Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF ENGINEERING
END TERM EXAMINATION - JUN 2023

Semester : Semester IV - 2021
Course Code : CSE2021
Course Name : Sem IV - CSE2021 - Data Mining
Program : B.Tech - All Programs

Date : 16-JUN-2023
Time : 9.30AM - 12.30PM
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 **(10 X 2 = 20M)**

1. Define what is “Apriori principle” and briefly discuss why Apriori principle is useful in association rule mining ?

(CO3) [Knowledge]

2. What is a dendrogram Explain with neat diagram ?

(CO5) [Knowledge]

3. Derive the relationship between covariance and correlation coefficient ?

(CO2) [Knowledge]

4. Enumerate the most common data mining techniques used in KDD Process ?

(CO1) [Knowledge]

5. Explain how regression technique can be used for data smoothing process ?

(CO2) [Knowledge]

6. Find the probability of dangerous fire when there is smoke. Let's the probability of dangerous fires are rare (1%) but smoke is fairly common (10%) due to barbecues, and 90% of dangerous fires make smoke ?

(CO4) [Knowledge]

7. How many phases are there in FP growth algorithm?

(CO3) [Knowledge]

8. What are the various types of cluster analysis methods?

(CO5) [Knowledge]

9. Define Pruning? Explain different approaches in Pruning?

(CO4) [Knowledge]

10. What is the need of Data Mining?

(CO1) [Knowledge]

PART B

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

11. Explain about Partationing clustering with algorithm and slove the problem by using K-Means clustering . where we have 2 group of visitors to a website using their age as follows: {16, 16, 17, 20, 20, 21, 21, 22, 23, 29, 36, 41, 42, 43, 44, 45, 61, 62, 66 }. 16 and 22 are the initial centroids for partitioning.

(CO5) [Comprehension]

12. Derive Naive Bayesian Theorem . Use Naive Bayes Classifier to estimate conditional probabilities of each attribute {color, legs, height, smelly} for the species classes: {M, H} using the data given in the table.Using these probabilities estimate the probability values for the new instance - (Color=Green, legs=2, Height=Tall, and Smelly=No).

No	Color	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

(CO4) [Comprehension]

13. Explain about FP-Growth algorithm . Consider the market basket transactions given in the following table. Let min_sup = 40% and min_conf = 40%. Generate the following .

Transaction ID	Items bought
T1	A,B,C
T2	A,B,C,D,E
T3	A,C,D
T4	A,C,D,E
T5	A,B,C,D

A) Find all the frequent item sets using Apriori algorithm.
B) Obtain significant of improving the of Efficiency of apriori .
C) Derive the FP-Growth Tree for the above transaction table.

(CO3) [Comprehension]

14. A) Describe about Knowledge discovery process in detail ?
B) Explain the major issues in data mining on Data Mining and Society.

(CO1) [Comprehension]

15. What is the need of dimensionality reduction? Explain any two techniques for dimensionality reduction ?

(CO2) [Comprehension]

PART C

ANSWER ALL THE QUESTIONS **(2 X 15 = 30M)**

16. "Perform Hierarchical Agglomerative clustering by considering Single linkage and average linkage for the given distance matrix. Also draw the dendrogram which shows all the steps".

S.No	A	B	C	D	E	F
A	0.00					
B	0.71	0.00				
C	5.66	4.95	0.00			
D	3.61	2.92	2.24	0.00		
E	4.24	3.54	1.41	1.00	0.00	
F	3.20	2.50	2.50	0.50	1.12	0.00

(CO5) [Application]

17. Following is the historical data of 14 people that records if they have purchased computer or not depending on age, income, student, credit rating. Apply Gini index to find out which among the four attributes is selected as root node for construction of decision tree and also find the splitting subset for the selected attribute.

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

(CO4) [Application]