Reg. No ..... Name **26U4157**

END SEMESTER EXAMINATION - MARCH 2026

SEMESTER 4 : INTEGRATED M.Sc. PROGRAMME COMPUTER SCIENCE-DATA SCIENCE

COURSE : **21UP4CRMCP11 : DATA MINING**

(For Regular 2024 Admission and Improvement/Supplementary 2023/2022/ 2021 Admissions)

Time : Three Hours Max. Weightage: 30

PART A

Answer any 8

1. State the function of ETL tool.
2. Define a closed itemset.
3. Notebook ⇒ {Pen, Pencil} [support = 3%]
   Write the meaning of the above association rule.
4. Define the concept - web usage mining.
5. Define data cube.
6. State any two drawbacks of decision tree induction algorithm.
7. List any two examples of a frequent itemset.
8. Define entropy in the ID3 algorithm.
9. List any two ways by which big data can be integrated.
10. List any two examples of density-based clustering algorithms.

(1 x 8 = 8 Weight)

PART B

Answer any 6

11. Assuming a specific application domain of your own, point out the research challenges of data mining.
12. Explain how the value of lift affects the correlation among itemsets.
13. According to the assumptions made, outlier detection methods can be classified into three types. Explain the types.
14. Suppose a group of 12 sales price records has been sorted as follows:
    5,10,11,13,15,35,50,55,72,92,204,215.
    Use histogram analysis for discretizing the above data.
15. “Data have quality if they satisfy the requirements of the intended use.” Discuss the major factors that affect the quality of data to be mined.
16. With an example, explain how confusion matrix helps in evaluating classifier performance.
17. Explain how dynamic itemset counting can improve the efficiency of Apriori algorithm.
18. Differentiate between posterior and prior probabilities with an example.

(2 x 6 = 12 Weight)

PART C

Answer any 2

19. Suppose that the data mining task is to cluster points (with *(x, y)* representing location) into three clusters, where the points are
    $A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$, $B_1(5,8)$, $B_2(7,5)$, $B_3(6,4)$, $C_1(1,2)$, $C_2(4,9)$.

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the k-means algorithm to show only
(a) The three cluster centers after the first round of execution.
(b) The final three clusters.

20. With an example, describe a method that mines the complete set of frequent itemsets without a costly candidate generation process.
21. Instead of using various techniques, data cleaning itself can be considered as a process or as a tool. Explain how this can be performed.
22. With an example, explain the Naive Bayesian classifier.

(5 x 2 = 10 Weight)