Department of Artificial Intelligence, SVNIT,SURAT
B.Tech-III ,SEM-V
Subject- Machine Learning(AI301)

LAB ASSIGNMENT-2

You need to implement all the preprocessing steps from scratch and can not use any standard function.

1. Use the dataset given in Assignment 1 and apply all pre-processing steps: Handling missing values (replace by 0 or mean/median of the column), Normalizing the data (Min-Max Scaling, Standard Scaling). Find the outliers in the data. Run the linear regression algorithm again and find out which combination of missing value handling and normalization produces the best fit line. Does the removal of outliers improve your results? Show these experimentally.

2. The goal of this assignment is to **preprocess structured data** using Python to understand the dataset and make it input ready to ML models.

This aims to:

- Load and preprocess a dataset containing missing/noisy structured data

The implementation is conducted using Python with key libraries such as `pandas`, `numpy`, `matplotlib`, `seaborn`, and `scikit-learn`.

2. **Problem Statement:**

Consider the dataset given as a csv file related to aviation incidents and accidents. Perform preprocessing on this data to extract following information

- Load and understand the dataset and its attributes.
- Convert date and time columns from strings to a single datetime object.
- Extract the following attributes from the dataset:

    1. Aircraft make name
    2. State name
    3. Aircraft model name
    4. Text information
    5. Flight phase
    6. Event description type
    7. Fatal flag

        - Create a new dataframe with only the required columns
        - Replace all Fatal Flag missing values with the required output

        - For the FLT_PHASE and ACFT_DMG_DESC columns, use the mode (the most frequent value) to fill in the missing data.
        - Verify if the missing values are replaced
        - Check the number of observations
- Drop the unwanted values/observations from the dataset. Remove all the observations where aircraft names are not available. Drop all columns that have more than 75 non-null values.
- Check the number of observations now to compare it with the original dataset and see how many values have been dropped
- Group the dataset by aircraft name
- View the number of times each aircraft type appears in the dataset
- Display the observations where fatal flag is “Yes”
- The ACFT_DMG_DESC column has multiple categories. Use one-hot encoding on this column to create separate binary columns for each category. Drop one of the resulting columns to avoid multicollinearity.

Advanced Task:

- Feature Engineering from Unstructured Text:
    - Extract Aircraft Phase: The FLT_PHASE column has a lot of missing data. However, the RMK_TEXT often contains keywords indicating the phase of flight, such as "LANDING", "TAKEOFF", "CRUISE", or "APPROACH".
    - Create a new feature FLIGHT_PHASE_TEXT and fill it by searching for these keywords within the RMK_TEXT.
    - If a keyword is found, assign the corresponding phase to the new column. If multiple are found, take the first one. If none are found, use 'UNKNOWN'.
    - Compare this newly created feature with the existing FLT_PHASE column to see how well you were able to infer the flight phase from the text.