Heart disease is the leading cause of death in the United States for men, women, and most racial and ethnic groups. About 1 in every 5 deaths in the United States is the direct result of heart disease, and someone dies of the condition once every 30-45 seconds. In economic terms, heart disease costs the country between 2 and 3 billion dollars every year on everything from healthcare services and medicines to lost productivity due to death. The effects are enormous, and heart disease identification and prevention remain one of the most important topics in the healthcare industry.

Increasingly, the ways that heart disease risk and identification get addressed are changing. Primary prevention models, once the gold standard in medicine, are not accurate enough to be effective. Low estimates put their accuracy at just under 60% while the highest ones place them around 80%. The goal of this project was to come up with a supervised learning model for heart disease that could outperform the American College of Cardiology / American Heart Association (ACC / AHA) risk score primary prevention model and predict heart disease with greater than 80% accuracy.

To do this, I used the UCI Heart Disease Dataset consisting of sixteen columns of data (one unique identifier, a location identifier, thirteen predictive features, and one target feature) and 920 observations. Four sub-datasets were merged to form this dataset, and each sub-dataset corresponds to a different location that the data was collected at (Cleveland, Hungary, Switzerland, and VA Long Beach). Three of the predictive features were missing a third or more of their values and about 6% of the patients were missing 40% or more of their data and were not included in the modeling process. Any missing data that remained was imputed. This left 865 patients and 10 predictive features (five continuous and five binary features) from which the model was built. The ten predictive features were as follows:

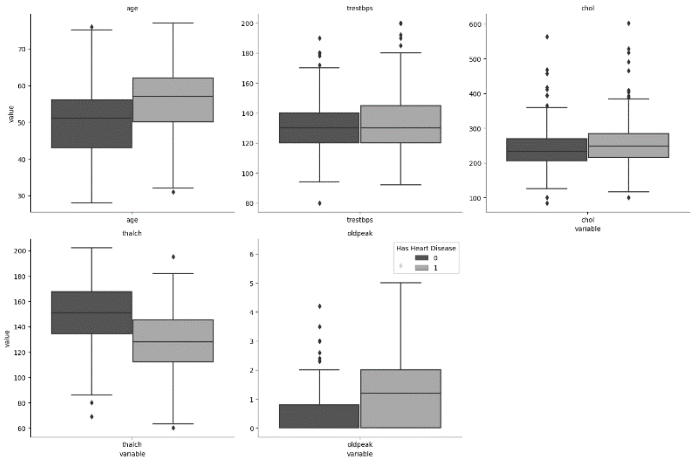
Continuous

- Age - Age of the patient in years
- Trestbps - Resting blood pressure (in mm Hg) on admission to the hospital
- Chol - Serum cholesterol (in mg/dl)
- Thalch - Maximum heart rate achieved
- Oldpeak - ST depression induced by exercise relative to rest

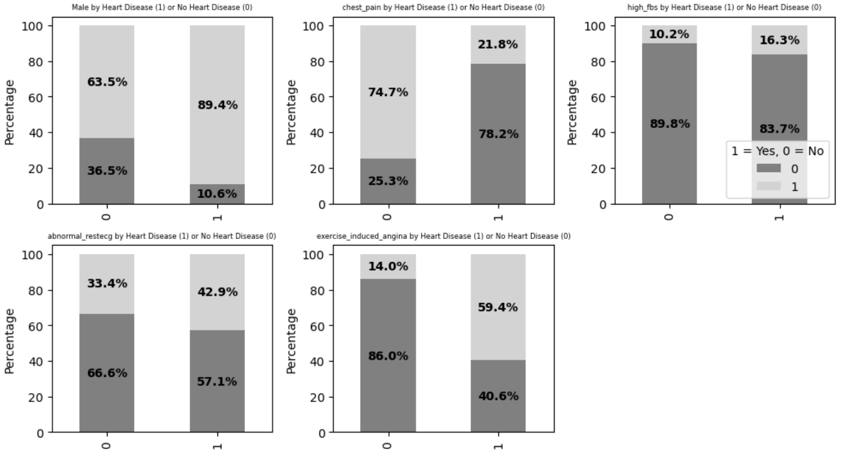
Binary

- Male - Whether the patient is a male
- Chest-pain - Whether the patient is experiencing chest pain
- High-fbs - Whether fasting blood sugar is above 1200 mg/dl
- Abnormal_restecg - Whether the electrocardiogram results were abnormal
- Exercise_induced_angina - Whether the patient had exercise induced angina

