PGD Data Science Sem 2
CASE STUDY :– Machine Learning
Marks :- 35
Date of Submission:- 30/04/2023

- **CONTEXT:** The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes
- **DATA DESCRIPTION:** The data concerns city-cycle fuel consumption in miles per gallon
- Attribute Information:
  1. mpg: continuous
  2. cylinders: multi-valued discrete
  3. displacement: continuous
  4. horsepower: continuous
  5. weight: continuous
  6. acceleration: continuous
  7. model year: multi-valued discrete
  8. origin: multi-valued discrete
  9. car name: string (unique for each instance)

- **PROJECT OBJECTIVE*:*** Goal is to cluster the data and treat them as individual datasets to train Regression models to predict ‘mpg’

Steps and tasks:

1. Import and warehouse data:
    - Import all the given datasets and explore shape and size.
    - Merge all datasets onto one and explore final shape and size.
    - Export the final dataset and store it on local machine in .csv, .xlsx and .json format for future use.
    - Import the data from above steps into python.
2. Data cleansing:
    - Missing/incorrect value treatment
    - Drop attribute/s if required using relevant functional knowledge
    - Perform another kind of corrections/treatment on the data.
3. Data analysis & visualisation:
    - Perform detailed statistical analysis on the data.
    - Perform a detailed univariate, bivariate and multivariate analysis with appropriate detailed comments after each analysis.

Hint: Use your best analytical approach. Even you can mix match columns to create new columns which can be used for better analysis. Create your own features if required. Be highly experimental and analytical here to find hidden patterns.

4. Machine learning:
    - Use K Means OR Hierarchical clustering to find out the optimal number of clusters in the data.
    - Share your insights about the difference in using these two methods.
5. Answer below questions based on outcomes of using ML based methods.
    - Mention how many optimal clusters are present in the data and what could be the possible reason behind it.
    - Use linear regression model on different clusters separately and print the coefficients of the models individually
    - How using different models for different clusters will be helpful in this case and how it will be different than using one single model without clustering? Mention how it impacts performance and prediction.
6. Improvisation:
    - Detailed suggestions or improvements or on quality, quantity, variety, velocity, veracity etc. on the data points collected by the company to perform a better data analysis in future.