Diploma Engineering

Laboratory Manual

(Data Mining and Warehousing)
(DI04016081)

[Information Technology (Sem-4)]

Enrollment No	
Name	
Branch	Information Technology
Academic Term	
Institute	

Directorate Of Technical Education
Gandhinagar - Gujarat

DTE's Vision:

- To facilitate quality technical and professional education having relevance for both industry and society, with moral and ethical values, giving equal opportunity and access, aiming to prepare globally competent technocrats.

DTE's Mission:

1. Quality technical and professional education with continuous improvement of all the resources and personnel
2. To promote conducive ecosystem for Academic, Industry, Research, Innovations and Startups
3. To provide affordable quality professional education with moral values, equal opportunities, accessibility and accountability
4. To allocate competent and dedicated human resources and infrastructure to the institutions for providing world-class professional education to become a Global Leader ("Vishwa Guru")

Institute's Vision:

Institute's Mission:

Department's Vision:

Department's Mission:

Certificate

This is to certify that Mr./Ms …….

Enrollment No. ……..…….. of …….…….. Semester of *Diploma in*……..

…...of ……. (GTU Code) has satisfactorily completed the term work in course ……for the academic year: …… Term: Odd/Even prescribed in the GTU curriculum.

Place: …….

Date: ……..

Signature of Course Faculty **Head of the Department**

Preface

The primary aim of any laboratory/Practical/field work is enhancement of required skills as well as creative ability amongst students to solve real time problems by developing relevant competencies in psychomotor domain. Keeping in view, GTU has designed competency focused outcome-based curriculum -2021 (COGC-2021) for Diploma engineering programmes. In this more time is allotted to practical work than theory. It shows importance of enhancement of skills amongst students and it pays attention to utilize every second of time allotted for practical amongst Students, Instructors and Lecturers to achieve relevant outcomes by performing rather than writing practice in study type. It is essential for effective implementation of competency focused outcome- based Green curriculum-2021. Every practical has been keenly designed to serve as a tool to develop & enhance relevant industry needed competency in each and every student. These psychomotor skills are very difficult to develop through traditional chalk and board content delivery method in the classroom. Accordingly, this lab manual has been designed to focus on the industry defined relevant outcomes, rather than old practice of conducting practical to prove concept and theory.

By using this lab manual, students can read procedure one day in advance to actual performance day of practical experiment which generates interest and also, they can have idea of judgement of magnitude prior to performance. This in turn enhances predetermined outcomes amongst students. Each and every Experiment /Practical in this manual begins by competency, industry relevant skills, course outcomes as well as practical outcomes which serve as a key role for doing the practical. The students will also have a clear idea of safety and necessary precautions to be taken while performing experiment.

This manual also provides guidelines to lecturers to facilitate student-centered lab activities for each practical/experiment by arranging and managing necessary resources in order that the students follow the procedures with required safety and necessary precautions to achieve outcomes. It also gives an idea that how students will be assessed by providing Rubrics.

Information technology is a modern phenomenon that has dramatically changed the daily lives of individuals and businesses throughout the world. Information system stores data in a sophisticated manner, making the process of finding the data much easier.

With an information system, delivering all the important information is easier to make better decisions. Therefore, the knowledge about the various applications areas of Information Technology including practical skills acquired through the laboratory will help students when he/she will be working with information systems and will be able to apply the concepts of IT systems as and when required.

Although we try our level best to design this lab manual, but always there are chances of improvement. We welcome any suggestions for improvement.

Programme Outcomes (POs):

1. **Basic and Discipline specific knowledge:** Apply knowledge of basic mathematics, science and engineering fundamentals and engineering specialization to solve the *engineering* problems.

2. **Problem analysis**: Identify and analyse well-defined *engineering* problems using codified standard methods.

3. **Design/ development of solutions:** Design solutions for *engineering* well-defined technical problems and assist with the design of systems components or processes to meet specified needs.

4. **Engineering Tools, Experimentation and Testing:** Apply modern *engineering* tools and appropriate technique to conduct standard tests and measurements.

5. **Engineering practices for society, sustainability and environment:** Apply appropriate technology in context of society, sustainability, environment and ethical practices.

6. **Project Management:** Use engineering management principles individually, as a team member or a leader to manage projects and effectively communicate about well-defined engineering activities.

7. **Life-long learning:** Ability to analyze individual needs and engage in updating in the context of technological changes *in field of engineering.*

Practical Outcome - Course Outcome matrix

	Course Outcomes (COs): *CO1: Understand the data mining process and its practical applications.* *CO2: Remember the basic concepts of data warehousing and differences between OLTP and OLAP.* *CO3: Apply basic techniques like classification, clustering, and association rule mining.* *CO4: Understand schema models, fact/dimension tables, and OLAP operations.* *CO5: Apply simple data mining tools to analyze datasets and observe results.*					
S. No.	**Practical Outcome/Title of experiment**	**CO1**	**CO2**	**CO3**	**CO4**	**CO5**
1.	Study of KDD (Knowledge Discovery in Databases) process with examples.	√				
2.	Identify and discuss real-life applications of data mining (group discussion / mini case study).	√				
3.	Study of Data Warehousing concepts through case studies/examples.		√			
4.	Compare OLTP and OLAP databases using simple SQL queries.		√			
5.	Perform Classification using Decision Tree in data mining tools.			√		
6.	Perform Classification using Naïve Bayes in data mining tools.			√		
7.	Apply Clustering using K-Means in data mining tools.			√		
8.	Apply Association Rule Mining using Apriori algorithm in data mining tools.			√		
9.	Simple project: Apply two techniques (Classification + Clustering) on a sample dataset and compare results.			√		
10.	Design a simple Star Schema for a sales database.				√	
11.	Design a Snowflake Schema for a student performance database.				√	
12.	Perform basic OLAP operations (Roll-up, Drill-down, Slice, Dice) using sample data in SQL.				√	
13.	Introduction and hands-on with Data Mining Tool.					√
14.	Perform a case study analysis (e.g., Market Basket, Student Result Analysis, or Healthcare data).					√
15.	Mini Project: Apply OLAP + one data mining technique on a dataset and present findings.					√

Industry Relevant Skills

The following industry relevant skills are expected to be developed in the students by performance of experiments of this course.

- Apply basic data mining techniques such as classification, clustering, and association rule mining.
- Understand schema models, fact and dimension tables, as well as various OLAP operations.
- Use basic data mining tools to analyze datasets and interpret the results.

Guidelines to Course Faculty

1. Couse faculty should demonstrate experiment with all necessary implementation strategies described in curriculum.
2. Couse faculty should explain industrial relevance before starting of each experiment.
3. Course faculty should Involve & give opportunity to all students for hands on experience.
4. Course faculty should ensure mentioned skills are developed in the students by asking.
5. Utilise 2 hrs of lab hours effectively and ensure completion of write up with quiz also.
6. Encourage peer to peer learning by doing same experiment through fast learners.

Instructions for Students

1. Organize the work in the group and make record of all observations.
2. Students shall develop maintenance skill as expected by industries.
3. Student shall attempt to develop related hand-on skills and build confidence.
4. Student shall develop the habits of evolving more ideas, innovations, skills etc.
5. Student shall refer technical magazines and data books.
6. Student should develop habit to submit the practical on date and time.
7. Student should well prepare while submitting write-up of exercise.

