Capstone Project Submission

Team Member’s Name, Email and Contribution:

1. Shubham Joshi (shubhkjoshi5@gmail.com

a. Data Wrangling
i. Data_Cardiovascular_risk
ii. Loading and Preprocessing
iii. Structuring Data
iv. Enriching Data
v. Data Validation
b. Data Mining
c. Data Analysis
d. Model Development
i. Decision Tree
ii. Random Forest
iii. Support Vector Machine
e. Visualizations
i. Box Plot
f. Segmentation
g. Summarization
h. Observations
i. Conclusions

2. Akshay Fasale (aksfasale99@gmail.com)

a. Data Wrangling
i. Data_Cardiovascular_risk
ii. Structuring Data
iii. Enriching Data
b. Data Mining
c. Data Analysis
d. Model Development
i. Logistic Regression
ii. K-Nearest Neighbor
iii. XGBoost Classifier
e. Visualizations
i. Dist. Plots and Sub Plots
f. Segmentation
g. Summarization
h. Observations
i. Conclusions

3. Kanika Kakra (kostubikakra11@gmail.com)

a. Data Wrangling
i. Data_Cardiovascular_risk
ii. Structuring Data
iii. Enriching Data
iv. Data Validation
b. Data Mining
c. Data Analysis
d. Model Development
i. Decision Tree
ii. Random Forest
iii. Support Vector Machine
- Visualizations
□ Count and Bar plots
- Segmentation
- Summarizations
- Observations
- Conclusions

GitHub Repo link.

Github Link:- https://github.com/11-Kani/CapstoneProjecIII_Cardiovascular_Risk_Prediction

Drive Link:- https://drive.google.com/drive/folders/1T_L8fWg0ccwAUYGyOYsOg-8eWcBKqwGz

Summary:-

Cardiovascular disease (CVD) is a general term for conditions affecting the heart or blood vessels. It's usually associated with a build-up of fatty deposits inside the arteries (atherosclerosis) and an increased risk of blood clots.It can also be associated with damage to arteries in organs such as the brain, heart, kidneys and eyes.CVD is one of the main causes of death and disability in the UK, but it can often largely be prevented by leading a healthy lifestyle.

Problem Statement and Objective :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. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

Exploratory data analysis or commonly known as EDA helps to explore data, and possibly formulate hypotheses that might cause new data collection and experiments. EDA build a robust understanding of the data, issues associated with either the info or process. It's a scientific approach to get the story of the data.

All machine learning algorithms use some input data to create outputs. This input data comprise features, which are usually in the form of structured columns. Algorithms require features with some specific characteristics to work properly. Here, the need for feature engineering arises.

Feature engineering mainly have two goals:

- Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
- Improving the performance of machine learning models.

When observation in one class is higher than the observation in other classes then there exists a class imbalance. We can clearly see that there is a huge difference between the data set. Solving this issue we use Synthetic Minority Oversampling Technique (SMOTE) technique.

This technique generates synthetic data for the minority class.

SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.

SMOTE algorithm works in 4 simple steps:

Choose a minority class as the input vector Find its k nearest neighbors (k_neighbors is specified as an argument in the SMOTE() function) Choose one of these neighbors and place a synthetic point anywhere on the line joining the point under consideration and its chosen neighbor Repeat the steps until data is balanced.

Finally we conclude we've noticed that XBG Classifier is the stand out performer among all models with an f1-score of 0.8397. it's safe to say that XGB Classifier provides an optimal solution to our problem.

In case of Logistic regression, We were able to see the maximum f1-score of 0.658.

KNN gave us recall of 86 % and with gridsearch CV it gave a recall of 93%

Out of the tree-based algorithms, the Random Forest Classifier was providing an optimal solution towards achieving our Objective. We were able to achieve an f1-score of 0.7703 We also noticed that in the case of Decision-tree Classifier, we were able to achieve an f1-score of **0.7034** for the test split.

For **SVM(Support Vector Machines)** Classifier, the f1-score lies around **0.7417**.

were able to achieve an f1-score of 0.7034 for the test split.

For SVM(Support Vector Machines) Classifier, the f1-score lies around **0.7417**.Finally , As in the medical domain (False negative values have importance we don’t want to mispredict a person safe when he has the risk) recall ha the most importance..KNN, XGB , Random Forest gave the best recall 0.86 ,0.80 ,0.81.