Problem 1

**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.

Problem 2

**Background of Problem Statement**

The dataset contains the charges billed using health insurance.

**Problem Objective:**

Create a linear regression model to predict the charges billed based on the data given.

**Dataset Info:**

Filename: insurance\_data.csv

* age: age of primary beneficiary
* sex: insurance contractor gender, female, male
* bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height
* children: Number of children covered by health insurance / Number of dependents
* smoker: Smoking
* region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
* charges: Individual medical costs billed by health insurance

**Perform following activities on the given dataset:**

* Perform exploratory data analysis and Data preprocessing (similar to problem 1)
* Feature engineering
* Train Test split(80-20)
* Use appropriate evaluation metrics(RMSE/R-squared) to validate if the model is performing well.
* Equation of the model.
* Provide names of significant and insignificant variables(If any insignificant variables and there, then check the impact on model after removing them).

Deliverables

* Submit separate jupyter notebooks for both problem statements
* Model pickle file(Final Model)

**Note: Please do not submit data set back with submission**