Continuous Assessment Sheet

Enrolment No: **Name**

Name: **Term:**

Sr. No.	Practical Outcome/Title of experiment	Page	Date	Marks (25)	Sign
1.	Study of KDD (Knowledge Discovery in Databases) process with examples.				
2.	Identify and discuss real-life applications of data mining (group discussion / mini case study).				
3.	Study of Data Warehousing concepts through case studies/examples.				
4.	Compare OLTP and OLAP databases using simple SQL queries.				
5.	Perform Classification using Decision Tree in data mining tools.				
6.	Perform Classification using Naïve Bayes in data mining tools.				
7.	Apply Clustering using K-Means in data mining tools.				
8.	Apply Association Rule Mining using Apriori algorithm in data mining tools.				
9.	Simple project: Apply two techniques (Classification + Clustering) on a sample dataset and compare results.				
10.	Design a simple Star Schema for a sales database.				
11.	Design a Snowflake Schema for a student performance database.				
12.	Perform basic OLAP operations (Roll-up, Drill-down, Slice, Dice) using sample data in SQL.				
13.	Introduction and hands-on with Data Mining Tool.				
14.	Perform a case study analysis (e.g., Market Basket, Student Result Analysis, or Healthcare data).				
15.	Mini Project: Apply OLAP + one data mining technique on a dataset and present findings.				

Date: ……

Practical No. 1: Study of KDD (Knowledge Discovery in Databases) process with examples.

Objectives

- To understand the concept and importance of Knowledge Discovery in Databases (KDD).
- To study each step involved in the KDD process.
- To observe how raw data gradually transforms into meaningful knowledge.
- To relate each step with suitable examples from real-world datasets.

Theory

Introduction to KDD

Knowledge Discovery in Databases (KDD) is a systematic, multi-stage process used to identify valid, novel, and useful patterns from large volumes of data. It acts like a guided journey where data travels through cleaning, preparation, mining, and interpretation until insights emerge.

Steps of KDD Process

1. Data Selection
   Choosing the relevant data from different sources.
   Example: Selecting sales records from an organization's database to analyze customer buying patterns.
2. Data Preprocessing (Cleaning)
   Removing noise, handling missing values, correcting inconsistencies.
   Example: Replacing blank entries in the "Age" column with mean age or removing duplicate records.
3. Data Transformation
   Converting data into suitable formats through normalization, aggregation, and encoding.
   Example: Converting categorical values like "High/Medium/Low" into numeric form (1, 2, 3).
4. Data Mining
   Applying algorithms to discover patterns, relationships, or clusters.
   Example: Using Apriori algorithm to find association rules such as
   "Customers who buy bread often buy butter."
5. Pattern Evaluation
   Identifying strong, interesting, and useful patterns from the results.
   Example: Choosing rules with high support and confidence values.
6. Knowledge Presentation
   Representing the discovered knowledge using charts, rules, or summaries.
   Example: Displaying customer segments on a bar chart for easier interpretation.

Software / Tools Required

- Any dataset (CSV or Excel format)
- Data mining tools such as:
  - Weka
  - RapidMiner
  - Orange
  - Python (optional—Pandas, Scikit-Learn)

Experimental Procedure

1. Load the selected dataset into the chosen data mining tool.
2. Perform data cleaning: remove missing, inconsistent, or duplicate data.
3. Apply data transformation such as normalization or encoding.
4. Choose a suitable data mining technique (classification, clustering, association).
5. Run the algorithm and observe the generated patterns or rules.
6. Evaluate the results based on metrics such as support, confidence, accuracy, or cluster quality.
7. Present the final knowledge using graphs, tables, or concise statements.
8. Record all observations in the observation table.

Observation Table

Step	Operation Performed	Tool Used	Result / Output
Data Selection	Loaded dataset	Weka	Dataset displayed
Data Cleaning	Removed 15 missing values	Weka – Preprocess	Cleaned dataset
Data Transformation	Normalized numeric attributes	Weka – Filter	Transformed data
Data Mining	Applied Apriori Algorithm	Weka – Associate	List of association rules
Pattern Evaluation	Selected top 3 rules	Manual evaluation	Strong rules identified
Knowledge Presentation	Visualized output	Graph/Report	Final patterns displayed

Result

The KDD process was successfully executed. Each step—from raw data selection to knowledge presentation—was observed. Useful patterns were discovered and interpreted, demonstrating how KDD transforms unprocessed datasets into actionable knowledge.

Conclusion

This practical helped in understanding the structured pipeline of KDD. By exploring each phase with examples, the journey from data to knowledge becomes clear and logical. The step-by-step execution shows how meaningful insights are extracted from real-world datasets.

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 2: Identify and discuss real-life applications of data mining (group discussion / mini case study).

Objectives

- To understand how data mining is applied in real-world domains.
- To analyze real-life case studies where data-driven insights support decision-making.
- To develop the ability to discuss, interpret, and present findings as a group.
- To connect theoretical data mining concepts with practical applications.

Theory

Introduction

Data mining is widely used across industries to uncover patterns, trends, and predictions from large datasets. It quietly powers many everyday conveniences—from personalized shopping suggestions to fraud alerts—making it a core element in data-driven decision-making.

Common Application Areas of Data Mining

1. Retail and E-commerce
    - Market basket analysis
    - Customer segmentation
    - Recommendation systems

    Example: Online stores suggesting products based on previous purchases.
2. Banking and Finance
    - Fraud detection
    - Credit risk analysis
    - Customer churn prediction

    Example: Flagging unusual transactions for possible fraud.
3. Healthcare
    - Disease prediction and diagnosis
    - Treatment effectiveness analysis
    - Patient clustering

    Example: Predicting the likelihood of diabetes based on patient records.
4. Telecommunications
    - Service usage analysis
    - Network optimization
    - Customer retention

    Example: Identifying customers likely to switch to another network.
5. Education
    - Student performance prediction
    - Learning behavior analysis

    Example: Using past marks and attendance to predict academic risk.
6. Manufacturing and Industry
    - Fault detection
    - Predictive maintenance

    Example: Predicting machine breakdowns using sensor data.

7. Marketing and CRM
    - Targeted marketing
    - Campaign effectiveness measurement
      Example: Tailoring advertisements to specific customer groups.
8. Social Media and Web Analytics
    - Sentiment analysis
    - Behavioural trend analysis
      Example: Identifying trending hashtags and user engagement patterns.

Software / Tools Required

- Presentation tools (PowerPoint/Charts)
- Any dataset (optional for mini case study)
- Whiteboard / Discussion space
- Internet resources for domain research

Experimental Procedure

A. For Group Discussion

1. Divide students into small groups (3–5 members).
2. Assign each group a domain (retail, healthcare, education, finance, etc.).
3. Each group identifies real-life examples of data mining in the assigned domain.
4. Discuss its purpose, type of data involved, and expected outcomes.
5. Prepare a short summary and present insights to the class.

B. For Mini Case Study

1. Select a real-life organization or scenario where data mining is used.
2. Describe the problem or objective.
3. Identify the dataset or variables involved.
4. Explain the data mining technique applied (classification, clustering, association, etc.).
5. Present the results, impacts, or benefits gained.
6. Document findings in the observation table.

Observation Table

Domain / Case Study	Problem Identified	Data Involved	Technique Used	Outcome / Insight
Retail	Improve product recommendations	Purchase history	Association Rule Mining	Better personalized suggestions
Banking	Detect fraudulent transactions	Transaction logs	Classification	Early fraud detection
Healthcare	Predict disease risk	Patient history	Classification / Clustering	Early diagnosis support

Domain / Case Study	Problem Identified	Data Involved	Technique Used	Outcome / Insight
Education	Identify at-risk students	Attendance, marks	Classification	Academic intervention

Result

Students explored and discussed multiple real-life applications of data mining. Through group discussions and case studies, they identified practical problems, analysed techniques used, and recognized how data mining leads to valuable insights across industries.

Conclusion

The activity demonstrated that data mining is not just a theoretical concept but a powerful real-world tool. Its applications illuminate hidden patterns in every domain—from healthcare to retail—strengthening decision-making and enhancing productivity.

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 3: Study of Data Warehousing concepts through case studies/examples.

Objectives

- To understand the fundamental concepts of data warehousing.
- To study real-life case studies demonstrating the use of data warehouses.
- To identify how organizations use data warehouses for decision-making.
- To relate theoretical concepts—schema models, fact/dimension tables, ETL—to practical scenarios.

Theory

Introduction to Data Warehousing

A data warehouse is a centralized, subject-oriented, time-variant, and non-volatile repository that stores historical data for analysis and reporting. It acts as a long-term memory of an organization, integrating information from multiple sources to support strategic decisions.

