Digital Assignment1:

Assignment Question: Machine Learning - Data Preprocessing and Exploratory Data Analysis (EDA)

Objective:
The goal of this assignment is to understand and apply various data preprocessing techniques and perform exploratory data analysis on a given dataset. You will learn how to clean, transform, and analyze data to uncover meaningful patterns and insights, which are essential steps before building any machine learning model.

Dataset:
The dataset chosen for your project/assignments

Assignment Tasks:

1. **Data Cleaning:**
    - Identify and describe the types of data quality issues present in the dataset. (e.g., missing values, duplicates, outliers).
    - Implement techniques to handle missing values and justify the chosen methods (e.g., imputation, removal).
    - Detect and remove duplicate entries in the dataset.
2. **Data Transformation:**
    - Perform data normalization or standardization where applicable and explain the rationale behind the transformation.
3. **Handling Categorical Data:**
    - Identify categorical variables in the dataset.
    - Convert categorical data into numerical formats using techniques like one-hot encoding or label encoding. Explain the impact of these transformations on the dataset.
4. **Feature Engineering:**
    - Create new features based on existing data that could potentially improve the performance of a machine learning model. Explain your thought process.
    - Analyze the importance of these new features using correlation or other statistical methods.
5. **Exploratory Data Analysis (EDA):**
    - Perform univariate analysis on at least three features, including both numerical and categorical features. Visualize and interpret the distributions.
    - Conduct bivariate analysis to explore relationships between different features. Use visualizations like scatter plots, heatmaps, or pair plots.
6. **Outlier Detection and Handling:**
    - Identify outliers in the dataset if any
    - Discuss the impact of outliers on your analysis and decide whether to remove or transform them.
7. **Data Visualization:**
    - Create meaningful visualizations (e.g., histograms, box plots, bar charts) to summarize the key findings from your EDA. Explain the insights gained from these visualizations.
8. **Correlation Analysis:**

- Calculate the correlation matrix for the numerical features in the dataset. Identify and discuss any strong correlations or lack thereof. How might this influence model selection?

9. **Summary and Insights:**
    - Summarize the key findings from your data preprocessing and EDA. Discuss how these findings will influence your approach to building a machine learning model on this dataset.
    - Suggest any further steps or considerations that should be taken before proceeding to model building.

Submission:

- Submit a detailed report with your answers to each question
- Include a Jupyter notebook or Python script with all the implemented code.

Submission deadline : 30th Aug 2024 12 midnight on TEAMS