Boot Camp on Artificial Intelligence

Practical Assignment - 1

Date of Submission: **Maximum Marks:**

Python Programming Assignment: Data Preprocessing

Objective: The objective of this assignment is to apply various data preprocessing techniques on a given dataset to clean and prepare it for further analysis.

Dataset: Titanic Dataset

You will be using the **Titanic dataset**, which contains information about passengers on the Titanic and whether they survived. You can download the dataset from this link or load it directly using seaborn:

https://www.kaggle.com/competitions/titanic/data?select=train.csv

---

Task 1: Load the Dataset

1. Import necessary libraries
2. Load the dataset
3. Display the first five rows of the dataset

Expected Output: The first five rows of the dataset.

Task 2: Handle Missing Values

1. Identify missing values in each column
2. Drop columns with too many missing values (threshold: more than 50% missing)
3. Fill missing numerical values with the median of the respective column
4. Fill missing categorical values with the most frequent value (mode)

Expected Output: A cleaned dataset without missing values.

Task 3: Handle Duplicate Data

1. Check for duplicate rows
2. Remove duplicate rows

Expected Output: The number of duplicate rows found and removed.

Task 4: Convert Categorical Features to Numeric

1. Convert categorical columns (sex, embark_town, class, etc.) using **one-hot encoding**
2. Convert Boolean columns (alone, who) to numeric (0 and 1)

Expected Output: The dataset with all categorical columns transformed into numeric values.

Task 5: Feature Scaling

1. Normalize numerical features (age, fare, etc.) using **Min-Max Scaling**
2. Standardize numerical features using **StandardScaler** and compare results

Expected Output: A scaled dataset where all numerical features are normalized/standardized.

Task 6: Outlier Detection using IQR Method

1. Compute the **Interquartile Range (IQR)** for numerical features (age, fare, etc.).
2. Identify outliers using the **1.5 * IQR** rule.
3. Remove or replace outliers with appropriate values (e.g., mean/median).

Expected Output: A dataset where outliers are handled using the IQR method.

Submission Instructions:

- Submit the Jupyter Notebook (.ipynb) or Python script (.py) with all completed tasks.
- Ensure that all code is well-commented, and outputs are displayed.
- Attach the cleaned dataset (CSV format) after preprocessing.
