**ASSIGNMENT 2: DATA SCIENCE PROJECT**

**REPORT**

**OUTLINE:**

1. Overview of the problem
2. Model building approach & how they work
3. Wrangling & aggregation methods used and why
4. Effectiveness of the model
5. Insights from the analysis
6. Significance of the results & conclusion
7. Limitations & future improvement

**OVERVIEW OF THE PROBLEM**

Develop models that predict the maximum daily energy use and pricing based on the provided weather data.

**MODEL BUILDING APPROACH & HOW THEY WORK**

Two datasets are provided:

dataset1: price\_demand\_data.csv

dataset2: weather\_data.csv

* **Read the two data frames**
* **Merge the two datasets on a common ‘Date’ column to apply the model and predict maximum energy usage and price**

**WRANGLING & AGGREGATION METHODS USED AND WHY**

To merge the two datasets, the same format of the date from both datasets has been extracted in a new column called 'Date\_Key’ with dd/mm/yyyy format.

* **Maximum demand per day**

Given: Demand for every half-hour period each day from January 2021 - August 2021

Extracted: Maximum demand per day in a new column TOTALDEMAND\_MAX

Why: We want to predict 'maximum' demand per day

* **Converting price: Categorical to numerical**

Since we want to predict the maximum daily price category, we convert price categories to numeric values, so as to choose the maximum out of the four values.

Replace the price categories with numeric values: 1, 2, 3, 4 corresponding to the categories low, medium, high & extreme, and create a new column 'PRICE\_NUMERIC'

* **The maximumum price per day**

Extracted the maximum price value corresponding to each date in a new column called PRICE\_NUMERIC\_MAX

* **Dropped duplicates in price\_demand\_data.csv & created a new Dataframe**

Since we had multiple entries for every date, we dropped the duplicates in Price\_demand\_data.csv and selected 4 columns: Region, Date, Maximum Demand, and Maximum Price in a new dataframe called 'dataset\_1'.

* **Merged on date: Data from 01/01/2021 - 31/08/2021**

Merged the two datasets: dataset\_1 & dataset2 on the 'Date\_Key' column using Inner join so as to get data for the desired duration.

* **Check if the merged dataset has values missing**

Applied .isna() and .isnull() over the dataset and summed it up to get the number of rows having values missing (viz., NaN in numeric arrays, None or NaN in object arrays, NaT in datetimelike)

* **Replaced missing values in the given columns**

Replaced missing values in the given columns by the mean value of the said column using .replace() and verified it again using .isna()

* **Calculated pairwise Pearson r Correlation matrix between chosen features**
* **Plotted the Heatmap of the chosen features**

**EFFECTIVENESS OF THE MODEL**

**PREDICT: Maximum daily energy usage based on weather data**

**MODEL APPLIED: Classification (KNN)**

**PREDICT: Maximum daily energy usage based on weather data**

**MODEL APPLIED: Classification (Decision Tree)**

**PREDICT: Maximum daily price category based on weather data**

**MODEL APPLIED: Classification (LINEAR REGRESSION)**