Data Warehousing and Mining - Question Paper Analysis

Mumbai University - Previous Year Papers (2022-2025)

Executive Summary

This document presents a comprehensive analysis of Mumbai University's Data Warehousing and Mining question papers from December 2022 to May 2025. The analysis identifies the **top 30 most important questions** based on frequency of occurrence and their importance in the curriculum. Additionally, three practice papers and one predicted paper for December 2025 are included.

Paper Format Overview

Examination Details:

- **Duration:** 3 hours
- **Maximum Marks:** 80
- **Paper Structure:**
  - Question 1 is **compulsory** (4 sub-questions × 5 marks = 20 marks)
  - Attempt **any 3 questions** from remaining 5 questions (Q2-Q6)
  - Each remaining question has 2 sub-parts (10 marks each)
  - Total: 20 + (3 × 20) = 80 marks

Topic Frequency Analysis

Based on analysis of 6 previous year papers (86 total questions), the following topics appear most frequently:

Topic	Frequency	Priority Level
Association Mining	19 times	**Very High**
Data Warehouse	10 times	**Very High**
Clustering	9 times	**High**
Data Preprocessing	8 times	**High**
Web Mining	7 times	**High**
Classification	6 times	**High**
Evaluation Metrics	4 times	**Medium**
Prediction	2 times	**Medium**
KDD Process	2 times	**Medium**

Top 30 Most Important Questions

A. Data Warehouse (Very High Priority)

1. **Explain features of Data Warehouse**
   - Repeated in: Dec 2022, Dec 2023, May 2025
2. **Explain Star Schema and Snowflake Schema with example**
   - Repeated in: May 2023, May 2025
3. **Explain different OLAP operations (Roll-up, Drill-down, Slice, Dice, Pivot)**
   - Repeated in: May 2024, Dec 2023, May 2025
4. **Explain ETL (Extract, Transform, Load) process in detail**
   - Repeated in: Dec 2024, Dec 2022, Dec 2023
5. **Explain three-tier data warehouse architecture**
   - Repeated in: Dec 2024
6. **Define Metadata and explain types of metadata**
   - Repeated in: Dec 2023
7. **Why does every data structure in data warehouse contain time element?**
   - Repeated in: May 2023

B. Association Mining (Very High Priority)

8. **Apply Apriori algorithm with given min-support and min-confidence**
   - **Most frequent question** - Appears in almost every paper
   - Different datasets and thresholds used
9. **What is Market Basket Analysis? Explain with example**
   - Repeated in: Dec 2024, May 2023, Dec 2022, May 2025
10. **Explain Support, Confidence, and Lift in association mining with formula**
    - Repeated in: Dec 2024, May 2024, May 2025
11. **Explain multilevel association mining with example**
    - Repeated in: Dec 2024, May 2023, Dec 2022
12. **Explain multidimensional association mining with example**
    - Repeated in: Dec 2023, May 2023, Dec 2022

C. Clustering (Very High Priority)

13. **Explain K-means clustering algorithm and apply on dataset**
    - **Second most frequent question** - Appears in 5 out of 6 papers
    - Different datasets: {2,4,10,12,3,20,30,11,25}, {6,14,18,22,1,40,50,11,25}
14. **Explain DBSCAN algorithm with appropriate diagram**
    - Repeated in: May 2023, Dec 2022, May 2025
15. **What is clustering? Explain hierarchical clustering (Agglomerative/Divisive)**
    - Repeated in: Dec 2023, May 2024
16. **Apply hierarchical clustering and draw dendrogram using single linkage**
    - Repeated in: Dec 2024, May 2024
17. **Explain BIRCH algorithm with diagram**
    - Repeated in: May 2024
18. **Explain CLARANS extension in web mining**
    - Repeated in: Dec 2024

D. Classification (Very High Priority)

19. **Explain Naive Bayesian classification with example and classify new tuple**
    - Repeated in: May 2023, Dec 2023
    - Usually given dataset with attributes: Age, Income, Student, Credit_rating
20. **Explain Decision Tree classification with Information Gain**
    - Repeated in: Dec 2024, May 2024
21. **Define classification and issues of classification**
    - Repeated in: Dec 2023

E. Data Preprocessing (High Priority)

22. **What is data preprocessing? Explain data cleaning techniques**
    - Repeated in: May 2023, Dec 2022
23. **Explain different data normalization techniques (Min-Max, Z-score, Decimal scaling)**
    - Repeated in: May 2025
24. **Explain different data sampling techniques with example**
    - Repeated in: May 2023, Dec 2022, May 2025
25. **Explain data integration phase methods**
    - Repeated in: Dec 2023

