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

B.E./ B. TECH.DEGREE EXAMINATIONS, MAY 2023

Fifth Semester

CS18502 – DATAMINING AND DATA WAREHOUSING

(Computer Science and Engineering)

(Regulation2018)

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 (10x2=20Marks)

(Answer all Questions)

		CO	RBT LEVEL
1.	Define Data Warehousing.	1	1
2.	Write the differences between Data warehouse, Data mart and Data lake.	1	2
3.	Compare and Contrast Bitmap and Join Indexing.	2	4
4.	Explain various data smoothing techniques.	2	2
5.	State Apriori property in frequent pattern mining.	3	1
6.	Point out the steps involved in association rule mining.	3	2
7.	List out the metrics for evaluating classifier performance.	4	2
8.	What is Silhouette coefficient?	4	1
9.	When would you prefer to use k-medoids than k-means clustering algorithm?	5	3
10.	What is Grid-based clustering method? List out various grid based and subspace clustering algorithms.	5	1

PART- B (5x 14=70Marks)

		Marks	CO	RBT LEVEL
11. (a)	Draw and Explain in detail about the data warehouse multitier architecture.	(14)	1	3
	(OR)			
(b)	Briefly explain about typical OLAP operations on multidimensional data with an example.	(14)	1	3

12. (a)	Compare and Contrast the OLAP server Architecture (ROLAP,MOLAP,HOLAP) and Evaluate the proximity measure for the following table (Nominal, Ordinal, Numeric)	(14)	2	4

S.NO	*TEST I*	*TEST II*	*TEST III*
1	Code A	Excellent	45
2	Code B	Fair	22
3	Code C	Good	64
4	Code A	Excellent	28

(OR)

(b)	Compare all six data transformation strategies with an example.	(14)	2	4
13. (a)	Explain the procedure to mine frequent itemsets efficiently using vertical data format. Apply ECLAT to mine frequent itemsets using the nine transactions given in the table. Consider minimum support as 2.	(14)	3	3

Transaction ID	Items
T1	{Bread, Butter, Jam}
T2	{Butter, Coke}
T3	{Butter, Milk}
T4	{Bread, Butter, Coke}
T5	{Bread, Milk}
T6	{Butter, Milk}
T7	{Bread, Milk}
T8	{Bread, Butter, Milk, Jam}
T9	{Bread, Butter, Milk}

(OR)

(b)	Which Patterns Are Interesting? Explain various Pattern Evaluation Methods in detail with examples.	(14)	3	3
14. (a)	Explain the Naive Bayesian Classifier in detail.	(14)	4	3

Consider the following training dataset.

S. No	Swim	Fly	Crawl	Class Label
1	Fast	No	No	Fish
2	Fast	No	Yes	Animal
3	Slow	No	No	Animal
4	Fast	No	No	Animal
5	No	Short	No	Bird
6	No	Short	No	Bird
7	No	Rarely	No	Animal
8	Slow	No	Yes	Animal
9	Slow	No	No	Fish
10	Slow	No	Yes	Fish
11	No	Long	No	Bird
12	Fast	No	No	Bird

Let X = (Slow, Short, Yes). Identify the class label that is more appropriate for X using Naive Bayesian Classifier.

(OR)

(b) Find the optimal separating hyperplane for the following data points using SVM algorithm. **(14) 4 3**

Positive labelled data points: {(3,1), (3,-1), (6,1),(6,-1)}
Negative labelled data points: {(1,0), (0,1), (0,-1),(-1,0)}

15. (a) Suppose that the data mining task is to cluster points (with *x*, *y* representing location) into three clusters, where the points are **(14) 5 3**

***A*1.(2,10), *A*2.(2,5), *A*3.(8,4), *A*4.(5,8), *A*5.(7,5), *A*6.(6,4), *A*7.(1,2), *A*8.(4,9).**

The distance function is Euclidean distance. Suppose initially we assign *A*1, *A4*, and *A*7 as the centre of each cluster, respectively. Use the *k-means* algorithm to show, The three cluster centres after the first round of execution, The final three clusters and write the K-Means Algorithm

(OR)

(b) Perform DBSCAN for the below mentioned table with €=2 and min points=2 and also find the core points, outliers and clusters. **(14) 5 3**

	X	Y
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5

PART- C (1x 10=10Marks)

(Q.No.16 is compulsory)

	Marks	CO	RBT LEVEL

16. Evaluate the agglomerative hierarchical clustering algorithm to build the dendrogram for the following dataset. And discuss in detail about Agglomerative Hierarchical Clustering with its pros and cons. **(10) 4 5**

	A	B	C	D	E
A	**0**	**9**	**3**	**6**	**11**
B	**9**	**0**	**7**	**5**	**10**
C	**3**	**7**	**0**	**9**	**2**
D	**6**	**5**	**9**	**0**	**8**
E	**11**	**10**	**2**	**8**	**0**

****