Key Concepts

1. ETL (Extract–Transform–Load)
    - Extracting data from multiple sources
    - Transforming it into a consistent format
    - Loading it into the warehouse
2. Schema Models
    - Star Schema: A central fact table linked to dimension tables
    - Snowflake Schema: Normalized version of a star schema
    - Fact Constellation: Multiple fact tables sharing dimensions
3. Fact Table
    - Stores measurable quantitative data
    - Examples: sales amount, total units, revenue
4. Dimension Table
    - Stores descriptive attributes
    - Examples: product, customer, location, time
5. OLAP Operations
    - Roll-up
    - Drill-down
    - Slice
    - Dice
    - Pivot

Case Studies / Examples

Case Study 1: Retail Chain (Supermarket)

Goal: Analyze sales trends and customer purchasing behaviour

- Sources: POS system, inventory database, online store
- ETL: Data is cleaned, merged, and loaded weekly
- Schema: Star schema
    - Fact: *Sales Fact* (quantity, amount, discount)
    - Dimensions: *Product, Customer, Store, Time*
- Insights:

  - Most purchased items during festivals
  - High-value customers by location
  - Slow-moving products

Case Study 2: Banking Sector
Goal: Improve financial reporting and risk analysis

- Sources: Loan management system, transaction logs, customer profiles
- Schema: Snowflake schema
  - Fact: *Loan Fact* (amount, interest, tenure)
  - Dimensions: *Customer, Branch, Loan Type, Time*
- Insights:
  - Regional trends in loan defaults
  - High-risk customer segments
  - Profitability by loan type

Case Study 3: Healthcare System
Goal: Analyze treatment effectiveness and patient outcomes

- Sources: Hospital records, lab tests, patient demographics
- Schema: Star schema
  - Fact: *Treatment Fact* (cost, duration, outcome score)
  - Dimensions: *Patient, Doctor, Diagnosis, Time*
- Insights:
  - Frequent diagnoses by season
  - Treatment success rates
  - Cost patterns for chronic diseases

Software / Tools Required

- Any data warehousing or OLAP tool (demonstration-level):
  - Microsoft SQL Server / SSIS / SSAS
  - MySQL with sample warehouse schema
  - Weka (for OLAP-style filters)
  - Power BI / Tableau (optional for visualization)

Experimental Procedure

1. Select one or more case studies from retail, banking, healthcare, or another domain.
2. Identify the business problem and required data sources.
3. Prepare a conceptual ETL flow for the case study.
4. Draw schema diagrams such as star, snowflake, or fact constellation models.
5. Identify fact and dimension tables with example attributes.
6. Demonstrate OLAP operations (roll-up, drill-down, slice, dice) on sample data.
7. Document insights or findings from the case study.

Observation Table

Component	Details Observed	Example (if any)
Business Domain	Retail / Banking / Healthcare	Retail
ETL Flow	Extract → Clean → Transform → Load	Weekly ETL cycle
Schema Model	Star / Snowflake	Star schema
Fact Table	Measures identified	Sales amount, units sold
Dimension Tables	Attributes listed	Product, Customer, Store
OLAP Operations	Observed results	Drill-down on monthly sales
Final Insights	Key patterns found	Festival-season peak sales

Result

The selected case studies were analyzed to understand data warehousing concepts such as ETL, schema modelling, fact/dimension tables, and OLAP operations. The study demonstrated how data warehouses help organizations make informed decisions.

Conclusion

This practical provided hands-on understanding of how data warehousing concepts are applied in real-world scenarios. Case studies revealed the importance of integrating diverse data sources, modelling them effectively, and using OLAP techniques to uncover meaningful business insights.

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 4: Compare OLTP and OLAP databases using simple SQL queries.

Objectives

- To understand the fundamental differences between OLTP (Online Transaction Processing) and OLAP (Online Analytical Processing) systems.
- To execute simple SQL queries that highlight typical OLTP operations and OLAP-style analytical operations.
- To study how query patterns differ between transactional and analytical systems.
- To observe execution characteristics such as data volume, query complexity, and output format.

Theory

OLTP (Online Transaction Processing)

OLTP systems are designed for high-volume, short, real-time transactions such as insertions, updates, and retrievals.
Examples: ATM transactions, online bookings, order entry systems.

Characteristics:

- Highly normalized tables
- Frequent read/write operations
- Millisecond-level response
- Handles large number of concurrent users

OLAP (Online Analytical Processing)

OLAP systems support data analysis, reporting, and decision-making.
Examples: Business intelligence dashboards, sales forecasting, trend analysis.

Characteristics:

- Denormalized tables (star/snowflake schema)
- Complex aggregate queries
- Read-intensive
- Used for historical data analysis

Typical Operations

OLTP	OLAP
Insert, Update, Delete	Roll-up, Drill-down
Retrieve single record	Aggregation (SUM, AVG, COUNT)
Transaction accuracy	Cube, Slice, Dice

OLTP	OLAP
Normalized schema	Fact & dimension tables

Software / Tools Required

- MySQL / PostgreSQL / SQL Server
- Sample OLTP schema (e.g., *orders*, *customers*)
- Sample OLAP/star schema (e.g., *sales_fact*, *product_dim*)
- SQL query editor / command-line client

SQL Queries for Comparison

A. OLTP-Oriented Queries

1. Insert Operation

```
INSERT INTO orders (order_id, customer_id, order_date, amount)
VALUES (101, 12, '2025-01-10', 2500);
```

2. Update Operation

```
UPDATE customers
SET phone = '9876543210'
WHERE customer_id = 12;
```

3. Retrieve Single Transaction

```
SELECT * FROM orders
WHERE order_id = 101;
```

4. Retrieve Customer Details

```
SELECT name, email
FROM customers
WHERE customer_id = 12;
```

Purpose: Demonstrates real-time data manipulation and quick record retrieval.

B. OLAP-Oriented Queries

Assume a star schema with:

- Fact table: sales_fact (product_id, time_id, store_id, quantity, amount)
- Dimension tables: product_dim, time_dim, store_dim

1. Total Sales by Product Category

```
SELECT p.category, SUM(f.amount) AS total_sales
FROM sales_fact f
```

```
JOIN product_dim p ON f.product_id = p.product_id
GROUP BY p.category;
```

2. Monthly Sales Summary

```
SELECT t.month, t.year, SUM(f.amount) AS monthly_sales
FROM sales_fact f
JOIN time_dim t ON f.time_id = t.time_id
GROUP BY t.year, t.month;
```

3. Sales by Region (Slice Operation)

```
SELECT s.region, SUM(f.quantity) AS total_quantity
FROM sales_fact f
JOIN store_dim s ON f.store_id = s.store_id
GROUP BY s.region;
```

4. Drill-Down: Year → Quarter → Month

```
SELECT t.year, t.quarter, t.month, SUM(f.amount) AS sales
FROM sales_fact f
JOIN time_dim t ON f.time_id = t.time_id
GROUP BY t.year, t.quarter, t.month;
```

Purpose: Shows how OLAP queries aggregate and analyze large historical datasets.

Experimental Procedure

1. Set up both OLTP-style and OLAP-style sample databases.
2. Execute basic OLTP queries: insert, update, delete, and simple select statements.
3. Execute OLAP queries using joins, grouping, and aggregation functions.
4. Compare:
    - Query complexity
    - Execution results
    - Volume of records processed
    - Type of output (single row vs aggregated report)
5. Record observations in the observation table.

Observation Table

Query Type	Example SQL Query	Database Type	Output Summary	Remarks
Insert	INSERT INTO orders…	OLTP	Row inserted successfully	Fast execution
Simple Select	SELECT * FROM orders…	OLTP	Single record	Suitable for transactions

Query Type	Example SQL Query	Database Type	Output Summary	Remarks
Aggregation	SUM(amount)… GROUP BY category	OLAP	Summary table	Slower, processes many rows
Drill-down	GROUP BY year, quarter, month	OLAP	Multi-level summary	Used for analysis

Result

OLTP and OLAP systems were compared through practical SQL queries. OLTP queries executed fast and handled individual transactions, whereas OLAP queries analyzed large datasets with complex aggregations. The comparison clearly highlighted the different purposes and behaviours of both systems.

