**Batch:-** D-2 **Roll No.:-**16010122151

**Experiment No. 1**

|  |
| --- |
| **Title: Implement data pre-processing using python on real world dataset** |

# Course Outcome:

# CO1 Understand basic concepts of data analytics to solve real-world problems

# Books/ Journals/ Websites referred:

(Students should write)

# Resources used:

(Students should write)

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

# Theory (About Data Preprocessing):

(Students should write)

# Program:

import pandas as pd

import numpy as np

**# Sample data**

data = {

'name': ['Alice', 'Bob', 'Charlie', 'Dave', 'Eve'],

'age': [25, np.nan, 30, 22, 35],

'gender': ['F', 'M', 'M', 'M', 'F'],

'income': [50000, 60000, 75000, np.nan, 80000]

}

df = pd.DataFrame(data)

**# Display the original data**

print("Original DataFrame:")

print(df)

**# User-defined function for discretization**

def discretize\_age(age):

if age < 30:

return 'Young'

elif age >= 30 and age < 40:

return 'Middle-aged'

else:

return 'Old'

**# Handling missing values (NaN)**

**# Fill missing values in 'age' with the mean age**

mean\_age = df['age'].mean()

df['age'].fillna(mean\_age, inplace=True)

**# Apply discretization function to 'age' column**

df['age\_category'] = df['age'].apply(discretize\_age)

**# Drop rows with missing values in any column**

df.dropna(inplace=True)

**# Convert categorical variables (gender) to numerical**

df['gender'] = df['gender'].map({'F': 0, 'M': 1})

**# Data normalization Min -Max**

**# Normalize 'income' column to range [0, 1]**

min\_income = df['income'].min()

max\_income = df['income'].max()

df['income\_normalized'] = (df['income'] - min\_income) / (max\_income - min\_income)

**# Display cleaned, preprocessed, and discretized data**

print("\nCleaned, Preprocessed, and Discretized DataFrame:")

print(df)

—-----------------------------------------------------------------------------------------------------------------------

**Task: Download the real time data set and implement data preprocessing techniques on the real time data set**

# Source of the dataset (URL):

# Platform used by the student:

# 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 if needed (Write user defined function)

# Normalization (Write user defined function)

# Discretization (Write user defined function)

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

# We are Using Data Set of Amazon Sales.

# 

# 

# Task 1:- Delete Rows and Columns

# 1)Deleting Rows :-

# 

# 

# Here we have deleted rows with index -0,1,2.

# 2) Deleting Columns:-

# 

# 

# We can see here that first we had 14 Columns and after deleting we have 12 Columns.

# The columns deleted are *“Region”* & *“Country”*.

# Task 2) Remove all NAN/ Null values

# 

# My data base does not contains any NAN values but it contains “####”

# So,

# Replacing all “####” values with NAN then using the df.dropna()

# Students need to write comments wherever needed

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

**Post lab questions:**

**Q.1 What are some common challenges encountered during data cleaning? How did you handle missing values in the provided dataset?**

**Q.2 Explain the importance of data normalization in the context of machine learning models. How does normalizing benefit the analysis?**

**Q.3 Discuss why it's essential to convert categorical variables like 'gender' into numerical representations.**