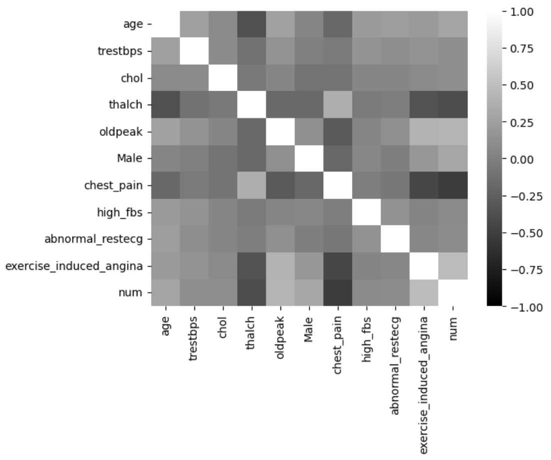
While some of the locations that the data was collected at provided the stage of heart disease the patient had, this project specifically focused on classifying patients as (1) “has heart disease” and (0) “does not have heart disease”. Bonferroni corrected t-tests conducted on the continuous predictive features showed significant differences between the “heart disease” and “no heart disease” groups at the .05 level for all five features (adjusted p-values were all .015 or less), and Bonferroni corrected chi-squared tests of the binary features showed significant differences at the .05 level for all but one of the features (high_fbs had a p-value of .06). All features were positively correlated with “has heart disease” with the exception of thalch (maximum heart rate achieved) and chest-pain (1 indicating has chest pain).



