



# Applications of Data Mining/Machine Learning in Business and Entrepreneurship

Case Studies and Exercise

06-12-2023

**Dr. Gemma Catolino**  
[g.catolino@tue.nl](mailto:g.catolino@tue.nl)



JADS

Jheronimus  
Academy  
of Data Science

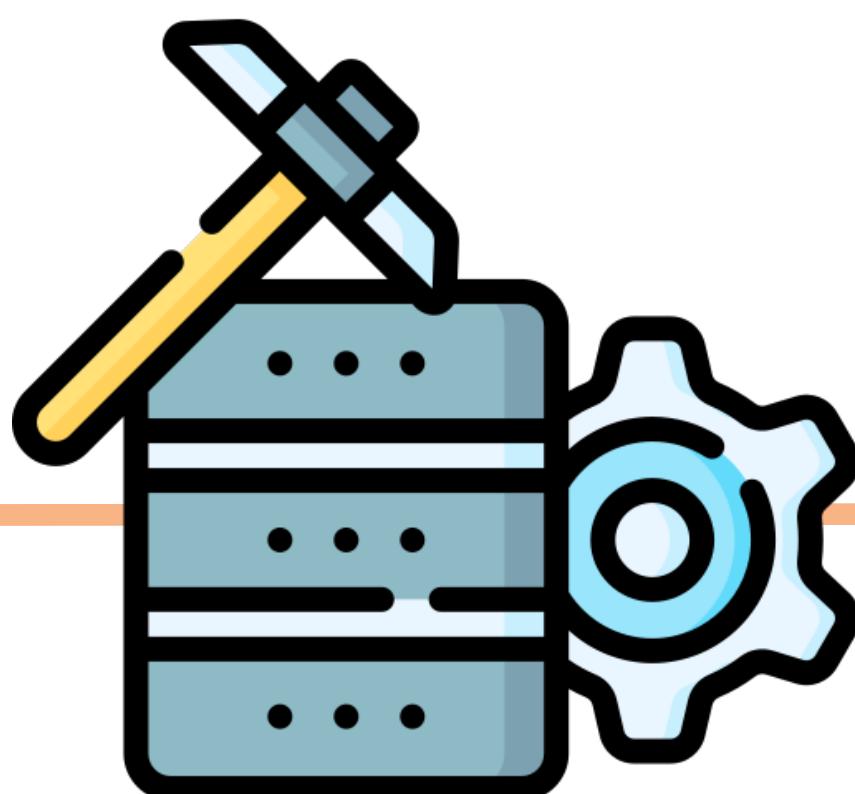


# Introduction to Data Mining

Data mining is the process of extracting useful information from large datasets to uncover patterns, relationships, and insights for decision-making.

Data mining techniques involve tasks such as data preprocessing, pattern recognition, and model building.

The goals of data mining include prediction, classification, clustering, and anomaly detection.



# Data Mining in a Nutshell

## DATA PREPROCESSING

---

Data preprocessing is an essential step in data mining that involves cleaning, transforming, and organizing the data to improve the accuracy and efficiency of the mining process.

## PATTERN RECOGNITION

---

One of the key techniques in data mining is pattern recognition, which involves identifying recurring patterns or relationships in large datasets.

## MODEL BUILDING

---

Model building is another important technique in data mining, where statistical and machine learning models are created and trained to make predictions or classify data based on patterns identified in the dataset.

# Data Mining in Healthcare

## Applications of Data Mining in Healthcare

- Identifying patterns in patient data to improve diagnosis and treatment
- Predicting disease outbreaks and epidemics
- Analyzing electronic health records to identify risk factors and improve patient outcomes
- Detecting fraud and abuse in healthcare systems



# Data Mining in Finance

## Applications of Data Mining in Finance



- Risk assessment and management: Data mining helps identify potential risks and predict market trends.
- Fraud detection: Data mining techniques can uncover patterns of fraudulent activities and assist in detecting financial fraud.
- Customer segmentation: By analyzing customer data, data mining can help identify different customer segments for targeted marketing strategies.
- Credit scoring: Data mining is used to develop credit scoring models that assess the creditworthiness of individuals and businesses.

