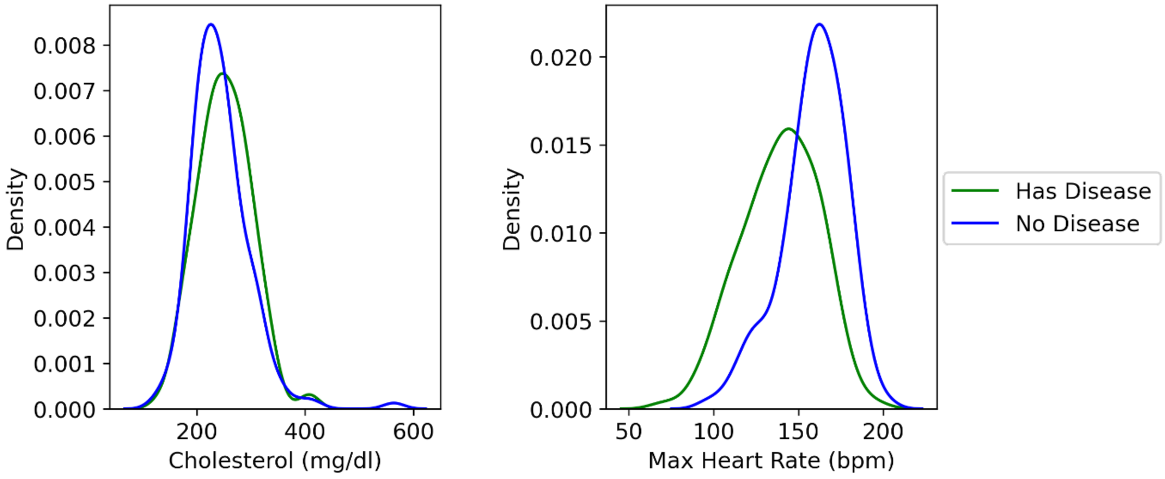


Figure 2: Cumulative Principal Components

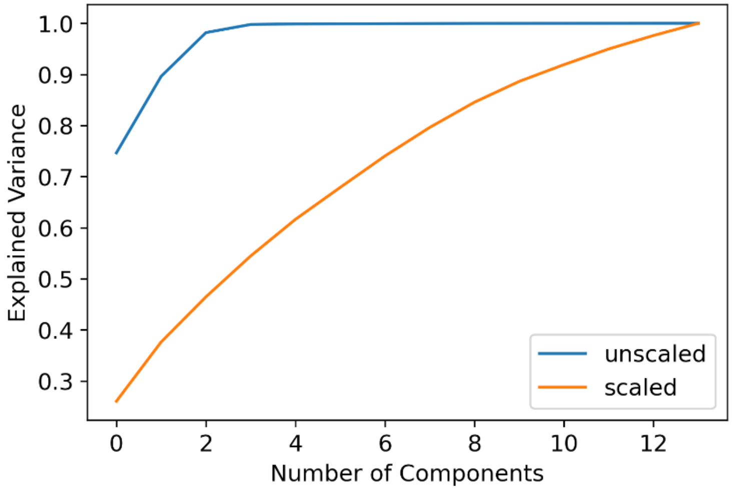
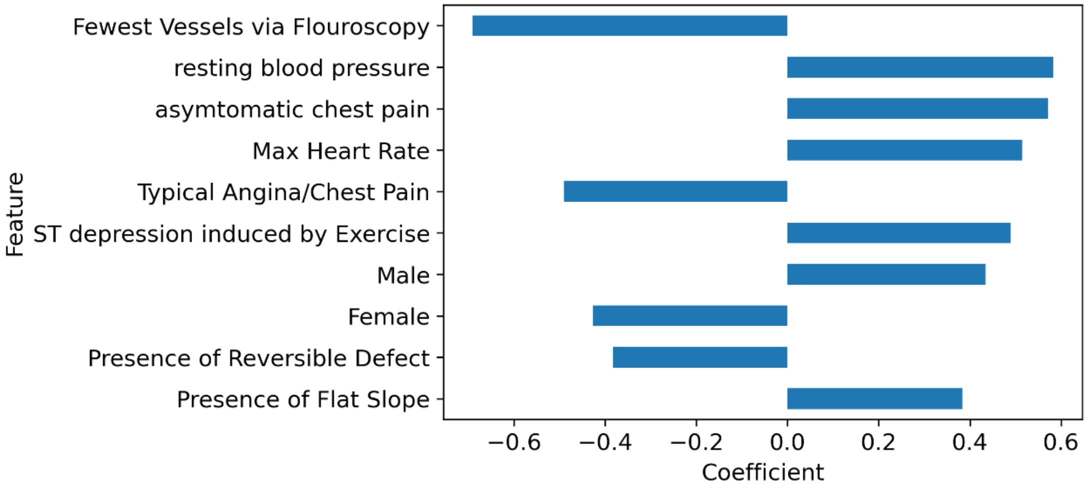


Figure 3: Logistic Regression Coefficients


Feature
Fewest Vessels via Flouroscopy
resting blood pressure
asymtomatic chest pain
Max Heart Rate
Typical Angina/Chest Pain
ST depression induced by Exercise
Male
Female
Presence of Reversible Defect
Presence of Flat Slope
Coefficient