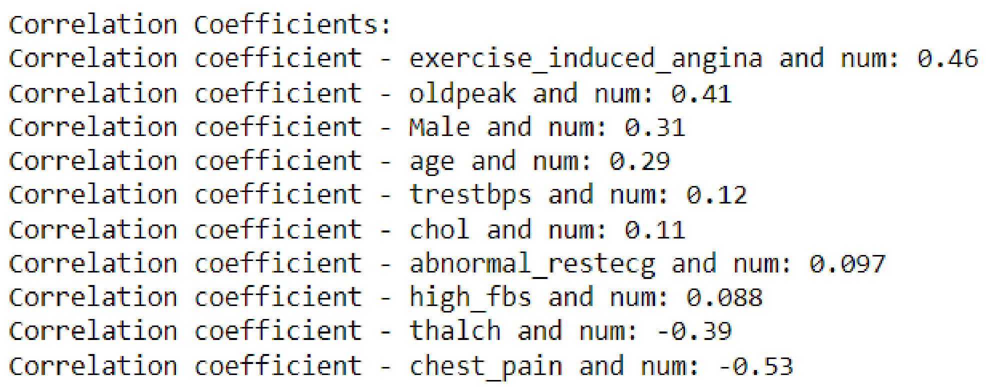
1. Continuous Features



2. Binary Features



3. Heatmap of Correlation Coefficients



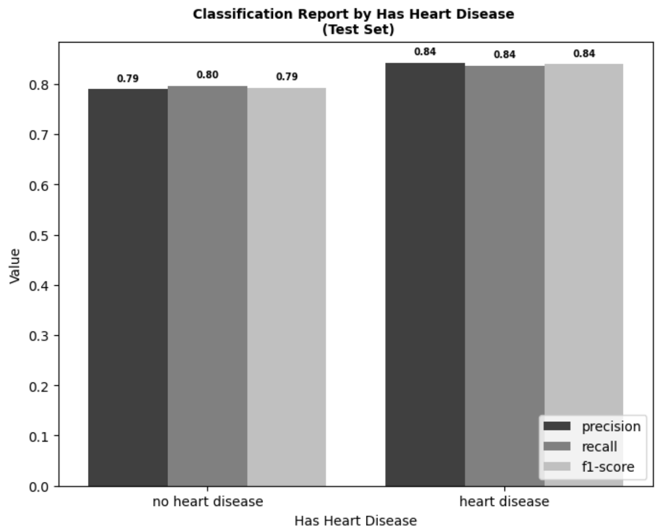
```
Correlation Coefficients:
Correlation coefficient - exercise_induced_angina and num: 0.46
Correlation coefficient - oldpeak and num: 0.41
Correlation coefficient - Male and num: 0.31
Correlation coefficient - age and num: 0.29
Correlation coefficient - trestbps and num: 0.12
Correlation coefficient - chol and num: 0.11
Correlation coefficient - abnormal_restecg and num: 0.097
Correlation coefficient - high_fbs and num: 0.088
Correlation coefficient - thalch and num: -0.39
Correlation coefficient - chest_pain and num: -0.53
```

4. Correlations Between Predictive Features and Outcome Variable

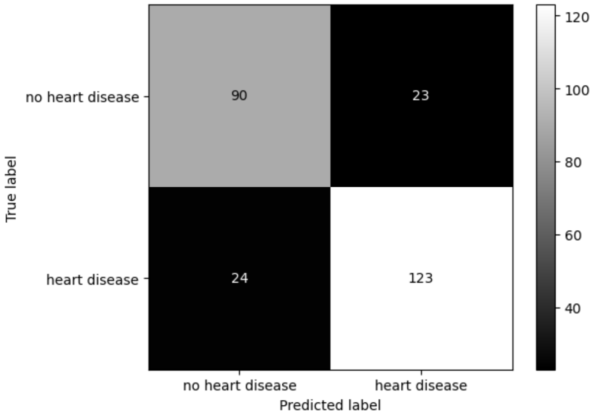
Five different models were evaluated (Logistic Regression, AdaBoost with a Decision Tree estimator, AdaBoost with an SVM Estimator, AdaBoost with a Logistic Regression Estimator, and a Random Forest) during training before a final model was selected. Prior to training the models, data was first split into training (70%) and testing (30%) sets and missing values were then imputed. Any scaling of the features that was being tested was also done after splitting the data into training and testing sets and before training the model. Hyperparameters were selected using GridSearchCV for the Logistic Regression models and RandomizedSearchCV (n_iter = 180) for the other models (to reduce training time). One of each type of model was selected for the final model selection process with the hyperparameters determined by those with the highest cross-validation F1 score.

For the final model selection, the model with the highest F1 score on the testing set was chosen as the best model. This was the Adaboost model with Decision Tree estimator (F1 score = 0.84), which narrowly outperformed the Logistic Regression model and other Adaboost models. The final model used min-max scaling and missing values were imputed using the median.

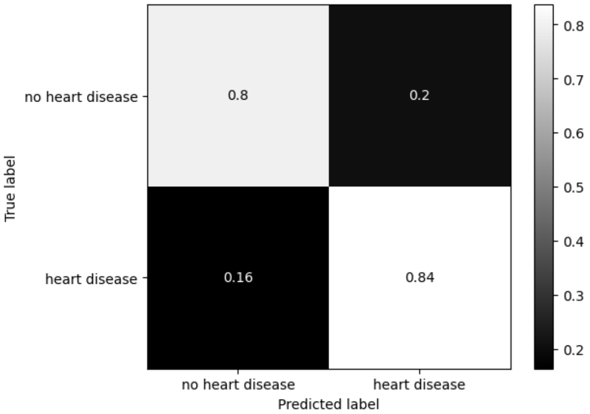
Further analysis of the testing set revealed that the model was better at classifying heart disease patients than those without the condition.



