**PG-DBDA, March2022 CDAC Mumbai**

**Practical Machine Learning**

**Marks: 30 Time: 2hrs.**

**Problem Statement 1:**

There is an automobile company Cars24X7 from India which aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

They want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Japanese market. Essentially, the company wants to know:

Which variables are significant in predicting the price of a car How well those variables describe the price of a car Based on various market surveys, the consulting firm has gathered a large dataset of different types of cars across the American market.

**Business Objectives:**

You as a Data scientist are required to apply some data science techniques for the price of cars with the available independent variables. That should help the management to understand how exactly the prices vary with the independent variables.

They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels.

Also build the Random Forest Regressor for optimal search of classifiers.

**Dataset:** data1.csv

Note : kindly build the project with following pointers

1.Data understanding and exploration

2.Data cleaning

3.Data preparation

4.Model building and evaluation

5.Result with error calculation

[15 Marks, 60 minutes]

**Problem Statement 2: For Question2 either use read\_csv if file is csv ....if file is excel then use read\_excel**

Chandrashekhar wants to purchase office-space. He does a detailed survey of the offices and corporate complexes in the area, and tries to quantify a lot of factors, such as the distance of the offices from residential and other commercial areas, schools and workplaces; the reputation of the construction companies and builders involved in constructing the apartments; the distance of the offices from highways, freeways and important roads; the facilities around the office space and so on.

Each of these factors are quantified, normalized and mapped to values on a scale of 0 to 1. Chandrashekhar then makes a table. Each row in the table corresponds to Chandrashekhar Observations for a particular house. If Chandrashekhar has observed and noted F features, the row contains F values separated by a single space, followed by the office-space price in dollars/square-foot. If Chandrashekhar makes observations for H houses, his observation table has (F+1) columns and H rows, and a total of (F+1) \* H entries.

Chandrashekhar does several such surveys and provides you with the tabulated data. At the end of these tables are some rows which have just F columns (the price per square foot is missing). Your task is to predict these prices. F can be any integer number between 1 and 5, both inclusive.

There is one important observation which Chandrashekhar has made.

The prices per square foot, are (approximately) a polynomial function of the features in the observation table. This polynomial always has an order less than 4

**Input Format**

The first line contains two space separated integers, F and N. Over here, F is the number of observed features. N is the number of rows for which features as well as price per square-foot have been noted.

This is followed by a table having F+1 columns and N rows with each row in a new line and each column separated by a single space. The last column is the price per square foot.

The table is immediately followed by integer T followed by T rows containing F columns.

**Constraints**

1 <= F <= 5

5 <= N <= 100

1 <= T <= 100

0 <= Price Per Square Foot <= 10^6 0 <= Factor Values <= 1

**Output Format**

[15 Marks, 60 minutes]

**Dataset:** dataset2.csv