JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -2 EXAMINATION- 2025

M. Tech-I Semester (CSE)

COURSE CODE (CREDITS): 22M11CI112 (3) MAX. MARKS: 25

COURSE NAME: INTRODUCTION TO DATA SCIENCE

COURSE INSTRUCTORS: 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 tasked with managing semi-structured data generated from an online survey platform. The data includes varying fields depending on the survey type, such as user responses, timestamps, device info, and optional nested metadata. (a) Explain which type of database would be most suitable for storing this semi-structured data? (b) Using Python and the pymongo library, write code snippets to: 1. Connect to a MongoDB database called "survey_db". 2. Insert a sample document representing a user response, including fields: user_id, survey_id, response, timestamp, and nested device information (type and OS). 3. Update the document for a specific user_id by adding a new field 'reviewed' with value True. 4. Delete all documents where the device.type is "tablet".	CO3	[2+5]
Q2.	Your team is building a search engine that indexes articles. The search engine is supposed to return articles based on keywords that users input. Users might search for terms like **"cars," "car,"** and **"automobiles"**. Would you apply stemming or lemmatization to reduce the word variations in the search query? Justify your choice and explain how it would improve the search engine results.	CO3	[3]
Q3.	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) Visualizing monthly sales trends of multiple products over the past year. b) Comparing the distribution and outliers of customer ages in a dataset. c) Showing the geographic concentration of customer purchases across different cities.	CO4	[5]

<table>
<tr><td></td><td>d) Displaying the relationship between two numerical variables to identify correlation.
e) Comparing the performance metrics (precision, recall, F1-score) of two classification models.</td><td></td><td></td></tr>
<tr><td>Q4.</td><td>A customer support center receives an average of 5 calls per hour. Each call has a 10% chance of being a technical support request.
(a) Calculate the probability that exactly 2 out of 5 calls received in an hour are technical support requests.
(b) Approximate the probability of exactly 2 technical support calls occurring in an hour.</td><td>CO5</td><td>[3]</td></tr>
<tr><td>Q5.</td><td>A dataset contains the following information about hours studied (x) and corresponding exam scores (y) for 5 students:

<table><tr><th>Hours Studied (x)</th><th>Exam scores (y)</th></tr><tr><td>1</td><td>50</td></tr><tr><td>2</td><td>55</td></tr><tr><td>3</td><td>65</td></tr><tr><td>4</td><td>70</td></tr><tr><td>5</td><td>75</td></tr></table>
(a) Using the least squares method, find the linear regression equation that predicts exam scores based on hours studied.

(b) Using your regression equation, calculate the predicted scores for each student and compute the Root Mean Squared Error (RMSE) of your model.

(c) Explain the difference between Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Which metric is generally preferred for evaluating regression models, and why?</td><td>CO6</td><td>[3+2+2]</td></tr>
</table>