Home assignment – Data Science

Background of Problem Statement:

Healthcare industry is massively investing in intelligent systems in order to make their services more efficient. One such way is by developing Medical Virtual Assistants. The US health organization did a survey on a group of people aged between 30 to 80 and collected a dataset. The dataset also serves as an input for project scoping and tries to specify whether a person has risk of heart attack or not.

Problem Objective:

Based on given information ML model should be able to predict if a person is likely to have heart attack or not.

Domain: Health Services

* **Use Random Seed = 42 everywhere**

1. **Load the data:**

- Read the "**US_Heart_Patients.csv**" file from the folder into the program.

2. **Perform the exploratory data analysis**

- Print the following information
  - First 10 rows of the data
  - 5-point summary
  - Information about the column (data types)
  - Number of outliers(extra points)
  - Any missing value
  - Correlation between variables
  - Distribution of the data
- Draw the charts and graphs for the above points if required

3. **Data Preprocessing**:

- Impute the missing values (if any).
- Outlier treatment (if any).
- Encode categorical features if needed.

4. **Split the dataset**:

- Split the data into 80% training dataset and 20% test dataset.

5. **Model preparation and evaluation**

- Run the following steps for Naïve Bayes, and Decision Tree:
  - Train the model and predict the output for both train and test data.
  - Calculate F1 score.
- Pick and explain the best model out of the two and explain its confusion matrix and classification report.