5. Classification Report (Test Set)

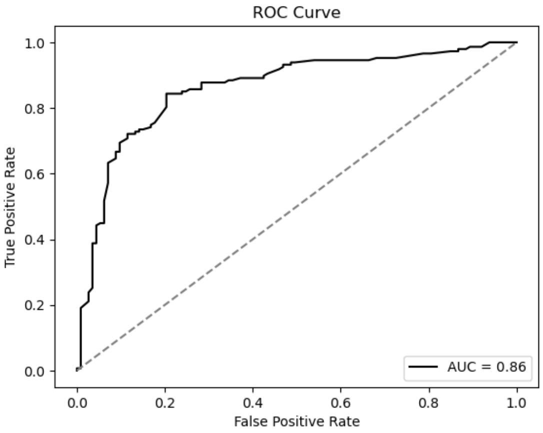


6. Confusion Matrix (Test Set)



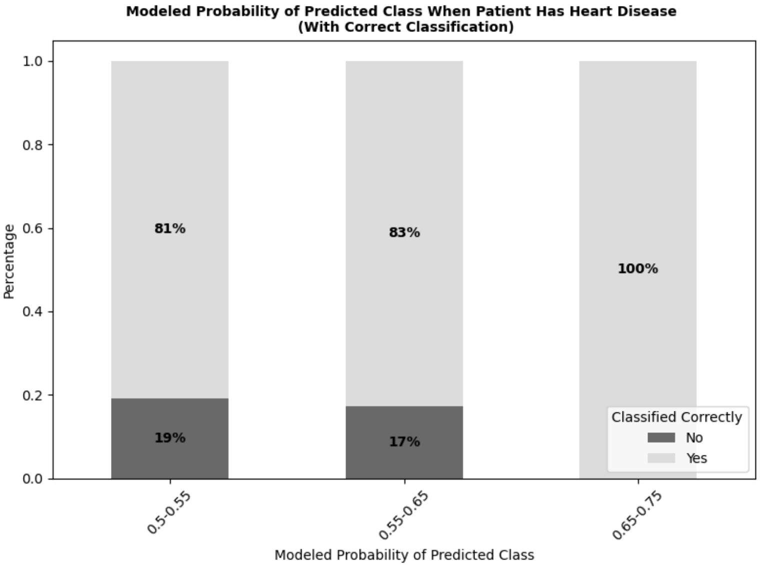
7. Confusion Matrix - Normalized (Test Set)

Plotting the true positive rate against the false positive rate, however, showed that the model was doing a really good job overall.

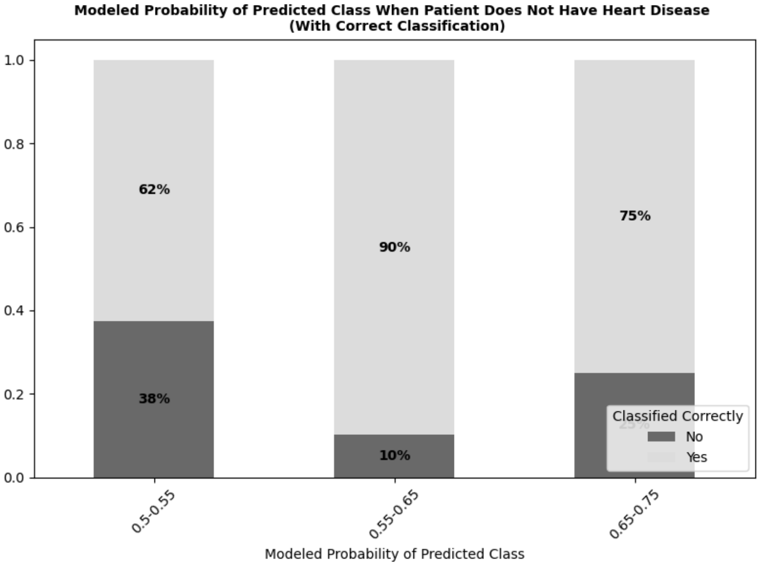


8. ROC Curve

As expected, the model tended to perform better when it was more confident in a prediction. This effect was more pronounced for instances where the patient had heart disease than in cases where the patient did not.

