**Batch: H1-3 Roll No.: 16010122083**

**Experiment 02**

|  |
| --- |
| **Title:**  Dataset pre-processing |

# Objective:

# 1. To learn how to prepare the dataset

# 2. To learn various steps in Data -Preprocessing

# Course Outcome:

# CO1: Learn how to locate and download datasets, extract insights from that data and present their findings in a variety of different formats.

# Books/ Journals/ Websites referred:

Google

W3School

https://pandas.pydata.org/

# Resources used:

Excel

Python  
VS Code for code editing

Pandas

sklearn.preprocessing

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Theory (About Data Preprocessing):

Data preprocessing is a crucial step in the data analysis and machine learning pipeline. It involves cleaning, transforming, and organizing raw data into a format that is suitable for analysis or modeling. Proper data preprocessing ensures that the data is accurate, consistent, and ready for effective use. Here are the key concepts in data preprocessing:

1. **Data Cleaning:** This step involves identifying and handling errors, inconsistencies, and missing values in the dataset. Methods include removing duplicates, filling missing values with statistical measures (mean, median, mode), and identifying outliers that might distort analysis.
2. **Data Transformation:** Data often needs to be transformed into a suitable format. Common transformations include:
   * **Normalization:** Scaling features to a common range (e.g., 0 to 1) to mitigate the impact of differing scales.
   * **Standardization:** Transforming features to have zero mean and unit variance, often useful for algorithms that assume normally distributed data.
   * **Log Transformation:** Applying logarithmic functions to skewed data to make it more normally distributed.
3. **Encoding Categorical Data:** Many algorithms require numerical input, so categorical variables must be encoded. Common techniques include:
   * **Label Encoding:** Assigning a unique numerical value to each category.
   * **One-Hot Encoding:** Creating binary columns for each category, indicating its presence or absence.

Some more uses more point for Data Processing are Future Engineering, Handling Imbalanced data, Dealing with Data and Time Data, Dimensionality Reduction, Data Splitting, Validation and Cross-Validation and Handling Text and Textual Data.

# Following points should be written by students

# Different steps in Data Preprocessing:

# Finding missing, null values

# Replacing missing, null values with statistical parameters

# Encoding categorical data

# Normalization

# Note: Student can use any technology like Tableau, Tableau-Prep, PowerBI, Google spreadsheet, excel, R programming, Python, Java any other technology for preprocessing.

# Platform used by the student:

# Working (Paste the code and Output for each Data Preprocessing task):

import pandas as pd

from sklearn.preprocessing import LabelEncoder, MinMaxScaler

# Sample Dataset to work upon

df = pd.read\_csv('sample\_dataset.csv')

# Find missing values

missing\_values = df.isnull().sum()

# Replace missing values with mean

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

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

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

# Encoding the Gender into 0 and 1

label\_encoder = LabelEncoder()

df['Gender'] = label\_encoder.fit\_transform(df['Gender'])

# Used for Normalization

scaler = MinMaxScaler()

df[['Income', 'Score']] = scaler.fit\_transform(df[[ 'Income', 'Score']])

df.to\_csv('preprocessed\_dataset.csv', index=False)

# Input:

# A screenshot of a spreadsheet Description automatically generated File: sample\_dataset.csv

# Output for each step:

# File: preprocessed\_dataset.csv

# A screenshot of a spreadsheet Description automatically generated

# After finding missing values replacing it with statistical data.

# Mean in Column age in null cell A5, Median in Column Income in null cell C6 and Mode in Column Score in null cell D9.

# A screenshot of a spreadsheet Description automatically generated

# Label Encoding the Column Gender, Male as 1 and Female as 0

# A screenshot of a spreadsheet Description automatically generated

# Normalization of Column Income and Score

# Conclusion (Students should write in their own words):

# Post Lab Question:

# Write the importance of Data Preprocessing

# Enhanced Model Performance: Proper preprocessing like scaling, normalization, and handling missing values ensures that data is in a consistent format, leading to improved model accuracy and performance.

# Noise Reduction: Cleaning noisy data, removing outliers, and handling inconsistent formats minimizes the impact of irrelevant or erroneous information on the analysis or model training.

# Feature Relevance: Feature selection and extraction techniques eliminate irrelevant or redundant features, focusing the model on the most informative attributes and reducing complexity.

# Addressing Missing Values: Dealing with missing data through imputation or removal prevents biased analysis and inaccurate predictions, contributing to more reliable results.

# Normalization of Scales: Standardizing different scales or units in features prevents certain features from dominating the model due to their larger values, leading to a more balanced influence on the model's outcome.

# Interpretability and Generalization: Proper preprocessing enables better understanding of data patterns and relationships, allowing models to generalize well to unseen data and increasing their interpretability.