Lab – 02: Data Preprocessing

1. Introduction

Data preprocessing is a crucial step in machine learning, as real-world datasets are often incomplete, inconsistent, or unformatted. Preprocessing ensures that the data is clean and suitable for ML algorithms.

Key preprocessing tasks include:

i.Handling missing values
Ii. Encoding categorical variables

2. Experiments and Observations

2.1 Handling Missing Values

Procedure:

1. Loaded a dataset with missing values.

2. Identified missing values using `isnull()` and `sum()`.

3. Handled missing values using methods such as:

○ Removing rows with missing data

○ Filling missing values with mean, median, or mode

○ Forward or backward filling

Code :

```
import pandas as pd
import numpy as np

# Creating a sample dataset with missing values
data = {
   "Name": ["Alice", "Bob", "Charlie", "David", "Eva"],
   "Age": [25, np.nan, 30, 28, np.nan],
   "Salary": [50000, 54000, np.nan, 58000, 60000]
}

df = pd.DataFrame(data)
print("Original Data:")
print(df)

# Check missing values
print("\nMissing Values:")
print(df.isnull().sum())
```

```
# Fill missing values with mean
df['Age'].fillna(df['Age'].mean(), inplace=True)
df['Salary'].fillna(df['Salary'].mean(), inplace=True)

print("\nData after filling missing values with mean:")
print(df)

# Alternatively, drop rows with missing values
# df.dropna(inplace=True)
```

Observation :

```
Original Data:
      Name   Age   Salary
0    Alice  25.0  50000.0
1      Bob   NaN  54000.0
2  Charlie  30.0      NaN
3    David  28.0  58000.0
4      Eva   NaN  60000.0

Missing Values:
Name      0
Age       2
Salary    1
dtype: int64

Data after filling missing values with mean:
      Name        Age   Salary
0    Alice  25.000000  50000.0
1      Bob  27.666667  54000.0
2  Charlie  30.000000  55500.0
3    David  28.000000  58000.0
4      Eva  27.666667  60000.0
```

2.2 Encoding Categorical Data

Procedure:

- Categorical variables cannot be used directly in ML models.
- Applied Label Encoding for ordinal data and One-Hot Encoding for nominal data.

Code :

```
from sklearn.preprocessing import LabelEncoder

# Sample dataset with categorical data
data = {
    "Name": ["Alice", "Bob", "Charlie", "David", "Eva"],
    "Department": ["HR", "IT", "Finance", "IT", "HR"]
}
df = pd.DataFrame(data)

print("Original Data:")
print(df)

# Label Encoding for Department (if ordinal)
le = LabelEncoder()
df['Dept_Label'] = le.fit_transform(df['Department'])
print("\nAfter Label Encoding:")
print(df)

# One-Hot Encoding for nominal data
df_onehot = pd.get_dummies(df, columns=['Department'])
print("\nAfter One-Hot Encoding:")
print(df_onehot)
```

Observation :

```
Original Data:
      Name Department
0    Alice         HR
1      Bob         IT
2  Charlie    Finance
3    David         IT
4      Eva         HR

After Label Encoding:
      Name Department  Dept_Label
0    Alice         HR           1
1      Bob         IT           2
2  Charlie    Finance           0
3    David         IT           2
4      Eva         HR           1

After One-Hot Encoding:
      Name  Dept_Label  Department_Finance  Department_HR  Department_IT
0    Alice           1               False           True          False
1      Bob           2               False          False           True
2  Charlie           0                True          False          False
3    David           2               False          False           True
4      Eva           1               False           True          False
```

3. Conclusion

- Missing values can be handled by filling with mean/median/mode or dropping rows, depending on dataset requirements.
- Categorical variables must be encoded for ML algorithms to process.
- Data preprocessing is essential to improve model accuracy and performance.