F. Web Mining (High Priority)

26. **What is web mining? Explain web content mining in detail**
    - Repeated in: Dec 2022, May 2024, May 2025
27. **Explain HITS (Hyperlink Induced Topic Search) algorithm with example**
    - Repeated in: Dec 2024, May 2023, Dec 2023
28. **Explain web structure mining in detail**
    - Repeated in: Dec 2023

G. KDD Process (High Priority)

29. **Draw and explain KDD (Knowledge Discovery in Databases) process with diagram**
    - Repeated in: May 2023, Dec 2022, May 2025
30. **Explain Data Mining Architecture with diagram**
    - Repeated in: May 2024, Dec 2023

H. Evaluation Metrics (High Priority)

31. **Calculate Accuracy, Recall, and Precision given TP, TN, FP, FN**
    - Repeated in: May 2023, May 2024
    - Example: TP=50, TN=20, FP=20, FN=10
32. **Explain confusion matrix with example**
    - Repeated in: Dec 2022, May 2025

I. Additional Important Questions (Medium Priority)

33. **What is prediction? Explain Linear Regression method**
    - Repeated in: Dec 2024, May 2024
34. **Compute dissimilarity between objects: Nominal and Asymmetric binary attributes**
    - Repeated in: Dec 2024, May 2024
35. **Calculate mean, median, mode, midrange, variance of data**
    - Repeated in: Dec 2023
36. **What is outlier? Explain different types of outliers**
    - Repeated in: May 2024

Study Strategy

Must-Know Topics (Study First):

1. **Apriori Algorithm** - Practice with different datasets
2. **K-means Clustering** - Must know the complete algorithm with numerical examples
3. **Data Warehouse Schemas** - Star and Snowflake
4. **OLAP Operations** - All 5 operations with examples
5. **ETL Process** - Complete understanding required
6. **Naive Bayesian Classification** - With probability calculations
7. **Decision Tree** - Information Gain formula
8. **Market Basket Analysis** - Support, Confidence, Lift formulas
9. **HITS Algorithm** - Hub and Authority scores
10. **KDD Process** - Complete diagram with all phases

Important Topics (Study Second):

- DBSCAN Algorithm
- Hierarchical Clustering with Dendrogram
- Data Preprocessing Techniques
- Web Content Mining
- Data Warehouse Features
- Three-tier Architecture
- Evaluation Metrics (Accuracy, Precision, Recall)

Good-to-Know Topics (If Time Permits):

- BIRCH Algorithm
- CLARANS
- Data Sampling Techniques
- Statistical Measures
- Outlier Detection
- Data Normalization

Tips for Examination

1. **Question 1 Strategy:**
   - Compulsory question with 4×5 marks
   - Usually covers 4 different topics
   - Keep answers concise (half page each)
   - Include diagrams where asked
2. **Main Questions Strategy:**
   - Each question worth 20 marks (2 parts of 10 marks each)
   - Choose questions where you can answer both parts well
   - Always include examples and diagrams
   - For algorithms, show step-by-step working
3. **Time Management:**
   - Question 1: 30-35 minutes
   - Each main question: 35-40 minutes
   - Reserve 15 minutes for review
4. **Scoring Tips:**
   - Algorithms: Show complete working with all iterations
   - Definitions: Write 2-3 lines minimum
   - Diagrams: Always label properly
   - Examples: Real-world examples score better
   - Formulas: Write and explain each component

Common Datasets Used in Papers

For K-means/Clustering:

- {2, 4, 10, 12, 3, 20, 30, 11, 25} with K=2
- {6, 14, 18, 22, 1, 40, 50, 11, 25} with K=2
- {2, 4, 10, 12, 3, 20, 11, 25, 56, 23} with K=2

For Apriori Algorithm:

- Transaction tables with 4-5 transactions
- Min-support: 30%, 40%, 50%, 60%
- Min-confidence: 60%, 70%, 75%, 80%

For Naive Bayesian:

- AllElectronics dataset with attributes:
  - Age (Youth, Middle-aged, Senior)
  - Income (Low, Medium, High)
  - Student (Yes, No)
  - Credit_rating (Fair, Excellent)
  - Class: Buys_computer (Yes, No)

Prepared by: Nitin Gupta