Cardiovascular Disease Prediction

INTRODUCTION :-

Cardiovascular disease or heart disease is the leading cause of death amongst women and men and amongst most racial/ethnic groups in the United States. Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease. From the CDC, roughly every 1 in 4 deaths each year are due to heart disease. The WHO states that human lifestyle is the main reason behind this heart problem. Apart from this there are many key factors which warn that the person may/may not get a chance of heart disease. The term heart disease is often used interchangeably with the term cardiovascular disease. Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke.

Cardiovascular disease, also called heart disease, is a class of diseases that involve the heart or blood vessels. Cardiovascular disease (CVD) includes:-
Atherosclerosis (plaque build up in walls of arteries)
Myocardial infarction (heart attack)
Stroke Heart failure Arrhythmia (abnormal heart rhythm)
Heart valve problems (valves not opening or closing enough)

DATA :-

Independent Features:-
Age (in days)
Gender (1-Female, 2-Male)
Height, Weight
Systolic BP, Diastolic BP
Cholesterol - (1 normal, 2 above normal, 3 well above normal)
Glucose - (1 normal, 2 above normal, 3 well above normal)
Smoking, Alcohol intake, Physical activity

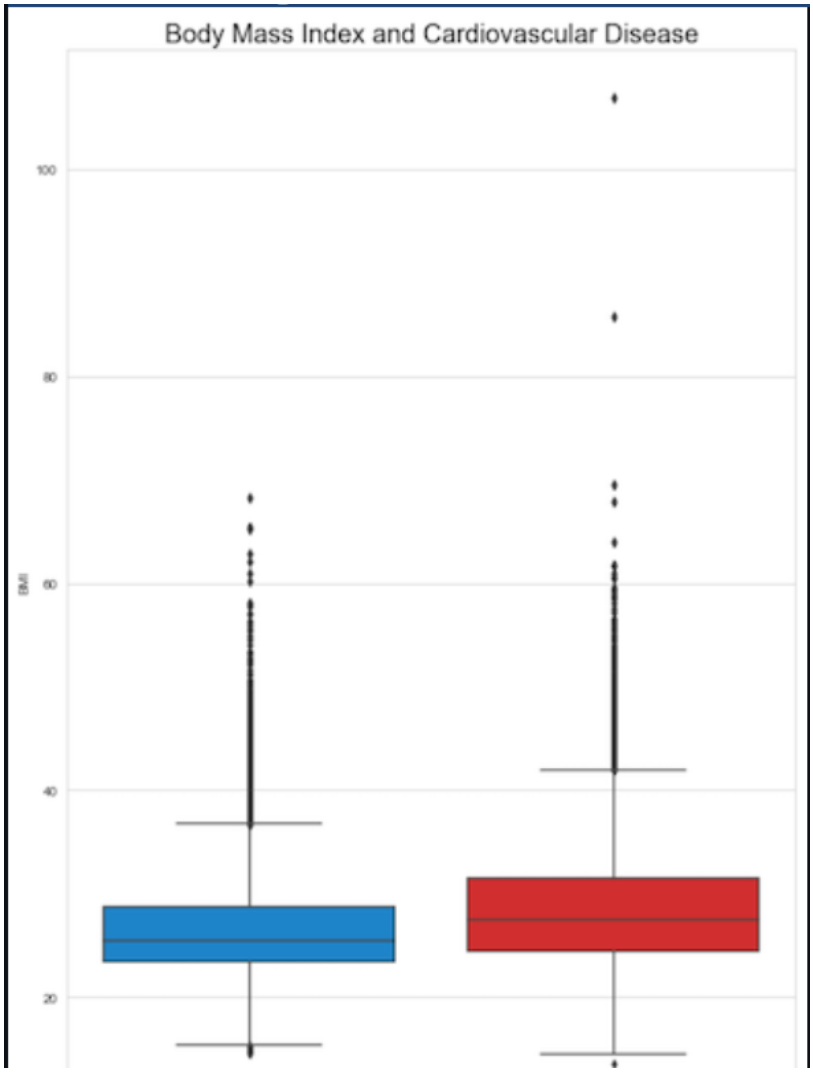
DATA CLEANING :-

Data was scrubbed by:

checking for null values and duplicates
dropping ID's
changing age from days to years
removing suspicious data
heights and weights that seemed too high or too low
systolic or diastolic pressure readings that were negative, unusually low, or diastolic was lower than systolic
changed columns names
Scrubbing data removed 1,488 data points from the original 70,000. Target variable was checked again to see that cleaning did not drastically alter class balance. Ratio for target features was still 50/50.

