**Capstone Project Submission**

|  |
| --- |
| **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 disease (CVD) is the leading cause of illness and death worldwide. Predicting the risk of cardiovascular disease is the key to primary prevention. Machine learning has attracted attention in analyzing increasingly large, complex healthcare data. 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 patients’ information. It includes over 4,000 records and 15 attributes. Variables Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.  The business use case of project is to consider a case in healthcare where a Machine Learning (ML) system is asked to predict the risk of a patient developing cardiovascular disease in a given time frame, say within the next one year. The output expected from the model is a risk score which if equal to or above a chosen threshold indicates that patient has a risk of developing cardiovascular disease and if it were below the threshold, indicates no risk.  The potential impact of your project is to correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time. ML model will help to predict future risk of cardiovascular diseases.  The approach to the problem statement is that, Firstly I started the study by handling missing values, treating outliers, our dataset was having many columns with null entities so removed null values. After that understanding the distribution of features both numerical and categorical through pie charts, histograms, bar plots etc.  By doing univariate, multivariate analysis we observe the features which are responsible of coronary heart disease. Feature engineering is done on our data set to improve machine learning model training, leading to better performance and greater accuracy  Using KNN Imputer and Simple imputer I've seen how smoking, systolic BP, diastolic BP, BMI, Heart rate, glucose, hypertensive, cholesterol, diabetes, etc. affects the person. Factors like Blood Pressure, Glucose Level, Age had created a huge impact on a person's heart condition.  We checked the correlations between the factors. Handled the class Imbalance using SMOTE (Synthetic Minority Oversampling Technique). SMOTE gave a good result of 50-50 class balanced data. Then I started building classification models. I started with Logistic regression with default parameters but I did not get a good score. Then I used a Support Vector Machine with various Kernel tricks. Also building tree-based model and ensembles like Gradient Boosting, XG Boosting, AdaBoost techniques with respective hyperparameters.  Evaluation techniques are used confusion matrices like Accuracy, Recall, Precision, and F1- Score. Decision Thresholds and Receiver Operating Characteristic (ROC) curve.  Then I tried the Random Forest Classifier and I got great Scores. Random Forest Classifier performed well. And at last, I've tried XGBoost Classifier and it had performed really well and I got the good scores with XGBoost Classifier as compared to other Models, so I conclude XG Boost is my optimal model for use and we can use this model for further in predicting Cardiovascular risk.  We plotted a graph after hyper parameter tuning. Even though the accuracy of some models were increased after hyper parameter tuning, still Ada boost and naive bayes are the best performing models.  Between those if we compare Naive bayes and Ada boost with the accuracy and recall, Naive Bayes is the best performance model. |
| **Please paste the GitHub Repo link.** |
| GitHub Link: https://github.com/j0kr97/Cardiovascular-risk-prediction  Drive link: - https://drive.google.com/drive/folders/1MIAdnpGBbibqntDh2ZxHZG94JdxdPcWd?usp=sharing |