Total number of printed pages: 25 **Programme(D)/6th/DCSE612**

2023

DATA MINING

Full Marks : 100

Time : Three hours

The figures in the margin indicate full marks for the questions.

*Answer **any five** questions.*

1.	**Answer the following questions:**		
	a)	**Match the following "Clustering approach" and "Clustering Methods":**	1x6=6

CLUSTERING APPROACH	CLUSTERING METHODS
Hierarchical approach	DBSCAN
Hierarchical approach	k-means
Density approach	Diana
Density approach	k-medoids
Partitioning approach	Agnes
Partitioning approach	OPTICS

	b)	**True or False:**	1x8=8
		(i) Database uses current data to take the decision. **(ii)** OLTP is a major task of DBMS. **(iii)** OLAP didn't process data analysis and decision making. **(iv)** The less detailed to more detailed data is retrieve in Roll-up operation. **(v)** High inter class similarity is not found in good clustering. **(vi)** High intra class similarity is found in good clustering. **(vii)** Ordinal variables cannot be continuous only. **(viii)** The summarized part of the cube is called a apex cuboid.	
	c)	Whether or not each of the following activities is a data mining task (**Yes/No**)	1x6=6
		(i) Dividing the customers of a company according to their gender. **(ii)** Dividing the customers of a company according to their profitability. **(iii)** Computing the total sales of a company. **(iv)** Sorting a student database based on student identification numbers. **(v)** Predicting the outcomes of tossing a (fair) pair of dice. **(vi)** Predicting the future stock price of a company using historical records.	
2.	**a)**	What is the need of data cleaning and data reduction in data-preprocessing?	8
	b)	Apply the Min-max normalization by setting min = 0 and max = 1 on the following group of data values 80, 150, 260, 360, 450, and 500.	6
	c)	Given price values (in pound) such as 28, 24, 4, 44, 18, 36, 16, 8, 45, 20, 30, 12, 32, 48, 38 and 40. Apply the all binning methods to partition the data.	6

3.	**a)**	What is data warehouse?	2
	b)	Discuss the Data matrix and Dissimilarity matrix.	4
	c)	Explain the Binary, Nominal and Ordinal variables.	6
	d)	Find out the three clusters using the k-means algorithm for the given data objects {2, 4, 10, 12, 3, 20, 30, 11, 25}. (Hint k=3 and keep continue till the no change in the formed clusters)	8
4.	**a)**	Explain the classification and prediction methods?	5
	b)	Describe the Frequent itemsets, Support and Confidence in Association Rule Mining?	7
	c)	A database has five transactions. Let min_sup = 60%. Find all frequent item sets using Apriori.	8

TID	Item sets
T100	{D, G, A, O, N}
T200	{B, G, A, O, N }
T300	{K, G, A, O, N }
T400	{J, G, A, O, N }

5.	Write short notes on the following (***any four***):		4x5=20
	a)	Data transformation	
	b)	Data Reduction	
	c)	Expected Information (I) for attribute selection measure	
	d)	Bayesian classifier with Bayes theorem (including formula)	
	e)	KDD process	
6.	Differentiate between the following (***any four***):		4x5=20
	a)	Unsupervised and Supervised learning	
	b)	Roll up and drill down	
	c)	Slice and Dice operation	
	d)	DIANA and AGNES	
	e)	Parametric and non-parametric methods in numerosity reduction	