Course Code:	CST359-4	

Sixth Semester B.E. (Computer Science and Engineering) Examination

Data Warehousing and Mining (Elective-II)

Time: 02 Hours]	[Max. Marks: 40

Instructions to Candidates:

1. All Questions carry marks as indicated.
2. Assume suitable data wherever necessary.

Que.		Description of Question	Marks	CO
1	(a)	What do you mean by ETL Process? What is the purpose of 'refresh' in ETL process?	02	CO1
	(b)	Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. (i) Draw a Star schema for the above scenario. (ii)Write an SQL query assuming the data is stored in a relational database with the schema Fee (day, month, year, doctor, hospital, patient, count, charge). List the total Fee collected by each doctor in 2020?	05	CO1
2	(a)	Given is the frequency of stop words in documents (The values are given in increasing order) : 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33,33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Apply the following methods and show the results :— (i) Use smoothing by bin means with a depth of 3. (ii) Use Min – Max normalization to transform the value 30 into the range 0.0 to 1.0. (iii) Use z – score normalization to transform the value 30 where the standard deviation of the above frequency is 12.94. (iv) Use normalization by decimal scaling to transform the value 30. (v) Plot an equi – width histogram of width 10 on graph paper.	05	CO2
	(b)	Data quality can be assessed in terms of accuracy, completeness, and consistency. Propose two other dimensions of data quality.	02	CO2
3	(a)	State the advantages of data partitioning in data – warehouse. Write a SQL query to create composite List – Range partitioning for the following scenario :	06	CO2

		Customer table having attributes cust_id, cust_name, cust_state and time_id. Perform list partitioning on state attributes and range partitioning on time –id. Partition definitions for list are as below : • Partition East should accept values ('WB', 'JK') • Partition South should accept values ('TN', 'AP') • Partition North should accept values ('UP', 'HP') • Partition Temp should accept any other state. Partition definitions for range are as below for the year 2020 : • Partition P1 should accept values for Jan, Feb, March, April. • Partition P2 should accept values for May, June, July, August. • Partition P3 should accept values for September, October, November, and December.		
4	**(a)**	State the Apriori Property. Using Apriori Algorithm find the final item set for the following dataset S. Where (min-Sup=50%, min_conf=70%).Generate all association rules and list the strongest rule. TID | Items Purchased 101 | Book ,Note, Pen 102 | Pencil ,Note ,Eraser 103 | Book, Pencil, Note, Eraser 104 | Pencil, Eraser	06	CO3
5	**(a)**	Use the dataset below to learn a decision tree which predicts if people pass Java Test (True or False), based on their previous GPA (High (H), Medium (M), or Low (L)) and whether or not they studied. GPA and Studied are two features. **Passed is the target function**. GPA | Studied | **Passesd?** L | F | F L | T | T M | F | F M | T | T H | F | T H | T | T Construct the decision tree using ID3 algorithm that would be learned for this dataset.	07	CO4

6	**(a)**	What do you mean by K-means clustering? List the drawbacks of k-means clustering algorithm. Draw the final clusters on graph paper. Assume the following dataset is given: (2,2),(4,4),(5,5),(6,6),(8,8),(0,4)(4,0). K-means is run with k=3, to cluster the dataset. Use Euclidean distance measure. K-means initial clusters C1,C2,and C3 are as follows: C1:{(2,2),(4,4),(6,6)} C2:{(0,4),(4,0)} C3:{(5,5),(8,8)}	07	CO4