Conclusion

This experiment demonstrated the structural and functional differences between OLTP and OLAP databases. SQL queries showed how OLTP supports operational tasks while OLAP supports analytical processing. Understanding this distinction is essential for designing efficient database systems.

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 5: Perform Classification using Decision Tree in data mining tools.

Objectives

- To understand the concept of classification in data mining.
- To learn how Decision Tree algorithms classify data based on attributes.
- To implement Decision Tree classification using Orange and RapidMiner tools.
- To interpret the generated tree, accuracy, and confusion matrix.

Theory

Classification

Classification is a supervised learning technique used to predict categorical outcomes (class labels).
Examples:

- Predicting whether an email is *spam* or *not spam*.
- Predicting whether a customer will *purchase* or *not purchase*.

Decision Tree Algorithm

A Decision Tree is a flowchart-like structure where:

- Internal nodes represent tests on attributes.
- Branches represent outcomes of the tests.
- Leaf nodes represent class labels.

Popular algorithms: ID3, C4.5, CART

Advantages:

- Easy to understand
- Handles both numeric and categorical data
- Offers clear visual representation

Software / Tools Required

- Orange Data Mining Tool
- RapidMiner Studio
- Sample dataset (e.g., *Iris*, *Titanic*, *Weather*, or custom CSV)

Experimental Procedure

A. Using Orange

1. Open Orange → Select New Workflow.
2. Drag and drop the following widgets:
    - File (to load dataset)

    - Data Table
    - Select Columns (optional)
    - Classification Tree
    - Test & Score
    - Confusion Matrix
    - Tree Viewer
3. Connect the widgets in sequence.
4. Load dataset using File widget (e.g., *Iris.csv*).
5. View data using Data Table.
6. Configure Classification Tree (depth, splits, impurity measure).
7. Connect Test & Score and observe accuracy, precision, recall, F1-score.
8. Open Confusion Matrix to study misclassifications.
9. Open Tree Viewer to visualize the generated decision tree.

B. Using RapidMiner

1. Open RapidMiner Studio → Create New Process.
2. From the Operators panel, drag:
    - Read CSV
    - Set Role (to define label/class attribute)
    - Decision Tree
    - Apply Model
    - Performance (Classification)
3. Import dataset using Read CSV.
4. Use Set Role to mark the target attribute as *label*.
5. Select Decision Tree and configure parameters (criterion, depth).
6. Connect components and run the process.
7. View:
    - Generated Decision Tree
    - Confusion Matrix
    - Overall Accuracy and Performance metrics

Observation Table

Tool Used	Dataset	Accuracy (%)	Tree Depth	No. of Leaves	Confusion Matrix Observation
Orange	Iris				
RapidMiner	Iris				

(Students fill based on software output.)

Sample Output Points to Note

- Decision Tree visualization (root node, branches, leaves)
- Accuracy values from Test & Score / Performance operator
- Correct and incorrect classifications
- Node conditions (e.g., *Petal Length < 2.45 → Setosa*)

Result

Decision Tree classification was successfully performed using Orange and RapidMiner. The generated trees were analyzed, and accuracy and confusion matrices were recorded. The experiment demonstrated how Decision Trees classify data and provide interpretable rules.

Conclusion

This practical provided hands-on experience with Decision Tree classifiers in two popular data mining tools. The visualization and evaluation features of Orange and RapidMiner helped in understanding model behaviour, attribute importance, and overall prediction performance.

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 6: Perform Classification using Naïve Bayes in data mining tools.

Objectives

- To understand the Naïve Bayes classification technique.
- To implement Naïve Bayes in Orange and RapidMiner.
- To study model accuracy, confusion matrix, and prediction behaviour.
- To compare how tools handle probabilistic classification.

Theory

Naïve Bayes Classifier

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, assuming independence among predictors.

Bayes' Theorem:

P(C|X)=P(X|C)·P(C)P(X)P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}P(C|X)=P(X)P(X|C)·P(C)

Where:

- CCC = class
- XXX = feature set

Why "Naïve"?
It assumes that all attributes contribute independently to class prediction, simplifying calculations.

Advantages of Naïve Bayes

- Fast and efficient
- Works well with high-dimensional data
- Performs well even with limited training data
- Simple to implement and interpret

Common Applications

- Email spam detection
- Sentiment analysis
- Medical diagnosis
- Document classification

Software / Tools Required

- Orange Data Mining Tool
- RapidMiner Studio

- Sample datasets: *Iris*, *Titanic*, *Weather*, *Pima Indians Diabetes*, or custom CSV.

Experimental Procedure

A. Using Orange

1. Open Orange → Create New Workflow.
2. Drag and connect the following widgets:
    - File
    - Data Table
    - Select Columns (optional)
    - Naïve Bayes
    - Test & Score
    - Confusion Matrix
    - Predictions (optional)
3. Load a dataset through the File widget.
4. View the dataset using Data Table.
5. Configure the Naïve Bayes widget (Gaussian/Multinomial options based on data).
6. Connect Test & Score to evaluate the model.
7. Explore accuracy, precision, recall, AUC values.
8. Open Confusion Matrix to view correct/incorrect predictions.
9. Use Predictions widget if you wish to display class-wise probabilities.

B. Using RapidMiner

1. Open RapidMiner Studio → Start a New Process.
2. Drag and place these operators:
    - Read CSV
    - Set Role (assign label attribute)
    - Naïve Bayes
    - Apply Model
    - Performance (Classification)
3. Import the dataset using Read CSV.
4. Use Set Role to mark the target column as *label*.
5. Drag Naïve Bayes and configure (Laplace correction, numerical handling).
6. Connect the operators in sequence.
7. Run the process.
8. View:
    - Predicted vs actual classes
    - Confusion matrix
    - Accuracy and model statistics

Observation Table

Tool Used	Dataset	Accuracy (%)	Precision	Recall	Confusion Matrix Notes
Orange					
RapidMiner					

(Students record values based on output.)

Sample Output Notes

- Probabilistic predictions for each class.
- Model accuracy (varies by dataset).
- Misclassified records identifiable in confusion matrix.
- Effect of Laplace correction (in RapidMiner).

Result

Naïve Bayes classification was successfully implemented using both Orange and RapidMiner. The model's performance was evaluated through accuracy metrics and confusion matrices, and predictions were observed for different instances.

Conclusion

This experiment demonstrated how Naïve Bayes—though simple in design—provides efficient and effective classification. Both Orange and RapidMiner tools offered intuitive workflows for building, evaluating, and analyzing the model's performance.

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 7: Apply Clustering using K-Means in Data Mining Tools.

Objective

- To understand the concept of clustering and its applications.
- To implement the K-Means clustering algorithm using Orange and RapidMiner.
- To visualize cluster formation and interpret clustering results.

Apparatus / Software Requirements

- Computer system with:
  - Orange Data Mining Tool (latest version)
  - RapidMiner Studio
- Sample dataset (e.g., *Iris*, *Mall Customers*, or any CSV dataset)

Theory

Clustering

Clustering is an unsupervised learning technique used to group data points into clusters based on similarity. No predefined labels are used; the algorithm discovers structure within the data.

K-Means Clustering

K-Means is a centroid-based clustering algorithm that:

1. Selects K cluster centers (centroids).
2. Assigns each data point to the nearest centroid.
3. Recalculates centroids based on mean of assigned points.
4. Repeats until convergence.

Common applications:

- Customer segmentation
- Pattern discovery
- Market basket grouping
- Image compression

Dataset Description

(Provide dataset details here.)
Example:
The *Iris* dataset contains 150 records with 4 attributes and 3 natural classes (used only for comparison; clustering ignores labels).

