Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SET B

SCHOOL OF ENGINEERING
END TERM EXAMINATION - JAN 2024

Semester : Semester III- 2022
Course Code : CSE2021
Course Name : Data Mining
Program : B.Tech.

Date : 08-JAN-2024
Time : 9:30AM - 12:30 PM
Max Marks : 100
Weightage : 50%

Instructions:
(i) Read all questions carefully and answer accordingly.
(ii) Question paper consists of 3 parts.
(iii) Scientific and non-programmable calculator are permitted.
(iv) Do not write any information on the question paper other than Roll Number.

PART A

ANSWER ALL THE QUESTIONS **5 X 2M = 10M**

1. Name the various kinds of data in data mining.

(CO1) [Knowledge]

2. Distinguish the relationship between covariance and correlation coefficient.

(CO2) [Knowledge]

3. Explain the concept of Support and Confidence in association rule mining.

(CO3) [Knowledge]

4. Distinguish Supervised and Unsupervised learning with an example.

(CO4) [Knowledge]

5. Name the different types of Clustering techniques.

(CO5) [Knowledge]

PART B

ANSWER ALL THE QUESTIONS **5 X 10M = 50M**

6. a. Explain the procedures of data mining as a process of knowledge discovery of data bases with a neat diagram.
b. Discuss about confluence of multiple disciplines in Data Mining.

(CO1) [Comprehension]

7. Given a set of samples S= (60,N),(75,N), (70,N),(90,Y),(85,Y),(95,Y),(100,N),(120,N),(125,N), (220,N) If S has to be partitioned in 2 intervals S1 & S2 for the split points 80 & 97. Determine the Best Split Point.

(CO2) [Comprehension]

8. Construct frequent association rules using apriori algorithm from the following data set: Minimum Support =3, Minimum Confidence=65%.

Transaction ID	Purchase Products
T1	Milk, Butter, Biscuits, Jam, Bread, Tea, Sugar
T2	Butter, Bread, Jam, Milk
T3	Chocolate, Biscuits, Tea, Sugar, Juice
T4	Jam, Bread, Milk, Tea, Butter
T5	Tea, Milk, Biscuits, Juice

(CO3) [Comprehension]

9. 1. The table above represent data set with two columns — Brightness and Saturation. Each row in the table has a class of either Red or Blue. Use KNN Classification for predicting the Class for a new sample when **Brightness=20 and Saturation=35.**

Let assume K=5.

Brightness	**SATURATION**	**CLASS**
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red
25	80	Blue

(CO4) [Comprehension]

10. Suppose you have a dataset of online shoppers with features such as the number of items purchased per session and the average session duration. Apply K-Means clustering to identify TWO Clusters of online shoppers based on their shopping behavior assuming initial centroid as (2,40) and (6,25).
The dataset is as follows:

Shopper	Items Purchased per Session	Average Session Duration (minutes)
A	5	20
B	2	40
C	8	15
D	1	60
E	6	25

(CO5) [Comprehension]

PART C

ANSWER ALL THE QUESTIONS **2 X 20M = 40M**

11. The given data set consists 14 customers historical data of purchased computer depending on age, income, student, credit rating. Apply ID3 algorithm to construct decision tree.

Age	Income	Student	Credit_Rating	Buys_Computer
<=30	High	No	Fair	No
<=30	High	No	Excellent	No
31 to 40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31 to 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 to 40	Medium	No	Excellent	Yes
31 to 40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

(CO4) [Application]

12. Apply Single linkage and Average linkage Agglomerative Clustering for the given distance matrix and visualize the clusters using dendogram.

	A	B	C	D	E	F
A	0					
B	1	0				
C	8	7	0			
D	5	4	3	0		
E	6	5	2	1	0	
F	4.5	3.5	3.5	0.5	1.5	0

(CO5) [Application]