Predicting Heart Disease

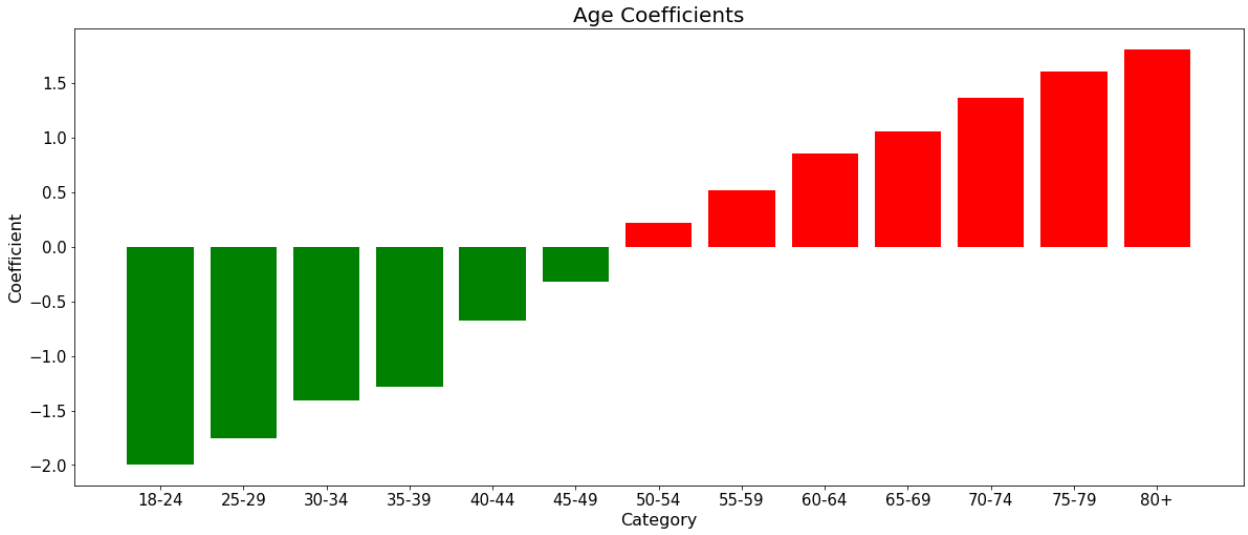
Introduction

Approximately 697,000 people die of heart disease in the United States every year. Additionally, heart disease is ranked #1 as the leading cause of death in the United States. Being able to accurately predict the factors that contribute to heart disease could help save lives and improve the health of many.

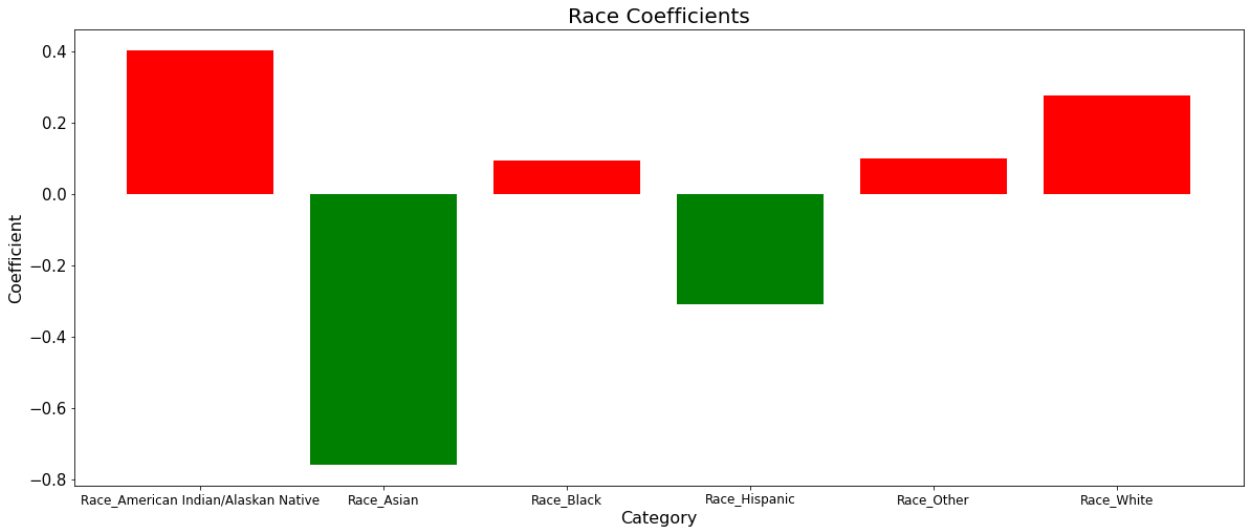
In this study, our team looked at different factors that could be used to predict heart disease. Using data from the 2020 annual CDC survey, we split up the factors into three categories: demographics, medical history, and external factors. The demographic factors included age, race, and gender. The medical history category included stroke, asthma, diabetes, kidney disease, and skin cancer. External factors were smoking, alcohol, BMI, sleep, physical activity, general health, and difficulty walking.

