**Instructions:**

**Capstone Project Submission**

1. Please ﬁll in all the required information.
2. Avoid grammatical errors.

**Summary and Conclusions:**

* We trained 7 Machine Learning models using the training dataset, and hyper parameter tuning was used in some models to improve the model performance.
* To build the models, missing values were handled, feature engineering and feature selection was performed, and the training dataset was oversampled using SMOTE to reduce bias on one outcome.
* Recall was chosen as the model evaluation metric because it was very important that we reduce the false negatives.
* Initial set of predictions were obtained using the baseline model, ie, logistic regression model, and other commonly used classification models were also build in search of better predictions.
* Predicting the risk of coronary heart disease is critical for reducing fatalities caused by this illness. We can avert deaths by taking the required medications and precautions if we can foresee the danger of this sickness ahead of time.
* It is critical that the model we develop has a high recall score. It is OK if the model incorrectly identifies a healthy patient as a high risk patient because it will not result in death, but if a high risk patient is incorrectly labelled as healthy, it may result in fatality.
* We were able to create a model with a recall of just 0.77 because of limitated data available and limited computational power available.
* A recall score of 0.77 indicates that out of 100 individuals with the illness, our model will be able to classify only 77 as high risk patients, while the remaining 33 will be misclassified.
* Future developments must include a strategy to improve the model recall score, enabling us to save even more lives from this disease. This includes involving more people in the study, and include people with different medical history, etc. build an application with better recall score.
* From our analysis, it is also found that the age of a person was the most important feature in determining the risk of a patient getting infected with CHD, followed by pulse pressure, prevalent hypertension and total cholesterol.
* Diabetes, prevalent stroke and BP medication were the least important features in determining the risk of CHD.

|  |
| --- |
| **Github Link:-**  **https://github.com/arunaherkar/Cardio-Vascular-Risk-Prediction\_capstone-project-III.git** |