Exam.Code:1005
Sub. Code: 35083

2125

M. E. (Information Technology)

First Semester

MEIT-1203: Data Mining and Analytics

Time allowed: 3 Hours | Max. Marks: 50

NOTE: Attempt five questions in all, including Question No. 1 which is compulsory and selecting two questions from each Unit.

x-x-x

I. Answer the following;-

(a) Define granularity in a data warehouse. Why does lower granularity increase storage cost?

(b) State one situation where tuple merging is preferred over attribute smoothing during data cleaning.

(c) For a numeric attribute with strong right skew, which discretization method is more suitable—equal-width or equal-frequency? Why?

(d) State one advantage of performing roll-up before mining descriptive statistics.

(e) Given Support(A)=0.3, Support(B)=0.2, and Support(A∪B)=0.06, compute the confidence of A→B.

(f) Why is naive Bayes often preferred when attributes are high-dimensional but sparse?

(g) State one real-world example of embedded data mining.

(h) What is the commercial advantage of using data marts instead of a large enterprise warehouse?

(i) In decision tree induction, when is Gini index preferred over entropy?

(j) Which preprocessing method is more appropriate for handling outliers: binning or normalization? Justify briefly. (10×1)

UNIT - I

II. A fact table stores Sales data with the following dimensions:

Time (Day → Month → Quarter → Year)

Store (Store → City → State)

Contd……P/2

Product (Item → Category)

Given the following aggregated values:

Year	State	Category	Total Sales (₹)
2024	Punjab	Electronics	18,00,000
2024	Punjab	Clothing	9,60,000
2024	Haryana	Electronics	12,50,000

(a) Draw a 3-D data cube structure for these dimensions.

(b) Perform the following OLAP operations:

(i) Roll-up from State → Region (North India)

(ii) Slice the cube for Category = "Electronics".

(c) Identify the type of schema (Star/Snowflake) if the dimension tables are fully normalized. Explain. (10)

III. A data scientist is preparing a large customer behavior dataset that contains missing values, duplicated records, varying scales across numerical attributes, and categorical attributes with inconsistent levels (e.g., "Delhi", "DEL", "New Delhi"). The dataset also needs to be reduced for faster mining and generalized for later descriptive analysis.

Explain, with justification, how the data scientist should design a four-stage preprocessing pipeline that includes: Data Cleaning, Data Integration, Data Transformation, and Data Reduction or Discretization.

Your answer must clearly describe:

a) The specific technique chosen at each stage (e.g., smoothing, tuple merging, schema alignment, normalization variant, PCA, binning, etc.)

b) Why that technique is the most suitable for the given dataset conditions?

c) How the final processed dataset becomes more suitable for mining tasks like classification or clustering? (10)

IV. A company collected the following summarized data for two customer segments:

Attribute	Segment X (n=40)	Segment Y (n=60)
Avg. Purchase (₹)	2,400	1,700
Std. Dev	350	420

(a) Perform analytical characterization for Segment X (mean, std, 5-number summary—assume min=1500, Q1=2000, median=2400, Q3=2600, max=3100).

(b) Perform class comparison between Segment X and Y highlighting two major differences.

(c) Compute the z-score for a purchase amount of ₹2,900 in Segment X. (10)

UNIT - II

V. Consider the following market basket dataset:

Transactions (TID):

T1: {Bread, Butter, Milk}

T2: {Bread, Eggs}

T3: {Milk, Eggs, Bread}

T4: {Butter, Milk}

T5: {Bread, Butter}

Let min-support = 40%, min-confidence = 60%.

(a) Using Apriori, compute all frequent 1-item sets and 2-item sets (show support counts).

(b) Generate one strong association rule from these item sets.

(c) Compute lift for the rule and interpret. (10)

VI. A dataset has 100 samples. A decision tree uses attribute "Credit Score" (Good/Bad). The class distribution is as follows:

Credit Score	Approved	Rejected	Total
Good	32	8	40
Bad	18	42	60
Total	50	50	100

(a) Compute information gain for the split on "Credit Score."

(b) Calculate Gini Index for the attribute Credit Score (Good/Bad) using the given class distribution. Compute Gini Index for each class and Weighted Gini After Split. (10)

VII. a) Explain any two commercial benefits that organizations gain by integrating Data Mining (DM) and Data Warehousing (DW) into their operations.

b) Describe any two real-world applications of data mining in modern business processes.

c) What is embedded data mining? Give one example.

d) Mention any two emerging research areas in data mining.

e) Why organizations increasingly rely on DM for strategic decision-making? (5x2)

x-x-x