# Data Mining in Digital Marketing



## Applications of Data Mining in Digital Marketing

- Data mining plays a crucial role in marketing by analyzing large datasets to uncover valuable insights and patterns.
- It helps businesses understand customer behavior, preferences, and purchase patterns.
- Data mining enables targeted marketing campaigns, personalized recommendations, and customer segmentation.
- By leveraging data mining techniques, marketers can optimize marketing strategies and improve ROI.

# Data Mining Applications Across Industries



# Amazon's Product Recommendations



## DATA

Amazon collects data on user browsing history, search queries, product views, past purchases, items added to wish lists, customer reviews, and ratings.

They also incorporate demographic information from user profiles, such as age, location, and gender.

Additional data sources may include social media activity and third-party data partnerships.

## TECHNIQUES USED

Collaborative filtering and item-based recommendation algorithms are employed to analyze the collected data, e.g., Top-k algorithm.

These algorithms identify patterns in user behavior to suggest products that are similar to those a user has shown interest in.

Machine learning models are used to continuously improve the accuracy of recommendations based on user feedback (reinforcement learning!!!).

# Google's Search Engine Ranking



## DATA

Google's search engine analyzes the content of web pages

User behavior such as click-through rates on search results and time on pages

External links pointing to web pages

Geographic location of users for localized search results

## TECHNIQUES USED

Google's search engine uses a complex algorithm that combines various ML techniques

Natural language processing (NLP) and neural networks for better understanding user intent

Providing relevant search results based on the quality and relevance of web pages



# Walmart's Supply Chain Optimization

## DATA

- Sales data
- Inventory levels
- Supplier performance metrics
- Weather data

## TECHNIQUES USED

- Demand forecasting
- Inventory optimization (Genetic algorithm)
- Route optimization algorithm



# Netflix's Movie Recommendation System

## DATA

Netflix collects data on user viewing history, ratings given to movies and TV shows, genre preferences, time spent watching, and other user behaviors.

They also analyze content metadata, e.g., movie and show descriptions and categories.

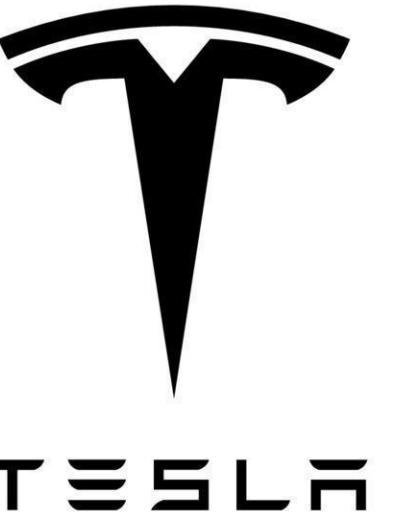
Netflix uses this data to create user profiles and make personalized recommendations.

## TECHNIQUES USED

Netflix employs a variety of recommendation algorithms, including collaborative filtering and content-based filtering.

They also use hybrid filtering techniques that combine multiple algorithms.

Deep learning techniques, such as neural networks, are used to analyze user behavior and content features for personalized recommendations.



# Tesla's Autonomous Driving

## DATA

Tesla's vehicles collect real-time data on road conditions, surrounding vehicles, lane markings, traffic signs, and vehicle performance.

The data includes information from cameras, radar, ultrasonic sensors, and other sensors.

This data is used to train the machine learning models for autonomous driving.

## TECHNIQUES USED

Tesla uses deep learning and neural networks to process sensor data for autonomous driving. Convolutional neural networks (CNNs) are used for image recognition, and recurrent neural networks (RNNs) are used for sequence prediction, enabling self-driving capabilities.



# Airbnb's Dynamic Pricing

## DATA

- Property location
- Time of year
- Local events and holidays
- Property features (e.g., number of bedrooms, amenities)

## TECHNIQUES USED

- ML models for dynamic pricing
- Regression analysis
- Time series forecasting

# Facebook's Ad Targeting



## DATA

- Demographic information (age, gender, location)
- User interests (liked pages, clicked links, shared content)
- Device information
- Social connections

## TECHNIQUES USED

- Decision trees
- Clustering algorithms
- Optimization models



# Uber's Surge Pricing

## DATA

Uber collects data on rider demand, driver availability, traffic conditions, historical pricing data, and geographic areas with high demand.

