QP Code: R251B035 | Reg. No | | | | | | | | | | | AY 24

GANDHI INSTITUTE OF ENGINEERING AND TECHNOLOGY UNIVERSITY, ODISHA, GUNUPUR

(GIET UNIVERSITY)

M.Tech. (First Semester) Regular Examinations, February – 2025

24MCSPE11021 – Data Mining and Data Warehousing

(CSE)

Time: 3 hrs | Maximum: 60 Marks

Answer ALL questions

(The figures in the right-hand margin indicate marks)

PART – A **(2 x 5 = 10 Marks)**

		CO #	Blooms Level
Q.1.	Answer ***ALL*** questions		
a.	Define data preprocessing steps.	CO1	K1
b.	Define Support and Confidence in Association Rule Mining.	CO2	K1
c.	Why is tree pruning useful in Decision Tree Induction?	CO3	K2
d.	Find the cosine similarity between the given two-term frequency vectors: X= [3,2,0,5,0,0,0,2,0,0] Y= [1,0,0,0,0,0,0,1,0,2]	CO4	K3
e.	Write short notes on outlier detection.	CO5	K2

PART – B **(10 x 5 = 50 Marks)**

		Marks	CO #	Blooms Level
	Answer *ALL* the questions			
2. a.	Describe data cleaning, integration, and transformation in data preprocessing.	10	CO1	K2
	(OR)			
b.	A company's HR department is analyzing employee salaries (in ₹1000s) across different departments. The salaries of employees in two departments, A and B, are given below: Salaries in Department A: 45, 50, 55, 60, 65, 70, 75, 80, 85, 90 Salaries in Department B: 48, 52, 58, 62, 68, 72, 78, 82, 88, 95 (i) Compute the Mean, Median, and Mode for each department separately. (ii) Compute the Combined Mean Salary of both departments. (iii)Analyze which department has a higher dispersion in salaries.	10	CO1	K3
3.a.	Given the following transaction database, find frequent itemsets and Generate Strong Association Rules using the Apriori algorithm (Minimum Support = 50% and Minimum Confidence= 60%). T1: {Bread, Milk, Butter} T2: {Bread, Diaper, Beer, Eggs} T3: {Milk, Diaper, Beer, Coke} T4: {Bread, Milk, Diaper, Beer} T5: {Bread, Milk, Diaper, Coke}	10	CO2	K3
	(OR)			
b.	A retail store wants to find frequent itemsets from its transaction database using the FP-Growth Algorithm. The store provides the following transaction data: T1: {A, B, D, E} T2: {B.C.E} T3: {A, B, C, E} T4: {B, E} Using minimum support = 2, perform the following: **(i)** Construct the FP-Tree. **(ii)** Extract frequent itemsets using the FP-Growth algorithm.	10	CO2	K3

4.a. Differentiate between OLAP and OLTP. 5 CO3 K4

b. Explain different types of normalization techniques with examples. 5 CO3 K2

(OR)

c. A retail company maintains a data warehouse to analyze sales data. The warehouse contains information about Products, Customers, Sales, and Time. The company wants to design a schema to efficiently store and analyze this data.

(i) Design a Star Schema for the given data warehouse with the following tables:

- Fact Table: Sales_Fact (Stores transactional sales data)
- Dimension Tables: 10 CO3 K6
  - Product_Dim (Product details)
  - Customer_Dim (Customer details)
  - Time_Dim (Date, Month, Year)
  - Store_Dim (Store location details)

(ii) Design a Snowflake Schema by normalizing at least one of the dimension tables.

5.a. (i) What is the Entropy of this collection of training samples with respect to target function classification?

(ii) What is the Information Gain of a1 and a2 relative to these training examples?

(iii) Draw a Decision Tree for the given Datasets.

Instance	Classification	a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

10 CO4 K3

(OR)

b. Given the following data points: (2, 3), (3, 3), (8, 8), (9, 9), (10, 10)
Apply the K-Means Algorithm with K = 2, using initial centroids as (2,3) and (8,8). Find the final clusters after two iterations. 10 CO4 K3

6.a. Differentiate between Agglomerative Clustering and Divisive Clustering. 5 CO5 K4

b. Explain any 2 of the following:
(i) Web Content Mining (ii) Web Structure Mining (iii) Web Usage Mining 5 CO5 K2

(OR)

c. A bank wants to classify loan applications as Approved or Rejected based on the following dataset:

Age	Income	Loan Status
Young	High	Approved
Young	Low	Rejected
Middle	High	Approved
Old	Medium	Approved
Old	Low	Rejected

10 CO5 K3

A new applicant has Age = Young and Income = Medium. Using the Naïve Bayes Classifier, predict whether the loan will be Approved or Rejected.

--- End of Paper ---