Course-End Project: Healthcare

Problem statement:

Cardiovascular disease 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.

Task to be performed:

1. Preliminary analysis:
    - Perform preliminary data inspection and report the findings on the structure of the data, missing values, duplicates, etc.
    - 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:
    - Get a preliminary statistical summary of the data and explore the measures of central tendencies and spread of the data
    - Identify the data variables which are categorical and describe and explore these variables using the appropriate tools, such as count plot
    - Study the occurrence of CVD across the Age category
    - Study the composition of all patients with respect to the Sex category
    - Study if one can detect heart attacks based on anomalies in the resting blood pressure (trestbps) of a patient
    - Describe the relationship between cholesterol levels and a target variable
    - State what relationship exists between peak exercising and the occurrence of a heart attack
    - Check if thalassemia is a major cause of CVD
    - List how the other factors determine the occurrence of CVD
    - 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.