**Capstone Project Submission**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Name: Souvik Bhattacharyya  Email: [souvik.ranchi@gmail.com](mailto:souvik.ranchi@gmail.com)  **Contribution:**   1. **Data loading** 2. **Data Wrangling** 3. **Exploratory Data Analysis** 4. **Hyperparameter Tuning** 5. **Resampling** 6. **Machine Learning Models training and testing** 7. **Conclusions** |
| **Please paste the GitHub Repo link.** |
| GitHub Link: https://github.com/Souvik0651/Cardiovascular-Risk-Prediction--classification-.git |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| Coronary heart disease is a type of heart disease where the arteries of the heart cannot deliver enough oxygen-rich blood to the heart. Many people have no symptoms, they do not know they have coronary heart disease until they have chest pain, blood flow to the heart is blocked, causing a heart attack, or the heart suddenly stops working, also known as cardiac arrest.  Look at the given dataset and find the distribution of each feature. Look for null values and outliers and replace or remove them if it is meaningful for our study. 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 3390 records and 17 attributes. Each attribute is a potential risk factor. There are both demographic and behavioral and medical risk factors. Use appropriate metrics to compare the performance of machine learning algorithms.  We have done null values replacement. We observed that we have imbalanced data set. So, we up-sampled our minority class data to make our dataset a balanced one. Then we trained machine learning models for both the original dataset and the modified dataset. We used the F1 score as a metric for comparing the performance of machine learning models.  Females are more than the number of males in our dataset but the number of males prone to heart disease is more compared to females.  Without using SMOTE all the models were not performing well as the dataset was imbalanced.  So, it is necessary to handle the imbalanced dataset.  After using multiple machine learning models, we found out that the SVM classifier is the best model for the prediction of CHD for this dataset compared to other classifiers. |