Code No.: DS504PC

R20 H.T.No.

		8	R						

CMR ENGINEERING COLLEGE: : HYDERABAD
UGC AUTONOMOUS
III–B.TECH–I–Semester End Examinations (Supply) - December- 2024
INTRODUCTION TO DATA MINING
(CSD)

[Time: 3 Hours] **[Max. Marks: 70]**

Note: This question paper contains two parts A and B.
Part A is compulsory which carries 20 marks. Answer all questions in Part A.
Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks.

PART-A (20 Marks)

1. a) List the methods of filling missing values. [2M]
b) What is the use of smoothing in data transformation? [2M]
c) What is the need of confidence measure in association rule mining? [2M]
d) Define maximal frequent item set. [2M]
e) What are the assumptions in Naïve Bayes classifier? [2M]
f) Why is tree pruning useful in decision tree induction? [2M]
g) Differentiate Agglomerative and Divisive hierarchical clustering. [2M]
h) What is density-based clustering? [2M]
i) List the applications of web usage mining. [2M]
j) What is web content mining? [2M]

PART-B (50 Marks)

2. Explain the need of data preprocessing and various forms of preprocessing. [10M]

OR

3. Explain how principal component analysis is carried out to reduce the dimensionality of data sets. [10M]

4. A database has five transactions. Let *min sup* is 60% and *min conf* is 80%. [10M]

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 algorithm.

OR

5. Suppose you have the set C of all frequent closed itemsets on a data set *D*, as well as the support count for each frequent closed itemset. Describe an algorithm to determine whether a given itemset *X* is frequent or not, and the support of *X* if it is frequent. [10M]

6. Explain Naïve-Bayes classification technique with an illustrative example. [10M]

OR

7. Discuss the methods for expressing attribute test conditions. [10M]

8. 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 [10M]
(i) The three cluster centers after the first round of execution.
(ii) The final three clusters.

OR

9.a) Provide the pseudocode of the object reassignment step of the PAM algorithm. [5M]
b) Illustrate the strength and weakness of *k-means* in comparison with *k-medoids*. [5M]

10. Discuss various kinds of patterns to be mined from web/server logs in web usage mining. [10M]

OR

11. Discuss the following [5M]
a) Text clustering. [5M]
b) Web structure mining.

****