Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name- Madhavi Mali **Email-madhavimali1996@gmail.com**
Please paste the GitHub Repo link.
GitHub Link https://github.com/madhavimali/Cardiovascular_Risk_Prediction_Classification_Madel
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Introduction:

Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. According to the WHO, an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Over three quarters of these deaths took place in low and middle-income countries. Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the most fatal. In the United States, for example, it is estimated that someone has a heart attack every 40 seconds and about 805,000 Americans have a heart attack every year (CDC 2019).

Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches. In this project we will be giving you a walk through on the development of a screening tool for predicting whether a patient has a 10-year risk of developing coronary heart disease (CHD) using different Machine Learning techniques.

Problem Statement: The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

Approach: Here first we imported data set and cleaned our data by removing null values or finding mean of the data to fill the null values. Then we performed EDA where we got valuable insights and further we Encoded the Categorical Columns, did Feature scaling and fitting into the models. At first we tried with basic logistic regression but soon realized we will need a much more complex model and so we then used a Decision Tree Classifier, Random Forest Classifier, XGB Model Classifier, KNN Classifier and SVM Classifier compared the results for improvement in the model fitting to understand the better results of the model as well as the metrics.

Conclusion:

The analysis is done with Cardio-Vascular data. Six classification techniques Logistic Regression, Decision Tree, Random Forest, XG Boosting, K Nearest Neighbors and Support Vector Machines are used to predict the Risk here. This statistical data analysis shows interesting outcomes in prediction methods and also in an exploratory data analysis.

The experimental results show that:

- In the given dataset we observe that Coronary heart disease increases from age 51 to 67 then decreases.
- We draw the count plot and observe that no. of male heart patients is more than female and also notice that male get early age heart diseases as compared to females.
- We observe the no. of heart patients who smoke more than as compared to those who won't and also notice that those who smoke get early heart disease as compared to those who won't.
- We draw the bar plot and observed that no. of cigsperday taken by male is more than female. So, male heart patients are more as compared to females.
- We draw the boxplot and observe that female BMI (The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2) is more than male BMI. that's leads to OVERWEIGHT and So, female CHD patients is more than male CHD patients.
- We draw the boxplot and observe that female Cholesterol is more than male Cholesterol. that's leads to OVERWEIGHT and So, in that case also female CHD patients is more than male CHD patients.
- We Observe that Female heart disease patients has more Heart Rate as compared to male heart disease patients.
- We also observe that male heart disease patients have more glucose level as compared to female heart disease patients.
- In the **Models Evaluation** Table (Testing set) our **auc-roc** score is more **0.80** except **Logistic regression** and **Decision Tree**. So we can say that our model predicted the classes in a good manner.
- **XGBClassifier** is performed the best with the best **Recall, Precision, F1-Score** and **Accuracy Score**.