Demographics

To begin, we looked at three main categories for this demographic section. We looked at age, race, and gender to see if there were any trends between those factors and heart disease. This first graph shows the coefficients in a logistic regression model where age is used to predict heart disease.

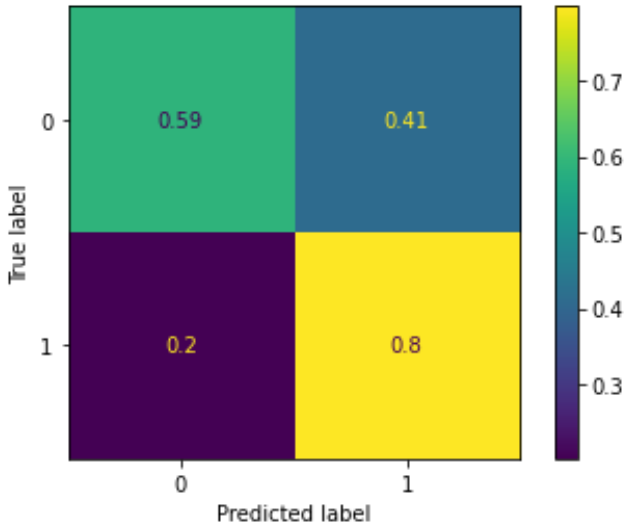


As you can see, there is a very clear correlation between higher age and a greater risk of heart disease. This is quite logical, as young people have a very low rate of heart problems. Race was also looked at in this manner and the results were interesting too.

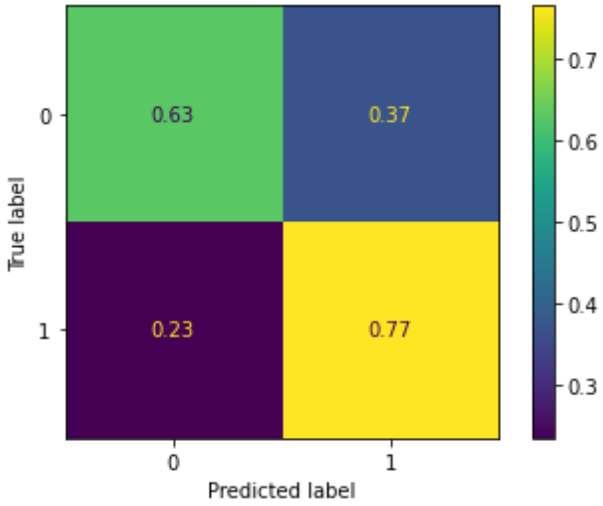


Here you can see that some races have a higher coefficient than others. Asian people tend to have the lowest rates of heart disease, and black and white people tend to have higher rates. This matches what CDC supports. We believe that American Indian/Alaskan Native is so high in this graph because it is a small group in this large dataset.

When creating our first logistic regression model with all the races, age groups, and genders, we got the following graph.



