(T) Roll No. ....283.....

PAPER ID—11134

B.Tech. EXAMINATION, 2024

(Fifth Semester)

COMPUTER SCIENCES

(Specialization Artificial Intelligence)

Data Mining and Warehousing

Time : 3 Hours *Maximum Marks : 70*

Before answering the question-paper candidates should ensure that they have been supplied to correct and complete question-paper. No complaint, in this regard, will be entertained after the examination.

Note : Attempt *Five* questions in all, selecting *one* question from each Unit. Q. No. **1** is compulsory. All questions carry equal marks.

(Compulsory Question)

1. Attempt all the subparts. Each subpart is of 2 marks : $2 \times 7 = 14$

   (a) How are data warehousing and data mining related ?

   (b) What is meant by data processing ?

   (c) What is market basket analysis and how can it benefit a supermarket ?

   (d) What are the advantages and disadvantages of decision tress over other classification methods ?

   (e) Name the methods used for dimensionality reduction.

   (f) Differentiate between precision and recall.

   (g) Explain anti-monotone property of support.

Unit I

2. (a) The frequency distribution of weight in grams of product of a given variety is given below. Calculate arithmetic mean and median.

Weight (grams)	Frequency
50-60	5
60-70	8
70-80	12
80-90	15
90-100	10

(b) A company is analyzing the relationship between the number of advertisements placed (X) and the corresponding sales revenue (Y) in thousands of dollars. The data collected over 8 months is as follows :

Number of Ads (X)	Sales Revenue (Y)
1	10
2	12
3	15
4	18
5	21
6	24
7	28
8	32

(i) Fit a regression line of the form $Y = a + bX$, where a is the intercept and b is the slope.

(ii) Predict the sales revenue if 10 advertisements are placed.

6+8=14

3. (a) The heights (in cm) of 8 students are : 150, 152, 155, 157, 160, 162, 165 and 168. Find the range, quartiles (Q1, Q2, Q3), and interquartile range (IQR).

(b) The following data are given for two companies combining the data for groups of male and female employees. Find out :

(i) The combined mean productivity for each company.

(ii) The combined variance for each company.

(iii) Which company has a higher average productivity per employee ?

(iv) Which company has more consistent productivity based on variance ?

Company	Group	Number of Employees	Mean Productivity	Variance
A	Male	50	80	16
	Female	30	85	25
B	Male	40	78	20
	Female	20	90	30

4+10=14

Unit II

4. (a) How can we handle missing values. Explain different techniques to handle missing values.

(b) Give brief description of the following :

(i) Binning

(ii) Outlier analysis

(iii) Generalization

(iv) Aggregation **6+8=14**

5. Differentiate between feature selection and feature extraction in data pre-processing. Explain different techniques of feature selection in detailed manner with suitable example. 14

Unit III

6. What is Association Rule Mining. A supermarket tracks customer purchases and has the following transaction data :

Transaction ID	Items Purchased
1	Bread, Milk, Butter
2	Bread, Milk
3	Milk, Butter
4	Bread, Butter
5	Bread, Milk, Butter, Cheese
6	Bread, Cheese
7	Milk, Butter, Cheese
8	Bread, Milk, Butter

Based on the transaction data and using Apriori algorithm to :

(i) Calculate the Support, Confidence, and Lift for the association rule : (Bread → Butter).

(ii) Determine if the rule is strong by comparing the confidence to a minimum threshold of 60%.

(iii) Suggest other additional strong rule (if any) based on the data, with support and confidence values above 40% and 60%, respectively.

14

7. (a) List out the differences between OLTP and OLAP.

(b) What is a data warehousing ? Explain various characteristics of data warehouse.

(c) Explain the process of combining and merging datasets. 4+6+4=14

Unit IV

8. Discuss in detail about ETL Phase 2 in data warehousing. 14

9. You are tasked with preparing a dataset for an ETL process. The dataset consist of certain issues like Missing values in key columns, Inconsistent date formats (e.g., MM/DD/YYYY and DD-MM-YYYY), Duplicated rows. Describe the step-by-step data wrangling process you would apply to clean this dataset before loading it into the data warehouse. Discuss the tools and techniques you would use and justify your choices. 14