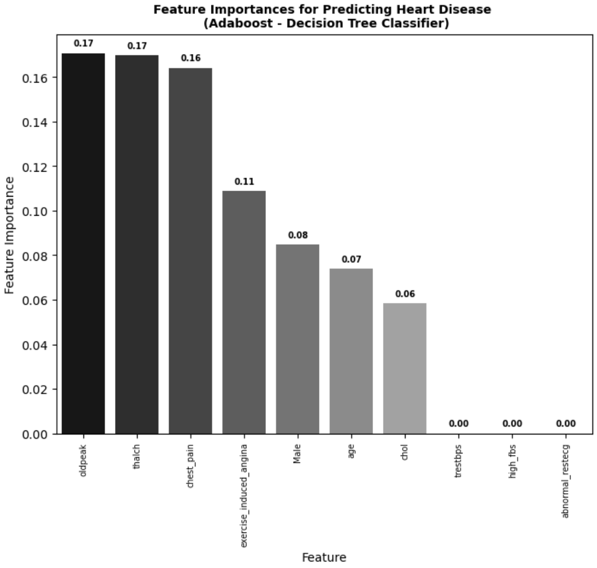


9. Probability of Predicted Class (Has Heart Disease)



10. Probability of Predicted Class (Does Not Have Heart Disease)

The Decision Tree estimator in the model gave some insight into which features were most important in classifying the patients. ST depression induced by exercise relative to rest (oldpeak) was the most important factor, followed by maximum heart rate achieved (thalch), and whether or not a patient had chest pain. Exercise induced angina, sex, age, and cholesterol levels were also important factors.



11. Feature Importances

Overall, the model was doing an excellent job at predicting heart disease and met the initial goal of achieving over 80% accuracy. The model was better at identifying instances where a patient had heart disease than when it did not (which is good in this case). It is safer to make a Type I error (false positive, saying that a patient has heart disease when they do not) than a Type II error (false negative, saying that a patient does not have heart disease when they do) since patients are better off taking precautionary measures with their lifestyle if they are a fringe case. Since the model also output predicted probabilities of each class (and the predicted classes were more accurate as the probabilities increased), these metrics could also be used to inform patients on the likelihood that they have heart disease.

Since two of the top 4 features of the model were related to exercise, future research might want to look at other variables related to exercise to see if they help inform the model even more. Adding more features (both related to exercise and not related to exercise) and data from other hospitals may also make the model more robust and help improve the accuracy of its predictions in a real-world setting. Three predictive features in this dataset were missing a third or more of their values and were not included, so collecting this information or trying to impute it might be a good place to start. Other ideas for future exploration are to see whether the predicted probabilities could be used to come up with a risk score for patients so that they could make educated decisions about their health. It might be more useful to have a heart disease risk score than a binary answer to the question of whether a patient has heart disease and future projects should explore this possibility.

By improving the ability to detect heart disease by just a small percentage, millions of dollars and thousands of lives can be saved. This project attempted to do that and future projects can expand on its work to make further improvements. Heart disease isn’t going away anytime soon, but by better identifying it, we can improve outcomes and work towards a healthier future.

References

Arshi, B., van den Berge, J. C., van Dijk, B., Deckers, J. W., Ikram, M. A., &amp; Kavousi, M. (2021, February 16). Implications of the ACC/AHA risk score for prediction of heart failure: The Rotterdam Study. BMC medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7885616/

Centers for Disease Control and Prevention. (2021, August 17). Health topics - heart disease - polaris. Centers for Disease Control and Prevention. https://www.cdc.gov/policy/polaris/healthtopics/heartdisease/index.html#:~:text=Heart%20disease%20costs%20the%20United,%2C%20medications%2C%20and%20premature%20death

Centers for Disease Control and Prevention. (2023, May 15). Heart disease facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm

Hutson, M. (2017, April 14). Self-taught artificial intelligence beats doctors at predicting heart ... Science. https://www.science.org/content/article/self-taught-artificial-intelligence-beats-doctors-predicting-heart-attacks

Kakadiaris, I. A., Vrigkas, M., Yen, A. A., Kuznetsova, T., Budoff, M., &amp; Naghavi, M. (2018, November 11). Machine learning outperforms ACC/AHA CVD Risk Calculator in Mesa. Journal of the American Heart Association. https://www.ahajournals.org/doi/10.1161/JAHA.118.009476