**Machine Learning**

Course-End Project Problem Statement

**Course-End Project: Healthcare**

**Problem statement:**

Cardiovascular diseases are the leading cause of death globally. It is therefore necessary to identify the causes and develop a system to predict heart attacks in an effective manner. The data below has the information about the factors that might have an impact on cardiovascular health.

**Dataset description:**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Age | Age in years |
| Sex | 1 = male; 0 = female |
| cp| | Chest pain type |
| trestbps | Resting blood pressure (in mm Hg on admission to the hospital) |
| chol | Serum cholesterol in mg/dl |
| fbs | Fasting blood sugar > 120 mg/dl (1 = true; 0 = false) |
| restecg | Resting electrocardiographic results |
| thalach | Maximum heart rate achieved |
| exang | Exercise induced angina (1 = yes; 0 = no) |
| oldpeak | ST depression induced by exercise relative to rest |
| slope | Slope of the peak exercise ST segment |
| ca | Number of major vessels (0-3) colored by fluoroscopy |
| thal | 3 = normal; 6 = fixed defect; 7 = reversible defect |
| Target | 1 or 0 |

**Note:**

Download **CEP 1\_ Dataset.xlsx** using the link given in the **Healthcare** project problem statement

**Task to be performed:**

1. Preliminary analysis:
   1. Perform preliminary data inspection and report the findings on the structure of the data, missing values, duplicates, etc.
   2. Based on these findings, remove duplicates (if any) and treat missing values using an appropriate strategy
2. Prepare a report about the data explaining the distribution of the disease and the related factors using the steps listed below:
   1. Get a preliminary statistical summary of the data and explore the measures of central tendencies and spread of the data
   2. Identify the data variables which are categorical and describe and explore these variables using the appropriate tools, such as count plot
   3. Study the occurrence of CVD across the Age category
   4. Study the composition of all patients with respect to the Sex category
   5. Study if one can detect heart attacks based on anomalies in the resting blood pressure (trestbps) of a patient
   6. Describe the relationship between cholesterol levels and a target variable
   7. State what relationship exists between peak exercising and the occurrence of a heart attack
   8. Check if thalassemia is a major cause of CVD
   9. List how the other factors determine the occurrence of CVD
   10. Use a pair plot to understand the relationship between all the given variables
3. Build a baseline model to predict the risk of a heart attack using a logistic regression and random forest and explore the results while using correlation analysis and logistic regression (leveraging standard error and p-values from statsmodels) for feature selection