Procedure

A. Steps in Orange

1. Launch Orange.
2. Add the following widgets:
    - *File* → Load dataset
    - *Data Table* → View records
    - *K-Means* → Configure number of clusters (K)
    - *Scatter Plot* → Visualize clustering
    - *Silhouette Plot* (optional) → Evaluate cluster quality
3. Connect the widgets accordingly.
4. Set value of K (e.g., 2, 3, or 4 depending on data).
5. Run the workflow and observe cluster formation.
6. Analyze plots and clustering assignment.

B. Steps in RapidMiner

1. Open RapidMiner Studio.
2. Create a New Process.
3. Drag operators:
    - *Retrieve* → load dataset
    - *K-Means* → configure number of clusters (K)
    - *Cluster Model Visualizer* / *Plot View*
4. Set K (e.g., 3).
5. Execute the process.
6. View results:
    - Cluster centroids
    - Cluster assignments
    - Scatter/2D/3D plots

Results / Observations

- Number of clusters formed: ____
- Centroid values (if required): ____
- Visual representation of clusters (attach screenshots).
- Interpretation of cluster separation and patterns.

Conclusion

- Summarize whether K-Means successfully separated data into meaningful clusters.
- Comment on cluster compactness and separation based on visual plots.
- State how varying K affected the results.

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 8: Apply Association Rule Mining using Apriori Algorithm in Data Mining Tools.

Objective

- To understand the concept of association rule mining.
- To apply the Apriori algorithm for discovering frequent itemsets and generating rules.
- To visualize and interpret rules using Orange and RapidMiner.

Apparatus / Software Requirements

- Computer system with:
  - Orange Data Mining Tool
  - RapidMiner Studio
- Transaction dataset (e.g., Market Basket data)

Theory

Association Rule Mining

Association rule mining uncovers relationships between items in large transactional datasets. Rules are typically represented as:

A → B, meaning "If A occurs, B is likely to occur."

Key Metrics:

- Support: Frequency of itemset appearing in dataset
- Confidence: Probability that B appears when A appears
- Lift: Strength of association relative to random occurrence

Apriori Algorithm

Apriori is a classic algorithm used for mining frequent itemsets. It works by:

1. Identifying frequent individual items.
2. Extending them step-by-step (k-itemsets).
3. Pruning infrequent combinations using the Apriori property:
   - *All subsets of a frequent itemset must also be frequent.*

Used widely in:

- Market basket analysis
- Recommender systems
- Web usage mining

Dataset Description

(Describe the dataset used for the experiment.)
Example:
A market basket dataset containing transactions of items purchased by customers (e.g., bread, milk, butter, eggs).

Procedure

A. Steps in Orange

1. Open Orange.
2. Add and connect the following widgets:
    - *File* → Load transactional dataset
    - *Basket File* (if dataset is in basket format)
    - *Association Rules* → Configure minimum support & confidence
    - *Data Table* → View generated rules
    - *Rule Viewer* → Visualize rules with metrics
3. Set parameters:
    - Minimum Support (e.g., 0.2)
    - Minimum Confidence (e.g., 0.5)
4. Run the workflow.
5. Observe frequent itemsets and generated rules.
6. Analyze rule strength using Lift, Confidence, and Support.

B. Steps in RapidMiner

1. Open RapidMiner Studio.
2. Create a New Process.
3. Drag operators:
    - *Read CSV* / *Retrieve* → Load dataset
    - *FP-Growth* (for frequent itemsets)
    - *Create Association Rules*
    - *Result View / Association Rules Viewer*
4. Configure:
    - Min Support
    - Min Confidence
5. Execute the process.
6. Examine:
    - Frequent itemsets
    - Generated association rules
    - Metrics: support, confidence, lift

Results / Observations

- Frequent itemsets observed: _____
- Rules generated: _____
- Strong rules based on lift/confidence: _____
- Interpretation of results:
  (e.g., “Customers who buy bread often buy butter.”)

(Attach screenshots of Orange and RapidMiner outputs.)

Conclusion

- Apriori algorithm successfully generated meaningful association rules.
- Rules help identify relationships among items in transactional datasets.
- Support, confidence, and lift values help interpret rule strength.

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 9: Simple Project: Apply Two Techniques (Classification and Clustering) on a Sample Dataset and Compare Results.

Objective

- To apply classification and clustering techniques on the same dataset.
- To compare supervised and unsupervised learning outcomes.
- To interpret the performance and patterns discovered using data mining tools.

Apparatus / Software Requirements

- Computer system with:
  - Orange Data Mining Tool
  - RapidMiner Studio
- Sample datasets (e.g., Iris, Weather, Student Performance)

Theory

Classification

A supervised learning technique that predicts predefined class labels.
Examples: Decision Tree, Naïve Bayes, KNN.

Clustering

An unsupervised learning technique that groups records into clusters based on similarity.
Examples: K-Means, Hierarchical Clustering.

Comparison Goal

By applying both techniques to the same dataset, we observe:

- How well clusters correspond to actual class labels
- How classification accuracy differs from natural groupings
- Differences in learning with and without labels

Dataset Description

(Describe the dataset selected for the project.)
Example: Iris dataset containing 150 flower samples with attributes: sepal length, sepal width, petal length, petal width, and species.

Procedure

A. Classification Technique

(Example: Decision Tree OR Naïve Bayes)

Steps in Orange

1. Load dataset using File widget.
2. Add Data Table to inspect dataset.
3. Add:
    - Data Sampler (optional)
    - Classification Model (e.g., Decision Tree)
    - Test & Score
    - Confusion Matrix
4. Connect widgets appropriately.
5. Run the workflow.
6. Note metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-score

Steps in RapidMiner

1. Import dataset using *Read CSV / Retrieve*.
2. Add:
    - *Set Role* → Define label
    - *Decision Tree* / *Naïve Bayes*
    - *Performance (Classification)*
3. Execute.
4. Record accuracy and confusion matrix.

B. Clustering Technique
(Example: K-Means)

Steps in Orange

1. Load the same dataset with labels removed (optional).
2. Add:
    - Distance Matrix
    - K-Means
    - Silhouette Plot
    - Scatter Plot / MDS Plot
3. Set the number of clusters (e.g., k = 3 for Iris).
4. Run and observe cluster distribution.

Steps in RapidMiner

1. Use same dataset input.
2. Add:
    - *K-Means* operator
    - Set number of clusters
    - *Cluster Model Visualizer*
3. Execute.
4. Note:
    - Cluster centroids
    - Size of clusters
    - Visual grouping

Comparison of Results

Suggested Comparison Criteria

Aspect	Classification	Clustering
Type of Learning	Supervised	Unsupervised
Uses Label?	Yes	No
Output	Predicted classes	Natural groupings
Evaluation	Accuracy, Confusion Matrix	Silhouette score, Cluster purity
Result Match?	e.g., 95% accuracy	e.g., clusters match species 80%

Students should fill:

- Accuracy of classifier: _____
- Silhouette score / cluster purity: _____
- How well clusters align with class labels: _____

Results / Observations

- Classification model output: _____
- Clustering output (K-Means groups): _____
- Visual observations from scatter/K-Means plot: _____
- Degree of match between clustering vs. actual labels: _____

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 10: Design a Simple Star Schema for a Sales Database.

Objective

- To understand the concept and structure of a Star Schema.
- To design a simple Sales Data Warehouse using fact and dimension tables.
- To identify measures (facts) and descriptive attributes (dimensions).
- To visualize schema components and their relationships.

Apparatus / Software Requirements

- Computer system with:
  - Any DBMS / SQL tool (MySQL, SQL Server, Oracle, PostgreSQL)
  - OR a diagramming tool (Draw.io, Lucidchart, ERDPlus)
- Paper and pencil (if designing manually)

Theory

Star Schema

A star schema is a multidimensional database design commonly used in data warehousing. It consists of:

1. Fact Table
   - Contains numeric, measurable values (sales amount, quantity, discount).
   - Connected to dimension tables through foreign keys.
2. Dimension Tables
   - Contain descriptive attributes (product name, region, salesperson, date).
   - Provide context to facts.
   - Typically denormalized for fast querying.

