Reg.No.:												

VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN
[AUTONOMOUS INSTITUTION AFFILIATED TO ANNA UNIVERSITY, CHENNAI]
Elayampalayam – 637 205, Tiruchengode, Namakkal Dt., Tamil Nadu.

Question Paper Code: 5008

B.E. / B.Tech. DEGREE END-SEMESTER EXAMINATIONS – MAY / JUNE 2024
Fifth Semester
Computer Science and Engineering
U19CSV31 – DATA WAREHOUSING AND DATA MINING
(Regulation 2019)

Time : Three Hours   Maximum : 100 Marks

Answer ALL the questions

Knowledge Levels (KL)	K1 – Remembering	K3 – Applying	K5 - Evaluating
	K2 – Understanding	K4 – Analyzing	K6 - Creating

PART – A

(10 x 2 = 20 Marks)

Q.No.	Questions	Marks	KL	CO
1.	How is a data warehouse different from a database? How are they similar?	2	K2	CO1
2.	Differentiate technical meta data and business meta data.	2	K2	CO1
3.	Write down the schemas used for multidimensional databases?	2	K2	CO2
4.	Define MOLAP and write its advantages and disadvantages.	2	K1	CO2
5.	Define KDD process and list the steps involved in the process?	2	K2	CO3
6.	How to handle Noisy Data?	2	K2	CO3
7.	With suitable example define monotone & anti – monotone property.	2	K2	CO4
8.	How the association rules are mined from large databases?	2	K2	CO4
9.	Compute the Euclidean distance between the two objects for the given tuples (22, 1, 42, 10) and (20, 0, 36, 8)	2	K3	CO5
10.	Write down the techniques used for outliers detection.	2	K2	CO5

PART – B

(5 x 13 = 65 Marks)

Q.No.		Questions	Marks	KL	CO
11.	a)	Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.			
		i. Enumerate three classes of schemas that are popularly used for modeling data warehouses.	3	K3	CO1
		ii. Draw a schema diagram for the above data warehouse using one of the schema classes listed in (a).	3		
		iii. Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?	4		
		iv. To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema (day, month, year, doctor, hospital, patient, count, charge).	3		
		(OR)			
	b)	Diagrammatically illustrate and explain the three tier data warehousing architecture.	13	K2	CO1
12.	a)	Explain the following in OLAP:			
		i. Roll up operation	3	K2	CO2
		ii. Drill Down operation	3		
		iii. Slice operation	3		
		iv. Dice operation	2		
		v. Pivot operation	2		
		(OR)			
	b)	List and discuss the basic features that are provided by reporting and query tools used for business analysis.	13	K2	CO2
13.	a)	Suppose a group of 1,500 people was surveyed. The gender of each person was noted. Each person was polled as to whether their preferred type of reading material was fiction or nonfiction. So they, have two attributes, gender and preferred reading. How the gender and preferred Reading are correlated?	13	K3	CO3

	male	female	Total
fiction	250 (90)	200 (360)	450
non -fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

(OR)

b) i. Write short notes on the following: 8 K2 CO3

a. No coupling

b. Loose coupling

c. Semitight coupling

d. tight coupling

ii. Explain in detail about Normalization in Data Transformation method with an example. 5

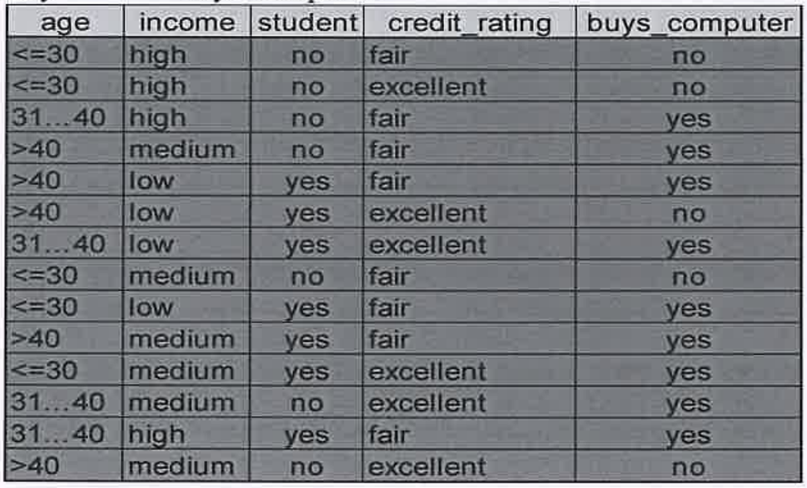
14. a) A database has five transactions. Let min sup = 60% and min conf = 80%. 13 K3 CO4

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

Find all frequent itemsets using Apriori algorithm.

(OR)

b) For the given dataset apply the Decision Tree classification to classify the label buys-computer: 13 K3 CO4



age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

15. a) Describe each of the following clustering algorithms in terms of the following criteria: (i) shapes of clusters that can be determined; (ii) input parameters that must be specified; and (iii) limitations. K2 CO5

(a) CLARA 3

(b) BIRCH 3

(c) ROCK 3

(d) Chameleon 2

(e) DBSCAN 2

(OR)

b) Cluster the following eight points (with (x, y) representing locations) into three clusters:
A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9).
The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1as the center of each cluster, respectively. Use the k-means algorithm to identify
(a) The three cluster centers after the first round execution 10 K2 CO5
(b) The final three clusters 3

PART – C

(1 x 15 = 15Marks)

Q.No.		Questions	Marks	KL	CO
16.	a)	The support vector machine (SVM) is a highly accurate classification method. However, SVM classifiers suffer from slow processing when training with a large set of data tuples. Discuss how to overcome this difficulty and develop a scalable SVM algorithm for efficient SVM classification in large datasets.	15	K3	CO4
		(OR)			
	b)	Consider the following data and divide the data into two clusters, i.e., k=2 Use K-Medoid Algorithm and device the cluster.	15	K3	CO5

S.No	X	Y
1	9	6
2	10	4
3	4	4
4	5	8
5	3	8
6	2	5
7	8	5
8	4	6
9	8	4
10	9	3