DATA EXPLORATION :-

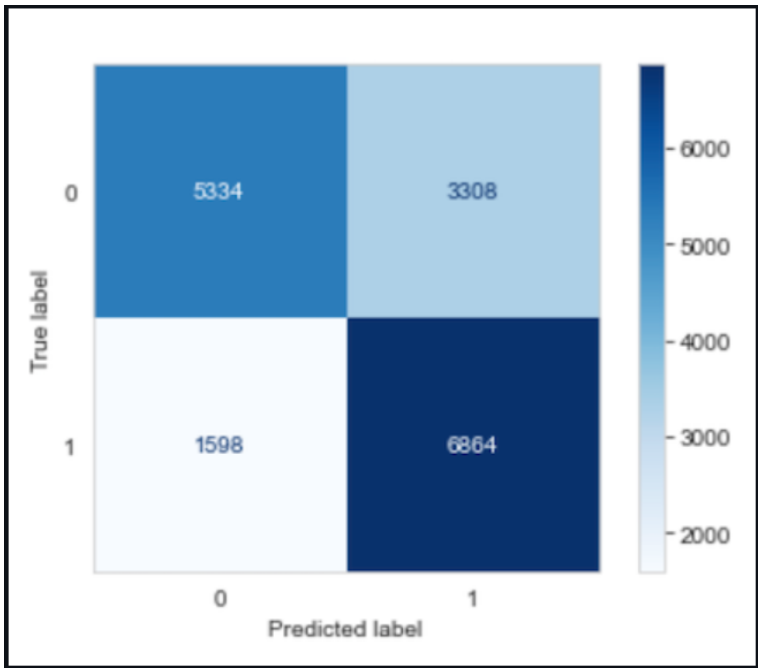
Different features were examined to determine if any had an individual correlation to cardiovascular disease.First, a feature was created to look at body mass index. Body mass index (BMI) is a measure of body fat based on height and weight that applies to adult men and women.Using height and weight, I was able to create a BMI column and see that individuals diagnosed with cardiovascular disease do have higher BMI than those that do not have cardiovascular disease.



Modeling :-

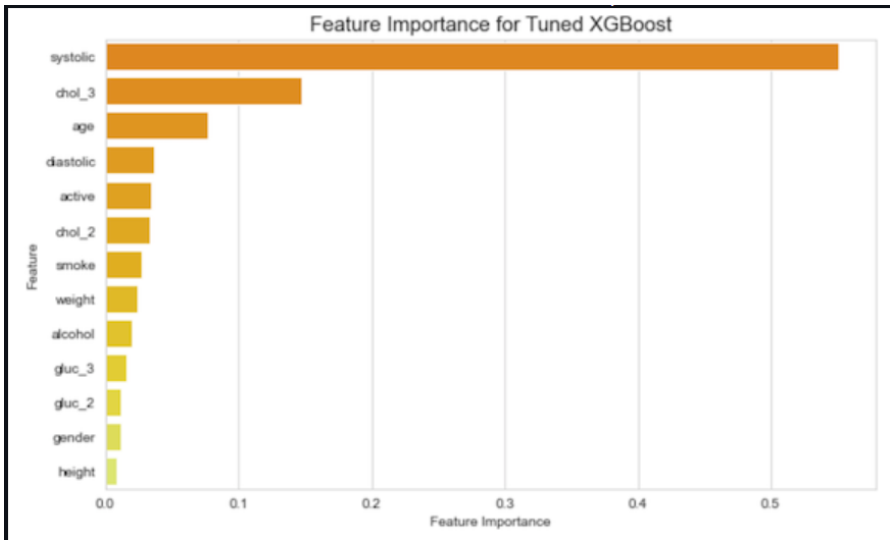
Again, the target feature here was whether an individual has cardiovascular disease or not. The data was split into training and testing data and the training data was scaled using Standard Scaler.The first models were built without any hyperparameter tuning. These models include:
Logistic Regression
Random Forest
Support Vector Machines
K-Nearest Neighbor
Decision Tree
XGBoost

The models were run and their results saved into a DataFrame.Then each model was tuned using hyperparameters to see if the accuracy, F1 score, and false negatives could be improved. That data would be again saved into a DataFrame to compare all models. A function was used to calculate the results after each model and save those results to the DataFrame. The function also displayed the classification report and a confusion matrix for the model.



Conclusion:-

Judging by the highest F1 score, lowest number of false negatives, highest number of true positives and greatest area under the curve (AUC) - the tuned XGBoost model performed the best.



Looking at the winning model - Randomized Search XGBoost - I evaluated the important features. By far, the most important feature was the systolic reading in blood pressure. Having cholesterol "well above normal" and age were the next largest important features.

When looking at the subjective features, activity was the most important feature followed by smoking then drinking alcohol.

These features indicated that when faced with a patient who has a high systolic reading, very elevated cholesterol levels, and is older, it would be prudent to run further tests and check for cardiovascular diseases.

It would also be important to coach these patients about a need to increase activity, improve diet, and decrease or stop smoking and drinking alcohol. A follow up should be done to check blood pressure and cholesterol levels again and consider prescribing a statin drug that would decrease cholesterol levels.