Cardiovascular Disease

Munirah Abdullah Alraqibah

Abstract:

Healthcare expenses are overwhelming national and corporate resources due to asymptomatic diseases, including cardiovascular diseases. Therefore, there is an imperative need for primeval discovery and treatment of such diseases. Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. The aim of this project is predicting heart disease of the patient through some predictors, listed in the description, by using classification techniques.

Description of the dataset:

The dataset contains about 70,000 observations "patients" with 13 Features each, those input features have 3 types: Objective: factual information; Examination: results of medical examination; and Subjective: information given by the patient. The predictors are Gender, Age, Height, Weight, Systolic blood pressure, Diastolic blood pressure, Cholesterol, Glucose, Smoking, Alcohol intake, and Physical activity. The response variable is a binary variable with a patient having a value of '0' if they do not have heart disease and a value of '1' if they do.

Data Source References: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Algorithms:

Feature Engineering:

- EDA
- Changing data type of string to dummy
- Plotting the correlation between target and features
- Cleaning data
- Dropping unnecessary features.
- Scaling data

Models:

In this project several classification techniques were performed. Logistic regression is the right algorithm to start with classification algorithms as it is considering a benchmark model, easy and fast model. It uses a logistic function to form binary output model. K-nearest neighbors is a non-parametric method used for classification and regression. It is one of the easiest machine learning techniques used, and it is comparatively slower than logistic regression. Moreover, it is a lazy learning model, with local approximation. Random Forest is a collection of decision trees and average vote of the forest is selected as the predicted output. Also, can handle high dimensional spaces as well as large number of training examples.

Model Evaluation and selection:

First, dataset of 70,000 observations was split into two parts, training, and test. About 80% of the data was used to fit the model and 20% for test.

Logistic Regression: Accuracy: 72.578%

KNNeighbors: Accuracy: 72.69%

Gradient Boosting Machines: Accuracy: 73.79%

Random Forest: Accuracy: 70.90%

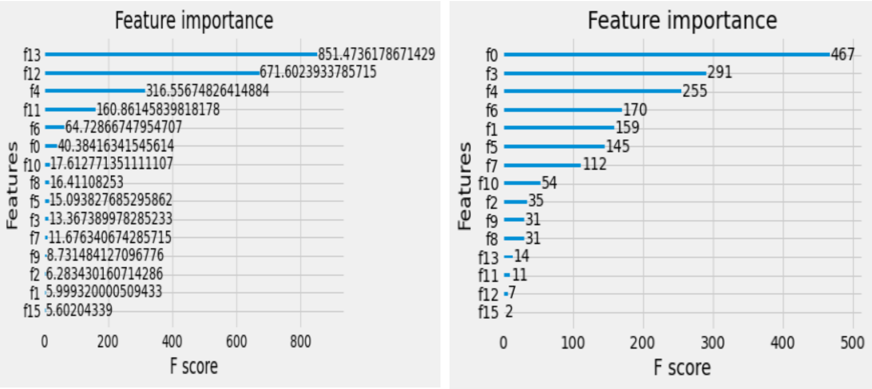
Naive Bayes: Accuracy: 80.27%

We can see that the highest accuracy rate was Naive Bayes with 13334 True positive. And False positive is high so in this case it may not seem to be the accurate model. Gradient Boosting Machines seems fairly to fit the model.

Tools:

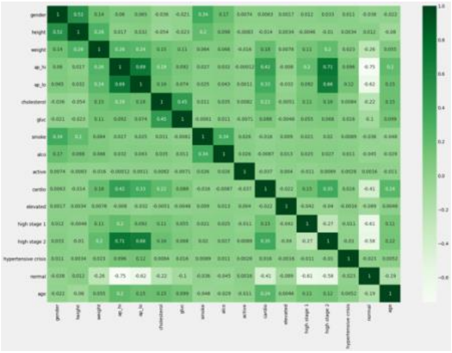
- Pandas and Numpy for data manipulation
- Seaborn and Matplotlib for plotting
- scikit-learn f and Xgboost for modeling

