Total number of printed pages: 02 **Programme(PG)/2nd/MCS214**

2025

DATA MINING and DATA WAREHOUSING

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) A data warehouse is based on a multidimensional data model. **(ii)** Different naming conventions in different sources lead to inconsistency. **(iii)** Clustering is used for data smoothing. **(iv)** K-means can handle the outliers. **(v)** Nominal variable can take more than two states. **(vi)** Ordinal variables can be continuous. **(vii)** AGNES is a Top down approach. **(viii)** Agglomerative approach iteratively merged together the clusters. **(ix)** PAM is efficient for large datasets. **(x)** OPTICS is a model based clustering. **(xi)** Density based clustering discover the arbitrary shape of cluster. **(xii)** Predicting the Covid-19 behavior is a data mining task. **(xiii)** Replacing the data by smaller representation in data reduction. **(xiv)** Removing the irrelevant attributes in data transformation.	
2.	**a)**	What are the data transformation methods?	4
	b)	Explain the KDD process with a diagram in details.	5
	c)	Apply the Z-score normalization on the following values 5, 10, 20, 30, 40 and 50 of attribute.	6
	d)	Discuss the OLAP operations in details.	5
3.	**a)**	What are the non-parametric methods in the numerosity reduction?	3
	b)	Find out the two clusters using the k-medoids algorithm for the given data objects {(2,4) (2,5) (3,6) (3,8) (4,5) (4,6)}. (Hint: k=2)	5
	c)	What do you mean by the good clustering?	2

d) Apply the Bayesian classification for predicting the buys_comp of the given test sample data, x= (31..40, MEDIUM, N, EXCELLENT) — 10

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 is association rule mining? — 2

b) What is the role of support and confidence in the association rule? — 4

c) Write down the Apriori Algorithm. — 6

d) Generate the frequent itemsets using the Apriori Algorithm with min support ≥ 50% and confidence ≥ 80%. — 8

TID	date	items_bought
T100	10/15/99	{K, A, D, B}
T200	10/15/99	{D, A, C, E, B}
T300	10/19/99	{C, A, B, E}
T400	10/22/99	{B, A, D}

5. Write short notes on the following (***any four***): — 4x5=20

a) Asymmetric and Symmetric binary variables

b) OPTICS

c) Confusion matrix

d) Information Gain

e) Overfitting

6. Differentiate between the following (***any four***): — 4x5=20

a) STING vs CLIQUE

b) Multi-layer neural network vs Backpropagation

c) Pre-pruning vs post-pruning

d) Lazy Learning vs Eager Learning

e) Star schema vs Snowflake Schema