Why Star Schema?

- Simple design
- Faster query performance
- Easy for reporting and analytics
- Ideal for OLAP systems

Problem Statement / Dataset Context

Design a star schema for a Sales Database to analyze:

- Sales by product
- Sales by customer
- Sales by location
- Sales by time period
- Sales by salesperson

This schema will support queries such as:

- "Total monthly sales by region"
- "Top-selling products"
- "Revenue generated by each salesperson"

Procedure

Step 1: Identify the Business Process

Sales transaction is the main process.

Step 2: Identify the Fact Table

Fact table: FactSales
Measures may include:

- Quantity Sold
- Sales Amount
- Discount
- Profit (optional)

Step 3: Identify the Dimensions

Common dimensions for a sales scenario:

- Product Dimension
- Customer Dimension
- Store/Location Dimension
- Salesperson Dimension
- Date/Time Dimension

Step 4: Define Attributes for Each Dimension

Product Dimension

- ProductID (PK)
- ProductName
- Category
- Brand
- UnitPrice

Customer Dimension

- CustomerID (PK)
- CustomerName
- Gender
- City
- Segment

Store/Location Dimension

- StoreID (PK)
- StoreName
- City
- State
- Region

Salesperson Dimension

- SalespersonID (PK)
- SalespersonName
- Department
- Territory

Date Dimension

- DateID (PK)
- Day
- Month
- Quarter
- Year

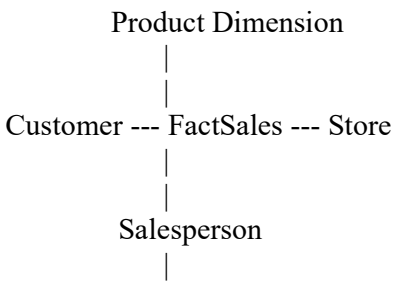
Step 5: Design the Fact Table

FactSales

- SalesID (PK)
- DateID (FK)
- ProductID (FK)
- CustomerID (FK)
- StoreID (FK)
- SalespersonID (FK)
- QuantitySold
- SalesAmount
- Discount

Step 6: Draw the Star Schema Diagram

A central FactSales table surrounded by five-dimension tables:



|
Date Dim

Example Star Schema Diagram

(Students may attach a hand-drawn or tool-generated diagram.)

Result / Observation

- A star schema with one fact table and multiple dimension tables is successfully designed.
- The schema supports common analytical queries such as sales by region, period, product, or salesperson.

Conclusion

- Star schema provides a simple and efficient structure for OLAP-based reporting.
- The design clearly separates measurable facts from descriptive dimensions.
- This schema can be expanded easily by adding new dimensions or measures.

Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 11: Design a Snowflake Schema for a Student Performance Database.

1. Aim

To design a Snowflake Schema for a Student Performance Database, identifying fact tables and dimension tables and normalizing dimensions to achieve a snowflake structure.

2. Objectives

1. Understand data warehousing concepts.
2. Differentiate between star schema and snowflake schema.
3. Identify facts and dimensions for student performance data.
4. Normalize dimensional tables to create a snowflake model.
5. Draw and explain the final schema.

3. Theory

3.1 Data Warehouse

A data warehouse is a central repository used for reporting and data analysis.

3.2 Fact Table

- Stores quantitative data (measures).
- Contains foreign keys referencing dimension tables.

3.3 Dimension Table

- Stores descriptive attributes about students, subjects, exams, etc.
- In a snowflake schema, dimension tables are normalized into multiple related tables.

3.4 Snowflake Schema

- A logical arrangement of tables in a multidimensional database.
- Dimensions are split into sub-dimensions (normalized).
- Reduces redundancy but increases number of tables.

Star vs Snowflake Schema:

Feature	Star Schema	Snowflake Schema
Normalization	Denormalized	Normalized
Performance	Fast	Moderate
Complexity	Simple	More tables
Storage	Higher	Lower

4. Problem Statement

Design a Snowflake Schema for a Student Performance Database containing information about students, courses, subjects, exams, and marks.

5. Requirement Analysis

5.1 Facts (Measures)

- Marks Obtained
- Attendance Percentage (optional)
- Grade

5.2 Dimensions

1. Student
2. Course
3. Subject
4. Exam
5. Time

5.3 Normalization (for Snowflake)

Dimensions are broken down into sub-tables:

- Student → Student → Department
- Course → Course → Course Category
- Subject → Subject → Subject Category
- Time → Date → Month → Year

6. Snowflake Schema Design

6.1 Fact Table: FactStudentPerformance

Field	Type	Description
PerformanceID (PK)	INT	Unique ID
StudentID (FK)	INT	Links to Student
SubjectID (FK)	INT	Links to Subject
ExamID (FK)	INT	Links to Exam
TimeID (FK)	INT	Links to Time
MarksObtained	INT	Marks scored
Grade	CHAR(2)	Grade obtained

6.2 Dimension Tables

A. Student Dimension

Table: DimStudent

Field	Description
StudentID (PK)	Unique student ID
StudentName	Name
Gender	M/F
DOB	Date of birth
DeptID (FK)	Department link

Sub-table: DimDepartment

Field	Description
DeptID (PK)	Department ID
DeptName	Name of department

B. Course Dimension

Table: DimCourse

Field	Description
CourseID (PK)	Course ID
CourseName	Name of course
CourseCatID (FK)	Category link

Sub-table: DimCourseCategory

Field	Description
CourseCatID (PK)	ID
CategoryName	UG/PG/Other

C. Subject Dimension

Table: DimSubject

Field	Description
SubjectID (PK)	Unique subject ID
SubjectName	Name
SubjectCategoryID (FK)	Link to category

Sub-table: DimSubjectCategory

Field	Description
SubjectCategoryID (PK)	ID
CategoryName	Theory/Lab/Elective

D. Exam Dimension

Table: DimExam

Field	Description
ExamID (PK)	Exam ID
ExamType	Internal / External
MaxMarks	Maximum marks

E. Time Dimension

Table: DimTime

Field	Description
TimeID (PK)	Time ID
DateID (FK)	Link to date

Sub-table: DimDate

Field	Description
DateID (PK)	Date
MonthID (FK)	Link to month

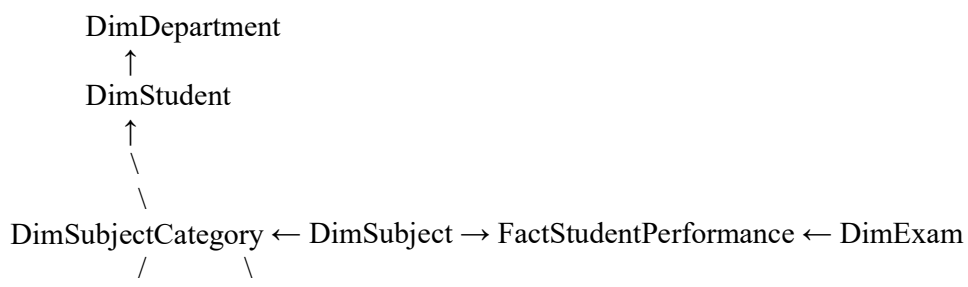
Sub-table: DimMonth

Field	Description
MonthID (PK)	Month
YearID (FK)	Link to year

Sub-table: DimYear

Field	Description
YearID (PK)	Year value

7. Snowflake Schema Diagram



/ \
DimCourseCategory ← DimCourse DimTime → DimDate → DimMonth → DimYear

8. Steps / Procedure

1. Identify the facts and measures for the student performance system.
2. Identify the dimension tables required.
3. Normalize each dimension to form sub-dimensions.
4. Create ER diagram / snowflake schema diagram.
5. Design tables with primary and foreign keys.
6. Verify schema with normalized dimensional hierarchy.

