Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

End - Term Examinations – MAY 2025

Date: 23-05-2025 **Time:** 01:00 pm –04:00 pm

School: SOIS	**Program:** BCD	
Course Code: CSA2021	**Course Name:** Data Warehousing and Data mining	
Semester: IV	**Max Marks:** 100	**Weightage:** 50%

CO - Levels	CO1	CO2	CO3	CO4	CO5
Marks	24	24	26	26	

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

Part A

Answer ALL the Questions. Each question carries 2marks. **10Q x 2M=20M**

1.	List out the applications of meta data.	**2 Marks**	**L1**	**CO1**
2.	Define Traditional approaches of Data Warehouse.	**2 Marks**	**L1**	**CO1**
3.	List out the typical OLAP operations.	**2 Marks**	**L1**	**CO2**
4.	Define set-grouping Hierarchy with one real time example.	**2 Marks**	**L1**	**CO2**
5.	Outline the role of two hyper planes in support vector machines.	**2 Marks**	**L1**	**CO3**
6.	Differentiate Clustering and Classification with an example.	**2 Marks**	**L2**	**CO3**
7.	State Back Propagation.	**2 Marks**	**L1**	**CO3**
8.	Outline the role of z-score in detecting outliers, and how it is calculated?	**2 Marks**	**L1**	**CO4**
9.	Define Proximity-based approaches.	**2 Marks**	**L1**	**CO4**
10.	A class of students took a math exam, and the majority of students scored between 60 and 85. However, a few students scored 120 and 5. Identify the outliers.	**2 Marks**	**L1**	**CO4**

Part B

Answer the Questions. **Total Marks 80M**

11.	**a.**	A large multinational company, XYZ Corp, has operations in multiple regions and countries. Each region maintains its own transactional databases for tracking sales, inventory, and customer information. The company wants to create a data warehouse to integrate data from all regions for centralized reporting and analysis. They also plan to use OLAP (Online Analytical Processing) for more complex queries. i) Describe the key components of a data warehouse architecture for XYZ Corp, explaining the role of each component in the data integration and analysis process. (10 marks) ii) Discuss how the ETL process (Extract, Transform, Load) can be implemented to handle data from multiple regions with different formats and structures. (10 marks)	**20 Marks**	**L2**	**CO1**
		Or			
12.	**a.**	A multinational company is integrating data from various regional offices, each using different database systems. To ensure consistency and smooth integration of data from these systems into the company's central data warehouse, the company decides to implement a **metadata management system**. **i)** Explain the role of **metadata** in this scenario, and how does it help in integrating data from different systems?(10 marks) **ii)** Discuss metadata management system. (10 marks)	**20 Marks**	**L2**	**CO1**
13.	**a.**	A company is analyzing its sales data, which includes dimensions such as time, region, and product category. The company wants to compute summary statistics like total sales, average sales, and number of units sold across different combinations of these dimensions for better decision-making. They decide to use a data cube for efficient computation of these summary statistics. i) Explain the concept of a data cube help in efficiently computing the summary statistics in this scenario? (10 marks) ii) Explain the company can compute a specific summary statistic, such as the total sales for the Electronics category in Region A for Q2 2025, using a data cube. (10 marks)	**20 Marks**	**L2**	**CO2**
		Or			
14.	**a.**	A company is evaluating the performance of its marketing campaigns using OLAP tools. The dataset contains information about **campaign type**, **duration**, **region**, and **sales generated**. **i)** Explain OLAP operation would be most useful to compare the sales generated by different campaign types across various regions? (10 marks)	**20 Marks**	**L2**	**CO2**

		ii) The company wants to see sales data only for **Campaign A** during the **first quarter of 2025** in the **East region**. Explain OLAP operation would help isolate this data? (10 marks)			

15.	**a.**	A dataset contains information about **customer purchase behavior**. The goal is to build a decision tree to predict whether a customer will buy a product (Yes/No). The features include **Age** (under 30, 30-50, over 50) and **Income** (low, medium, high). Here is the sample dataset: **CUSTOMER ID** / **AGE** / **INCOME** / **PURCHASE OR NOT** 101 / 25 / LOW / YES 102 / 38 / LOW / NO 103 / 29 / MEDIUM / YES 104 / 56 / MEDIUM / YES 105 / 72 / HIGH / NO 106 / 44 / HIGH / YES 107 / 62 / LOW / NO Sketch a Decision tree by using Gini index method.	**20 Marks**	**L3**	**CO3**
		Or			
16.	**a.**	A company is using a **Naive Bayes classifier** to predict whether an email is **Spam** or **Not Spam** based on the following features: • **Feature 1: Presence of the word "free"** (Yes/No) • **Feature 2: Presence of the word "offer"** (Yes/No) The dataset is as follows: **E_MAIL NO** / **WORD "FREE"** / **WORD "OFFER"** / **SPAM(YES/NO)** 1 / YES / YES / YES 2 / NO / YES / YES 3 / YES / NO / NO 4 / NO / NO / NO 5 / YES / YES / NO 6 / YES / NO / YES **i)** Calculate the **prior probabilities** for the two classes (Spam and Not Spam). (10 marks) **ii)** Given an email with the words "free" and "offer" present, compute the **posterior probability** for each class (Spam and Not Spam) using the **Naive Bayes theorem**. (10 marks)	**20 Marks**	**L3**	**CO3**

17.	**a.**	Explain in detail about Proximity-based approaches.	**20 Marks**	**L2**	**CO4**
		Or			
18.	**a.**	A bank is analyzing the **monthly spending patterns** of its credit card customers. Most customers spend between **₹10,000 and ₹50,000** per month. However, one customer's record shows a monthly spend of **₹8,00,000**. Explain in detail about data mining technique should the bank use to determine whether this value is an outlier, and why?	**20 Marks**	**L2**	**CO4**