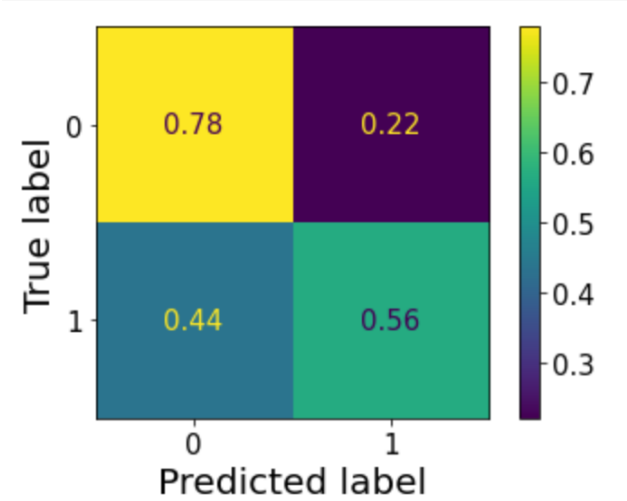
By just looking at demographic data we got an 80 percent true positive rate for predicting heart disease. However, accuracy and our true negative are pretty bad in this model. We did some more model building and made a decision tree model, a random forest model, and ensemble those to create our best model for just demographics.



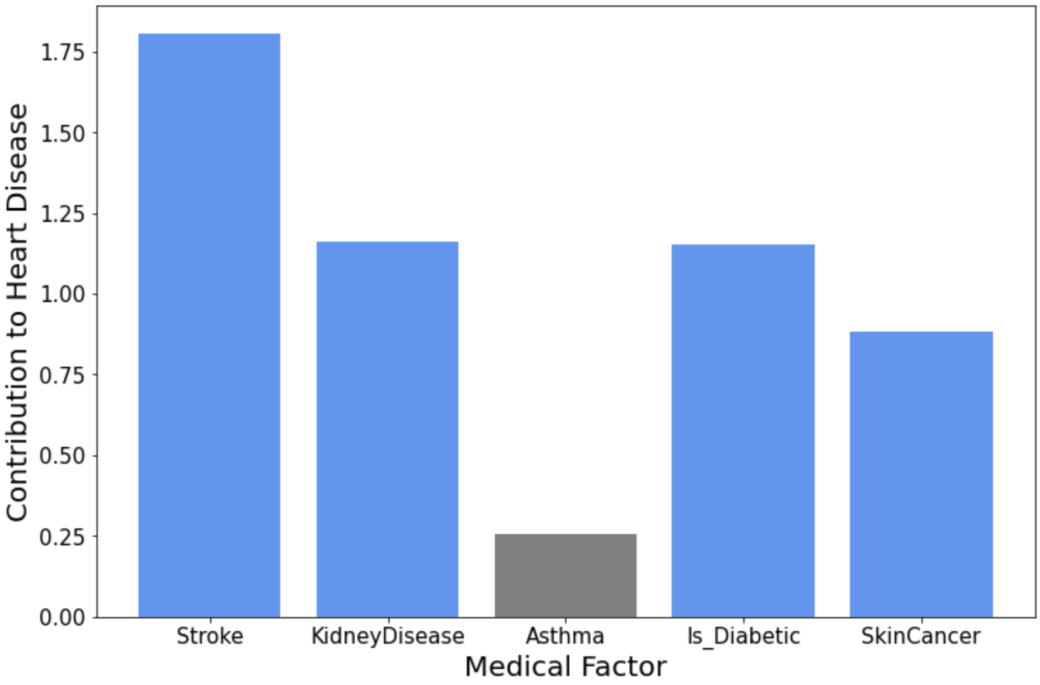
Here we have higher accuracy and a more balanced model. We sacrificed a few percentage points for the true positive rate, but our true negative is higher.

Medical History

In order to analyze medical factors, we began by creating a logistic regression model that only used medical factors - stroke, asthma, diabetes, kidney disease, and skin cancer.

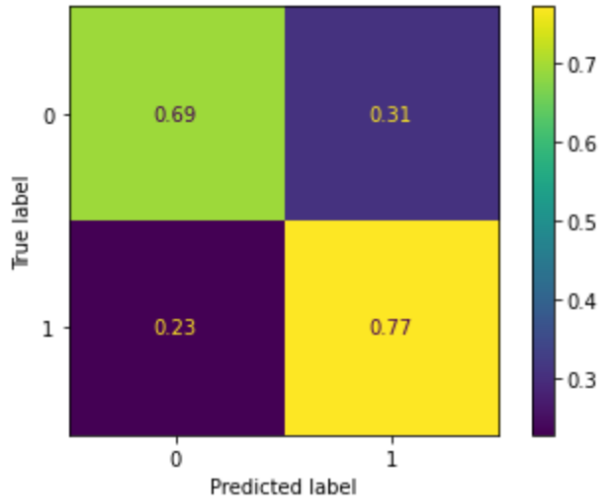


As demonstrated by the confusion matrix above, this base model was fairly good at predicting true negative values (those who did not have heart disease) but did a poor job at predicting those who did have heart disease. However, this model did provide valuable information regarding the coefficients for each variable, which is illustrated in the graph below.



