4a	Consider the following dataset that contains three attributes: Age, Income, and Credit Score. **ID** | **Age** | **Income** | **Credit Score** 1 | 25 | 50000 | 700 2 | 45 | 90000 | 800 3 | 35 | 75000 | 650 4 | 50 | 48 | 720 5 | ? | 40000 | 600 6 | 29 | 58000 | ? 697 a) Replace the missing value in the "Income", Credit Score" and "Age" column and update the above table. b) Discretize the "Age" attribute into three bins: "Young" (18-30 years), "Middle-aged" (31-45 years), and "Senior" (46-60 years). Draw a Histogram for that.	1.5	CO2
4b	Normalize the "Income" and "Credit Score" attributes using Min-Max normalization to scale the values between 0 and 1 after handling the missing values in the table above.	1.5	

5a	Given the following star schema for a retail company:
	• Fact Table: Sales
	  ○ Attributes: DateID, ProductID, StoreID, SalesAmount
	• Dimension Tables:
	  ○ Date Dimension:
	  ▪ Attributes: DateID, Year, Quarter, Month, Day
	  ○ Product Dimension:
	  ▪ Attributes: ProductID, ProductName, Category, Brand
	  ○ Store Dimension:
	  ▪ Attributes: StoreID, StoreName, City, State, Region
	Draw the star schema for the retail company, labelling the fact and dimension tables.
		2	CO2

5b	Construct a data cube that aggregates SalesAmount by Year, Product Category, and Region for the above star schema.	1	

B-Tech 5th Semester

END-SEMESTER EXAMINATION, NOV-DEC, 2024

Course Title: Data Mining | Duration: 03 Hours

Course Code: COCSC16/CDCDC16 | Max. Marks: 40

Note: - Attempt all questions in the given order only. Missing data/information (if any), maybe suitably assumed & mentioned in the answer.

Q. No.	Question	Marks	CO
Q 1	**Attempt any 2 parts of the following**		
1a	What is the significance of Data Mining in discovering patterns and insights from large datasets? Consider an example of sales data for predicting customer purchasing behavior. Also discuss the process of converting raw data into useful knowledge for Data mining tasks.	4	1,2
1b	Discuss the challenges of applying Data Mining to unstructured data, such as text from social media posts or news articles. Provide an example of using statistical methods in data mining. For example, calculate the mean, median, or mode in a dataset and explain its relevance to data mining.	4	1,2
1c	Compare and contrast Supervised Learning and Unsupervised Learning techniques in Data Mining. Provide examples of algorithms for each technique and discuss when each is best applied.	4	1,2
Q 2	**Attempt any 2 parts of the following**		
2a	Suppose a retail company wants to analyze its sales data across various dimensions such as time, product, and region. Explain how an OLAP cube would be structured to answer queries like "What were the sales for each product category last quarter in each region?"	4	2,3

2b	Given a relational database for a university (e.g., tables for students, courses, enrollment), design a simple data warehouse star schema for analyzing student performance over time.	4	2,3
2c	(i)Given a continuous dataset (e.g., age or income), demonstrate how you would apply binning as a discretization technique. (ii)For a dataset with hierarchical categories (e.g., product hierarchy: electronics > smartphones > brand), describe the process of creating a concept hierarchy.	4	1,2, 3
Q 3	**Attempt any 2 parts of the following**		
3a	Explain key terms related to association rules, such as support, confidence, and lift. Describe how each of these metrics is calculated and why they are important for evaluating the strength and relevance of association rules in a dataset.	4	1,4
3b	Suppose we have the following dataset that has various transactions. Find the frequent itemsets and generate the association rules using the Apriori algorithm: TID | ITEMSETS T1 | A, B T2 | B, D T3 | B, C T4 | A, B, D T5 | A, C T6 | B, C T7 | A, C T8 | A, B, C, E T9 | A, B, C **Given: Minimum Support= 2, Minimum Confidence= 50%** Discuss the limitations of the Apriori algorithm, and mention scenarios where it might not be the most efficient method for mining association rules.	4	2,4
3c	Discuss the key components of FP-Growth, including the **FP-tree** and **conditional pattern base**, and explain how they help in mining frequent itemsets. Generate the *frequent pattern* from the following data set using ***FP growth***, where *minimum support* 3.	4	2,4

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

Q 4 **Attempt any 2 parts of the following**

4a

Color	Legs	Height	Smelly	Species
White	3	Short	Yes	M
Green	2	Tall	No	M
Green	3	Short	Yes	M
White	3	Short	Yes	M
Green	2	Short	No	H
White	2	Tall	No	H
White	2	Tall	No	H
White	2	Short	Yes	H

Using the above data with Naive Bayes Classification, identify the species of an entity with the following attributes.

X={Color=Green, Legs=2, Height=Tall, Smelly=No}

4 2,4

4b

How Does the K-Nearest Neighbors Algorithm Work?

Brightness	Saturation	Class
40	20	Red
50	50	Blue
60	90	Blue
10	15	Red
70	70	Blue
60	10	Red
25	80	Blue

4 2,4

	Use KNN to find **class** when Brightness is 20 and Saturation is 35 (K=5, use ***Euclidean Distance***)		
4c	Imagine you are analyzing a medical diagnostic model for cancer detection with a confusion matrix as follows: | | Predicted: Positive | Predicted: Negative | | --- | --- | --- | | Actual: Positive | 70 | 30 | | Actual: Negative | 20 | 80 | (i)Calculate all relevant metrics (Accuracy, Precision, Recall, F1 Score). (ii)Discuss the potential implications of this model's performance in a healthcare setting. What are the risks of high false positives and false negatives in this context?	4	2,3, 4
Q 5	**Attempt any 2 parts of the following**		
5a	Apply the K-Means algorithm with K=2 on the dataset: (10, 5), (12, 7), (8, 6), (13, 8), (9, 5), (11, 6). Start by selecting the first two points as initial centroids. Perform two iterations by assigning each point to the nearest centroid, updating centroids after each iteration. Round Euclidean distances to the nearest whole number and show calculations and final clusters. Explain all the steps in detail	4	2,5
5b	Explain Density-based clustering algorithms. Using DBSCAN, identify the core points, boundary points and outliers in the following dataset: Points: **P(2, 3), Q(3, 4), R(4, 5), S(5, 4), T(3, 3), U(5, 5), V(6, 3)** Take Eps = 2 and MinPts = 3.	4	2,5
5c	Given the following one-dimensional data set representing the ages of individuals: {20, 25, 30, 35, 40, 45, 50}. Form a dendrogram using single as well as complete linkage agglomerative hierarchical clustering and explain the process of clustering, showing the step-by-step merging of clusters. What is the significance of a dendrogram in hierarchical clustering?	4	3,5