BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: B.TECH.
BRANCH: CS

SEMESTER: V/ADD
SESSION: MO/2025

SUBJECT: IT335 DATA MINING CONCEPTS AND TECHNIQUES

TIME: 3 Hours

FULL MARKS: 50

INSTRUCTIONS:

1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
4. Before attempting the question paper, be sure that you have got the correct question paper.
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

			CO	BL
Q.1(a)	Compare and contrast descriptive and predictive data mining tasks with suitable real-life examples. Given the following price dataset: 8, 16, 18, 30, 42, 42, 48, 50, 52, 56, 58, 68, and a bin size of 3; perform data discretization by creating equal-depth bins. Then apply smoothing using bin means and bin boundaries.	[5]	1	3
Q.1(b)	(i) How do attribute selection and dimensionality reduction differ from each other and in what ways are they similar? (ii) A market researcher collected data from 100 customers to examine whether **'gender'** is associated with **'purchase decision'** for a new product. Out of 40 males, 30 purchased the product. Whereas, out of 60 females, 40 did not purchased the product. Using a **chi-square test of independence** at a **5% significance level** (α = **0.05**) and **1 degree of freedom**, determine whether gender and purchase decision are correlated, justify your answer. Given, chi-square value at 0.05 significant level for 1 degree of freedom is 3.841.	[2+3 = 5]	1	2, 3
Q.2(a)	Discuss the three-tier data warehouse architecture with appropriate diagram.	[5]	2	2
Q.2(b)	GlobalTech Retail is a company that sells electronics and home appliances through different sales channels. The company wants to build a **data warehouse** to analyze its sales performance. The management wants to understand how sales are affected by **product categories**, **promotions**, and **sales channels**. You are given the following information: • The facts or the measures in the data warehouse are Total Revenue, Quantity Sold, Unit Cost. • The dimension Products follows a hierarchy Department -> Category -> Product • Each level has its own attributes such as: Department (Dept_ID, Dept_Name), Category (Catg_Id, Catg_Name, Description, Dept_ID), Product (Prod_ID, Prod_Name, Brand, Weight, Catg_Id) • The dimension Promotion has the attributes (Promo_Id, Promo_Name, Promo_Type, Start_Date, End_Date) & dimension Date has the attributes (Date_Id, Day, Month, Quarter, Year) • The dimension Channel has attributes (Channel_Id, Channel_Name). If Channel_Name is Online Store then a Sub-Dimension Shipping_Method (Shipping_Key, Shipping_Method) is associated to the dimension Channel. (i) State the type of suitable data warehouse schema to represent the above mentioned scenario to reduce data redundancy and better data integrity and justify your answer over the other schemas. (ii) Draw the best suitable schema diagram for the given description.	[2+3=5]	3, 5	3, 5
Q.3(a)	How can partitioning method improve the efficiency of Apriori based association mining? Which statistical measurement can be used to relate correlation and confidence of two item sets X, Y and how it can be verified?	[3+2=5]	3	2

PTO

Q.3(b)	Consider a small grocery shop, during the market-basket analysis 8 transactions (T1-T8) are considered. Use FP Growth algorithm to answer the questions given below. Consider the minimum support count nearly 40% of the total transaction. The transactions are given below: T1: {Milk, Bread, Butter} T2: {Bread, Diapers, Soap, Eggs} T3: {Milk, Diapers, Soap, Cola} T4: {Bread, Milk, Diapers, Soap} T5: {Bread, Milk, Diapers, Cola} T6: {Bread, Butter} T7: {Milk, Bread, Diapers, Soap} T8: {Bread, Cola} (i) Find out the frequent item sets of size 1. Create the FP-Tree following the FP Growth algorithm. (ii) From the FP-Tree find out the Conditional Pattern Bases for Cola and Soap.	[5]	3, 4	3, 5
Q.4(a)	Compare and contrast Regression and Classification approaches with appropriate examples. A node has 80% of samples from class A and 20% from class B. What is the Gini impurity and what does the value imply?	[3+2=5]	3	3, 4
Q.4(b)	(i) What is the key assumption behind the Naïve Bayes classifier? What problem does Laplace smoothing solve in Naïve Bayes classification? (ii) Consider a tuple t1 in the dataset 'Websites' and class labels are given as 'Phishing' or 'Not Phishing'. Given, P(Phishing) = 0.3, P(url_link | Phishing) = 0.9, P(url_link | Not Phishing) = 0.2. Classify whether the tuple t1 is Phishing or Not Phishing?	[3+2=5]	3, 4, 5	2, 3
Q.5(a)	(i) The convergence criteria of k-Means clustering algorithm is represented by the squared-error criterion which is mathematically represented as: $$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$ Explain the expression and each term used in the context of k-Means clustering algorithm. (ii) Given a partially constructed dendrogram where points A and B merge at a height 1; point C joins the merged cluster {A, B} at height 4. Points D and E merged at a height 2. Finally, the cluster {A, B, C} merges with the cluster {D, E} at height 7. Draw the dendrogram based on the given information and explain whether A and C are more similar or D and E are more similar.	[2+3=5]	3,4	2, 3
Q.5(b)	You are given the following data set in a 2-D feature space: A1(2,10), A2(2,5), A3(8,4), A4(5,8), A5(7,5), A6(6,4), A7(1,2), A8(4,9) Assume that there are three groups in the feature space with initial centroids at A1, A4 and A7. Use Manhattan distance (taxicab) as a proximity measure. Show the full procedure for **two complete iterations** of k-Means algorithm. Show the detailed steps for each iteration along with the recomputing cluster centroids. State the final cluster assignments and the centroids.	[5]	3,4	3, 4

:::::24/11/2025:::::M