**Name**: Uffan Mehmood Khan (22L-7947)

**Section**: BS(SE)-7C

**Course**: Generative AI

**Assignment 01**

**Section I:** *The Preprocessing Pipeline for Tabular Data*

**Q1. Why is preprocessing essential before applying machine learning algorithms to tabular datasets? Give two examples of what could go wrong without preprocessing.**

**A1**. Preprocessing is essential because raw tabular data may often contain errors, inconsistencies, missing values, and features that may mislead machine learning algorithms and make them useless within practical usage. Ensuring a secure and valid preprocessing pipeline, by handling outliers, errors and improving quality ensures that the dataset enhances a models performance.

* **Skewed or Inaccurate Analyses from Missing Data Problem:**

Datasets often contain missing values because of errors during data entry, sensor malfunctions, or incomplete data collection results in a ML model to produce inaccurate results or simply fail. For example, if you are analyzing customer purchasing habits and have many missing age values, your customer segments will be incomplete and inaccurate.

* **Misleading Insights from Inconsistent or Noisy Data Problem:**

Data can be inconsistent due to different formats, units, or descriptors from various sources. This is known as "noisy data", leading to misleading insights. For example, a dataset that includes temperature measurements in both Celsius and Fahrenheit without conversion will provide inaccurate results for any analysis that relies on consistent numerical values.

Without preprocessing, models may overfit, converge slowly, or misinterpret relationships in the dataset.

**Q2. Differentiate between standardization and normalization. In which scenarios would you prefer one over the other?**

**A2**. Standardization transforms features to have a mean of zero and a standard deviation of one, while normalization scales values into a fixed range, usually [0,1].

Standardization is preferred when features follow a Gaussian distribution or when algorithms assume standardized data, such as in Support Vector Machines or Linear Regression. Normalization is useful in scenarios where features have varying units and need to be compared directly, such as in neural networks or image data where pixel intensities must be scaled uniformly.

For example, in medical data, if "blood pressure" and "cholesterol level" are used together, standardization helps remove bias from different measurement units, whereas normalization is critical for image datasets where pixel values range from 0 to 255.

**Q3. Discuss the advantages and disadvantages of Label Encoding vs One-Hot Encoding. Provide an example where Label Encoding may lead to incorrect results.**

**A3.** Label Encoding assigns a unique integer to each category, while One-Hot Encoding creates binary vectors representing the presence or absence of each category.

Label Encoding is memory-efficient but can incorrectly imply ordinal relationships where none exist. For example, if "red," "blue," and "green" are encoded as 0, 1, and 2, a decision tree might incorrectly assume that "green > blue > red."

One-Hot Encoding avoids this problem by treating categories independently but increases dimensionality, which may be inefficient for datasets with many categories. A practical example where Label Encoding fails is in a dataset of countries: assigning "USA = 0, UK = 1, China = 2" would imply China > UK > USA, which has no real-world meaning, leading to biased predictions.