(86)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -2 EXAMINATIONS- 2026

B.Tech.-VI Semester (CSE/IT)

COURSE CODE (CREDITS): 25B1WCI642 (3) MAX MARKS: 25

COURSE NAME: Introduction to Data Analytics and Visualization

COURSE INSTRUCTOR: Dr Nancy Singla MAX. TIME: 1 Hour 30 Min

Note: (a) All questions are compulsory.

(b) The candidate is allowed to make suitable numeric assumptions wherever required for solving problems

(c) Use of calculator is allowed.

Q. No	Question	CO	Marks
Q1	You are working on a sentiment analysis project for customer reviews of an e-commerce website. Before training your machine learning model, you need to preprocess the text data. Discuss the role of stemming and lemmatization in improving text preprocessing for this task. Demonstrate with a short code snippet how both methods transform sample words like "running," "studies," and "better."	CO3	[5]
Q2	(a) Explain how a box plot helps in identifying outliers during data processing. (b) You are given the following dataset of employee salaries (in ₹): [25000, 27000, 28000, 30000, 31000, 32000, 33000, 100000] Identify any outliers using the IQR rule. (c) Draw a suitably labelled box plot for this data, clearly indicating any outliers.	CO3	[2+2+1]
Q3	A company is testing a new security system. Each login attempt has a 97% chance of being successful. (a) Which probability distribution is suitable for modeling each of the following attempts? Explain briefly why. (i) a single login attempt? (ii) the number of successful logins out of 50 attempts? (b) Using the appropriate distribution(s): 1. What is the probability that a single login attempt fails? 2. What is the probability that exactly 48 out of 50 login attempts succeed? 3. Write Python code using `scipy.stats` to compute both probabilities.	CO5	[2+3]

<table>
<tr><td>Q4</td><td>A researcher is testing whether a new teaching method improves student performance. The null hypothesis is that the method has no effect. A sample of 40 students is tested, and the statistical test yields a p-value of 0.03.
(a) Define the p-value in hypothesis testing.
(b) Explain what a p-value of 0.03 means in the context of this study if the significance level is α=0.05.</td><td>CO5</td><td>[2]</td></tr>
<tr><td>Q5</td><td>A hospital is studying the relationship between number of hours of physical therapy per week (X) and mobility score improvement (Y) in patients recovering from knee surgery. The data for 6 patients is shown below:
<table><tr><th>Hours of Therapy (X)</th><th>Mobility Improvement Score (Y)</th></tr><tr><td>2</td><td>15</td></tr><tr><td>4</td><td>25</td></tr><tr><td>6</td><td>35</td></tr><tr><td>8</td><td>50</td></tr><tr><td>10</td><td>55</td></tr><tr><td>12</td><td>60</td></tr></table>(a) Derive the best-fit regression line (Y = a + bX).
(b) Compute the Mean Squared Error (MSE) for the fitted line.
(c) Predict the mobility improvement score for a patient receiving 9 hours of therapy per week.</td><td>CO5</td><td>[2+2+1]</td></tr>
<tr><td>Q6</td><td>For each of the following scenarios, identify the most suitable type of data visualization to effectively represent the data and briefly explain your choice:
(a) Showing the geographic concentration of customer purchases across different cities.
(b) Displaying the relationship between two numerical variables to identify correlation.
(c) Comparing the performance metrics (precision, recall, F1-score) of two classification models.</td><td>CO4</td><td>[3]</td></tr>
</table>