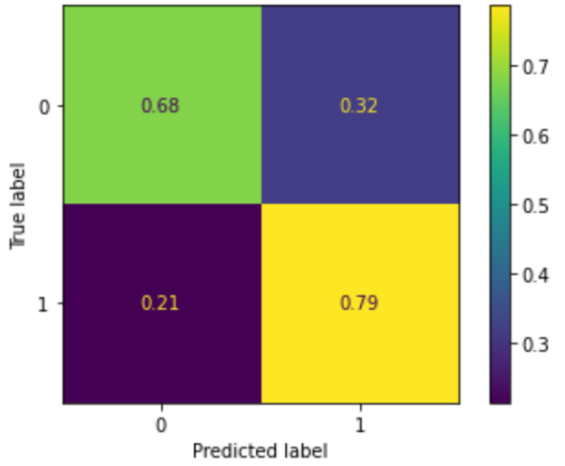
Stroke, kidney disease, diabetes, and skin cancer all had a significant impact on whether someone had heart disease. However, the graph demonstrates that asthma is not a good predictor for heart disease. For this reason, asthma was not used in any subsequent models.

This medical factor analysis was then combined with the demographics model in order to form another logistic regression model that used age, race, gender, stroke, kidney disease, diabetes, and skin cancer. This model had an accuracy of 69% and a ROC AUC score of 0.73.



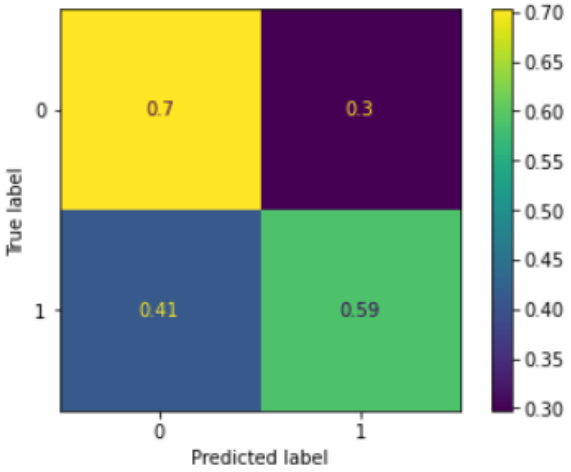
Using this same combination of demographic and medical factors, we tested several other models, including a decision tree classifier, nearest neighbors, and random forest classifier. Out of these three models, the random forest performed the best with a true positive rate of 0.79 and

a true negative rate of 0.68. This model also had an accuracy of 68% and a ROC AUC score of 0.73.

