UNIT-1

A retail company has collected customer transaction data from multiple stores. However, the dataset contains irrelevant attributes, duplicate records, and inconsistent formats. Analyze how different data preprocessing techniques such as data cleaning, data reduction, and data transformation can improve the dataset's quality. Support your answer with suitable examples.

A dataset collected from various sources contains missing values, redundant features, and inconsistent data formats. As a data scientist, analyze the role of data cleaning, data reduction, data transformation, and data discretization in preparing the dataset for machine learning. Provide relevant examples to justify your analysis.

UNIT-2

Apply Exploratory Data Analysis (EDA) techniques to visualize the distribution of a real-world dataset and interpret the results.

Analyze how histograms and box plots complement each other in understanding numerical distributions. Provide examples.

UNIT-3

Evaluate the effectiveness of A/B testing in real-world decision-making by designing a hypothetical experiment and interpreting results.

Demonstrate the use of ANOVA in comparing multiple categories of data. Explain the interpretation of results with a case study.

UNIT-4

An e-commerce company wants to estimate the average delivery time for orders. Since tracking delivery times for all 500,000 orders per month is impractical, they select a random sample of 200 orders.

Questions:
1.Explain how the Central Limit Theorem (CLT) allows the company to estimate the true mean delivery time, even if individual delivery times vary widely and are not normally distributed.
2.If the standard deviation of delivery times is 2.5 days, what is the expected standard error for a sample of 200 orders?
3.If the sample mean delivery time is 4.2 days, construct a 95% confidence interval for the true mean delivery time.

4.If the company increases the sample size from 200 to 800 orders, how will this affect the sampling distribution and the precision of the estimate?
5.How can accurate estimation of delivery times using CLT help the company optimize logistics and improve customer satisfaction?

UNIT-5

Compare the visualization capabilities of Altair and RapidMiner in terms of handling large datasets.

Analyze how Tableau can be used for real-time data visualization in business analytics. Provide an example of its application.