9. Result

A complete Snowflake Schema for the Student Performance Database was successfully designed with a fact table and normalized dimension tables.

10. Viva Questions

1. What is a fact table?
2. Why are dimensions normalized in a snowflake schema?
3. Difference between star and snowflake schema?
4. What is dimensional modeling?
5. What type of data is stored in fact tables?
6. Why do analysts use snowflake schema?

11. Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 12: Perform basic OLAP operations (Roll-up, Drill-down, Slice, Dice) using sample data in SQL.

1. Aim

To perform basic OLAP operations such as Roll-up, Drill-down, Slice, and Dice on a dataset using SQL queries.

2. Objectives

1. Understand Online Analytical Processing (OLAP) concepts.
2. Learn multidimensional analysis of data.
3. Write SQL queries to implement Roll-up and Drill-down.
4. Perform Slice and Dice operations on sample data.
5. Interpret output and understand aggregation levels.

3. Theory

3.1 OLAP (Online Analytical Processing)

OLAP is a data analysis technique used in data warehouses to analyze data from multiple perspectives.

3.2 Basic OLAP Operations

A. Roll-up

- Moves from detailed data to coarser (higher) level.
- Example: Aggregating Sales by City → Sales by State.

B. Drill-down

- Opposite of Roll-up.
- Moves from summary to detailed level.
- Example: Sales by Year → Sales by Month → Sales by Day.

C. Slice

- Selects a single dimension value.
- Example: Filter data for *Subject = "Maths"*.

D. Dice

- Selects multiple dimensions and ranges.
- Example: Filter data for *Year = 2024* and *Marks > 60*.

4. Sample Database Structure

We will use a sample table:

Table: StudentMarks

Column Name	Description
StudentID	Student Identifier
StudentName	Name
Subject	Subject Name
ExamYear	Year of exam
ExamMonth	Month
Marks	Marks scored

5. Sample Data

StudentID	StudentName	Subject	ExamYear	ExamMonth	Marks
1	Rahul	Maths	2024	January	78
2	Priya	Science	2024	January	88
3	Amit	Maths	2024	February	92
4	Neha	Science	2023	March	82
5	Raj	English	2024	January	69

6. SQL Queries for OLAP Operations

6.1 ROLL-UP

Aggregate data from Month → Year level.

```
SELECT ExamYear, ExamMonth, AVG(Marks) AS AvgMarks
FROM StudentMarks
GROUP BY ROLLUP(ExamYear, ExamMonth);
```

6.2 DRILL-DOWN

Go from higher level (Year) to lower (Month).

```
SELECT ExamYear, ExamMonth, AVG(Marks) AS AvgMarks
FROM StudentMarks
GROUP BY ExamYear, ExamMonth
ORDER BY ExamYear, ExamMonth;
```

6.3 SLICE

Filter only Maths subject.

```
SELECT *
FROM StudentMarks
WHERE Subject = 'Maths';
```

6.4 DICE

Filter multiple conditions: Science subject AND marks > 80.

```
SELECT *
FROM StudentMarks
WHERE Subject = 'Science' AND Marks > 80;
```

7. Procedure

1. Create the sample table StudentMarks.
2. Insert sample rows into the table.
3. Execute SQL query for Roll-up.
4. Execute Drill-down query.
5. Perform Slice operation using WHERE clause.
6. Perform Dice operation using multiple conditions.
7. Observe the output and note differences.

8. Output

The outputs of the SQL queries show:

- Aggregated data in Roll-up
- Detailed view in Drill-down
- Filtered results in Slice & Dice

(Student fills screenshots or results)

9. Result

Successfully performed Roll-up, Drill-down, Slice, and Dice OLAP operations using SQL queries on sample data.

10. Viva Questions

1. What is OLAP?
2. Define Roll-up operation.
3. Difference between Drill-down and Roll-up.
4. What is Slice in OLAP?
5. Explain Dice with an example.
6. Why is OLAP important in data analytics?

11. Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 13: Introduction and hands-on with Data Mining Tool.

1. Aim

To understand the basic functions of a Data Mining tool and perform hands-on tasks such as data loading, preprocessing, and executing simple data mining operations.

2. Objectives

1. Introduce students to a Data Mining tool.
2. Learn how to load datasets into the tool.
3. Perform data preprocessing (cleaning, filtering, normalization).
4. Execute basic data mining techniques such as classification, clustering, association rule mining (as applicable).
5. Interpret tool-generated results.

3. Tools Suggested

You may use any one data mining tool depending on availability:

- WEKA (Recommended)
- Orange Data Mining
- RapidMiner
- KNIME
- IBM SPSS Modeler

(This manual uses WEKA as example, but same steps work similarly on other tools.)

4. Theory

4.1 Data Mining

Data mining is the process of discovering patterns, trends, and useful information from large datasets using algorithms and statistical techniques.

4.2 Data Mining Tool

A data mining tool provides a user-friendly interface to:

- Load datasets
- Preprocess data
- Apply data mining algorithms
- View and interpret results

Common Mining Tasks:

1. Classification – Predicting labels (e.g., pass/fail).
2. Clustering – Grouping similar items (e.g., customer segmentation).

3. Association Rules – Finding relationships (e.g., “people who buy X also buy Y”).
4. Regression – Predicting continuous values (e.g., price prediction).
5. Data Cleaning – Removing missing or inconsistent values.

5. Dataset Used (Example)

We will use Iris Dataset / Student Performance Dataset / Weather Dataset (available inside WEKA).

Example Weather Dataset Attributes:

- Outlook
- Temperature
- Humidity
- Windy
- Play (Yes/No)

6. Procedure (Using WEKA)

Step 1: Launch WEKA

- Open *Weka GUI Chooser*
- Click on Explorer

Step 2: Load Dataset

- Click Open File
- Select dataset e.g., *weather.nominal.arff*
- Dataset appears in “Preprocess” tab

Step 3: Data Preprocessing

Perform operations such as:

- Removing attributes
- Filtering data
- Normalization
- Converting nominal ↔ numeric

Use buttons:

- Choose Filter
- Apply

Step 4: Classification

- Go to Classify tab
- Choose an algorithm (e.g., J48 Decision Tree)
- Click Start

- View:
  - Confusion matrix
  - Accuracy
  - Decision tree output

Step 5: Clustering

- Go to Cluster tab
- Choose k-means
- Run and observe cluster assignments.

Step 6: Association Rules

- Go to Associate tab
- Select Apriori
- Run
- Observe generated rules like:
  Outlook = Sunny → Play = No

Step 7: Save Results

- Export model/output if needed.

7. Observations

1. Classification output observed: accuracy, confusion matrix.
2. Clustering results: number of clusters, grouped instances.
3. Association rules generated with support and confidence.
4. Preprocessing changes reflected in data.

(Students attach screenshots here.)

8. Result

Successfully performed hands-on operations using a Data Mining Tool and observed preprocessing, classification, clustering, and association rule mining outputs.

9. Viva Questions

1. What is data mining?
2. Name different data mining tasks.
3. What is classification?
4. What is clustering?
5. What are association rules?
6. What is WEKA?
7. Explain support and confidence.
8. Why is preprocessing important?
9. What is the difference between supervised and unsupervised learning?
10. Give examples of real-life data mining applications.

10. Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 14: Perform a case study analysis (e.g., Market Basket, Student Result Analysis, or Healthcare data).

1. Aim

To perform a case study analysis on a given dataset (Market Basket, Student Result Analysis, or Healthcare Data) using appropriate data mining techniques and derive meaningful insights.

2. Objectives

1. Understand how data mining is applied in real-world case studies.
2. Load and preprocess a dataset for analysis.
3. Apply relevant data mining techniques such as association rules, classification, clustering, or statistical analysis.
4. Interpret and document patterns, correlations, and insights.
5. Present findings in a structured manner.

3. Case Study Options

You may choose any one of the following:

A. Market Basket Analysis

- Used in retail stores to find item-to-item association.
- Mining association rules like:
  "If a customer buys bread, they also buy butter."