The data is used to implement surge pricing, adjusting prices in real-time during peak demand periods, thus balancing supply and demand efficiently.

Example: In New York City, Uber uses data on traffic congestion and historical demand patterns to determine surge pricing during rush hour.

## TECHNIQUES USED

ML models are used to predict and optimize pricing during peak demand.

Series analysis and regression analysis are employed to determine surge pricing.

Example: Uber uses time series analysis to forecast demand and adjust prices accordingly during major events like concerts or sports games.

# Consideration on Data Mining Applications

Data mining applications are widely used in various industries, including e-commerce, search engines, retail, entertainment, transportation, and social media.

These applications leverage data to enhance customer experiences, optimize operations, improve decision-making, and increase revenue.

Machine learning techniques, such as collaborative filtering, regression analysis, and neural networks, are commonly employed in data mining applications.

# Conclusion: The Future of Data Mining

Data mining will continue to play a crucial role in extracting valuable insights from large datasets.

Advancements in technology, such as machine learning and artificial intelligence, will enhance the capabilities of data mining.

Ethical considerations and privacy concerns will become increasingly important in the future of data mining.

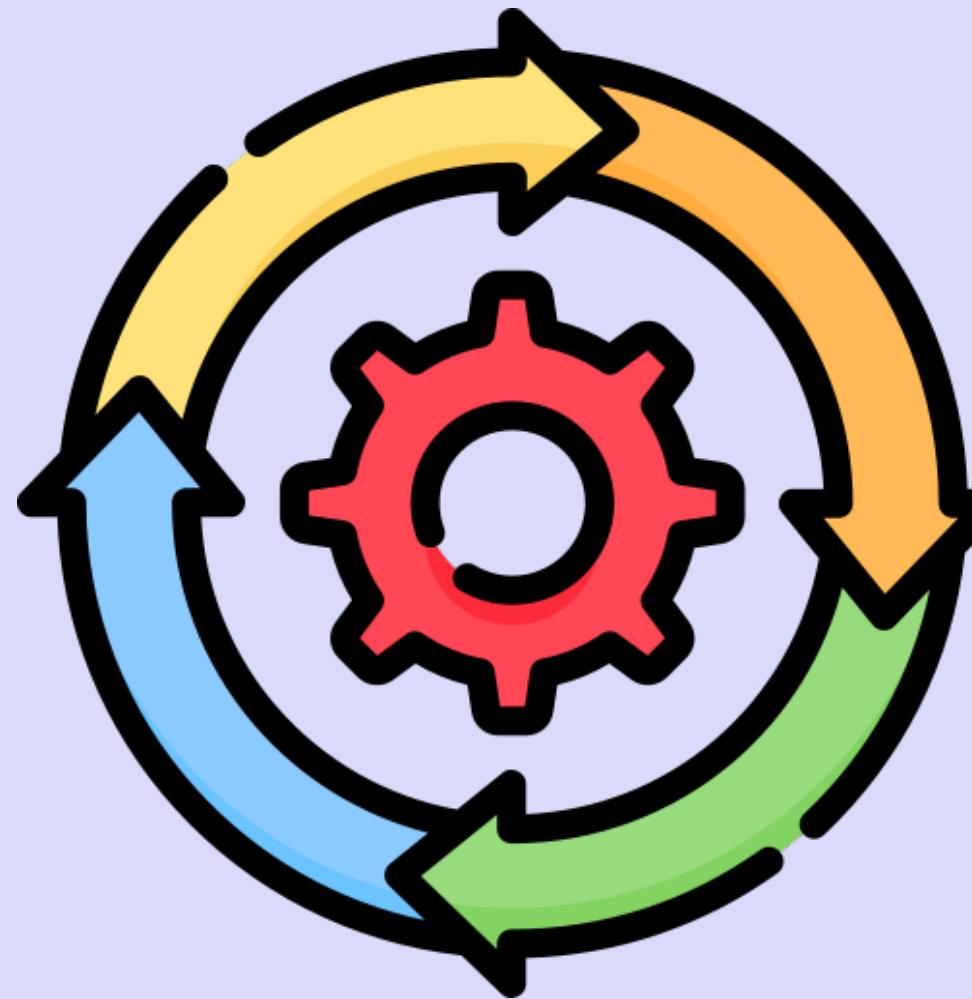


## SCENARIO

You are hired as an external consultant team by a regional hospital company to improve Hospital Patient Care.

