Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Team Member's Role: - ❖ Kirtesh Verma(kirteshverma12345@gmail.com) Contribution: ○ Data understanding ○ Handling null and missing values ○ Performing EDA ○ Removing outliers ○ Logistic regression ○ Support vector machine ○ Hyperparameter tuning on SVM ❖ Pravin Bejjo(praveen.bejo.pb@gmail.com) Contribution: ○ Data understanding ○ Data visualization ○ Multivariate analysis ○ Handle imbalance data using SMOTE technique ○ Random forest ○ XGBoost classifier ○ XGBoost (feature importance) ❖ Sahil Pardeshi(8623879021.sp@gmail.com) Contribution: ○ Data understanding ○ Data visualization ○ Bivariate analysis ○ Decision tree classifier ○ K-nearest neighbors classifier ○ Hyperparameter tuning on KNN
Please paste the GitHub Repo link.
Github Link:- https://github.com/praveenbejo95/Cardiovascular-risk-prediction-ML-classification-project

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Heart disease is one the major cause of morbidity and mortality globally. A heart attack happens when the flow of oxygen-rich blood to a section of heart muscle suddenly becomes blocked and the heart can't get oxygen. If blood flow isn't restored quickly, the section of heart muscle begins to die.

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 3,390 records and 17 attributes. Variables Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

Doctors and Scientists across the globe have started to look into Machine Learning Techniques to develop screening tools. In this project, we shall 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) based on their present health conditions using different Machine Learning Techniques.

The first step in the exercise involved exploratory data analysis where we tried to dig insights from the data in hand. It included univariate and multivariate analysis in which we identified certain trends, relationships, correlation and found out the features that had some impact on our dependent variable. The second step was to clean the data and perform modifications. We checked for missing values and outliers and removed irrelevant features. We also do Feature Engineering and one hot encoding for the categorical features. The third step was handled the imbalanced data using SMOTE technique. The fourth step was to try various machine learning algorithms on our split and standardized data. We tried different algorithms namely; Logistic regression, Decision tree, Random Forest, K-Nearest Neighbour, XGBoost, Support Vector Machine. We did hyperparameter tuning and evaluated the performance of each model using various metrics. The best performance was given by the XGBoost and Random Forest model with accuracy 93% and 89%, and F1_score was 93% and 89%, Precision was 96% and 91%, and Recall was 90% and 87% respectively.

The most important features who had a major impact on the model predictions were; Age, Heartrate, totchol, BMI, Education, avgBP, and Glucose. Men are more likely to have heart disease compared to women. factor. Elderly age group people are more at risk to CHD then young and middle age group people. Heart disease is a severe problem and though with age the chances of getting heart related problem increases, a good balance between your diet and physical exercise will reduce the risk factor and help to live longer.