JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST-3 EXAMINATION- 2026

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) Calculator is allowed.

Q. No.	Question	CO	Marks
Q1.	a. What does a subject-oriented data warehouse signify? Give an example. b. Compare the star, snowflake, and fact constellation schemas in terms of their efficiency in supporting analytical query processing. c. Compute the Cosine similarity for the vectors, x = (0, 1, 0, 1), y = (1, 0, 1, 0). d. Find the correlation coefficient for the vectors x = (0, 1, 1, 1), y = (1, 1, 1, 0). e. Find the Supremum distance between x= (22, 1, 42, 10), and y = (20, 0, 36, 8).	1,2,3	[5]
Q2.	a. What is a dendrogram in hierarchical clustering? How to get the optimal number of clusters using a dendrogram? b. Consider the following distance matrix for the data points P1, P2… P8. Label these as core, border and noise points to form clusters using DBSCAN algorithm. Take Epsilon = 3.5 and MinPts = 3.	6	[2] [4]

	P1	P2	P3	P4	P5	P6	P7	P8
P1	0.00							
P2	4.24	0.00						
P3	4.47	5.10	0.00					
P4	3.16	4.00	1.41	0.00				
P5	2.00	5.83	4.00	3.16	0.00			
P6	1.00	3.61	5.00	3.61	3.00	0.00		
P7	6.08	3.61	3.61	3.61	6.71	6.00	0.00	
P8	2.00	3.16	2.83	1.41	2.83	2.24	4.12	0.00

Q3.	a. Explain how AdaBoost updates the weights of training samples to improve model performance. b. Consider a multi-class classification problem with n classes. Let N_i denote the total number of actual instances belonging to class i, where $i = 1, 2, 3, \ldots, n$. For the i^{th} class, let $TP_i, TN_i, FP_i, and\ FN_i$ denote the numbers of True Positives, True Negatives, False Positives, and False Negatives, respectively. Using the One-vs-Rest approach, derive the mathematical expression for the Weighted Average Precision of the classifier.	4	[2] [3]

Q4.	a. Consider two clusters of 2-D data points: Cluster 1 containing (1,2), (2,3), and (3,3), and Cluster 2 containing (6,7), (7,8), and (8,8). Using Manhattan distance as the similarity measure, determine the Dunn Index for the given cluster formation. Based on the obtained value, critically examine whether the clusters are compact and well separated, and justify the effectiveness of the clustering solution. b. Suppose the data mining task is to cluster the following eight data points (where (x, y) represents location) into three clusters: P1(2, 8), P2(3, 6), P3(7, 7), P4(8, 5), P5(6, 3), P6(5, 4), P7(1, 2), P8(4, 9). Apply the k-Means algorithm to form three clusters. Use P1, P4, and P7 as the initial cluster centers. Compute the new cluster centers after the first iteration using Euclidean distance.	6	[3] [4]
Q5.	Construct the FP-Tree for the given transaction database and use it to identify all frequent itemsets. Assume the minimum support count is 3. **Transaction ID** | **Items** T1 | I1, I2, I3 T2 | I2, I3, I4 T3 | I4, I5 T4 | I1, I2, I4 T5 | I1, I2, I3, I5 T6 | I1, I2, I3, I4	5	[6]
Q6.	A supermarket manager wants to analyze customer purchasing patterns to improve product placement, combo offers, and cross-selling strategies. The following transaction database represents grocery items purchased by customers. Apply the Apriori algorithm to the given dataset using a minimum support threshold of 33.34% and a minimum confidence threshold of 60% to determine the frequent itemsets and generate association rules. Based on the discovered strong association rules, suggest suitable marketing strategies or product bundling recommendations for the supermarket. **Transaction ID** | **Items** T1 | HotDogs, Buns, Ketchup T2 | HotDogs, Buns T3 | HotDogs, Coke, Chips T4 | Chips, Coke T5 | Chips, Ketchup T6 | HotDogs, Coke, Chips	5	[6]