B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

June / July 2025 Semester End Main Examinations

Programme: B.E.　　　　　　　　　　　　　　　　　　　**Semester: V**

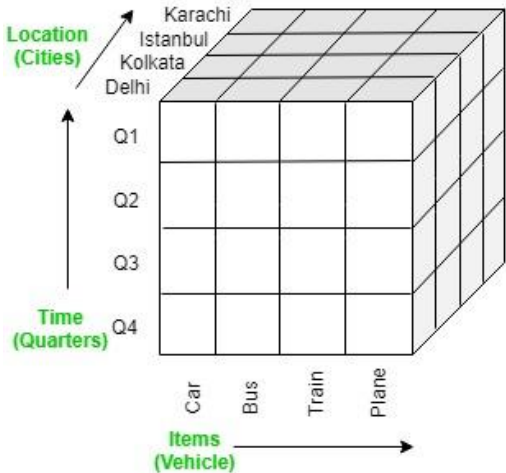**Branch: Artificial Intelligence and Machine Learning**　　**Duration: 3 hrs.**

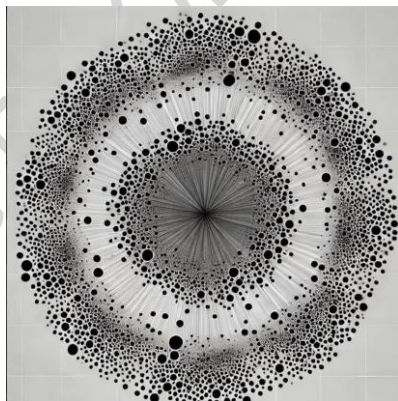**Course Code: 24AM5PEKDD**　　　　　　　　　　　　　**Max Marks: 100**

Course: Knowledge Discovery

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
　　　　　　　2. Missing data, if any, may be suitably assumed.

<table>
<tr><td colspan="3" align="center">UNIT - I</td><td><i>CO</i></td><td><i>PO</i></td><td>Marks</td></tr>
<tr><td>1</td><td>a)</td><td>Data preprocessing is necessary before datamining process. Justify the statement.</td><td><i>CO1</i></td><td><i>PO 2</i></td><td>05</td></tr>
<tr><td></td><td>b)</td><td>Outline various steps of knowledge discovery with a neat diagram.</td><td><i>CO2</i></td><td><i>PO1</i></td><td>10</td></tr>
<tr><td></td><td>c)</td><td>Given two points in 3D space A=(2,3,5), B=(6,7,9)
1. Calculate the Euclidean distance between A and B.
2. Calculate the Manhattan distance between A and B.
3. Represent A and B as vectors and compute the Cosine similarity between them.</td><td><i>CO2</i></td><td><i>PO1</i></td><td>05</td></tr>
<tr><td colspan="3" align="center">OR</td><td></td><td></td><td></td></tr>
<tr><td>2</td><td>a)</td><td>Normalize the following group of data: 200,300,400,600,1000.
(a) min-max normalization by setting min = 0 and max = 1
(b) z-score normalization
(c) z-score normalization using the mean absolute deviation instead of standard deviation
(d) normalization by decimal scaling</td><td><i>CO2</i></td><td><i>PO3</i></td><td>08</td></tr>
<tr><td></td><td>b)</td><td>A retail company is analyzing its customer database to identify buying patterns and improve marketing strategies. Suggest techniques or tools that can be used for handling missing values of the data for pattern identification.</td><td><i>CO2</i></td><td><i>PO2</i></td><td>05</td></tr>
<tr><td></td><td>c)</td><td>Use a flowchart to summarize the following procedures for attribute subset selection:
(a) stepwise forward selection
(b) stepwise backward elimination
(c) a combination of forward selection and backward elimination</td><td><i>CO2</i></td><td><i>PO1</i></td><td>07</td></tr>
</table>

<div style="writing-mode: vertical">**Important Note:** Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.</div>

			UNIT - II			
3	a)		Illustrate different schemas for multidimensional data models with a neat diagram for each.	CO1	PO1	**10**
	b)		Describe the **ETL (Extract, Transform, Load)** process in the context of a data warehouse with diagram.	CO1	PO2	**05**
	c)		Differentiate between MOLAP (Multidimensional OLAP), ROLAP (Relational OLAP).	CO2	PO1	**05**
			OR			
4	a)		Perform any four OLAP operations on the given data cube. Support your answer with pictorial representation for each.	CO3	PO2	**10**

	b)		Differentiate between the following data warehouse models:
			   i. Enterprise warehouse
			   ii. Data mart
			   iii. Virtual warehouse
				CO1	PO2	**06**

	c)		Explain the role of **metadata** in a data warehouse.	CO1	PO1	**04**
			UNIT - III			
5	a)		Apply apriori algorithm on the dataset given support count =2 and confidence = 70%.	CO3	PO3	**10**
	b)		Write FP growth algorithm for frequent itemset mining.	CO1	PO1	**06**

				CO1	PO1	04
		c)	How does association rule can be measured? Explain any two.			

<div align="center">OR</div>

6	a)	Apply partitioning algorithm on the given dataset for the minimum support count=2.	CO2	PO3	10

TID	List of Items
T1	I1, I2, I4
T2	I2, I4
T3	I3, I4
T4	I1, I2, I5
T5	I1, I4
T6	I1, I3, I5
T7	I1, I2 ,I3, I5
T8	I1, I2, I3

	b)	Describe the application of frequent pattern mining in **market basket analysis**.	CO3	PO1	05
	c)	Write Spade algorithm for vertical sequence mining.	CO1	PO1	05

<div align="center">UNIT - IV</div>

7	a)	Examine the diagram below and determine which clustering algorithm would be most suitable for this dataset. Justify your choice with proper reasoning.	CO3	PO2	04

	b)	Summarize the working of BIRCH algorithm. Apply the same on given dataset having maximum radius of a cluster in CF Tree=2. P1(1,2),P2(2,2),P3(2,3),P4(8,8),P5(8,9),P6(25,80)	CO2	PO3	10
	c)	Describe any two major tasks to evaluate clustering algorithms.	CO1	PO1	06

<div align="center">OR</div>

8	a)	Given the points A(3, 7), B(4, 6), C(5, 5), D(6, 4), E(7, 3), F(6, 2), G(7, 2) and H(8, 4), Find the core points and outliers using DBSCAN. Take Eps = 2.5 and MinPts = 3.	CO3	PO3	10

		b)	Explain the working and applications of probabilistic hierarchical clustering with an example.	*CO1*	*PO1*	**10**
			UNIT - V			
	9	a)	How does Probabilistic Model-Based Clustering work for clustering and specify their advantages over traditional algorithms?	*CO2*	*PO2*	**06**
		b)	Describe the applications and challenges of Clustering Graph and Network Data.	*CO2*	*PO2*	**06**
		c)	Write SCAN algorithm for cluster analysis on graph data.	*CO1*	*PO2*	**08**
			OR			
	10	a)	Describe the problems, challenges and methodologies in clustering high dimensional data.	*CO1*	*PO1*	**10**
		b)	Illustrate the types of Biclusters and highlight the possible ways of mining them.	*CO1*	*PO1*	**10**

******