Reg. No ..... Name **24U418**

END SEMESTER EXAMINATION - MARCH 2024

SEMESTER 4 - INTEGRATED M.Sc. PROGRAMME COMPUTER SCIENCE

COURSE : **21UP4CRMCP11 - DATA MINING**

(For Regular -2022 Admission and Improvement/Supplementary - 2021 Admission)

Time : Three Hours Max. Weightage: 30

PART A

Answer any 8

1. List any two ways by which big data can be integrated.
2. List any two ensembling methods.
3. Define constraint-based clustering.
4. To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property is used to reduce the search space. Define the Apriori property.
5. Name any two algorithms that are used for classification and prediction.
6. List any four steps involved in data pre-processing.
7. List any two examples of a frequent itemset.
8. Find the midrange from the following data:
   89, 77, 88, 91, 88, 93, 99, 79, 87, 84, 86, 82, 88, 89, 78
9. State the reason why clustering is also called data segmentation.
10. State how a market basket is represented for analysis.

(1 x 8 = 8 Weight)

PART B

Answer any 6

11. Classify and explain the outlier detection methods on the basis of supervision provided on the data.
12. Use these methods to normalize the following group of data: 200,300,400,600,1000
    (a) z-score normalization
    (c) normalization by decimal scaling
13. Write the procedure in Python / R that implements the Apriori Algorithm.
14. Discuss how binary representation helps in association analysis.
15. Discuss briefly the impacts of data mining in the society.
16. Differentiate between supervised learning and unsupervised learning.
17. Discuss briefly the k-fold cross validation technique of improving the classification accuracy.
18. Explain how data cube technology can be used as a data reduction technique.

(2 x 6 = 12 Weight)

PART C

Answer any 2

19. With an example, explain the induction of a decision tree using information gain.
20. With an example, explain how Chi-square test helps in data integration.

21. A database has five transactions as given below. Let min_sup = 60% and min_conf = 80%.

TID	items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

Find all frequent itemsets using Apriori and list all the strong association rules that hold.

22. 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.

(5 x 2 = 10 Weight)