Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

❖ **Kirtesh Verma** Email : kirteshverma12345@gmail.com

Contribution:

1. Data understanding
2. Handling null or missing values
3. Performing EDA
4. Removing outliers
5. Logistic Regression
6. Support Vector Machine and Hyperparameter tuning

❖ **Pravin Bejjo** Email: praveen.bejo.pb@gmail.com

Contribution:

1. Data understanding
2. Data visualization
3. Multivariate analysis
4. Handling imbalanced data using SMOTE
5. Random Forest
6. XGBoost Classifier
7. Feature importance

❖ **Sahil Pardeshi** Email: 8623879021.sp@gmail.com

Contribution:

1. Data understanding and visualization
2. Bivariate analysis
3. Decision Tree Classifier
4. K-Nearest Neighbour Classifier and Hyperparameter tuning

Please paste the GitHub Repo link.

GitHub Link: https://github.com/KirteshVerma/Cardiovascular_Risk_Prediction

Drive Link: - https://drive.google.com/drive/folders/1JzdN9bDhzeWTYgzbgTzGhDOd7yrUoWH2

PROJECT SUMMARY:

CARDIOVASCULAR RISK PREDICTION

Coronary Heart Disease (CHD) is the term that describes what happens when your heart's blood supply is blocked or interrupted by a build-up of fatty substances in the coronary arteries. Over time, the walls of your arteries can become furred up with fatty deposits. This process is known as atherosclerosis and the fatty deposits are called atheroma. The most important behavioural risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It has 17 attributes and 3390 entries.

Our key objective is to predict 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. For this we first performed EDA to understand the features which increases the chances of getting CHD.

In the second step we performed Data cleaning. There where null values present in the columns like education, cigsPerday, BPMeds, totChol, BMI, heartRate, and glucose column. We filled the null values and with respective median and mode. There were outliers in the dataset but we didn't remove as it will result in the loss of information. We featured engineered a column named 'avgBP'. We also did the label encoding for the categorical features.

In the third step we tried different machine learning on the clean and standardised data. We used different types of models namely; logistic regression, decision tree, random forest, Support Vector Machine (SVM), and K-Nearest Neighbour (KNN). We did hyperparameter tuning of the features and evaluated models through different evaluation matrices. The best model turned out to be Random Forest with an F1 score of 0.87 followed by KNN with hyperparameter tuning and an F1 score of 0.82.

We did a thorough analysis of the data and understood the affects and increases the chances of getting CHD. We concluded that the people who are on BP medication or have diabetes, people who previously had a stroke, who are hypertensive have higher chances of coronary heart disease (CHD). Age is an important 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.