**Project Name : Healthcare**

**Project task to find out Cardiovascular disease (CVD)**

**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 ( As given )**

|  |  |
| --- | --- |
| **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 |

**Task has been performed**

1. Preliminary analysis:
   1. Perform preliminary data inspection and report the findings on the structure of the data, missing values, duplicates, etc. => **Done**
   2. Based on these findings, remove duplicates (if any) and treat missing values using an appropriate strategy => **Done**
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 => **Done**
   2. Identify the data variables which are categorical and describe and explore these variables using the appropriate tools, such as count plot => **Done**
   3. Study the occurrence of CVD across the Age category

**(Min Age is 29 & Max age is 77)**

**Make group of 25 to 35, 36 to 45, 46 to 55, 56 to 65, 66 to 75, 76 to 85** => **Done**

* 1. Study the composition of all patients with respect to the Sex category => **Done**
  2. Study if one can detect heart attacks based on anomalies in the resting blood pressure (trestbps) of a patient => **Done**
  3. Describe the relationship between cholesterol levels and a target variable => **Done**
  4. State what relationship exists between peak exercising and the occurrence of a heart attack => **Done**
  5. Check if thalassemia is a major cause of CVD => **Done**
  6. List how the other factors determine the occurrence of CVD => **Done**
  7. Use a pair plot to understand the relationship between all the given variables => **Done**

1. 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 stats models**) for feature selection

Note: Not understand above bold text matter from point no 3