data mining project

BOOK RECOMMENDATION SYSTEM AND USAGE PATTERN ANALYSIS

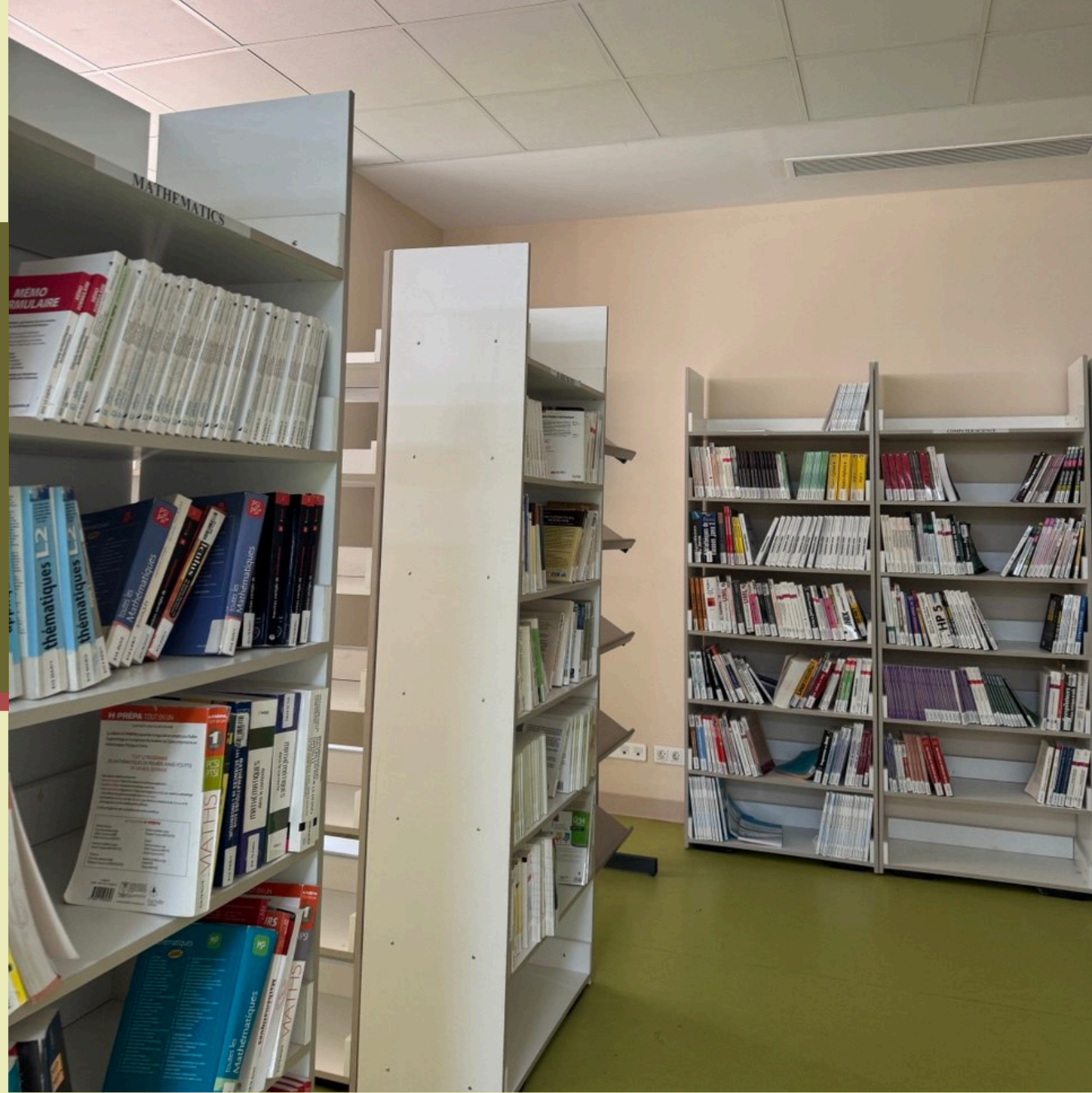
Team:

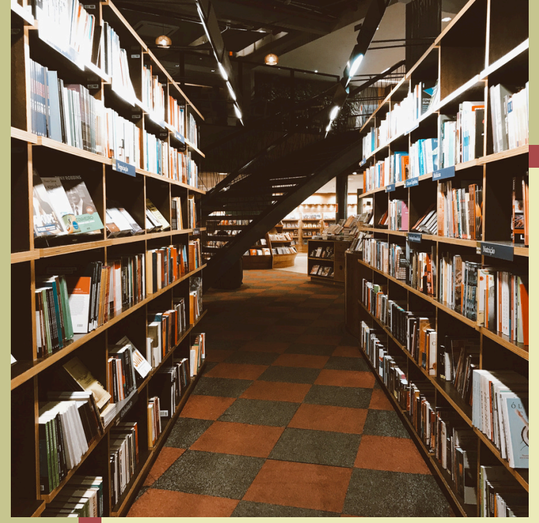
Guettaf Temam Selsabila

Bensabra Ala

Sekfali Lina

Chouimet Noussayba Om Salama





Overview

1. Introduction & Context
2. Problem Statement
3. Project Objectives
4. Methodology
5. Data Analysis & Results
6. Conclusion & Impact
7. Future Work

Introduction

Modern libraries continuously collect data through borrowing transactions, inventory updates, and catalog records. However, without proper analysis, this data does not contribute to improving library services.

