Lab 1

Data Preprocessing

In this lab, we will explore **Data Preprocessing**, a critical step in the data mining process that prepares raw data for analysis by cleaning, reducing, normalizing, and discretizing it.

1 Introduction

The *Adult Census Income* dataset, also known as the *Census Income dataset* or the *Adult dataset*, is sourced from the U.S. Census Bureau Database and is commonly used in classification tasks. The goal of this dataset is to predict if an individual's annual income exceeds $50,000 based on demographic characteristics.

The dataset contains 48,842 records and 14 attributes (features) as follows:

Attribute	**Description**
`age`	The age of the individual.
`workclass`	The type of employment (e.g., Private, Government, Self-employed, etc.).
`fnlwgt`	The final weight of the survey.
`education`	The level of education.
`education-num`	The number of years of education, in integer form.
`marital-status`	Marital status (single, married, divorced, etc.).
`occupation`	Occupation (Managerial, Clerical, Service, etc.).
`relationship`	Relationship to the household (spouse, child, other relative, etc.).
`race`	Race.
`sex`	Gender.
`capital-gain`	Income from capital (not from wages).
`capital-loss`	Loss from capital (from investments).
`hours-per-week`	The number of hours worked per week.
`native-country`	Country of origin.

The dataset also contains a target variable **income** with two values, including `<=50K` and `>50K`, representing the individual's income.

To get the dataset, please visit: Adult - UCI Machine Learning.

Students will learn to apply various data preprocessing techniques on the dataset mention above using essential **Python** libraries such as `pandas`, `numpy`, `matplotlib`, etc. (no `sklearn`). The result will be presented in a **Jupyter Notebook**.

2 Description

2.1 Data Cleaning

The data is dirty in the real world with lots of potential incorrect data from human mistakes, computing errors, or transmission problems. And that's why data needs preprocessing to ensure the accuracy and reliability of later analyses.

To perform data cleaning tasks, let's answer the questions and address the problems below:

- **Assessment of Missing Data**: Is there any missing data present? If so, what actions should be taken and why?
  - Should the affected records be excluded from the dataset?
  - Should the missing values be imputed? If so, which imputation methods will be applied (e.g., global constant, mean, etc.)?
- **Identification of Duplicate Records**: Are there any duplicate records present in dataset? If duplicates exist, keep only one of them.
- **Additional Data Cleaning Methods**: Are there any further steps, methods that can be implemented to make the data to be more cleaner?

2.2 Features Selection

Complex data analysis can take a significant amount of time to execute on the complete dataset. Higher dimensions make data more sparse and create exponentially more subspace combinations, but these combinations often lack meaning. Therefore, features selection is an important step in data preprocessing that simplifies the dataset while retaining its most essential characteristics, allowing for more efficient modeling and visualization.

To perform features selection, let's address the following questions:

- **Feature Selection**: Which features (attributes) are the most informative in the data? Why? Which features should be kept in the dataset?
  - Are there any features that are redundant or highly correlated with others? If so, which methods (e.g., correlation thresholding, Variance Inflation Factor) will be used to eliminate these features?
  - How many features should be retained to ensure minimal information loss?

2.3 Data Normalization

Data normalization is a critical step in data preprocessing, where the values of features are scaled to a common range. This ensures that no single feature disproportionately influences the analysis.

To perform data normalization, let's address the following questions:

- **Need for Normalization**: Does the data contain features with different scales?
- **Normalization Techniques**: Which normalization techniques should be applied? Why?
  - **Min-Max Scaling**:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \times (X_{\text{max\_new}} - X_{\text{min\_new}}) + X_{\text{min\_new}}$$

  - **Decimal Scaling**:

$$X_{\text{scaled}} = \frac{X}{10^j}$$

  - **Z-score Standardization**:

$$X_{\text{normalized}} = \frac{X - \mu}{\sigma}$$

  - Are there any other normalization methods (e.g., **Robust Scaling**) that can handle outliers effectively?

2.4 Data Discretization

Data discretization is a preprocessing technique that transforms continuous data into intervals. This step is useful for simplifying data patterns and is often used in machine learning algorithms that work better with categorical data.

To perform data discretization, let's address the following questions:

- **Discretization Techniques**: Which techniques should be applied for discretization?
  - **Binning**: Equal-width (distance) partitioning, Equal-depth (frequency) partitioning.
  - **Histogram analysis**.
  - Are there any other methods?

3 Report (Jupyter Notebook)

The source code, result will be reported in a Jupyter Notebook with the following requirements:

- Student information (Student ID, full name, etc.).
- Self-evaluation of the assignment requirements.
- Detailed explanation of each step. Illustrative images, diagrams and equations are required.
- Each processing step must be fully commented, and results should be printed for observation.
- The report needs to be well-formatted.
- Before submitting, re-run the notebook (`Kernel` → `Restart & Run All`).
- References (if any).

4 Assessment

No.	**Details**	**Score**
1	Data Cleaning	25%
2	Features Selection	25%
3	Data Normalization	25%
4	Data Discretization	25%
	Total	**100%**

5 Notices

Please pay attention to the following notices:

- This is an **INDIVIDUAL** assignment.
- Duration: about 2 weeks.
- Any plagiarism, any tricks, or any lie will have a 0 point for the course grade.

The end.