QP Code: RN22BTECH249 Reg. No | | | | | | | | | | | AR 22

Gandhi Institute of Engineering and Technology University, Odisha, Gunupur (GIET University)

B. Tech (Fifth Semester - Regular) Examinations, November – 2024

22BCSPC35002/22BCDPC35002-Data Mining and Data Warehousing

(CSE, CSE(DS))

Time: 3 hrs Maximum: 70 Marks

(The figures in the right hand margin indicate marks)

PART – A **(2 x 5 = 10 Marks)**

		CO #	Blooms Level
Q.1.	Answer ***ALL*** questions		
a.	Let c be a candidate itemset in C_k generated by the Apriori algorithm. How many length-(k-1)subsets do we need to check in the prune step ?	CO2	K6
b.	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30,33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Using this above data answer the following question Use min-max normalization to transform the value 35 for age on to the range [0.0,1.0]	CO1	K4
c.	Define support and confidence in Association Rule Mining.	CO3	K2
d.	Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?	CO3	K6
e.	Explain the principle of hierarchical clustering.	CO4	K3

PART – B **(15 x 4=60 Marks)**

<u>Answer ***ALL*** the questions</u> Marks CO # Blooms Level

2. a. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results — 8 CO1 K2

age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
% fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

i. Calculate the median, and standard deviation of age and %fat.

ii. Draw the box plots for age and identify outlier.

iii. Draw a quantile plot based on these two variables

b. Suppose that a data warehouse for GIET University consists of the four dimensions, *student, course, semester, and instructor* and two measures *count* and *avg_grade*. At the lowest conceptual level(e.g., for a given *student, course, semester, and instructor combination*), the *avg_grade* measure stores the actual course grade of a student. At higher conceptual levels, *avg_grade* stores the average grade for the given combination. — 7 CO2 K1 K3 K5

i. Draw a snowflake schema diagram for the data warehouse.

ii. Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of CS courses for each GIET University student.

(OR)

c. What are the different types of data warehouse architecture? Explain Three-tier architecture of data warehouse with suitable example. — 8 CO2 K1 K3

d. Demonstrate computation of the following measures for similarity/dissimilarity among data: 7 CO1/CO4 K1 K2

i. Cosine measure
ii. Euclidean distance
iii. Manhattan measure.

3.a. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30,33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. 8 CO3 K3 K5

i. Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.
ii. What other methods are there for data smoothing?

b. Explain how Support-Confidence Rule is sometimes misleading with example. Define the two correlation measures. Consider given contingency table of sales transactions for computer games and videos and find correlation using **lift and χ^2** : 7 CO3 K3 K5

	game	$\overline{game}$
video	4000 (4500)	3500 (3000)
$\overline{video}$	2000 (1500)	500 (1000)

(OR)

c. Consider a transactional database where I1, I2, I3, I4, I5, I6, I7 are the different items and T1,T2,T3,T4,T5 are the transactions. 8 CO3 K3 K5 K6

T1 {I1, I2, I3, I5},T2 {I1, I2, I3, I4, I5}, T3 {I1, I2, I3, I7}, T4 {I1, I3, I6},
T5 {I1, I2, I4, I5, I6}.

Suppose the minimum support is 60% and minimum confidence = 50%. Find all frequent itemsets and the association rules using Apriori Algorithm.

d. Given two objects represented by the tuples (22,1,42,10) and (20,0,36,8). Compute Euclidean distance, Manhatton distance and Minkowski distance(q=3) between these two given objects. 7 CO1/CO4 K2,K3 K4

4.a. i. "Closed frequent itemset is best choice for containing complete information regarding frequent itemsets than maximal itemset." - Justify the statement with appropriate example. **(4 Marks)** (4+4) CO3 K1,K3,K4

ii. Given the following contingency database:

TID	T1	T2	T3	T4	T5
Items	M, O, N, K, E, Y	D, O, N, K, E, Y	M, A, K, E	M, U, C, K. Y	C, O, K, I, E

Calculate the relative support and confidence for 2-itemset frequent itemset of the given database. **(4 Marks)**

b. Why is naive Bayesian classification called "naive"? Briefly outline the major ideas of naive Bayesian classification. Using the above data set and Naive-Bayes classification identify the species of an entity with the following attributes. 7 CO3 K2,K3,K5

X={Color=Green, Legs=2, Height=Tall, Smelly=No}

Sl. No.	**Color**	**Legs**	**Height**	**Smelly**	**Species**
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

(OR)

c. What does splitting criterion mean in decision tree induction? 8 CO3 K2,K3,K5

For the following Medical Diagnosis Data, create a decision tree.

Sore Throat	**Fever**	**Swollen Glands**	**Congestion**	**Headache**	**Diagnosis**
YES	YES	YES	YES	YES	**Strep Throat**
NO	NO	NO	YES	YES	Allergy
YES	YES	NO	YES	NO	Cold
YES	NO	YES	NO	NO	Strep Throat
NO	YES	NO	YES	NO	Cold
NO	NO	NO	YES	NO	Allergy
NO	NO	YES	NO	NO	Strep Throat
YES	NO	NO	YES	YES	Allergy
NO	YES	NO	YES	YES	Cold
YES	YES	NO	YES	YES	Cold

d. Determine the equation of hyperplane that divides the data points into two classes Positively labelled data points (3,1)(3,-1)(6,1)(6,-1) and Negatively labelled data points (1,0)(0,1)(0,-1)(-1,0) 7 CO3 K3,K5

5.a. 8 CO4 K1,K3,K5

Object Identifier	Test-1 (categorical)	Test-2 (ordinal)	Test-3 (ratio-scaled)
1	Code-A	Excellent	445
2	Code-B	Fair	22
3	Code-C	Good	164
4	Code-A	Excellent	1210

Consider the above dataset and calculate dissimilarity matrix for Test-1, Test-2 and Test-3.

b. Use K-means algorithm and Euclidean distance to cluster the following 10 points into 3 clusters 7 CO4 K3,K5

A1(2,10),A2(9,4),A3(8,4),A4(9,4),A5(5,8),A6(7,5),A7(6,4),A8(1,2),A9(4,9),A10(6,10).

Suppose the initial centers are A1,A4 and A9,run the K-means algorithm for 3 iterations at the end of each iterations show the cluster centers and also the final three clusters.

(OR)

c. Define medoid. Create two clusters for given data set using k-medoid clustering for the given data set. Let, the initial cluster medoids are C1 -(4, 5) and C2 -(8, 5) respectively. Execute the process upto 2 iterations or when the stopping condition is met, which one is earlier. 8 CO4 K1,K4,K5

Id	0	1	2	3	4	5	6	7	8	9
X	8	3	4	9	8	5	7	8	7	4
Y	7	7	9	6	5	8	3	4	5	5

d. What do you mean by clustering? Apply Agglomerative nesting clustering on the given distance matrix. 7 CO4 K3,K5

Draw single linkage and max linkage dendogram for representing the cluster.

	P1	P2	P3	P4	P5
P1	0				
P2	9	0			
P3	3	7	0		
P4	6	5	9	0	
P5	12	9	2	8	0

--- End of Paper ---