Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF INFORMATION SCIENCE
END TERM EXAMINATION - JUN 2023

Semester : Semester IV - 2021
Course Code : CSA2021
Course Name : Sem IV - CSA2021 - Data Warehousing and Data Mining
Program : BSD

Date : 12-JUN-2023
Time : 1.00PM - 4.00PM
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. What are the advantages of implementing a data warehouse in an organization?
(CO1) [Knowledge]
2. Explain the Disadvantages of Fuzzy Clustering.
(CO3) [Knowledge]
3. Meta data can be classified into how many types and mention those types
(CO2,CO1) [Knowledge]
4. What is Prior Probability and Posterior Probability?
(CO3,CO4) [Knowledge]
5. How Outliers are different from Noise?
(CO4) [Knowledge]
6. What are the three factors that should be considered in partial materialization of cuboids or sub cubes?
(CO2) [Knowledge]
7. What are the three steps that are followed by Data warehouse designer?
(CO1) [Knowledge]
8. What are the Pros and Cons of Semi-Supervised Methods outlier detection?
(CO4) [Knowledge]
9. List out the three choices for data cube materialization given a base cuboid.
(CO2) [Knowledge]
10. What is Clustering?
(CO3) [Knowledge]

PART B

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

11. Suppose you are a risk analyst working for an insurance company. Can you explain the concept of Bayesian Belief Networks (BBNs) and provide a scenario-based explanation of how BBNs can be used to assess and predict the likelihood of insurance claims related to natural disasters, such as hurricanes or earthquakes, based on various factors like location, property type, and historical data?

(CO3) [Comprehension]

12. Imagine you're a data scientist working for a financial institution. Can you explain the concept of Support Vector Machines (SVM) and provide a scenario-based explanation of how SVM can be applied to classify and detect fraudulent credit card transactions based on transaction patterns, customer behavior, and historical fraud data?

(CO3) [Comprehension]

13. Let's consider a scenario where you are working as a data analyst for a large telecommunications company. Can you explain the concept of join indexing in the context of data warehousing and data mining, and provide an example of how join indexing can be applied to improve query performance in analyzing customer data for marketing campaigns?

(CO2) [Comprehension]

14. Imagine you're tasked with designing a data warehouse for a retail company. Could you outline and explain the nine-step method for designing a data warehouse and provide a brief explanation of each step within the context of this scenario?

(CO1) [Comprehension]

15. Imagine you're a data analyst working for a manufacturing company. Can you explain the different types of outliers that can occur in manufacturing data, and provide a scenario-based explanation of how each type of outlier can impact the analysis of production data, such as detecting faulty equipment or identifying quality control issues in the manufacturing process?

(CO4) [Comprehension]

PART C

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

16. **Table 3.3** A 3-D view of sales data for *AllElectronics*, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

	location = *"Chicago"*				***location*** = *"New York"*				***location*** = *"Toronto"*				***location*** = *"Vancouver"*			
	item				***item***				***item***				***item***			
time	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

a) Consider the above given data, analyze the data and draw the 3D cube.
b) Suppose that you would now like to view the above sales data with an additional fourth dimension, such as supplier (Assume there are Three suppliers SUP1, SUP2 and SUP3) draw the 4D cube.
c) How can data warehousing provide significant benefits to an organization in terms of data management and decision-making processes?

(CO2,CO1) [Application]

17. a) Imagine you are a data scientist working for a social media company. One of your tasks is to develop a system to classify user comments as either spam or legitimate. Can you explain the concept of Naïve Bayesian Classifier and provide a scenario-based explanation of how it can be applied to automatically classify user comments as spam or legitimate based on their textual content and other relevant features?

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

Consider the above data and

b) Estimate the conditional probabilities of P (A |+), P(B|+), P (C |+), P (A |-), P (B |-) and P (C |-)
c) Use the conditional probability estimates and predict the class label for the test sample P (A = 0, B = 1, C = 0)

(CO4,CO3) [Application]