PVP 20

Code: 20IT4501E

III B.Tech - I Semester – Regular / Supplementary Examinations
NOVEMBER 2024

DATA MINING
(INFORMATION TECHNOLOGY)

Duration: 3 hours Max. Marks: 70

Note: 1. This paper contains questions from 5 units of Syllabus. Each unit carries 14 marks and have an internal choice of Questions.
2. All parts of Question must be answered in one place.
BL – Blooms Level CO – Course Outcome

			BL	CO	Max. Marks
UNIT-I					
1	Explain the different types of data that can be mined in data mining. Discuss in detail the various patterns that can be discovered through data mining techniques. Include examples for each type of pattern to illustrate their practical applications.		L2	CO1	14 M
OR					
2	Discuss the major technologies used in data mining and how they contribute to the process of extracting valuable information from large datasets. Additionally, analyze the key applications of data mining across different industries and the major issues and challenges faced in the field.		L2	CO1	14 M
UNIT-II					
3	Analyze the different methods used to measure data similarity and dissimilarity. Discuss the		L3	CO2	14 M

		significance of these measures in tasks like clustering, and compare techniques such as Euclidean distance, Manhattan distance, cosine similarity and Jaccard similarity with examples.			
OR					
4	a)	Given the following two data vectors representing two documents A and B in a term-document matrix: Document A: [3,0,2,5] Document B: [1,1,3,2] Calculate the cosine similarity between Document A and Document B. Interpret the result in terms of the similarity between the two documents.	L3	CO2	4 M
	b)	Illustrate major tasks in data pre-processing and discuss issues to consider during data integration.	L3	CO2	10 M
UNIT-III					
5		Given the following transaction dataset, use the Apriori algorithm to find all frequent itemsets with a minimum support threshold of 50%. Then, generate all possible association rules from these frequent itemsets with a minimum confidence threshold of 60%.	L3	CO3	14 M

Transaction ID	Items Purchased
1	Bread, Milk, Butter
2	Bread, Milk, Cheese
3	Milk, Cheese
4	Bread, Butter
5	Milk, Butter, Cheese
6	Bread, Milk, Butter, Cheese

OR

6	Discuss the concepts of frequent pattern mining, association rule mining, and pattern growth approaches. Explain how the Apriori algorithm and the FP-Growth algorithm differ in their methodology for finding frequent itemsets. Additionally, analyze the strengths and weaknesses of both algorithms and under what circumstances one might be preferred over the other.	L3	CO3	14 M
	UNIT-IV			
7	Discuss the concept of decision tree induction in classification. How does the decision tree learning process work, and what are the key criteria used for splitting the data at each node? Additionally, mention the strengths and weaknesses of decision tree classifiers.	L3	CO3	14 M
	OR			
8	Given the following dataset, construct a decision tree using the Information Gain criterion. Show each step of the process, including the calculation of entropy and information gain at each node. ID Weather Temperature Humidity Windy Play 1 Sunny Hot High False No 2 Sunny Hot High True No 3 Overcast Hot High False Yes 4 Rainy Mild High False Yes 5 Rainy Cool Normal False Yes 6 Rainy Cool Normal True No 7 Overcast Cool Normal True Yes 8 Sunny Mild High False No 9 Sunny Cool Normal False Yes 10 Rainy Mild Normal False Yes	L3	CO3	14 M

	11	Sunny	Mild	Normal	True	Yes			
	12	Overcast	Mild	High	True	Yes			
	13	Overcast	Hot	Normal	False	Yes			
	14	Rainy	Mild	High	True	No			
UNIT-V									
9	Illustrate the principles and differences between partitioning methods and hierarchical methods in cluster analysis. How does the K-means algorithm work, and what are its strengths and weaknesses compared to hierarchical clustering methods such as agglomerative clustering?						L3	CO3	14 M
OR									
10	Given the following data points, perform K-means clustering with k=3. Show each step of the algorithm, including the initial assignment of centroids, the assignment of data points to clusters, the recalculation of centroids, and the final clusters. Data Point X Y A 1 2 B 1 4 C 3 2 D 5 8 E 6 6 F 8 8 G 7 6 H 9 7						L3	CO3	14 M