**Capstone Project Submission**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. AKSHAY FASALE (aksfasale99@gmail.com)  * Data Wrangling   Data\_Cardiovascular\_risk  Loading and Preprocessing  Structuring Data  Enriching Data   * Data Mining * Data Analysis * Model Development * Logistic Regression * K-Nearest Neighbor * XGBoost Classifier * Visualizations * Dist. Plots and Sub Plots Debugging * Segmentation * Observations * Summarization * Conclusions  1. Shubham Joshi(shubhkjoshi5@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 * Visualizations * Box Plot * Debugging * Segmentation * Observations * Summarization * Conclusions  1. Kanika Kakra(kostubikakra11@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 * Visualizations * Count and Bar plots * Debugging * Segmentation * Observations * Summarization * Conclusions  1. Rishikesh Damale(rishikesh.13021999@gmail.com)  * Data Wrangling   Data\_Cardiovascular\_risk  Structuring data  Enriching data  Data Validation   * Data Mining * Data Analysis * Model Development * Logistic Regression * K-Nearest Neighbor * XGBoost Classifier * Visualizations * Historical Bars * Debugging * Segmentation * Observations * Summarization * Conclusions |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/aksfasale/cardiovascular-risk-prediction>  Drive Link:- |
| **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. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age. The most important behavioral 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.  **Problem statement**  ***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. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.*** |
| 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. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors  Modeling:  We performed several modeling over the data   * Logistic Regression Classifier * Decision Tree Classifier * Random Forest Classifier * XGB Classifier * K-Nearest Neighbour(KNN Classifier) * Support Vector Machine(SVM Classifier).   **Conclusion:**  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**.  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**.  Finally, as in the medical domain (False negative values have importance we don’t want to mis predict 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. |