Capstone Project Submission

Team Member's Name, Email and Contribution:

1. Arvind Krishna (killerdude.arvind@gmail.com)
    - Data Wrangling
        - Data_Cardiovascular_risk
        - Loading and Preprocessing
        - Structuring Data
        - Enriching Data
        - Data Validation
    - Data Mining
    - Data Analysis
    - Model Development
        - Decision Tree
        - Random Forest
        - Support Vector Machine
        - Neural Network
    - Visualizations
        - Box Plot
    - Segmentation
    - Summarization
    - Observations
    - Conclusions

2. Sahil Ahuja (s.ahuja38@gmail.com)
    - Data Wrangling
        - Data_Cardiovascular_risk
        - Structuring Data
        - Enriching Data
    - Data Mining
    - Data Analysis
    - Model Development
        - Logistic Regression
        - K-Nearest Neighbour
        - Naïve Bayes Classifier
        - XGBoost Classifier
    - Visualizations
        - Dist Plots and Sub Plots
    - Segmentation
    - Summarization
    - Observations
    - Conclusions

3. Keshav Sharma (keshav1506sharma@gmail.com)
    - Data Wrangling
        - Data_Cardiovascular_risk
        - Structuring Data
        - Enriching Data
        - Data Validation
    - Data Mining
    - Data Analysis
    - Model Development
        - Decision Tree
        - Random Forest
        - Support Vector Machine
        - Neural Network

- Visualizations
  - Count and Bar plots
- Segmentation
- Summarizations
- Observations
- Conclusions

4. Jayesh Panchal (jaypan290497@gmail.com)
   - Data Wrangling
     - Data_Cardiovascular_risk
     - Structuring Data
     - Data Validation
   - Data Mining
   - Data Analysis
   - Model Development
     - Logistic Regression
     - K-Nearest Neighbour
     - Naïve Bayes Classifier
     - XGBoost Classifier
   - Visualization
     - Historical Bars
   - Segmentation
   - Summarization
   - Observations
   - Conclusion

Please paste the GitHub Repo link.

Github Link:-

https://github.com/Keshav1506/Cardiovascular-Risk-Prediction

Drive Link:-

https://drive.google.com/drive/folders/13MZStuxx1KRx4-b3-mDob4HcR-J6qC0b?usp=sharing

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

Cardiovascular diseases, also called CVDs, are the leading cause of death globally, causing an estimated 17.9 million deaths each year. *CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions.* The most important behavioural risk factors of heart disease and stroke are *unhealthy diet, physical inactivity, tobacco use* and *harmful use of alcohol.* The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity.

The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD) based on their health statistics and information about their tobacco usage.

As a first step we explored through all the columns in the given dataset to understand the rubrics and content of dataset that we have. There were few null values which we removed as the medical data of any person cannot be simply assumed. Then we tried to check the correlation among various features and found some linear relationship between all

independent variables and the target variable '10YearCHD'.

In this dataset, there was a class imbalance which we corrected using the SMOTE technique and predicted the results based on it. Upon implementation of various algorithms, We've noticed that *XBG Classifier* is the stand out performer among all models with an f1-score of 0.828. It is by far the second highest score we have achieved. Hence, 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 *0.656*. 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 *0.838* for the test split, which is higher than any other model (excluding NN).

We have also implemented an *experimental Neural Network,* however the results were ambiguous and non-conclusive. Currently we were able to see accuracy as high as *0.850*, but the accuracy fluctuates within a continuous range.

Finally, we can conclude that Random forest and XGB classifier might provide the best results for this particular problem, moreover we can optimize these models using Grid Search CV(cross validation) and hyper-parameter tuning to get better results.