□ Experiment 2: Data Preprocessing – Handling Missing Values and Normalization

Title:

Data Preprocessing using pandas and scikit-learn

Aim:

To clean and prepare raw data by handling missing values and normalizing numerical data for machine learning models.

Objectives:

- Learn how to identify and handle missing values.
- Perform data normalization and scaling.
- Understand why preprocessing is essential before training a model.

Theory:

Data preprocessing is the first step in the ML pipeline. It ensures that data is clean, consistent, and suitable for model training.
Common steps include:

- **Handling missing values:** Filling, removing, or imputing missing data.
- **Normalization:** Scaling values between 0 and 1 (Min-Max scaling).
- **Standardization:** Converting data to have mean = 0 and std = 1.

Algorithm / Steps:

1. Import libraries (pandas, NumPy, scikit-learn).
2. Create or load a dataset with missing values.
3. Detect missing values using `isnull()`.
4. Handle missing data using `fillna()` or by dropping rows.
5. Normalize numerical data using MinMaxScaler.
6. Display preprocessed data.

Sample Python Code:

```
# Experiment 2: Data Preprocessing

import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler

# 1. Create a sample dataset with missing values
data = {'Age': [22, 25, np.nan, 28, 30],
        'Salary': [40000, 50000, 45000, np.nan, 60000]}
df = pd.DataFrame(data)
print("Original Data:\n", df)

# 2. Handle missing values
df['Age'].fillna(df['Age'].mean(), inplace=True)
df['Salary'].fillna(df['Salary'].median(), inplace=True)
print("\nAfter Handling Missing Values:\n", df)

# 3. Normalize data
scaler = MinMaxScaler()
df[['Age', 'Salary']] = scaler.fit_transform(df[['Age', 'Salary']])
print("\nAfter Normalization:\n", df)
```

Expected Output:

- Missing values replaced with mean (Age) and median (Salary).
- Data scaled between 0 and 1.

Example Output:

```
Original Data:
     Age   Salary
0  22.0  40000.0
1  25.0  50000.0
2   NaN  45000.0
3  28.0      NaN
4  30.0  60000.0

After Handling Missing Values:
     Age   Salary
0  22.0  40000.0
1  25.0  50000.0
2  26.25 45000.0
3  28.0  47500.0
4  30.0  60000.0

After Normalization:
        Age    Salary
0  0.000000  0.000000
```

```
1  0.428571  0.666667
2  0.642857  0.333333
3  0.857143  0.500000
4  1.000000  1.000000
```

Result:

The experiment demonstrated data cleaning and normalization techniques essential for ML model preparation.

Viva Questions:

1. Why is data preprocessing important?
2. What is the difference between normalization and standardization?
3. What functions are used to fill missing values in pandas?
4. What happens if we don’t handle missing data before training?
5. What is the role of MinMaxScaler in preprocessing?