Pre-TASK

The lifecycle of the machine learning experiment/project is as follows:

1- **Problem definition**. Define the issue and solve the problem. Also, we can determine whether machine learning is the appropriate solution!

2- **Data Collection**. In this stage, we gather relevant data that will be used to train and evaluate the machine-learning model. Ensure the data is clean and contains the necessary features.

3- **Data preprocessing**. This means, cleaning and preprocessing the data to handle missing values, outliers, and other issues. Moreover, normalize or standardize features to ensure consistent scales.

Based on the previous stages:
we start to think of our first task (Preprocessing):

4- **Model selection**. Choose the appropriate algorithm based on the nature of the problem (Classification, regression, clustering, ..).

5- **Model training**. Train the selected model on the training dataset.

6- **Monitoring and maintenance**. Model the model's performance in the production environment. Also, update the model as needed to adapt changes.

Based on the previous stages:

we start to think of our first task (Preprocessing):

1- The problem> in NLP is the huge number of texts in a file which needs to be cleared out. Stopwords (e.g., "the," "is," "and" "an," "a" …etc) which is a little meaningful information in the English language. The file is city data where we have several international cities and their descriptions. The description has a huge amount of Stopwords and we want to remove them.

2- Data collection> We use the existing dataset. You can download it from the following link

https://drive.google.com/file/d/1u94O5B3RmlU4ubaHNVcqCd4CihkOrq73/view?usp=sharing

3- Data preprocessing> We remove Stopwords because they carry little meaningful information.

Previous Hands-on:

1- First, in Google Colab, start by listing the experiments by sequence by clicking on the icon +Text from the top bar. Organising is essential.

2- Import the suitable libraries for such a project.

3- Load the given dataset.

4- Run the following code:

```
import numpy as np
import pandas as pd
from nltk.corpus import stopwords
import nltk
nltk.download('stopwords')
df = pd.read_csv('city_data_1.csv')  ##USE your OWN NAME/PATH
def clear(city):
    city = city.lower()
    city = city.split()
    city_keywords = [word for word in city if word not in stopwords.words('english')]
    merged_city = " ".join(city_keywords)
    return merged_city
for index, row in df.iterrows():
    clear_desc = clear(row['description'])
    df.at[index, 'description'] = clear_desc
updated_dataset = df.to_csv('city_data_cleared33.csv') ##Use your own name/path
```

5- Make sure to modify the dataset's name and path based on your own.

6- A new csv file 'city_data_cleared33.csv', or any other name you may suggest, will be produced on the path you choose and will show the removed Stopwords.

7- Once you run and observe the modification in the new dataset, make a new text within the same notebook to start the next experiment.

Now, this is the given TASK!

Download the dataset from the following link:

https://drive.google.com/file/d/1kuBKRSbPsulEGGL-4b2RXX9EqPLPT5BM/view?usp=sharing

Answer the following questions:

1. Can you describe the dataset's structure and format? What are the dimensions (rows and columns)?

2. Did you identify any missing or null values in the dataset? If so, how did you handle them?

3. What are the different types of variables or features present in the dataset?

4. Have you performed any exploratory data analysis (EDA)? If yes, what insights did you gain from it?

5. What ML algorithms or models would you consider using for this dataset? Why?

6. Have you split the dataset into training and testing sets? If yes, what percentage did you use for testing?

7. Which performance metrics would you use to evaluate the model's accuracy or success?

8. What are the limitations or challenges you encountered while working with this dataset?

****Note**. Send the Google Colab notebook using the extension ipynb.

****For clarifications and answers**, use Google Docs or Microsoft Word and attach it in the email.