**COM624 MACHINE LEARNIG LAB WEEK 1**

**Lab Activity: Data Cleaning and Preprocessing with the Titanic Dataset**

**Module:** Level 6 Software Engineering **Duration:** 2 Hours **Lab Title:** Data Cleaning with Python – Titanic Dataset **Learning Outcomes:**

* Understand the importance of data cleaning in machine learning
* Apply Python techniques to identify and handle missing or inconsistent data
* Prepare a dataset for analysis and modeling

**Introduction**

In this lab, you will work with the Titanic dataset—a well-known dataset used in data science education. The goal is to clean and prepare the data for future analysis or machine learning tasks. Data cleaning is a critical first step in any data-driven project, ensuring that the dataset is accurate, complete, and usable.

**Task Overview**

You will:

1. Load the Titanic dataset using Python
2. Explore the dataset and identify missing values
3. Apply data cleaning techniques
4. Encode categorical variables
5. Save the cleaned dataset for future use

**Dataset Access**

You may access the dataset in two ways:

* **GitHub URL:** https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv
* **Local Path (if downloaded):** C:\Users\YourName\Documents\Datasets\titanic.csv *(Replace with your actual path)*

**Step-by-Step Instructions**

**Step 1: Load the Dataset**

python

import pandas as pd

# Load from GitHub

df = pd.read\_csv('https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv')

# OR load from local path

# df = pd.read\_csv('C:/Users/YourName/Documents/Datasets/titanic.csv')

df.head()

**Step 2: Explore the Dataset**

python

# View basic structure

df.info()

# Check for missing values

df.isnull().sum()

**Explanation:** This helps you understand which columns have missing data and the overall structure of the dataset.

**Step 3: Drop Irrelevant Columns**

python

# Drop columns not useful for analysis

df.drop(['PassengerId', 'Name', 'Ticket', 'Cabin'], axis=1, inplace=True)

**Explanation:** These columns may not contribute meaningfully to analysis or prediction.

**Step 4: Handle Missing Values**

python

# Fill missing 'Age' with median

df['Age'].fillna(df['Age'].median(), inplace=True)

# Fill missing 'Embarked' with mode

df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

**Explanation:** Filling missing values helps maintain dataset integrity without removing valuable rows.

**Step 5: Encode Categorical Variables**

python

# Convert 'Sex' to numeric

df['Sex'] = df['Sex'].map({'male': 0, 'female': 1})

# Convert 'Embarked' to dummy variables

df = pd.get\_dummies(df, columns=['Embarked'], drop\_first=True)

**Explanation:** Machine learning models require numerical input, so categorical variables must be encoded.

**Step 6: Save the Cleaned Dataset**

python

# Save cleaned dataset

df.to\_csv('titanic\_cleaned.csv', index=False)

Explanation: This allows you to reuse the cleaned data in future analysis or modeling tasks.

**Expected Outcome**

By the end of this lab, you should have:

* A cleaned version of the Titanic dataset
* A clear understanding of how to handle missing data and encode variables
* A saved CSV file ready for analysis

**Extension Task (Optional)**

* Use seaborn or matplotlib to visualize missing data before and after cleaning
* Write a short reflection on the cleaning decisions you made and why

**Upload Your Lab activity week 1 on your GITHUB share your link with me. Make sure it is made public**