Q. Code: 515873

Reg. No. | | | | | | | | | | | | | |

B.E. / B.TECH. DEGREE EXAMINATIONS, MAY 2024

Fifth Semester

CS18502 – DATA MINING AND DATA WAREHOUSING

(Computer Science and Engineering)

(Regulation 2018/2018A)

TIME: 3 HOURS **MAX. MARKS: 100**

COURSE OUTCOMES	STATEMENT	RBT LEVEL
CO 1	Students will be able to understand data warehouse concepts, architecture, business analysis and tools.	**2**
CO 2	Students will be able to understand data pre-processing and data visualization techniques.	**2**
CO 3	Students will be able to study algorithms for finding hidden and interesting patterns in data using association algorithms.	**3**
CO 4	Students will be able to apply various classification and clustering techniques using tools.	**3**
CO 5	Students will be mastering the data mining techniques in various applications like social, scientific and environmental context.	**3**

PART- A (10 x 2 = 20 Marks)

(Answer all Questions)

		CO	RBT LEVEL
1.	Sketch the recommended approach for data warehouse development.	**1**	**2**
2.	What are OLAP and OLTP systems?	**1**	**2**
3.	Identify four real time applications in data mining.	**5**	**3**
4.	List out the major tasks in data preprocessing.	**2**	**2**
5.	Determine the closed and maximal frequent item sets from the given set of transactions.	**3**	**3**

Transaction ID	Items
T1	A, B, C, D
T2	A, B, C
T3	A, C, D

6.	What is Association Rule Mining in Market Basket Analysis?	**3**	**2**
7.	Give few applications where decision tree classifiers are popular.	**4**	**3**
8.	Differentiate classification and clustering.	**4**	**2**
9.	List the various clustering methods.	**4**	**2**
10.	Name the various distance measures used in clustering?	**4**	**2**

PART- B (5 x 14 = 70 Marks)

			Marks	CO	RBT LEVEL
11. (a)		With proper sketch, explain in detail about the multitier architecture of a data warehouse.	**(14)**	**1**	**2**
		(OR)			
(b)		With proper sketch, explain in detail about the schemas for multidimensional data models.	**(14)**	**1**	**2**
12. (a)		List the various kinds of patterns can be mined using data mining technology. Can such patterns be generated alternatively by data query processing or simple statistical analysis?	**(14)**	**2**	**2**
		(OR)			
(b)	**(i)**	Briefly outline how to compute the dissimilarity between objects described by the following: (a) Nominal attributes (b) Ordinal attributes (c) Numeric attributes (d) Term Frequency Vectors	**(8)**	**2**	**2**
	(ii)	With suitable example, construct a flowchart to summarize the procedure for backward elimination attribute subset selection.	**(6)**	**2**	**2**
13. (a)	**(i)**	A database consists of five transactions. Consider minimum support to be 3. Identify the frequent itemsets using FP growth algorithm and generate strong association rules with the minimum confidence of 60%.	**(14)**	**3**	**3**

Transaction ID	Items
T1	{E,K,M,N,O,Y}
T2	{D,E,K,N,O,Y}
T3	{A,E,K,M}
T4	{C,K,M,U,Y}
T5	{C,E,I,K,O,O}

(OR)

	(ii)	A database consists of nine transactions. Consider minimum support to be 2. Identify the frequent itemsets using apriori algorithm.	**(14)**	**3**	**3**

Transaction ID	Items
T1	A, B
T2	B, D

T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

14. (a) Estimate conditional probabilities of each attributes {color, legs, height, smelly} for the Species classes: {M, H} for the following data. **(14) 4 3**

S. 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

Use Naive Bayesian classifier, classify the new instance with {Color=Green, Legs=2, Height=Tall and Smelly=No}.

(OR)

(b) For the given data, predict the class of a person with height =170 cm and weight =57 kg using KNN classifier. Take K=5. **(14) 4 3**

Height (cm)	**Weight (kg)**	**Class**
167	51	Underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Underweight
169	58	Normal
173	57	Normal
170	55	Normal

15. (a) For the given data, compute two clusters using K-means algorithm for clustering where initial cluster centers are (1.0, 1.0) and (5.0, 7.0). Execute for two iterations. **(14) 4 3**

Record Number	**A**	**B**
R1	1.0	1.0
R2	1.5	2.0

R3	3.0	4.0
R4	5.0	7.0
R5	3.5	5.0
R6	4.5	5.0
R7	3.5	4.5

(OR)

(b) Use the distance matrix in the below table to perform agglomerative hierarchical clustering for single linkage and complete linkage. Show your results by drawing a dendogram. The dendogram should clearly show the order in which the points are merged. **(14) 4 3**

	P1	**P2**	**P3**	**P4**	**P5**
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

PART- C (1 x 10 = 10 Marks)

(Q.No.16 is compulsory)

		Marks	**CO**	**RBT LEVEL**
16.	A database has four transactions.	**(10)**	**3**	**3**

Transaction ID	Items bought
T100	{A, C, D, F, G, I, M, P}
T200	{A, B, C, F, I, M, O}
T300	{B, F, H, J, O}
T400	{B, C, K, S, P}
T500	{A, C, E, F, L, M, N, P}
T600	{B, D, L, H, E, G}
T700	{F, G, H, I, L, M}

(a) Find all frequent itemsets using ECLAT algorithm.

(b) List all the strong association rules that satisfies minimum support and minimum confidence.

Let min_sup=70% and min_conf=80%.

****