**Capstone Project Submission**

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. Puneet Suthar ([7sutharpuneet@gmail.com](mailto:7sutharpuneet@gmail.com))  * Understanding Data   Understanding different column  Having overview of data   * Data Cleaning * Imputing null values using knn * Dropping unused column. * Exploratory Data Analysis * Univariate analysis  1. Dist plot with skewness 2. Box plot 3. Imputing the outliers with knn  * Bivariate analysis  1. Line plot 2. Bar plots  * Multivariate analysis  1. Correlation plot 2. Pair plot  * Feature engineering * Encoding of columns (sex, is smoking, education') * Removal of class imbalance using SMOTE * Model Building * Created a instance of all models. * Hyperparameter tuning of all models simultaneously. * Created a function which compares all the models on these (confusion matrix, classification report and roc curve) metrics and returns a data frame containing all information of all metric for every model . * Created pickle file of model and scaler used in the project to replicate the same results in the web app. * Web application for prediction of rental bike count * Collected all data form the user as input, and applied required transformations * Used pickled models * Done the prediction. * Conclusions |
| **Please paste the GitHub Repo link.** |
| Github Link:- https://github.com/Puneet-Suthar/Cardiovascular\_Risk\_Prediciton  Drive Link:- https://drive.google.com/drive/folders/115HIeQRavYUt7s3UsIbp\_32dnXytusCF?usp=share\_link |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| My Capstone project aims to build a predictive model to identify patients at risk of developing Coronary Heart Disease (CHD) based on various demographic, behavioral, and medical risk factors. The project involves various steps such as EDA, data preprocessing, model selection, training, and deployment. During EDA, we explored the dataset using various plots such as distribution, box plot, regression plot, correlation heatmap, and pair plot to gain insights into the data. We identified several observations and concluded that the dataset had a high degree of imbalance, several null values, and some interesting trends that suggest higher risk factors among males, elderly patients, and patients with morbidities such as diabetes and hypertension.  We then proceeded to preprocess the data by performing label encoding, train-test splitting, and dealing with data imbalances by using the SMOTE technique. Finally, we trained several models such as Logistic Regression, Decision Tree Classifier, and Random Forest Classifier, evaluated their performance metrics, and selected the best model. We chose the Random Forest Classifier as our final model due to its superior performance. Finally, we deployed our model as a web app using streamlit. I hope it can help clinicians identify at-risk patients and implement preventative measures. |