## GOAL

Your primary goal is to enhance patient care and reduce hospital readmissions. Your task is to analyze available data to identify patients at risk of readmission, develop intervention strategies, and evaluate their effectiveness. You have access to various data sources to address this challenge.



# PROCESS

## **1. Group Forming**

## **2. Data Source Selection:**

Brainstorm and list potential data sources you would consider to achieve the goal of improving patient care and reducing readmissions. Consider both traditional and non-traditional sources.

## **3. Data Extraction, Machine Learning Strategy, and Need Analysis:**

Select the data you want to extract from the chosen data sources and explain why you consider this data relevant for improving patient care and reducing readmissions. You must consider also ethical privacy components (see next slide). Finally you also decide if machine learning is needed. If so, you should specify whether supervised or unsupervised machine learning is more appropriate and why.

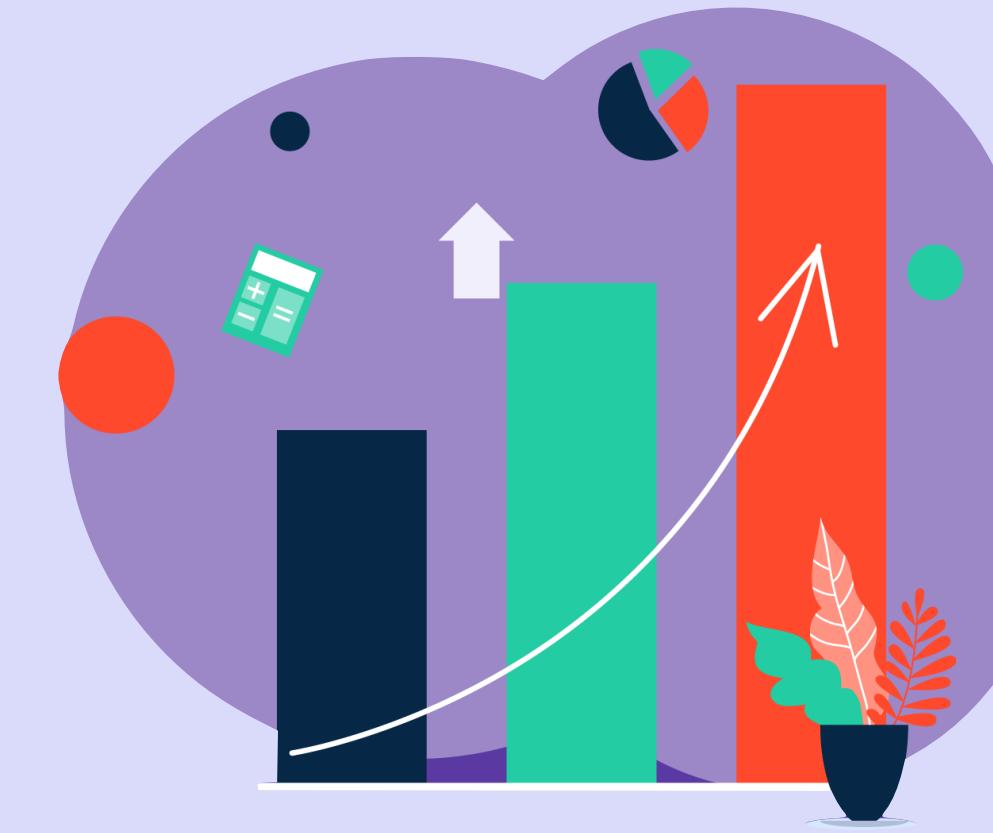
Example:

Need Analysis: Does machine learning play a role in identifying at-risk patients, or can rules-based approaches be sufficient?

