**Cardiovascular Risk Prediction Analysis using ML Approach**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Name: Debanjan Ganguly  Email: [officialdeba10@gmail.com](mailto:officialdeba10@gmail.com)  Project Contribution:   * Data Exploration * Exploratory Data Analysis * Feature Selection * Oversample dataset using SMOTE * Model Preparation – Defining Pipeline, Hyperparameter Tuning, Evaluation of Models using Metrics used for Classification * Model Selection and Predict Risk |
| **GitHub Repo & Drive Link:** |
| GitHub Link: -  <https://github.com/awesomedeba10/Cardiovascular-Risk-Prediction---AlmaBetter-Capstone-Project>  Drive Link: -  <https://drive.google.com/drive/folders/18mofco9uAKY1l0wMd0_Rgt7zJBor-s5d?usp=sharing>  Colab Notebook Link: -  <https://colab.research.google.com/drive/1zZoXjVoTHs2xDjI78VQUEd5ODZgWRTTH?usp=sharing> |
| **Project Summary:** |
| Heart disease is the leading cause of death worldwide, accounting for one third of deaths in 2019. Heart disease cases nearly doubled over the period, from 271 million in 1990 to 523 million in 2019, and the number of heart disease deaths rose from 12.1 million to 18.6 million. Over three quarters of these deaths took place in low- and middle-income countries.  Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the most fatal. The silver lining is that heart attacks are highly preventable and simple lifestyle modifications (such as reducing alcohol and tobacco use; eating healthily and exercising) coupled with early treatment greatly improves its prognosis.  In this project, the classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD) or not. The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It includes over 3,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.  The dataset is an imbalance dataset, so SMOTE has been used to balance the dataset for both classes. Top features have been selected and normalized before training. Many machine learning algorithms like Logistic Regression, Random Forest, LinearSVC etc. has been used to train and evaluate with both imbalanced and oversampled dataset. Out of which, random forest is proved to be best in performance with a f1 score of 0.85. Recall is far more important metrics than accuracy or precision. As this is a sensitive domain, and we can’t put any risk of falsely identify a potentially risked patient. However, the maximum recall we’ve achieved is 0.8549. |
| **Reference:**  *Declaration: The references mentioned below are not the direct copy of code but used for learning purpose of various techniques or metrics or ideas that have been adapted in this project after learning and understanding the concepts thoroughly from these mentioned articles.*  <https://towardsdatascience.com/simple-example-using-boruta-feature-selection-in-python-8b96925d5d7a>  <https://christophm.github.io/interpretable-ml-book/shap.html#shap-feature-importance>  <https://towardsdatascience.com/5-feature-selection-method-from-scikit-learn-you-should-know-ed4d116e4172>  <https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb> |