Total number of printed pages: 35 **Programme(D)/6th/DCSE612**

2024

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 "Pre-Processing approach" with their "tasks":**	1x6=6
		Pre-Processing approach / **Tasks** Data Cleaning, Data Transformation, Data Reduction / Data Compression, To fill the missing values, To handle high dimensionality, Concept hierarchy, To resolve inconsistencies, Normalization	
	b)	**True or False:**	1x8=8
		(i) Missing data may be due to equipment malfunction. **(ii)** Binning is used to handle noisy data. **(iii)** K-means handles outlier well. **(iv)** AGNES is a top down approach. **(v)** In DIANA iteratively clusters are merged together. **(vi)** STING is a grid based clustering approach. **(vii)** There will an arbitrary shape of cluster in density based clustering. **(viii)** The more detailed to less detailed data is retrieve in drill down operation.	
	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)** Finding the name of patients having blood pressure in database. **(iii)** Handwritten Digit Recognition. **(iv)** Fraudulent and Abusive Data. **(v)** Anomaly detection. **(vi)** Predicting the Covid-19 behavior.	
2.	**a)**	What is the need of data cleaning and data reduction in data-preprocessing?	4
	b)	Apply the Min-max normalization by setting min = 0 and max = 1 on the following group of data values 60, 100, 200, 300, 400, and 500.	8
	c)	Find out the three clusters using the k-medoids algorithm for the given data objects {4, 8, 12, 18, 3, 22, 32, 10, 24}. (Hint k=3 and keep continue till the no change in the formed clusters)	8
3.	**a)**	What are the weakness and strength of Neural Network?	4
	b)	How a multi-layer neural network works?	4

	c)	Apply the Bayesian classification for predicting the buys_comp of the given test sample data, x= ((31…40, HIGH, Y, EXCELLENT)	12

Age	Income	Student	Credit_rating	Class:Buys_comp
<=30	HIGH	N	FAIR	N
<=30	HIGH	N	EXCELLENT	N
31.....40	HIGH	N	FAIR	Y
>40	MEDIUM	N	FAIR	Y
>40	LOW	Y	FAIR	Y
>40	LOW	Y	EXCELLENT	N
31.....40	LOW	Y	EXCELLENT	Y
<=30	MEDIUM	N	FAIR	N
<=30	LOW	Y	FAIR	Y
>40	MEDIUM	Y	FAIR	Y
<=30	MEDIUM	Y	EXCELLENT	Y
31....40	MEDIUM	N	EXCELLENT	Y
31....40	HIGH	Y	FAIR	Y
>40	MEDIUM	N	EXCELLENT	N

4.	**a)**	What do you mean by Association rules?	3
	b)	What are the types of Association rules?	3
	c)	Describe the frequent itemsets.	2
	d)	A database has five transactions. Let minimum support is 50%. Find all frequent item sets using Apriori algorithm.	12

TID	Item sets
T100	{1, 2, 3, 4}
T200	{1, 2, 3, 5, 6, }
T300	{2, 3, 4, 5 }
T400	{3, 4, 5, 6 }

5.	Write short notes on the following (***any four***):		4x5=20
	a)	Decision Tree	
	b)	Data Warehousing	
	c)	Dendogram	
	d)	Data Cube	
	e)	OLAP operations	
6.	Differentiate between the following (***any four***):		4x5=20
	a)	Classification and Clustering	
	b)	Data Matrix and Dissimilarity matrix	
	c)	Support and Confidence (ARs)	
	d)	OLAP and OLTP	
	e)	Binary variables and Nominal variables	