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: 23AM5PEKDI **Max Marks: 100**
Course: Knowledge Discovery

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, and appeal to evaluator will be treated as malpractice.

			UNIT - I	*CO*	*PO*	Marks
1	a)		Differentiate the following: i) Discrimination and classification ii) Characterization and clustering iii) Classification and regression	*CO1*	*PO1*	**08**
	b)		Consider a group of 12 sales price records 5,10,11,13,15,35,50,55,72,92,204,215. Partition them into three bins by smoothing and find the following: i) Equal-frequency (equal-depth) partitioning ii) Equal-width partitioning iii) Clustering	*CO1*	*PO2*	**06**
	c)		Describe various methods for handling noisy data during preprocessing.	*CO1*	*PO1*	**06**
			OR			
2	a)		Define Data mining. Describe how data can be mined?	*CO1*	*PO2*	**05**
	b)		i. Illustrate the steps involved in knowledge discovery from data. ii. Given the data set: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34. Apply binning methods for data smoothing using equi-depth bins (Bin size = 3).	*CO1*	*PO3*	**10**
	c)		Summarize the different forms of data preprocessing with a neat diagram.	*CO1*	*PO1*	**05**
			UNIT - II			
3	a)		Differentiate OLTP and OLAP Systems in data warehousing.	*CO1*	*PO1*	**06**
	b)		Illustrate the three-tier data warehouse architecture by incorporating a schematic diagram.	*CO2*	*PO1*	**08**
	c)		Describe the key components and goals of the ETL (Extraction, Transformation, and Loading) process with a neat sketch.	*CO2*	*PO1*	**06**
			OR			
4	a)		Illustrate the recommended approach for developing data warehouse models with diagram.	*CO1*	*PO2*	**10**
	b)		Define a data warehouse and explain its key characteristics.	*CO2*	*PO1*	**05**
	c)		Elucidate the working of database schema design by visually	*CO2*	*PO2*	**05**

			presenting a Snowflake schema for a comprehensive sales data warehouse that encompasses both sales and shipping information. Provide graphical representation of Snowflake schema.			
			UNIT - III			
	5	a)	Explain how measures are computed in Data warehouse models	*CO1*	*PO1*	**06**
		b)	Assume a data warehouse consists of three dimensions [time, doctor, patient], and the two measures [count, charge]. Based on the given scenario answer the following: i) Draw a schema diagram for the above data warehouse fact constellation. ii) Apply the specific OLAP operation.	*CO1*	*PO2*	**06**
		c)	Define Fact Constellation, Fact table and Dimension table. Plot Fact Constellation for the following data • **Placement** is a *fact table* having attributes: (Stud_roll, Company_id, TPO_id) with facts: (Number of students eligible, Number of students placed). • **Workshop** is a *fact table* having attributes: (Stud_roll, Institute_id, TPO_id) with facts: (Number of students selected, Number of students attended the workshop). • **Company** is a *dimension table* having attributes: (Company_id, Name, Offer_package). • **Student** is a *dimension table* having attributes: (Student_roll, Name, CGPA). • **TPO** is a *dimension table* having attributes: (TPO_id, Name, Age). • **Training Institute** is a *dimension table* having attributes: (Institute_id, Name, Full_course_fee)	*CO2*	*PO2*	**08**
			OR			
	6	a)	Provide definitions and examples for the following terms: i. Data cube ii. Base cuboid iii. Apex cuboid iv. Dimensions v. Facts	*CO3*	*PO1*	**05**
		b)	Explain the process of database schema design by visually presenting a Snowflake schema for a comprehensive sales data warehouse and a Fact Constellation schema for an integrated data warehouse that includes both sales and shipping information.	*CO2*	*PO1*	**10**
		c)	Outline the unique views regarding the design of a data warehouse.	*CO1*	*PO1*	**05**
			UNIT - IV			
	7	a)	Strong association rules may be negatively correlated. Justify the statement with an example.	*CO1*	*PO2*	**06**
		b)	Analyze the methods to improve the efficiency of Apriori-based Algorithm.	*CO1*	*PO2*	**06**

		c)	Find the frequent itemsets and generate association rules for below dataset. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%). Transaction ID | Items T1 | Hot Dogs, Buns, Ketchup T2 | Hot Dogs, Buns T3 | Hot Dogs, Coke, Chips T4 | Chips, Coke T5 | Chips, Ketchup T6 | Hot Dogs, Coke, Chips	*CO2*	*PO2*	**08**
			OR			
	8	a)	Prove that strong rules may not always be interesting in pattern evaluation methods with an example.	*CO3*	*PO2*	**06**
		b)	Write Apriori algorithm to discover frequent itemsets in dataset	*CO1*	*PO1*	**06**
		c)	Generate a FP tree for the following transaction dataset [minimum_support=30%]. Transaction ID | Items 1 | E,A,D,B 2 | D,A,C,E,B 3 | C,A,B,F 4 | B,A,D 5 | D 6 | D,B 7 | A,D,E 8 | B,C	*CO3*	*PO2*	**08**
			UNIT - V			
	9	a)	Describe the different approaches in statistical-based outlier detection in brief.	*CO1*	*PO1*	**06**
		b)	Apply Single linkage agglomerative hierarchical clustering algorithm for the following set of data points: 18,22,25,27,42,43. Use Euclidean distance to calculate distance between two data points.	*CO1*	*PO1*	**08**
		c)	Illustrate the working of DBSCAN with suitable example.	*CO3*	*PO1*	**06**
			OR			
	10	a)	Describe the following clustering approaches with an example for each. i) Partitioning ii) hierarchical iii) density-based iv) grid-based methods.	*CO1*	*PO1*	**10**
		c)	Apply k-means clustering on following data points: {5, 7, 16, 18, 24, 26, 34, 38}until the centroids converge given that, i) K=3 ii) The three initial centroids are (15,25,31) respectively.	*CO1*	*PO1*	**10**

****