School of Computer Science Engineering and Technology

Lab:1

Course-B. Tech.	**Type-** Core
Course Code- CSET301	**Course Name-** Artificial Intelligence and Machine Learning
Year- 2025	**Semester-** Odd
Date- 25/07/2025	**Batch-** 2023-2027

CO-Mapping

	CO1	CO2	CO3	CO4	CO5
Q1		√	√		

AI/ML Lab – Tabular Data Preprocessing

Objective:

This lab aims to introduce students to fundamental data preprocessing techniques. Students will learn to clean, transform, and prepare tabular data for analysis using Python tools like pandas, numpy, and sklearn.

Problem Statement

You are provided with the famous **Titanic passenger dataset**. Your task is to perform basic preprocessing operations on the data to make it suitable for analysis and machine learning models.

The dataset can be loaded from the following link:
https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv

You can directly load the Titanic dataset in your Colab notebook using the following line of code:

```
url =
'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.
csv'
```

This URL points to a publicly available CSV file containing Titanic passenger data. You can use it with `pd.read_csv(url)` to load the dataset into a pandas DataFrame.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Instructions

Use your understanding of data preprocessing to carry out the following tasks in your own way:

- Load the dataset.
- Identify and handle missing values.
- Deal with duplicate data
- Convert categorical columns into numerical form.
- Normalize appropriate numerical features.
- Apply sorting and filtering logic.
- Engineer at least one new column based on your logic.
- (Optional) Create one or more simple visualizations to understand patterns in the data.

You may use tools such as **pandas**, **numpy**, **scikit-learn (sklearn)**, **matplotlib**, or **seaborn**. Make sure to include clear comments in your code to explain your approach.

Solution:

```
# Import necessary libraries
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder, StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns

# Load the Titanic dataset
url =
'https://raw.githubusercontent.com/datasciencedojo/datasets/master/tita
nic.csv'
df = pd.read_csv(url)
df.head()

# Check for missing values
print("Missing values in each column:")
print(df.isnull().sum())

# Print the number of duplicate rows
print(f"Number of duplicate rows: {df.duplicated().sum()}")

# Drop duplicate rows
df = df.drop_duplicates()

# Fill missing values
df['Age'] = df['Age'].fillna(df['Age'].mean())
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])

# Encode categorical columns
le = LabelEncoder()
df['Sex'] = le.fit_transform(df['Sex'])
df['Embarked'] = le.fit_transform(df['Embarked'])
df.head()

# Scale numeric features
scaler = StandardScaler()
df[['Age', 'Fare']] = scaler.fit_transform(df[['Age', 'Fare']])

# Sort by Fare
df_sorted = df.sort_values(by='Fare', ascending=False)
df_sorted.head()

# Filter passengers who paid more than average fare
high_fare = df[df['Fare'] > df['Fare'].mean()]
high_fare.head()
```

```
# Create a new column (symbolic AgeGroup based on scaled Age)
df['AgeGroup'] = pd.cut(df['Age'], bins=[-np.inf, -1, 0.0, 0.8, 1.6,
3],
                        labels=['Error', 'Child', 'Teen', 'Adult',
'Senior'])
df.head()

# Countplot of Survival
sns.countplot(data=df, x='Survived')
plt.title('Survival Count')
plt.show()

# Histogram of Fare
sns.histplot(df['Fare'], kde=True)
plt.title('Fare Distribution')
plt.show()
```