DS321 – Phase 3 Project Assignment
Title: Data Preprocessing and Data Quality Enhancement

Objective

In this phase, you will transform your raw dataset into a clean, consistent, and analysis-ready version. Building on **Phase 1 (Dataset Selection)** and **Phase 2 (Introduction, Related Work, and EDA)**, you will now focus on **data preprocessing** as a critical step in the data science pipeline.

You must:

1. Diagnose and document data quality issues (missing values, outliers, inconsistent formats, duplicates, etc.).
2. Apply appropriate preprocessing techniques to address these issues.
3. Justify your choices and compare the dataset **before and after** preprocessing.

Use the **same IEEE template on e-Learning** for all formatting and citations.

Tasks and Deliverables

1 – Data Quality Diagnosis

In a subsection titled **"Data Quality Assessment"**, you must:

- Briefly remind the reader of your dataset (1 short paragraph):
  - Domain, source, and main prediction/analysis goal.
- Systematically describe data quality problems, including at least:
  - **Missing values**: Which features? How many? (percentages per column).
  - **Outliers**: Which numerical features show extreme values? (Use boxplots or IQR-based summary).
  - **Inconsistent / mixed types**: e.g., strings instead of numbers, mixed date formats.
  - **Duplicates**: Number of duplicate rows (if any).
- Summarize your findings in at least **one table**, e.g.:

Issue Type	Feature(s)	Description	Evidence (count/%)
Missing Values	Age, Income	12% missing in "Income"	120 / 1000 rows
Outliers	Purchase Amount	Extreme right tail, IQR > 3	35 suspected rows
Duplicates	All features	Repeated rows in transactions table	18 duplicates

(You can adapt the columns to fit your dataset.)

2 – Data Preprocessing Pipeline

In a subsection titled **"Data Preprocessing Methods"**, you must:

Implement and document a **clear preprocessing pipeline** using Python (pandas, NumPy, etc.). At a minimum, your pipeline should include:

1. **Handling Missing Values**
    - Choose suitable strategies (e.g., deletion, mean/median imputation, mode, KNN).
2. **Encoding Categorical Variables**
    - Apply suitable encoding (e.g., one-hot encoding, label encoding, ordinal encoding).
    - Justify why this encoding is suitable for your future ML models (classification/regression).
3. **Outlier Treatment**
    - Clearly state how you detect outliers (e.g., IQR rule, z-score).
    - Decide to **keep, cap, or remove** outliers, and explain your reasoning (e.g., true extreme behavior vs. data errors).
4. **Feature Scaling (if relevant)**
    - Apply standardization or normalization for numerical features if you plan to use models sensitive to scale (e.g., KNN, SVM, logistic regression).
5. **Dealing with Imbalanced Classes (if classification)**
    - Check class distribution (e.g., target variable counts).
    - If heavily imbalanced, apply techniques such as:
        - Under-sampling / over-sampling,
        - SMOTE or similar.
    - Describe what you did and show **before/after** class distribution.

You must present your pipeline as a **logical sequence of steps**, either as:

- A **numbered list in the paper**, and/or
- A **simple flowchart figure**.

3 – Before/After Comparison and Justification

In a subsection titled **"Impact of Preprocessing"**, you must:

1. Provide **summary statistics before vs. after** preprocessing for key variables (e.g., means, medians, min, max, missing counts).
2. Include at least **one figure** that illustrates the effect of preprocessing, such as:
    - Histogram before vs. after handling outliers.
    - Boxplot before vs. after scaling.
    - Bar chart of class distribution before vs. after balancing.

Each figure must include:

- Title
- Axis labels
- Figure caption (e.g., "Figure 3: Distribution of 'Age' Before and After Outlier Treatment").

3. Write **1–2 paragraphs** reflecting on your decisions:
    - Which preprocessing step had the biggest impact on the dataset?
    - What trade-offs did you make (e.g., removing rows vs. keeping more data with imputation)?
    - How does this cleaned dataset better support future modeling in Phase 4?

Code Requirements

You must submit a **Python code file (.ipynb or .py)** that:

- Loads the original dataset (same as Phases 1 and 2).
- Performs **all preprocessing steps** described in your report.
- Produces the **tables and figures** included in your paper.
- Uses clear comments explaining each major step (e.g., `# Handle missing values in 'Age' using median`).

Submission Requirements

You must submit **both**:

1. **IEEE-formatted file**
    - Use the **same IEEE template** from e-Learning.
    - Add a section titled **"Data Preprocessing"** with subsections for:
        - Data Quality Assessment
        - Data Preprocessing Methods
        - Impact of Preprocessing
2. **Python Code File**
    - `.ipynb` (Jupyter/Colab) or `.py`.
    - File name suggestion: `Phase3_Preprocessing_YourID.ipynb`.

Upload both files to e-Learning **before the deadline**.

Notes

- Work must be **original**. **Plagiarism = zero grade**, **AI-generated text = zero grade** (same policy as previous phases).
- You must continue using the **same dataset** selected in **Phase 1** and analyzed in **Phase 2**.
- Your preprocessing should be **consistent** with your research goal (e.g., do not drop important variables just to simplify).