EXPERIMENT 3C

Data Preprocessing – Create A Data Set

Aim:

To understand the importance of data preprocessing in data science.

Algorithm:

1. Import the required libraries — pandas and numpy.

2. Create a dataset containing student details with possible missing, invalid, and duplicate values.

3. Convert the dataset into a pandas DataFrame and display it.

4. Fill missing values in Age, Math\_Score, and Science\_Score with their respective mean values.

5. Replace negative Age values with the mean age.

6. Correct invalid Attendance(%) values greater than 100 by setting them to 100.

7. Remove duplicate records using drop\_duplicates().

8. Normalize Math\_Score and Science\_Score by dividing each by 100.

9. Display the processed and cleaned dataset.

Program:

import pandas as pd

import numpy as np

data = {

"Student\_ID": [101, 102, 103, 104, 105, 106, 106], # Duplicate ID

"Name": ["Alice", "Bob", "Charlie", "David", "Eva", "Frank", "Frank"],

"Age": [20, 21, np.nan, 23, -19, 22, 22],

"Math\_Score": [85, 90, np.nan, 70, 65, 88, 88],

"Science\_Score": [78, 82, 79, 60, 95, None, 80],

"Attendance(%)": [92, 85, 88, 45, 110, 80, 80]

}

df = pd.DataFrame(data)

df.to\_csv("students.csv", index=False)

print("Original Dataset\n")

print(df)

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

df["Math\_Score"].fillna(df["Math\_Score"].mean(), inplace=True)

df["Science\_Score"].fillna(df["Science\_Score"].mean(), inplace=True)

df.loc[df["Age"] < 0, "Age"] = df["Age"].mean()

df.loc[df["Attendance(%)"] > 100, "Attendance(%)"] = 100

df.drop\_duplicates(inplace=True)

df["Math\_Score\_Normalized"] = df["Math\_Score"] / 100

df["Science\_Score\_Normalized"] = df["Science\_Score"] / 100

print("\nProcessed Dataset:\n", df)

A screenshot of a computer

AI-generated content may be incorrect.Output:

Result:

Hence the python program for data preprocessing is written and executed successfully.