Data mining provides tools to:

- Discover hidden patterns in large datasets
- Understand user behavior
- Support strategic and operational decisions

This project demonstrates how data mining techniques can transform raw library data into meaningful knowledge.

Problem Statement

The analysis addresses several key challenges observed in the library:

- A large portion of books are rarely or never borrowed
- A small number of books dominate borrowing activity
- Reader engagement is low and inconsistent
- There is no automated system to recommend books

These issues lead to inefficient resource allocation and missed opportunities to improve user experience.

Project Objectives

The main objectives of this project are to:

- Analyze borrowing behavior and usage patterns
- Identify popular and underutilized books
- Segment readers based on their activity levels
- Segment books based on popularity and circulation behavior
- Discover relationships between books borrowed together
- Build a data-driven book recommendation system

Datasets Used

The project uses three interconnected datasets:

- Borrowings Dataset (422 records):
- Captures interactions between readers and books, including loan and return dates.
- Catalogue Dataset (617 books):
- Contains book metadata such as titles, authors, and number of copies.
- Inventory Dataset (4,105 copies):
- Tracks individual physical copies and their availability.

Together, these datasets provide a complete view of library usage and resources.

Methodology

The project follows a structured data mining pipeline:

1. **Data Preprocessing**
2. **Exploratory Data Analysis (EDA)**
3. **Feature Engineering**
4. **K-Means Clustering**
5. **Association Rule Mining (Apriori)**

This methodology ensures data quality, meaningful analysis, and reliable results.

Data Preprocessing

Data preprocessing was essential to ensure accuracy and consistency.

Main steps included:

- Removing duplicate records
- Handling missing and corrupted values
- Standardizing book titles across datasets
- Converting dates and numerical fields to correct formats

After preprocessing, the datasets were clean, consistent, and suitable for analysis

Exploratory Data Analysis (EDA)

Exploratory analysis revealed important usage patterns:

- Average loan duration: 16 days
- 75% of books are returned within the allowed period
- Only 21.6% of books were borrowed at least once
- The top 3 books account for 40% of all borrowings

These results highlight a strong imbalance between highly popular and rarely used books.

Feature Engineeringis

Feature engineering transformed raw data into meaningful variables.

Book Features Included:

- Total number of borrowings
- Average loan duration
- Number of unique readers
- Number of available copies

Reader Features Included:

- Total borrowings
- Activity duration
- Book diversity

These features enable effective clustering and pattern discovery.

Book Clustering (K-Means)

Books were clustered using PCA combined with K-Means.

- Optimal number of clusters: 2

Cluster Results:

- Cluster 0 (97.7%): Regular books with low circulation
- Cluster 1 (2.3%): Bestsellers with very high demand

Bestsellers were borrowed 28 times more than regular books and were mainly mathematics textbooks.

Reader Clustering (K-Means)

Readers were also clustered using PCA and K-Means.

- Optimal number of clusters: 9

The clustering revealed different reader profiles, including:

- One-time readers
- Occasional readers (largest group)
- Highly engaged readers

The average reader borrowed 1.56 books, indicating a reader retention challenge.

Association Rule Mining

Association rule mining was applied using the Apriori algorithm.

Results:

- 8 strong association rules identified
- Highest lift value: 4.028

Example Rule:

Readers who borrowed Probability also borrowed Functions with 84.6% confidence.

These rules reflect meaningful learning and borrowing relationships.

Recommendation System

Based on the association rules, a book recommendation system was designed.

The system:

- Recommends books frequently borrowed together
- Uses real borrowing behavior
- Helps readers discover relevant materials

This approach improves book visibility and reader engagement.

Key Insights

The main insights of the project include:

- Book usage is highly concentrated in a small number of titles
- Reader behavior is diverse and can be segmented effectively
- Strong subject-based borrowing relationships exist
- Data mining reveals patterns not visible through manual analysis

Impact on Decision-Making

The results support several data-driven decisions:

- Increase copies of highly demanded books
- Reduce or reevaluate unused books
- Apply personalized recommendations
- Design targeted reader engagement strategies

These actions improve efficiency and user satisfaction.

Conclusion

This project demonstrates how data mining techniques can transform raw library data into actionable knowledge.

Key achievements include:

- Effective segmentation of books and readers
- Discovery of strong borrowing associations
- Development of a practical recommendation system

The project highlights the value of data-driven approaches in modern library management.

Future Work

Possible extensions of this work include:

- Real-time recommendation systems
- Predictive models for book demand
- Interactive dashboards for librarians
- Integration of digital resources

Thank You

Thank you for your attention.
Questions?