JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST-3 EXAMINATION-2025

B.Tech-VI Semester (BI)

COURSE CODE (CREDITS): 18B1WBI632 (3) MAX. MARKS: 35

COURSE NAME: Data Warehousing and Mining for Bioinformatics

COURSE INSTRUCTORS: Ekta Gandotra MAX. TIME: 2 Hrs.

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.	a. Compare the star, snowflake, and fact constellation schemas in terms of design complexity and their efficiency in supporting analytical query processing.	1, 2	3
	b. Given the following dataset, determine the five-number summary and draw a boxplot to visually represent the distribution. 12, 14, 18, 19, 21, 23, 24, 26, 27, 28, 29, 30, 32, 33, 34, 36, 37, 38, 40, 41, 42, 45, 47, 50, 55.		2
	c. Describe any two techniques for detecting outliers in a dataset.		2
Q2.	a. Evaluate the usefulness of the lift metric in association rule mining. How effective is it in measuring the strength and relevance of discovered patterns between itemsets?	5	3
	b. Apply the Apriori algorithm on the following transaction dataset to find the frequent patterns and generate the association rules. Use a minimum support of 3 and a minimum confidence of 60%.		4

TID	Itemsets
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, B, C, E
T6	A, B, C, D
T7	A, C
T8	B, C, E
T9	A, B, E

Q. No.	Question	CO	Marks
Q3.	a. Given a feedforward neural network with an input layer of 3 neurons, one hidden layer with 4 neurons using ReLU activation, and an output layer with 2 neurons using softmax activation, how many weights and biases are there in total?	4	3

b. Given the following training dataset, which predicts whether a student passes a course (Yes/No) based on their CGPA level (High, Medium, Low) and whether they studied (Yes/No), apply the C4.5 algorithm to determine the root node of the decision tree. 4

	1	2	3	4	5	6
CGPA	L	L	M	M	H	H
Studied	No	Yes	No	Yes	No	Yes
Passed	No	Yes	No	Yes	Yes	Yes

Q4. a. Apply the DBSCAN algorithm to the following dataset and label each data point as Core, Border, or Noise. 6 3

A(3, 7), B(4, 6), C(5, 5), D(6, 4), E(7, 3), F(6, 2), G(7, 2), H(8, 4). Use the parameters Epsilon (ε) = 2, Minimum Points (minPts) = 3, and the following distance matrix.

	A	B	C	D	E	F	G	H
A	0	1.41	2.83	4.24	5.66	5.83	6.40	5.83
B		0	1.41	2.83	4.24	4.47	5.00	4.47
C			0	1.41	2.83	3.16	3.61	3.16
D				0	1.41	2.00	2.24	2.00
E					0	1.41	1.00	1.41
F						0	1.00	2.83
G							0	2.24
H								0

b. Using the distance matrix given in the above question, perform agglomerative hierarchical clustering using the single linkage method. Illustrate each step of the clustering process and represent the final result using a dendrogram. Also find the optimal number of clusters. 4

Q5. a. Given the following two clusters of 2D points: 6 3

Cluster 1: (1, 2), (2, 3), (3, 3)

Cluster 2: (6, 7), (7, 8), (8, 8)

Using Manhattan distance, compute the Dunn Index for these two clusters. Also, analyze the quality of the clustering based on the value of the Dunn Index obtained.

b. In a binary classification task with 3 input features, a Bagging-based ensemble is configured with max_features = 2 and n_estimators = 3. Each base model has a 70% accuracy on the test set, and predictions are made via majority voting. 4

i. What is the maximum accuracy the ensemble can achieve under ideal conditions?

ii. What is the minimum accuracy the ensemble might achieve in the worst-case scenario?

Justify your answers.