Total number of printed pages: 02 **Programme(D)/6th/DCSE612**

2025

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 "Pre-Processing approach" with the "To handle the challenges":**	1x6=6

Pre-Processing approach	To handle the challenges
Data Cleaning, Data Transformation, Data Reduction	Noisy, Missing values, Different attributes, Same value expressed differently, Huge amount of data, Range of attributes

	b)	**True or False:**	1x14=14
		(i) Missing data can be handled by data reduction. **(ii)** Discrepancies in codes or names due to inconsistency. **(iii)** Binning is used for data smoothing. **(iv)** K-medoids cannot handle the outlier. **(v)** Nominal variable cannot take more than two states. **(vi)** Ordinal variables can be continuous. **(vii)** AGNES is a bottom up approach. **(viii)** Divisive approach iteratively merged together the clusters. **(ix)** PAM is efficient for small datasets. **(x)** OPTICS is a grid based clustering. **(xi)** There will be fixed shape of cluster in density based clustering. **(xii)** Predicting the Covid-19 behavior is not a data mining task. **(xiii)** Replacing the data by smaller representation in data reduction. **(xiv)** Removing the irrelevant attributes in data normalization.	
2.	**a)**	Why data cleaning approach is required in the data-preprocessing?	2
	b)	Apply the Min-max normalization by setting min=0 and max=1 on the following group of data values 50, 150, 250, 350, 450 and 500.	7
	c)	What are the weakness and strength of K-means clustering?	4
	d)	Find out the two clusters using the k-means algorithm for the given data objects {4, 6, 10, 12, 14, 16, 20, 24, 30, 32, 36, 38}. (Hint: k=2)	7
3.	**a)**	What is a decision tree?	2
	b)	What is process to do classification based on the decision tree?	3
	b)	What is backpropagation neural network algorithm?	3

	c)	Apply the Bayesian classification for predicting the buys_comp of the given test sample data, x= ((>40, MEDIUM, N, FAIR) **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	12
4.	**a)**	What do you understand by association rules?	3
	b)	Explain the support and confidence with the formula.	6
	c)	How to differentiate the itemset and frequent itemset?	3
	d)	A database has five transactions. Let minimum support is 50%. Find all frequent item sets using Apriori algorithm. TID | Item sets T100 | {1, 3, 4} T200 | {1, 3, 4, 6, } T300 | {3, 4, 5 } T400 | {1, 2, 3, 4, 6 }	8
5.	Write short notes on the following (***any four***):		4x5=20
	a)	Dissimilarity matrix	
	b)	Multilayer Neural Network	
	c)	Dendogram	
	d)	data warehouse modeling based on star schema	
	e)	OLAP operations	
6.	Differentiate between the following (***any four***):		4x5=20
	a)	Model based Clustering and Grid based Clustering	
	b)	Data normalization and data transformation	
	c)	BIRCH and DIANA	
	d)	OLTP and OLAP	
	e)	Nominal variables and Categorical variables	