# **Data Analyst: Capstone project**

# **Domain: Healthcare**

Problem Statement:

Cardiovascular diseases are one of the leading causes of deaths globally. To identify the causes and develop a system to predict potential heart attacks in an effective manner is necessary. The data presented has all the information about relevant factors that might have an impact on cardiovascular health. The data needs to be studied in detail for further analysis.

There is one dataset data that has 14 attributes with more than 4000 data points

Expected Deliverables:

You are required to determine and examine the factors that play a significant role in increasing the rate of heart attacks. Also, use the findings to create and predict a model.

Data Snapshot:

A screenshot of a medical report

Description automatically generated

A screenshot of a computer

Description automatically generated

Project Tasks:

Importing, Understanding, and Inspecting Data :

1. Perform preliminary data inspection and report the findings as the structure of the data, missing values, duplicates, etc.

2. Based on the findings from the previous question, remove duplicates (if any) and treat missing values using an appropriate strategy.

3.Get a preliminary statistical summary of the data. Explore the measures of central tendencies and the spread of the data overall

Performing EDA:

1. Identify the data variables which might be categorical in nature. Describe and explore these variables using appropriate tools. For example: count plot.

2. Study the occurrence of CVD across different ages.

3. Can we detect heart attack based on anomalies in resting blood pressureof the patient?

4. Study the composition of overall patients w.r.t . gender.

Performing EDA and Modeling:

1.Describe the relationship between cholesterol levels and our target variable.

2.What can be concluded about the relationship between peak exercising and occurrence of heart attack?

3. Is thalassemia a major cause of CVD? How are the other factors determining the occurrence of CVD?

4.Use a pair plot to understand the relationship between all the given variables.

5. Perform logistic regression, predict the outcome for test data, and validate the results by using the confusion matrix

Tasks and observations:

Task 1: Importing, Understanding, and Inspecting Data :

A screenshot of a computer

Description automatically generated

* Categorical Variables: sex, cp, fbs, restecg, exang, slope, ca, thal, target
* Continous Variables: age, trestbps, chol, thalach, oldpeak
* Using df.info() -checked for null values.
* Also checked and dropped duplicates as shown in the following screenshots

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

Task 2: Exploratory Data Analysis:

A graph of a number of numbers

Description automatically generated with medium confidence

* The above chart represents the CVD across all age groups
* We can not the count is after the age of 50 and the risk reduces after the age of 70
* The below chart shows the overall composition of with respect to gender

A graph with blue squares

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

* We can identified the anomalies using the boxplot as show in the above image and fixed the anomalies.

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Task 3: EDA and modelling:

* Relation between cholesterol and target

A screenshot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

* relationship between peak exercising and occurrence of heart attack

A screenshot of a computer screen

Description automatically generated

Pair plot

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated