(149)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -2 EXAMINATIONS- 2026

B.Tech-IV Semester (CSE/IT/BT/ Mathematics and Computing)

COURSE CODE (CREDITS): L-25B11CI413 (3) MAX MARKS: 25

COURSE NAME: Artificial Intelligence and Machine Learning COURSE INSTRUCTOR: Dr. Ravindara, Dr. Aman, Dr. Monika, Ms. Vani 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 problem (c) Use of calculator is allowed*

Q.No.	Question	CO	Marks
Q1	**A.** **[2 Marks]** A student claims: 'Adding more data points always improves a linear regression model.' Do you agree? Under what specific condition can adding a single data point drastically worsen the model? Name this phenomenon. **B.** **[5 Marks]** You are given **six** data points mapping temperature X (°C) to ice-cream sales Y (units). The first five points are well-behaved. The sixth point is a suspicious entry logged during a system error: Point | P1 | P2 | P3 | P4 | P5 | P6 X (Temp °C) | 20 | 25 | 30 | 35 | 40 | 42 Y (Sales) | 100 | 130 | 160 | 190 | 220 | 10 **a)** **[2 Marks]** Fit a regression line using ALL six points. Compute m and b using Least Squares. **b)** **[1 Marks]** Now fit a regression line using only the first FIVE points. **c)** **[2 Marks]** Compare the two slopes. Calculate percentage change in slope. What does this tell you about P6?	4	7
Q2	A bank develops a machine learning model to detect fraudulent transactions. In the dataset, fraudulent cases constitute only 1% of the total data, making it highly imbalanced. The model predicts "not fraud" for almost all transactions and achieves an accuracy of 99%. **a)** **[1 mark]** In this context, explain why accuracy is a misleading performance metric. **b)** **[1 mark]** Suggest more appropriate evaluation metrics for this problem and provide definitions for these metrics. **c)** **[2 mark]** If the model is too simple and misses fraud cases, what problem is occurring? Suggest techniques to overcome this problem.	2	4
Q3	**A.** **[2 marks]** Explain the concept of **Information Gain (IG)** and **Entropy (H)** in the context of decision tree learning. Why do we want to maximize Information Gain at each split? **B.** **[5 marks]** Consider the following dataset predicting whether a student **Passes** (Yes/No) an exam based on **Hours Studied** (High, Low) and **Attendance** (Good, Poor) and **Slept Well** (Yes, No). Assume $\log_2 3 = 1.6$. **a)** **[1 mark]** Find H(Passes).	4	7

b) **[2 marks]** Compute IG for all three attributes and identify the root node.

c) **[2 marks]** Draw the complete decision tree.

	1	2	3	4	5	6	7	8
Hours	H	H	H	L	L	L	L	H
Attendance	G	G	P	G	G	P	P	P
Slept Well	Y	N	Y	Y	N	Y	N	N
Passes	Y	Y	Y	Y	N	N	N	N

Q4	**A.** **[2 marks]** K-Means clustering always converges, yet it is *not* guaranteed to find the globally optimal solution. Explain this apparent contradiction precisely, and state one condition under which two different random initializations can yield completely different final cluster assignments even on the same dataset. **B.** **[5 marks]** A bank runs K-Means (k=3) on 10,000 credit card transactions using two features: Transaction Amount (₹) and Time of Day (hour, 0–23). Raw (un-normalized) values are used. After convergence the clusters are: Cluster 1: Low amount, daytime — labelled *"Normal"* Cluster 2: High amount, daytime — labelled *"Premium"* Cluster 3: High amount, night — labelled *"Suspicious"* The model flags Cluster 3 transactions for manual review. The bank reports WCSS = 48,000 and calls the model a success. However, the fraud team reports that 60% of actual fraudulent transactions are being missed.	3	7

	Cluster 1	Cluster 2	Cluster 3
Amount centroid (₹)	800	45,000	44,000
Hour centroid	13	14	2
Size	7200	2600	200

a) **[2 marks]** Without normalisation, Euclidean distance is dominated by one feature. Identify which feature dominates and calculate how many times larger its range is compared to the other feature. Hence explain why Cluster 3 captures so few transactions.

b) **[1 mark]** A fraudulent transaction occurs at ₹900, 3 AM. Compute its Euclidean distance to all three centroids using the raw values above, and state which cluster it is assigned to. Is this correct?

c) **[1 mark]** The analyst argues "WCSS is low, so the model is good." Construct a precise counter-argument using the concept of what K-Means actually optimises vs. what fraud detection requires.

d) **[1 mark]** Suggest two concrete changes — one to pre-processing and one to the algorithm choice — that would make this fraud detection system more reliable.