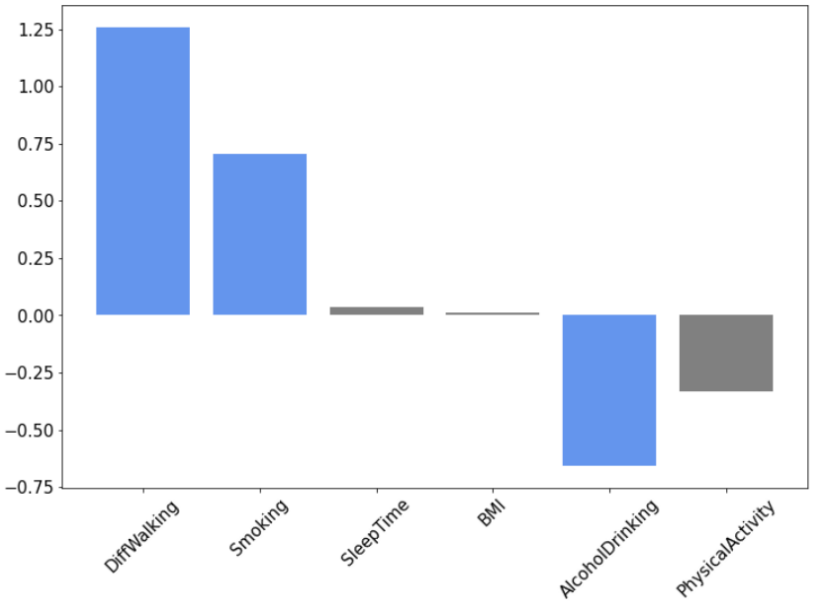


External Factors

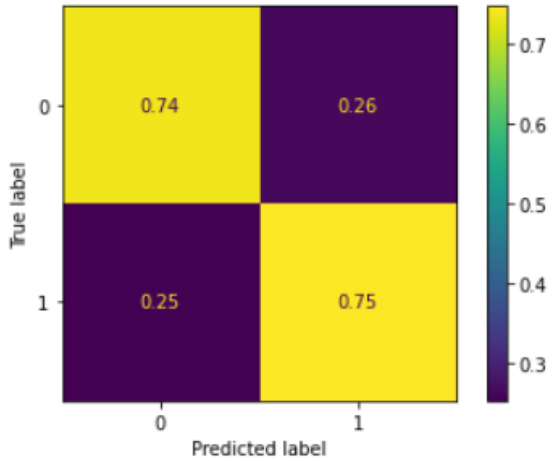
External factors that were provided in this study were difficulty walking (yes or no), smoking (have they smoked at least 100 cigarettes in their lifetime), drinking alcohol (have they had at least 5 drinks in a week), BMI, sleep (average hours of sleep a night), and physical activity (whether they have been active in the last 30 days). Similarly to medical factors, we created a base model of a logistic regression model with only external factors.



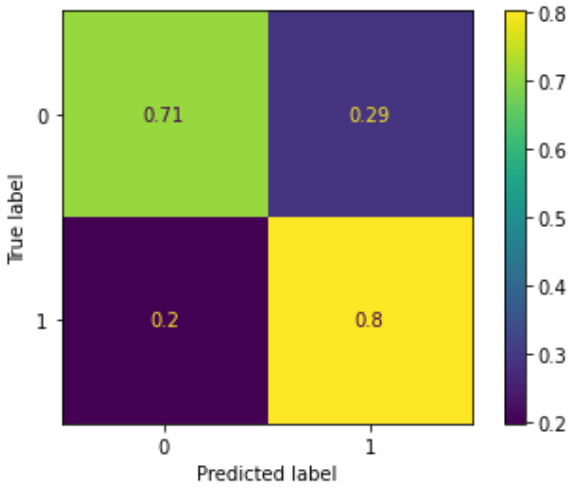
The base model for external factors had a fair true positive rate, but a bad true negative rate. This model did better at predicting people who have heart disease than predicting people who do not have heart disease. Next, we looked at the coefficients for external factors to give us a better idea of what factors we should use.



As you can see, difficulty walking, smoking, and drinking alcohol had the highest impact on our model. Sleep, BMI, and physical activity had very little impact so we decided to take those out of our model. After we analyzed which external factors to use we combined these factors with our medical factors and demographic factors to form another logistic regression model. Below you can see that our model did fairly well, with a true positive rate of 0.74 and a true negative rate of 0.75. The accuracy of our model was 74% with a ROC AUC of 0.74.



Again, we used the same combination of factors to perform on different models. Out of the decision tree classifier, nearest neighbors, and random forest classifier models the decision tree classifier did the best at predicting whether or not someone would have heart disease. After making these models and building upon each factor, the highest accuracy we were able to get was with a decision tree classifier. We received an accuracy of 71 percent and a true positive rate of 80 percent.



Conclusion

In this study, we were able to determine that the model that produced the best accuracy of having heart disease used the factors of age, race, gender, stroke, diabetes, kidney disease, skin cancer, difficulty walking, smoking, and drinking alcohol. With these factors combined, we used a decision tree classifier, nearest neighbors, and random forest classifier models. The highest accuracy we were able to produce was with a decision tree classifier with 71% accuracy and a fairly high true positive rate and true negative rate.