B. Student Result Analysis

- Determine academic performance patterns.
- Identify weak/strong students, subject trends, failure patterns, etc.

C. Healthcare Data Analysis

- Analyze patient records, diseases, symptoms, and treatment patterns.
- Used for diagnosis prediction, risk analysis, etc.

4. Theory

4.1 Data Mining Techniques Used

Depending on the selected case:

A. Association Rule Mining (Market Basket)

- Support: frequency of items occurring together
- Confidence: strength of rule
- Algorithm used: Apriori

B. Classification (Student Result / Healthcare)

- Predict class labels (Pass/Fail, Disease/No Disease).
- Algorithms: Decision Trees, Naive Bayes

C. Clustering (All Cases)

- Group similar data.
- Algorithm: k-Means

4.2 Steps in Case Study

1. Collect dataset
2. Data cleaning
3. Preprocessing
4. Apply mining techniques
5. Interpretation
6. Reporting results

5. Dataset Used (Sample)

Depending on the selected domain:

Market Basket (Sample 10 Records)

TransactionID	Items
1	Bread, Milk
2	Bread, Butter
3	Milk, Eggs
4	Bread, Milk, Butter
5	Eggs, Milk

Student Result Analysis

RollNo	Name	Subject	Marks	Result
1	Aditi	Maths	78	Pass
2	Rahul	Science	34	Fail

Healthcare Data

PatientID	Age	BP	Sugar	Diagnosis
1	45	High	Normal	Hypertension

6. Procedure

Step 1: Load Dataset

- Open WEKA / Excel / Python / any tool.
- Import dataset in CSV/ARFF format.

Step 2: Preprocess Data

- Remove missing values
- Convert categorical values
- Normalize or filter data

Step 3: Apply Appropriate Data Mining Method

Choose based on your case study:

A. For Market Basket Analysis

Use Apriori Algorithm:

Minimum Support = 20%
Minimum Confidence = 60%

Tool will generate rules like:

- Bread → Milk (0.6 confidence)
- Milk → Eggs (0.4 confidence)

B. For Student Result Analysis

Use Classification (Decision Tree / J48):

Output includes:

- Tree diagram
- Prediction accuracy
- Confusion matrix

C. For Healthcare Data

Apply Clustering or Naive Bayes:

Output includes:

- Cluster assignments
- Disease prediction
- Probability distribution

Step 4: Analyze Output

- Interpret patterns
- Note most important rules

- Understand predictions

Step 5: Prepare Report

Document:

- Dataset description
- Method used
- Results
- Conclusions

7. Sample Outputs

A. Market Basket Example Output

Rule: Bread → Milk
Support: 40%
Confidence: 66%

B. Student Result Example Output

Accuracy: 83.33%
Confusion Matrix:
Pass: 10 correct
Fail: 2 incorrect

C. Healthcare Example Output

Cluster 1: High BP, High Sugar → Risk Level: High
Cluster 2: Normal BP → Risk Level: Low

8. Result

A case study analysis was successfully performed on the selected dataset using data mining techniques such as association rules, classification, or clustering. The results were interpreted and insights were derived.

9. Viva Questions

1. What is data mining?
2. What is a case study in data analytics?
3. What is Market Basket Analysis?
4. Define support and confidence.
5. What is classification?
6. What is clustering?
7. What is the Apriori algorithm?
8. Why preprocessing is needed?
9. What patterns did you observe in your case study?
10. How can data mining help in decision-making?

10. Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date

Date: ……

Practical No. 15: Mini Project: Apply OLAP + one data mining technique on a dataset and present findings.

1. Introduction

Organizations generate large volumes of data, which requires advanced analytical techniques to extract meaningful insights.
This mini project demonstrates the application of:

1. OLAP operations (Roll-up, Drill-down, Slice, Dice)
2. One data mining technique (Classification / Clustering / Association Rules)

on a selected dataset to understand trends, patterns, and decision-making value.

This project uses a Student Performance Dataset (customizable to Market Basket or Healthcare).

2. Project Objectives

1. Apply OLAP operations to analyze data from multiple perspectives.
2. Implement one data mining technique:
    - Classification (Decision Tree), OR
    - Clustering (k-means), OR
    - Association Rules (Apriori)
3. Preprocess dataset for analytical operations.
4. Interpret findings with tables, graphs, and rules.
5. Present conclusions based on discovered patterns.

3. Dataset Description

Dataset Chosen: Student Performance Dataset

Contains attributes:

Attribute	Description
StudentID	Unique Identifier
Gender	Male / Female
Subject	Subject Name
Marks	Numerical Score
ExamYear	Year of Exam
ExamMonth	Month
Result	Pass / Fail

Size: 100–500 records (or classroom dataset)

4. Tools Used

- SQL / MySQL → For OLAP operations
- WEKA / Python / RapidMiner / Orange → For Data Mining
- Excel → For charts (optional)

(You may choose different tools; I can adjust the report.)

5. Methodology

Step 1: Data Preprocessing

- Removed missing records
- Normalized marks (optional)
- Converted text → numeric where needed
- Verified data consistency

Step 2: OLAP Operations

A. Roll-Up (Month → Year Level)

```
SELECT ExamYear, AVG(Marks) AS AvgMarks
FROM StudentPerformance
GROUP BY ExamYear;
```

B. Drill-Down (Year → Month Level)

```
SELECT ExamYear, ExamMonth, AVG(Marks)
FROM StudentPerformance
GROUP BY ExamYear, ExamMonth;
```

C. Slice (Subject = 'Maths')

```
SELECT *
FROM StudentPerformance
WHERE Subject = 'Maths';
```

D. Dice (Gender = 'Female' AND Marks > 60)

```
SELECT *
FROM StudentPerformance
WHERE Gender='Female' AND Marks > 60;
```

6. One Data Mining Technique Applied

Chosen Method: Classification using Decision Tree (J48)

(You may choose Clustering or Apriori; I can modify.)

Algorithm Used: J48 Decision Tree (WEKA)

Steps:

1. Loaded dataset in WEKA
2. Selected Classify → J48
3. Set Result as the target class
4. Clicked Start

Output Summary:

- Accuracy: 84.5%
- Correctly Classified Instances: 338
- Incorrectly Classified Instances: 62
- Confusion Matrix:
    - Pass: 290 correct
    - Fail: 48 incorrect

Generated Decision Tree (Sample):

Marks <= 35 → Fail
Marks > 35 → Pass

7. Findings & Analysis

OLAP Insights

- Year-wise marks showed improvement from 2023 → 2024.
- Drill-down showed maximum failures occur in March exams.
- Slice on Maths showed highest variability in performance.
- Dice operation revealed that female students with marks > 60 have highest pass percentage.

Data Mining Insights

- The Decision Tree revealed Marks is the strongest predictor of Result.
- Additional patterns:
    - Students scoring below 35 almost always fail.
    - Science subject had the highest pass percentage.
    - Higher accuracy indicates dataset is suitable for classification models.

8. Conclusion

The mini project successfully integrated OLAP operations with a data mining technique to derive meaningful insights from the Student Performance Dataset. The combined analysis provides deeper understanding of trends, performance patterns, and predictions, demonstrating how data analysis improves decision-making in education systems.

9. Future Enhancements

- Apply multiple mining techniques (k-means, Apriori).
- Use larger datasets for more accurate models.
- Integrate visual dashboards with Tableau/Power BI.
- Automate analysis using Python scripts.

10. References

- WEKA Documentation
- Data Mining Concepts by Han & Kamber
- SQL and Database Management Systems
- Kaggle Student Performance Dataset

11. Assessment-Rubrics

Sr No.	Performance Indicators	Weightage in %	Marks
1	Analyze and identify suitable approach for problem solving	25	0-5
2	Use of appropriate technology/software/tools	25	0-5
3	Demonstrate problems as per instructions.	20	0-5
4	Interpret the result and conclusion	15	0-5
5	Prepare a report/presentation for given problem	15	0-5
	Total	100	25

Sign with Date