Communication

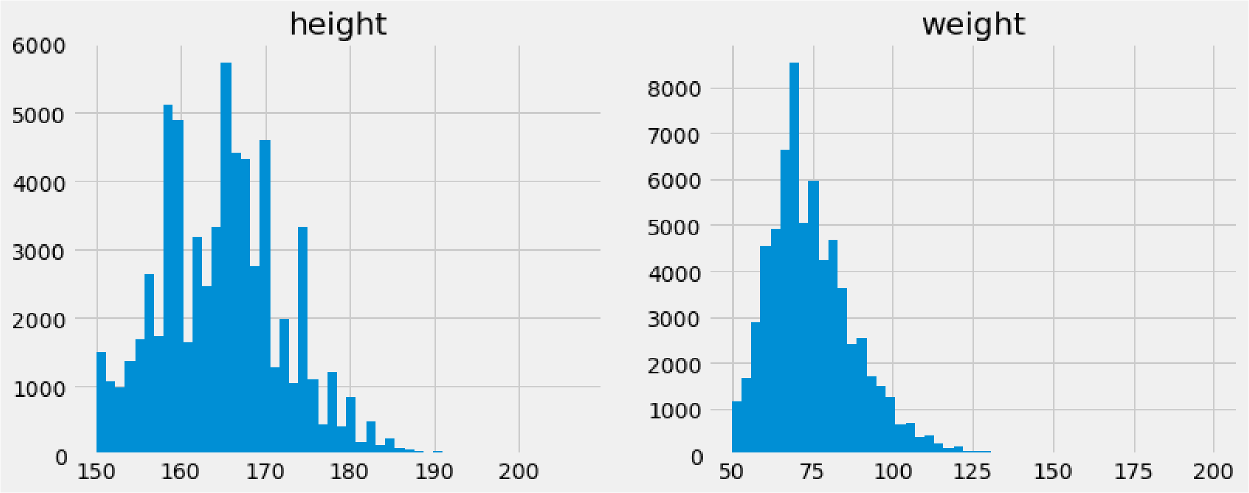


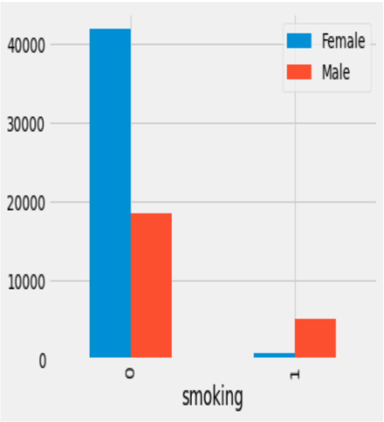
-

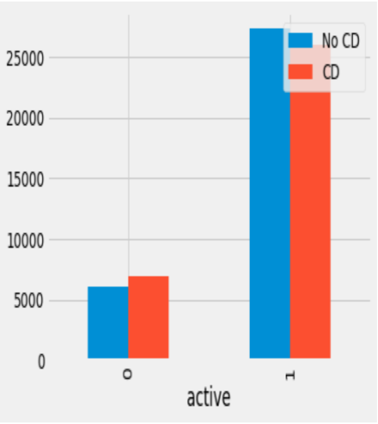
Gain views that normal blood pressure is the most important featuer while in frequency is the least important.

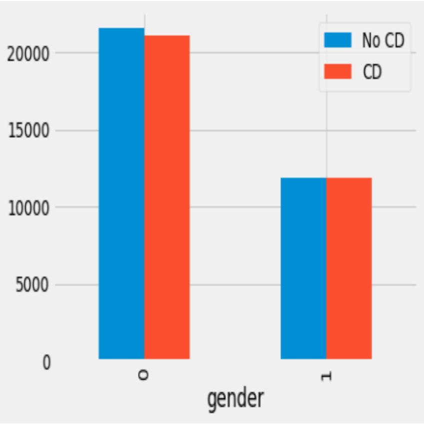


Looking into the correlation heatmap, and it seems like our target has either positive or negative higher correlation with weight,ap_hi,ap_lo, cholesterol, high stage 1, high stage 2, normal, age.









- from the first plot we can see that the number of Males smoker more than female's smoker.
- As we can see from the second plot that the number of active patients either do have CD or not is triple times the number of patients who are not active for both cases as well.
- according to the third plot and by ignoring that the number of female patients is slightly more than the number of male paitents, we can see that male in both cases of having or not having a heart disease is almost the same value while female who has a heart disease is slightly lower than females who have it. also, by comparing between of gender, female patients are almost double than male patients.