Q1.) Explain Data preprocessing in Details.

ANS.
- Data Preprocessing is a critical step in Data Analysis and Machine Learning.
- It involves transforming the raw data into a clean and structured format to enhance its qualities and Usability.

STEPS:-

(1.) Data cleaning.
- Handle the missing values using fill and remove.
- Removes Duplicates.
- Correct the inconsistent data values.

(2.) Data Transformation.
- Normalize or scale data using Min-Max scaling.
- Encode categorical data using label Encoding and One-hot Encoding.

(3.) Data Reduction.
- Dimensionality reduction using PCA.
- Feature selection to eliminate irrelevant variables.

(4.) Data Integration.
- Merge Data from different sources.

→ Importance:-

i.) Ensures Accuracy.
ii.) Ensures Consistency.
iii.) Ensures Efficiency.

Q2.) Explain Data Frame with a suitable example?

ANS.
- A DataFrame is a two-dimensional Data structure with Python's pandas Library, similar to a spread sheet or a SQL Table.
- It consists of rows and columns, where each column can hold a different Datatype.

For Example:-

```
import pandas as pd
data = {"Name": ["Alice", "Bob", "Charlie"],
        "Age": [25, 30, 35],
        "City": ["Pune", "Mumbai", "Nagpur"]
       }
df = pd.DataFrame(data)
print(df)
```

OUTPUT:-

	NAME	AGE	CITY
0	Alice	25	Pune
1	Bob	30	Mumbai
2	Charlie	35	Nagpur

Features:

i.) Easy Data Manipulation.
ii.) Provides statistical Methods like mean(), sum(), describe(), etc.
iii.) Allows Filtering and Slicing.

Q3.) What is the Limitation of the Label Encoding Method?

Ans. Label Encoding assigns an unique numeric values to each category in a column.

Limitations:-

1. It may introduce unintended ordinal relations between categories.
2. It may misinterpret encoded values as having importance ~~of~~ or ranking, leading to incorrect predicition
3. If a categorical feature has many unique values then encoding may lead to inefficient computer.

Q4) What is the Need of Data Normalisation?

ANS. o Data Normalization scales data to a specific range like 0-1, etc.

Need of Normalization:-

① It ensures that features with larger magnitude do not dominate the Model.

② It speeds up the training.

③ It prevents biased in distance-based Algorithm like KNN and SVM

④ It ~~features when~~ is useful when Features have different different units like age in years, income in Dollars.

Common Techniques:-

1. Min-max scaling.
2. Z-score Normalization.

Q5) What are the different Techniques for Handling the Missing Data?

ANS. Handling Missing Data ensures Accurate Analysis.

Common Techniques:-

(1) Deletion Methods.

(i) Listwise Deletion:-
Remove rows with missing values

(ii) Pairwise Deletion:-
Use available data without deleting rows.

(2) Imputation Methods:-

(i) Mean/Median Imputation
- Replace missing values with mean or median.

(ii) Mode Imputation.
- Use the most Frequent (precise) value.

(iii) Forward / Backward

. Fill/will previous or next value.

(3.) Prediction Based Methods

Use the Machine Learning Algorithms to predict the missing values.

(4.) Indicator Variable

Create a Binary variable indicating whether data is missing.

(5.) Dropping Entire Column.

Used if a column has too many missing values.

16/11/25