Capstone Project Submission

Team Member's Name, Email and Contribution:

Name : kamalnayan kumar **Email: nayankumar7099@gmail.com**

Contribution: Individual

1) Data cleaning and preprocessing :-
   - Here I checked and dealt with missing and duplicate variables from the data set as these can grossly affect the performance of different machine learning algorithms (many algorithms do not tolerate missing data).
2) Exploratory Data Analysis :-
   - Here I wanted to gain important statistical insights from the data and the things that I checked for were the distributions of the different attributes, correlations of the attributes with each other and the target variable and I calculated important odds and proportions for the categorical attributes.
   -
3) Feature Engineering :-
   - - Feature Selection: Since having irrelevant features in a data set can decrease the accuracy of the models applied, I used Tree-based: SelectFromModel which is an embedded method that uses algorithms that have built-in feature selection methods which were later used to build different models.

4) Model Implementation :-
   - Fitting various models on our data and optimizing them via using GridsearchCV
   - Using these models to make predictions on test and train data. The Models implemented are :-
     1. Logistic Regression
     2. Random Forest
     3. XGBoost
     4. Support Vector Machine

5) Data Visualization :-
   - Using several kinds of charts like histogram, heatmap, pair plot, countplot etc. to better visualize data and understand correlation and trends.
6) Model performance comparison :-
   - Comparison of all implemented models using various Classification evaluation metrics like Accuracy, Precision, Recall, F1 Score, AUC.
7) Conclusion :-
   - Drawing some insights from the data and the predictions made by our various predictive models on unseen (test) data.

GitHub Repo links.

Github Link:- https://github.com/nayankr77/Cardiovascular_Risk_Prediction

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

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 patient's information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

Conclusions :-

- The number of people who have Cardiovascular heart disease is almost equal between smokers and non-smokers.

- The top features in predicting the ten year risk of developing Cardiovascular Heart Disease are 'age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose'.

- The Support vector machine with the radial kernel is the best performing model in terms of accuracy and the F1 score and Its high AUC-score shows that it has a high true positive rate.

- Balancing the dataset by using the SMOTE technique helped in improving the models' sensitivity.

- With more data(especially that of the minority class) better models can be built.

……..