**­­­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**](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.

A screenshot of a computer

Description automatically generated

**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.