**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Name: Konda Vidyasagar  Email: [kondasagar236@gmail.com](mailto:kondasagar236@gmail.com)  **Contribution:**   1. **Data loading** 2. **Data Wrangling** 3. **Univariate Analysis** 4. **Bivariate Analysis** 5. **Resampling** 6. **Machine Learning Models training and testing** 7. **Conclusions** |
| **Please paste the GitHub Repo link.** |
| GitHub Link: https://github.com/Kondasagar/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)** |
| **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.**  **Problem Statement:**  **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 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 and behavioral and medical risk factors. Use appropriate metrics to compare the performance of machine learning algorithms.**  **Approach:**  **We have done null values replacement. Then we did univariate analysis and bivariate analysis. We observed that we have imbalanced data set. So, we up sampled our minority class data to make our dataset into balanced one. Then we trained machine learning models for both original dataset and modified dataset. We used F1 score as metric for comparing the performance machine learning models.**  **Conclusions:**  **• Females are more than number of males in our dataset but number of males prone to heart disease are more compared to females.**  **• Higher education people are less but all education level people having equal share of heart disease prone.**  **we have imbalanced dataset. So, we built machine learning algorithms in two scenarios. They are**  **• without up sample (no resample)**  **• with up sample (resample)**  **No Resampling on dataset:**  **F1 Score on test dataset:**  **• Naive Bayes Classifier: 0.57**  **• Random Forest Classifier: 0.46**  **• XG Boost Classifier: 0.51**  **Resampling dataset case:**  **F1 score on test dataset:**  **• Logistic Regression: 0.66**  **• Naive Bayes Classifier: 0.51**  **• Decision Tree Classifier: 0.67**  **• KNN Classifier: 0.78**  **• XG Boost Classifier: 0.85**  **• Random Forest Classifier: 0.7**    **Git Link: https://github.com/Kondasagar/Cardiovascular-Risk-Prediction** |