U.S.N.										

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

Autonomous Institute Affiliated to VTU

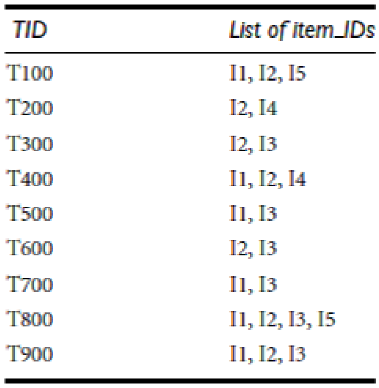
January / February 2025 Semester End Main Examinations

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.

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.

<table>
<tr><th></th><th></th><th>UNIT - I</th><th><i>CO</i></th><th><i>PO</i></th><th>Marks</th></tr>
<tr><td>1</td><td>a)</td><td>Define data mining? Explain the process of Knowledge discovery in database (KDD) With a neat diagram.</td><td>CO1</td><td>PO1</td><td><b>06</b></td></tr>
<tr><td></td><td>b)</td><td>Describe the common types of data quality issues that need to be addressed during data cleaning, and how missing data can be handled in datasets?</td><td>CO1</td><td>PO2</td><td><b>07</b></td></tr>
<tr><td></td><td>c)</td><td>Describe the common types of data transformations used in the KDD process in brief. Explain the role of dimensionality reduction in data transformation.</td><td>CO1</td><td>PO2</td><td><b>07</b></td></tr>
<tr><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td>a)</td><td>Consider the given Data Matrix representing 2-dimensional points of a line
<table>
<tr><th>point</th><th>attribute1</th><th>attribute2</th></tr>
<tr><td><i>x1</i></td><td>1</td><td>2</td></tr>
<tr><td><i>x2</i></td><td>3</td><td>5</td></tr>
<tr><td><i>x3</i></td><td>2</td><td>0</td></tr>
<tr><td><i>x4</i></td><td>4</td><td>5</td></tr>
</table>
Calculate the Dissimilarity Matrix using the following methods:
1. Euclidean distance.
2. Manhattan Distance.
3. Minkowski Distance with h=3.</td><td>CO2</td><td>PO3</td><td><b>10</b></td></tr>
<tr><td></td><td>b)</td><td>The company's income spans from $20,000 to $100,000. Normalize the income of a newly added employee with salary $25,000 using:
1. Min-Max normalization to scale it to the range [0, 1].
2. Z-score normalization if the mean(μ) of the dataset is 55,000 and the standard deviation((σ) is 10,000.
3. Decimal Scaling Normalization with j=5.
Explain each method and provide the transformed income values after applying each normalization technique.</td><td>CO3</td><td>PO3</td><td><b>10</b></td></tr>
</table>

			UNIT - II			
	3	a)	Discuss the role of metadata in the data integration process. List the pros and cons of using Extract, Transform, Load (ETL) tools for data integration.	CO1	PO1	**08**
		b)	Define a data cube and explain its significance in data warehousing. Describe the different techniques used for computing data cubes.	CO2	PO1	**08**
		c)	Compare exploratory and predictive data cube analysis.	CO2	PO1	**04**
			OR			
	4	a)	Explain the following with suitable example: 1. Fact table 2. Dimension table 3. 0-D(apex) cuboid 4. Starnet query model	CO1	PO1	**04**
		b)	Define OLAP. Illustrate the main types of OLAP operations.	CO2	PO1	**08**
		c)	Explain the multi-tier architecture of a Data Warehouse with the help of a clear and well-labelled diagram.	CO2	PO2	**08**
			UNIT - III			
	5	a)	Consider the Transactional data of AllElectronics branch given below. Find frequent patterns with association rules using Apriori algorithm with minimum confidence threshold=70% and minimum support=2.  *TID* — *List of item_IDs* T100 — I1, I2, I5 T200 — I2, I4 T300 — I2, I3 T400 — I1, I2, I4 T500 — I1, I3 T600 — I2, I3 T700 — I1, I3 T800 — I1, I2, I3, I5 T900 — I1, I2, I3 **Transactional data of AllElectronics branch**	CO3	PO3	**10**
		b)	Illustrate the process of frequent itemset generation using Transaction Reduction method in datamining with an example.	CO2	PO2	**10**
			OR			
	6	a)	Construct Frequent Pattern (FP) Tree for the Transaction Dataset given below. The given data is a hypothetical dataset of transactions with each letter representing an item. Consider the minimum support as 3.	CO3	PO3	**10**

Transaction ID	Items
T1	{E,K,M,N,O,Y}
T2	{D,E,K,N,O,Y}
T3	{A,E,K,M}
T4	{C,K,M,U,Y}
T5	{C,E,I,K,O,O}

		b)	How is the Apriori property effectively employed in the algorithm? Illustrate the Apriori algorithm for finding frequent itemsets by confined Candidate Generation.	*CO3*	*PO3*	**10**
			UNIT - IV			
	7	a)	Provide Balanced Iterative Reducing and Cluster using Hierarchies (BIRCH) Algorithm. Apply the same to cluster the given data: D= {(4,5), (3,4), (2,6), (3,8), (6,2), (7,2), (7,4), (8,4), (7,9)} Consider Max branch =2, Threshold (T)<1.5.	CO3	PO3	**10**
		b)	Explain the concept of density-based clustering and how it differentiates clusters from noise.	*CO2*	*PO2*	**5**
		c)	Write Probabilistic Hierarchical Clustering Algorithm using Gaussian Distribution.	*CO2*	*PO2*	**5**
			OR			
	8	a)	How does the DBSCAN algorithm identify and group data points based on density in a spatial dataset, and what are the key parameters that influence the clustering results?	*CO3*	*PO3*	**10**
		b)	Explain the working and applications of probabilistic hierarchical clustering with an example.	*CO1*	*PO2*	**10**
			UNIT - V			
	9	a)	Define probabilistic clustering. How does it differ from traditional clustering methods?	*CO1*	*PO2*	**06**
		b)	What makes clustering graph and network data different from clustering traditional datasets?	CO2	PO1	**06**
		c)	Explain How does Principal Component Analysis (PCA) support clustering?	CO2	PO2	**08**
			OR			
	10	a)	Illustrate the types of Biclusters and highlight the possible ways of mining them.	CO3	PO2	**10**
		b)	Describe the problems and challenges associated with clustering high-dimensional data, as well as the methodologies used to address them.	CO3	PO2	**10**

******