8. (a) Compare and contrast centroid-based clustering and density-based clustering methods with examples. 4

(b) Explain the concept of Clustering Feature (CF) in the BIRCH algorithm 4

(c) What are outliers? Discuss two methods for outlier detection in data mining. 6

UNIT-V

9. Apply the K-Means clustering algorithm to the following points: P1(3, 9), P2(1, 4), P3(7, 3), Q1(6, 8), Q2(8, 5), Q3(5, 3), R1(2, 1), and R2(5, 9). The distance function to be used is Euclidean distance. Initially, P1, Q1, and R1 are selected as the centers of the three clusters. Perform K-Means clustering iteratively and find the final clusters. 14

10. Apply the DBSCAN clustering algorithm to the following points: A1(1, 2), A2(2, 2), A3(2, 3), B1(8, 7), B2(8, 8), B3(25, 80), and C1(24, 79). Use Euclidean distance as the distance measure. Set the parameters as Eps = 1.5 and MinPts = 3. Identify the core points, border points, outliers, and determine the final clusters formed. 14

B.Tech Even Semester Examination, May, 2025

Computer Science & Engineering

(6th Semester)

Course No: CSE-603(06)

(Data Mining)

Full Marks: 70

Pass Marks: 28

Time: 3 hours

Note:

1. Attempt one question taking one from each unit.
2. Begin each answer in a new page
3. Answer parts of a question at a place
4. Assume reasonable data wherever required
5. The figure in the right margin indicates full marks for the question
6. All the mathematical symbols and abbreviations have their usual meanings.

UNIT-I

1 Explain in detail the steps involved in the KDD (Knowledge Discovery in Databases) process. 14

2. (a) In a box of 12 bulbs, 4 are defective. Two bulbs are selected at random without replacement. What is the probability that both are non-defective? 4

(b) A standard deck of 52 cards is shuffled. Find the probability that:i.The first card is an ace.ii. The first two cards are both aces.iii.At least one ace appears in the first two cards. 2+3+5

UNIT-II

3. (a) Given the following data: [50, 80, 100, 150, 200], normalize it using

i. Min-Max Normalization to the range [0,1].

ii. Z-Score Normalisation 3+3

(b) Explain the following:

i. Missing Data

ii. Noisy Data

iii. Inconsistent

iv. Binning 2+2+2+2

4. (a) Find the frequent itemsets using the Apriori algorithm with a minimum support count of 3 using the following data: 7

Transaction ID	Items
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread
T4	Milk, Bread, Butter
T5	Bread, Butter

(b) Using the data given in Q4(a), generate association rules with minimum confidence of 70%. 7

UNIT-III

5. (a) Consider the following dataFit a simple linear regression line Y=a+bX and predict Y when X = 6. 7

X	1	2	3	4	5
Y	2	4	5	4	5

(b) Derive the expression for logistic regression. 7

6. (a) Create a decision tree by finding the best attribute to split first using Gini Index using the following data: 10

Age	Income	Student	Credit Rating	Buys Computer
<=30	High	No	Fair	No
<=30	High	No	Excellent	No
31-40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31-40	Low	Yes	Excellent	Yes
<=30	Medium	No	Fair	No
<=30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
<=30	Medium	Yes	Excellent	Yes
31-40	Medium	No	Excellent	Yes
31-40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

(b) Explain entropy and information gain in detail. 4

UNIT-IV

7. (a) Discuss the concepts of hard clustering and soft clustering with suitable examples 5

(b) What is the Minkowski distance? How do Manhattan and Euclidean distances fit into it? 5

(c) Explain the requirements for good clustering in data mining. How does the similarity measure impact clustering quality? 4