Predicted Question Paper - December 2025

Data Warehousing and Mining

Mumbai University - BE Computer Engineering/AIDS - Semester V

Time: 3 hours
Max. Marks: 80

Instructions:

1. Question 1 is compulsory.
2. Attempt any 3 questions out of the remaining.
3. Assume suitable data if required.

Q.1 (Compulsory - Attempt All) [20 Marks]

(a) Explain the features of Data Warehouse. Why is it important for business intelligence? [5 Marks]

(b) What is Market Basket Analysis? Explain with a real-world example from e-commerce. [5 Marks]

(c) Draw and explain the Knowledge Discovery in Databases (KDD) process with a detailed diagram showing all phases. [5 Marks]

(d) Calculate Accuracy, Precision, Recall, and F1-Score for the following confusion matrix: [5 Marks]

	Predicted Positive	Predicted Negative
Actual Positive	65	15
Actual Negative	20	50

Q.2 [20 Marks]

(a) Explain Star Schema and Snowflake Schema with suitable examples. A retail company wants to analyze sales data across different stores, products, time periods, and customers. Design an appropriate schema. [10 Marks]

(b) Explain the following OLAP operations with examples on a sales data cube: [10 Marks]

1. **Roll-up** (aggregation)
2. **Drill-down** (detailed view)
3. **Slice** (selection)
4. **Dice** (sub-cube)
5. **Pivot** (rotation)

Q.3 [20 Marks]

(a) Consider the following transaction database from an online grocery store: [10 Marks]

Transaction_ID	Items Purchased
T001	Milk, Bread, Eggs, Butter
T002	Milk, Bread, Cheese, Yogurt
T003	Bread, Butter, Jam
T004	Milk, Eggs, Butter, Cheese
T005	Milk, Bread, Butter, Yogurt
T006	Eggs, Cheese, Yogurt
T007	Milk, Bread, Eggs, Cheese
T008	Bread, Butter, Jam, Cheese

Apply **Apriori Algorithm** with minimum support = 40% and minimum confidence = 70% to:

1. Find all frequent itemsets (show L1, L2, L3...)
2. Generate strong association rules with their support and confidence values

(b) What is data preprocessing? Why is it essential before data mining? Explain any three data cleaning techniques with examples. [10 Marks]

Q.4 [20 Marks]

(a) What is clustering? Explain the K-means clustering algorithm with a flowchart. Apply K-means clustering on the following customer age dataset to segment them into 3 groups (K=3): [10 Marks]

Customer Ages: {22, 25, 28, 32, 35, 38, 42, 45, 48, 52, 55, 58}

Initial Centroids: C1 = 22, C2 = 42, C3 = 58

Show all iterations with:

- Distance calculations (Euclidean distance)
- Cluster assignments
- New centroid calculations
- Continue until convergence

(b) The following dataset contains information about loan applicants: [10 Marks]

ID	Age	Income	Credit_Score	Has_Loan	Default
1	Young	High	Good	No	No
2	Young	High	Good	Yes	No
3	Middle	High	Good	No	No
4	Senior	Medium	Good	No	Yes
5	Senior	Low	Poor	No	Yes

ID	Age	Income	Credit_Score	Has_Loan	Default
6	Senior	Low	Poor	Yes	No
7	Middle	Low	Good	Yes	No
8	Young	Medium	Poor	No	Yes
9	Young	Low	Poor	Yes	No
10	Senior	Medium	Poor	Yes	Yes
11	Young	Medium	Good	Yes	No
12	Middle	Medium	Good	No	No
13	Middle	High	Poor	Yes	No
14	Senior	Medium	Good	No	No

Using **Naive Bayesian Classification**, predict whether a new applicant will default:
New Applicant: (Age = Young, Income = Medium, Credit_Score = Good, Has_Loan = Yes)

Show all probability calculations step by step.

Q.5 [20 Marks]

(a) Explain the Decision Tree Induction algorithm for classification. Discuss how Information Gain is calculated and used to select the best attribute for splitting. [10 Marks]

For the dataset below, calculate Information Gain for each attribute:

Outlook	Temperature	Humidity	Wind	Play_Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes

Which attribute should be selected as the root node?

(b) What is web mining? Explain web content mining and web structure mining in detail with real-world applications. [10 Marks]

Q.6 [20 Marks]

(a) Explain the HITS (Hyperlink Induced Topic Search) algorithm for web page ranking. Given the following simplified web graph, calculate hub and authority scores for all pages after 2 iterations: [10 Marks]

Web Graph:

- Page A links to: B, C
- Page B links to: C, D
- Page C links to: A, D
- Page D links to: A

Initial hub score = 1, Initial authority score = 1 for all pages.

Show complete calculations.

(b) Explain the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm with an appropriate diagram. Discuss its advantages over K-means clustering and explain the role of parameters Eps (epsilon) and MinPts. [10 Marks]

Q.7 [20 Marks]

(a) Explain the ETL (Extract, Transform, Load) process in detail with a comprehensive diagram. Discuss various transformation operations performed during the ETL process. [10 Marks]

(b) Explain the following concepts with examples: [10 Marks]

1. **Multilevel association mining** - Show how rules can be discovered at different concept hierarchy levels
2. **Multidimensional association mining** - Explain inter-dimension and hybrid-dimension association rules
3. **Data normalization techniques** - Min-Max normalization and Z-score normalization

END OF PAPER

Important Note: This predicted paper is based on the analysis of previous year question papers (2022-2025) and covers the most frequently asked topics. Focus on:

- Apriori Algorithm (appears in almost every paper)
- K-means Clustering (very high frequency)
- Naive Bayesian Classification
- Data Warehouse Schemas
- OLAP Operations
- ETL Process
- HITS Algorithm
- Decision Trees
- KDD Process

Prepared by: Nitin Gupta