Strategy: If machine learning is needed, is supervised learning suitable for predicting readmissions or unsupervised learning for clustering similar patient profiles?

## **4. Final Consideration Together:**

Class discussion to compare different approaches, considering the necessity of machine learning, and highlighting the importance of data selection in addressing healthcare challenges.



# DATA SOURCE

## 1. Electronic Health Records (EHR):

Data Available: Patient demographic information (age, gender, ethnicity, location), medical history, vital signs, lab results, medication records.

## 2. Insurance Claims Data:

Data Available: Claims history, billing codes (ICD-10, CPT, HCPCS), costs, provider information.

## 3. Patient Surveys:

Data Available: Patient satisfaction surveys, health-related lifestyle surveys, social determinants of health (e.g., socioeconomic status, living conditions).

## 4. Post-Discharge Follow-up:

Data Available: Post-treatment checkup details, readmission data, patient-reported outcomes (health-related quality of life, symptoms).

## 5. External Data Sources:

Data Available: Social media data related to patients' health discussions, local environmental data (e.g., pollution levels, weather conditions), academic research findings and treatment guidelines.

## 6. You can decide to provide new data to extract.

## **Electronic Health Records (EHR):**

**Description:** Electronic Health Records contain comprehensive patient health information, including demographics, medical history, diagnosis and treatment records, vital signs (e.g., blood pressure, heart rate, temperature), laboratory results (e.g., blood tests, imaging reports), and medication history (prescriptions, dosages, adherence). EHRs provide a detailed patient health profile, capturing their entire medical journey within the hospital.

## **Insurance Claims Data:**

**Description:** Insurance claims data comprises detailed records of medical services and procedures provided to patients, along with associated billing codes (e.g., ICD-10 (**International Classification of Diseases, 10th Edition**) for diagnoses, CPT (**Current Procedural Terminology**) for procedures, HCPCS (**Healthcare Common Procedure Coding System**) for supplies and services). This data also includes cost breakdowns for medical expenses and reimbursements. Additionally, it contains information on the healthcare providers who rendered the services, offering insights into the type and quality of care received by patients.

## **Patient Surveys:**

**Description:** Patient surveys encompass various types of data related to patient experiences and lifestyles. These may include patient satisfaction surveys that collect feedback on the healthcare experience, health-related lifestyle surveys capturing information about diet, exercise, smoking habits, and mental health. Patient surveys may also include data on social determinants of health, such as socioeconomic status, living conditions, and support networks, which can provide a broader understanding of patients' well-being.

## **Post-Discharge Follow-up:**

**Description:** Post-discharge follow-up data tracks the progress of patients after their treatment, including details of post-treatment checkup appointments, information about any readmissions to the hospital, and patient-reported outcomes. Patient-reported outcomes encompass health-related quality of life assessments, tracking of symptoms, and evaluations of overall satisfaction with the treatment and care received.

## **External Data Sources:**

**Description:** External data sources encompass additional information beyond the hospital's immediate records. Social media data includes patients' discussions and concerns related to their health, potentially offering insights into their conditions and needs. Local environmental data includes factors like pollution levels, weather conditions, and disease outbreaks, which can have an impact on patients' health. Academic research findings and treatment guidelines provide up-to-date information on the latest medical knowledge and best practices in patient care.



# ETHICAL AND PRIVACY CONSIDERATION

## **Patient Privacy and Consent:**

Discuss the importance of patient consent and data anonymization. You should consider how to protect patient privacy and obtain necessary permissions before using sensitive healthcare data.

## **Data Security and Compliance:**

Emphasize the importance of data security and compliance with regulations. You should consider encryption, secure data storage, and adherence to legal requirements.

## **Bias and Fairness:**

Discuss the potential for bias in healthcare data and machine learning models. Think about fairness, transparency, and strategies to mitigate bias when making decisions that affect patient care.



# Applications of Data Mining/Machine Learning in Business and Entrepreneurship

Case Studies and Exercise

06-12-2023

**Dr. Gemma Catolino**  
[g.catolino@tue.nl](mailto:g.catolino@tue.nl)