

Data Mining Lab Manual

Department of Artificial Intelligence
B.Tech in Artificial Intelligence
Kathmandu University

Contents

1	Unit 1: Introduction to Data Mining	3
1.1	Lab Objective	3
1.2	Background Theory	3
1.3	Experiment	3
1.4	Outcome	3
2	Unit 2: Exploratory Data Analysis	4
2.1	Lab Objective	4
2.2	Background Theory	4
2.3	Experiment	4
2.4	Outcome	4
3	Unit 3: Data Preparation	5
3.1	Lab Objective	5
3.2	Background Theory	5
3.3	Experiment	5
3.4	Outcome	5
4	Data Warehousing and OLAP	6
4.1	Lab Objective	6
4.2	Problem Statement	6
4.3	Background Theory	6
4.4	Experiment	7
4.5	Outcome	7
5	Unit 5: Supervised Data Mining Methods	8
5.1	Lab Objective	8
5.2	Background Theory	8
5.3	Experiment	8
5.4	Outcome	8

6 Unit 6: Unsupervised Data Mining Methods	9
6.1 Lab Objective	9
6.2 Background Theory	9
6.3 Experiment	9
6.4 Outcome	9
7 Unit 7: Data Mining and Ethics	10
7.1 Lab Objective	10
7.2 Background Theory	10
7.3 Experiment	10
7.4 Outcome	10
8 Unit 8: Applications and Case Studies	11
8.1 Lab Objective	11
8.2 Background Theory	11
8.3 Experiment	11
8.4 Outcome	11

1 Unit 1: Introduction to Data Mining

1.1 Lab Objective

To introduce the concepts of data mining, its goals, popular methodologies, and related tools and technologies.

1.2 Background Theory

- **Data Mining:** The process of discovering patterns, correlations, trends, or useful information from large datasets.
- **Goals:** Prediction, classification, clustering, summarization, and anomaly detection.
- **Methodologies:**
 - CRISP-DM: Cross-Industry Standard Process for Data Mining.
 - KDD: Knowledge Discovery in Databases.
 - SEMMA: Sample, Explore, Modify, Model, Assess (by SAS).

1.3 Experiment

- Study various data mining tools like Weka, Orange, RapidMiner, and Scikit-learn.
- Describe and compare CRISP-DM, SEMMA, and KDD with real-world examples.

1.4 Outcome

Students will be able to describe data mining concepts, goals, tools, and methodologies with examples.

2 Unit 2: Exploratory Data Analysis

2.1 Lab Objective

To explore and visualize data using statistical and graphical techniques.

2.2 Background Theory

- **Data Types:** Numerical, categorical, temporal, text-based, etc.
- **Univariate Analysis:** Histograms, box plots, bar charts.
- **Bivariate Analysis:** Scatter plots, correlation matrices.
- **Multivariate Analysis:** Pair plots, parallel coordinates, heatmaps.

2.3 Experiment

1. Load a dataset (e.g., Titanic or Iris dataset).
2. Use `pandas` and `matplotlib/seaborn` to perform:
 - Descriptive statistics (mean, median, mode, std).
 - Graphical summaries for each type of analysis.

2.4 Outcome

Students will understand how to explore data using statistical summaries and visualizations.

3 Unit 3: Data Preparation

3.1 Lab Objective

To prepare raw data for mining by cleaning, transforming, and reducing it.

3.2 Background Theory

- **Data Cleaning:** Handling missing, noisy, and inconsistent data.
- **Integration:** Combining data from multiple sources.
- **Transformation:** Normalization, standardization, encoding.
- **Reduction:** Feature selection, PCA, binning.
- **Discretization:** Converting continuous data to categorical.

3.3 Experiment

1. Load dataset with missing and noisy values.
2. Perform:
 - Imputation techniques (`mean()`, `median()`, etc.)
 - Feature encoding (label encoding, one-hot).
 - Normalization (min-max) or standardization (z-score).
 - PCA for dimensionality reduction.

3.4 Outcome

Students will be capable of cleaning and preparing data for further analysis or mining.

4 Data Warehousing and OLAP

4.1 Lab Objective

To design a simple data warehouse for a retail business, create a data cube, and perform OLAP operations such as roll-up, drill-down, slice, and dice while exploring efficient cube computation, access methods, and query processing techniques.

4.2 Problem Statement

A retail company wants to analyze its sales data across three dimensions:

- **Product**
- **Time**
- **Location**

The key measure is **Total Sales Amount**. Students will:

- Design a star schema for the data warehouse.
- Populate it with sample data.
- Build and query a data cube using SQL.
- Perform OLAP operations for analysis.

