First Step in Data Mining

What You Need to Know About Data Mining

From Basics → Classification → Clustering → Parallel Computing with MPI

Author:
ADEL AZZI

Edition: 2025

Institution / University

Student, University of Science and Technology Houari Boumediene (USTHB)

Index

- **Chapter 12: Hierarchical Clustering**
    - **Agglomerative (AGNES) vs. Divisive (DIANA) Approaches**
    - **Dendrogram Visualization**
    - **Linkage Methods: Single, Complete, and Average**
- **Chapter 13: DBSCAN and Density-Based Clustering**
    - **Density Concepts and Parameter Tuning (Epsilon, MinPts)**
    - **Anomaly and Noise Detection**

Part IV: Advanced Concepts & High-Performance Data Mining

- **Chapter 14: Introduction to Parallel Computing**
    - **Introduction**
    - **What is Parallel Computing?**
    - **Why Parallel Computing?**
    - **Types of Parallelism**
    - **Parallel Architectures**
    - **Challenges in Parallel Computing**
- **Chapter 15: Introduction to MPI (Message Passing Interface)**
    - **Fundamentals: MPI_Send, MPI_Recv, and Collective Operations**
    - **Communicators, Rank, and Size**

- **Challenges**
- **Useful URLs**

{ Introduction to the Book }

Introduction to the Book

Welcome to this comprehensive guide on **Data Mining and High-Performance Clustering Techniques**. Whether you are a student, researcher, or data enthusiast, this book is designed to help you build a strong theoretical foundation while also mastering practical skills through hands-on examples.

To ensure a well-rounded learning experience, **each chapter** in this book is divided into **three structured sections**:

🔹 1. Course Part (Theoretical Concepts)

This section provides clear and concise **theoretical explanations** of the topic. You will explore key definitions, formulas, algorithms, and examples — designed to help you understand the core principles behind each method or model.

🔹 2. TD Part (Directed Exercises)

The TD (Travaux Dirigés) section includes a series of **guided exercises** to reinforce the theoretical knowledge. These problems range from simple to advanced levels, helping you apply what you've learned and prepare for exams or interviews.

🔹 3. TP Part (Practical Labs / Code Implementation)

The TP (Travaux Pratiques) section focuses on **hands-on programming** using tools like **Python**, **scikit-learn**, **mpi4py**, and more. You will learn how to implement algorithms, visualize results, and evaluate performance through real datasets.

Structure of the Book

The book is divided into **four main parts**:

- ❖ **Part I**: Fundamentals of Data Mining
- ❖ **Part II**: Supervised Learning
- ❖ **Part III**: Unsupervised Learning
- ❖ **Part IV**: Parallel and High-Performance Data Mining

Each part builds progressively, helping you move from basic concepts to advanced implementations using parallel architectures such as **MPI**.

Objective of the Book

- To provide a solid understanding of **data mining theory**
- To bridge the gap between **mathematical models** and **real-world coding**
- To introduce **parallel computing principles** applied to data science
- To prepare you for academic projects, research, or industrial roles

Target Audience

- Master's and engineering students in computer science, data science, or HPC
- Educators and researchers seeking a structured, practical resource
- Developers or analysts transitioning into machine learning or big data

We hope this book empowers you to **think critically**, **code confidently**, and **analyze data at scale**.

Let's begin your journey into data mining!