4.3 Background Theory

- **Data Warehouse:** A centralized repository that stores integrated data from multiple sources for querying and analysis.
- **Star Schema:** A dimensional model with a central fact table and related dimension tables.
- **Data Cube:** A multi-dimensional representation of data used for OLAP.
- **OLAP Operations:**
 - **Roll-up:** Aggregating data along a dimension.
 - **Drill-down:** Navigating from summary to detailed data.
 - **Slice:** Selecting a single dimension.
 - **Dice:** Selecting two or more dimensions to filter data.

4.4 Experiment

1. Create the following schema in PostgreSQL:
 - fact_sales(product_id, location_id, time_id, amount)
 - dim_product(product_id, product_name)
 - dim_location(location_id, location_name)
 - dim_time(time_id, month, year)
2. Populate each table with at least 5 rows of sample data.
3. Write SQL queries to:
 - Compute total sales by product.
 - Perform a roll-up by year.
 - Slice by a specific location.
 - Dice for specific product and year.

4.5 Outcome

Students will learn to model, implement, and query a simple data warehouse and understand basic OLAP operations and cube computation techniques.

5 Unit 5: Supervised Data Mining Methods

5.1 Lab Objective

To understand and implement supervised learning algorithms for classification and prediction tasks.

5.2 Background Theory

- **Classification:** Assigning labels to data points based on input features.
- **Prediction:** Estimating continuous values from features.
- **Algorithms:**
 - Decision Trees, Random Forests
 - Bayesian Networks
 - K-Nearest Neighbors
 - Linear and Logistic Regression
 - Support Vector Machine (SVM)
 - Artificial Neural Networks (ANN)
- **Model Evaluation:**
 - Training vs Testing
 - Cross-validation, Holdout, Bootstrap
 - Accuracy, Precision, Recall, F1-score
 - ROC and Precision-Recall curves
 - Loss functions (e.g., MSE, Log loss)

5.3 Experiment

1. Load dataset (e.g., Titanic or Iris).
2. Implement classification algorithms using `scikit-learn`.
3. Train, test and evaluate using metrics and visualizations.

5.4 Outcome

Students will be able to implement and evaluate various classification and prediction techniques.

6 Unit 6: Unsupervised Data Mining Methods

6.1 Lab Objective

To apply unsupervised learning techniques like clustering and association rule mining.

6.2 Background Theory

- **Association Rules:** Discovering interesting relationships among items.
 - Apriori algorithm
 - Interestingness Measures: Support, Confidence, Lift
- **Clustering:** Grouping data based on similarity.
 - K-Means, K-Medoids
 - Expectation Maximization (EM)
 - DBSCAN
 - Agglomerative and Divisive Clustering
 - Validation: Intrinsic (Silhouette), Extrinsic (Adjusted Rand Index)

6.3 Experiment

1. Implement Apriori using `mlxtend`.
2. Apply K-means and DBSCAN on datasets.
3. Visualize and validate the results.

6.4 Outcome

Students will understand how to discover associations and perform cluster analysis effectively.

7 Unit 7: Data Mining and Ethics

7.1 Lab Objective

To examine ethical aspects of data mining including privacy, fairness, and transparency.

7.2 Background Theory

- **Privacy and Security:** Anonymization, encryption, consent.
- **Social Impact:** Discrimination, bias in models.
- **Accountability:** Transparent model building and decision making.
- **Bias:** In data collection, sampling, and labeling.

7.3 Experiment

- Analyze a biased dataset and demonstrate fairness issues.
- Compare outputs of models trained on biased vs debiased data.
- Discuss ethical case studies (e.g., COMPAS recidivism algorithm).

7.4 Outcome

Students will understand the importance of ethical design in data mining projects.

8 Unit 8: Applications and Case Studies

8.1 Lab Objective

To explore real-world data mining applications through case studies and outlier detection.

8.2 Background Theory

- **Outlier Detection:** Identifying anomalies in data.
 - Types: Point, contextual, collective
 - Methods: Z-score, IQR, DBSCAN, Isolation Forest
- **Applications:**
 - Educational Data Mining
 - Business Intelligence
 - Spatial Data Mining
 - Time Series and Text Mining

8.3 Experiment

1. Use Isolation Forest or Z-score to detect outliers in real datasets.
2. Select a case study and perform end-to-end mining.
3. Document insights and business value.

8.4 Outcome

Students will apply data mining in practical scenarios and learn to extract meaningful insights.