

# Data Mining Process, Techniques, Tools & Examples

## What is Data Mining?

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.

It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology.

The insights derived via Data Mining can be used for marketing, fraud detection, and scientific discovery, etc.

Data mining is also called as Knowledge discovery, Knowledge extraction, data/pattern analysis, information harvesting, etc.

## Types of Data

Data mining can be performed on following types of data

- Relational databases
- Data warehouses
- Advanced DB and information repositories
- Object-oriented and object-relational databases
- Transactional and Spatial databases
- Heterogeneous and legacy databases
- Multimedia and streaming database
- Text databases
- Text mining and Web mining

# Data Mining Implementation Process



Let's study the Data Mining implementation process in detail

## Business understanding:

In this phase, business and data-mining goals are established.

- First, you need to understand business and client objectives. You need to define what your client wants (which many times even they do not know themselves)
- Take stock of the current data mining scenario. Factor in resources, assumption, constraints, and other significant factors into your assessment.
- Using business objectives and current scenario, define your data mining goals.
- A good data mining plan is very detailed and should be developed to accomplish both business and data mining goals.

## Data understanding:

In this phase, sanity check on data is performed to check whether its appropriate for the data mining goals.

- First, data is collected from multiple data sources available in the organization.
- These data sources may include multiple databases, flat filer or data cubes. There are issues like object matching and schema integration which can arise during Data Integration process. It is a quite complex and tricky process as data from various sources unlikely to match easily. For example, table A contains an entity named cust\_no whereas another table B contains an entity named cust-id.

- Therefore, it is quite difficult to ensure that both of these given objects refer to the same value or not. Here, Metadata should be used to reduce errors in the data integration process.
- Next, the step is to search for properties of acquired data. A good way to explore the data is to answer the data mining questions (decided in business phase) using the query, reporting, and visualization tools.
- Based on the results of query, the data quality should be ascertained. Missing data if any should be acquired.

## **Data preparation:**

In this phase, data is made production ready.

The data preparation process consumes about 90% of the time of the project.

The data from different sources should be selected, cleaned, transformed, formatted, anonymized, and constructed (if required).

Data cleaning is a process to "clean" the data by smoothing noisy data and filling in missing values.

For example, for a customer demographics profile, age data is missing. The data is incomplete and should be filled. In some cases, there could be data outliers. For instance, age has a value 300. Data could be inconsistent. For instance, name of the customer is different in different tables.

Data transformation operations change the data to make it useful in data mining. Following transformation can be applied

## **Data transformation:**

Data transformation operations would contribute toward the success of the mining process.

**Smoothing:** It helps to remove noise from the data.

**Aggregation:** Summary or aggregation operations are applied to the data. I.e., the weekly sales data is aggregated to calculate the monthly and yearly total.

**Generalization:** In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.

**Normalization:** Normalization performed when the attribute data are scaled up or scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.

**Attribute construction:** these attributes are constructed and included the given set of attributes helpful for data mining.

The result of this process is a final data set that can be used in modeling.

## Modelling

In this phase, mathematical models are used to determine data patterns.

- Based on the business objectives, suitable modeling techniques should be selected for the prepared dataset.
- Create a scenario to test check the quality and validity of the model.
- Run the model on the prepared dataset.
- Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

## Evaluation:

In this phase, patterns identified are evaluated against the business objectives.

- Results generated by the data mining model should be evaluated against the business objectives.
- Gaining business understanding is an iterative process. In fact, while understanding, new business requirements may be raised because of data mining.
- A go or no-go decision is taken to move the model in the deployment phase.

## Deployment:

In the deployment phase, you ship your data mining discoveries to everyday business operations.

- The knowledge or information discovered during data mining process should be made easy to understand for non-technical stakeholders.
- A detailed deployment plan, for shipping, maintenance, and monitoring of data mining discoveries is created.
- A final project report is created with lessons learned and key experiences during the project. This helps to improve the organization's business policy.

## Data Mining Techniques



## **1. Classification:**

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

## **2. Clustering:**

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

## **3. Regression:**

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

## **4. Association Rules:**

This data mining technique helps to find the association between two or more items. It discovers a hidden pattern in the data set.

## **5. Outer detection:**

This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

## **6. Sequential Patterns:**

This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

## 7. Prediction:

Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

## Challenges of Implementation of Data mine:

- Skilled Experts are needed to formulate the data mining queries.
- Overfitting: Due to small size training database, a model may not fit future states.
- Data mining needs large databases which sometimes are difficult to manage
- Business practices may need to be modified to determine to use the information uncovered.
- If the data set is not diverse, data mining results may not be accurate.
- Integration information needed from heterogeneous databases and global information systems could be complex

## Data mining Examples:

### Example 1:

Consider a marketing head of telecom service provides who wants to increase revenues of long-distance services. For high ROI on his sales and marketing efforts customer profiling is important. He has a vast data pool of customer information like age, gender, income, credit history, etc. But its impossible to determine characteristics of people who prefer long distance calls with manual analysis. Using data mining techniques, he may uncover patterns between high long distance call users and their characteristics.

For example, he might learn that his best customers are married females between the age of 45 and 54 who make more than \$80,000 per year. Marketing efforts can be targeted to such demographic.

### **Example 2:**

A bank wants to search new ways to increase revenues from its credit card operations. They want to check whether usage would double if fees were halved.

Bank has multiple years of record on average credit card balances, payment amounts, credit limit usage, and other key parameters. They create a model to check the impact of the proposed new business policy. The data results show that cutting fees in half for a targeted customer base could increase revenues by \$10 million.

## **Data Mining Tools**

Following are 2 popular Data Mining Tools widely used in Industry

### **R-language:**

R language is an open source tool for statistical computing and graphics. R has a wide variety of statistical, classical statistical tests, time-series analysis, classification and graphical techniques. It offers effective data handling and storage facility.

### **Oracle Data Mining:**

Oracle Data Mining popularly knowns as ODM is a module of the Oracle Advanced Analytics Database. This Data mining tool allows data analysts to generate detailed insights and makes predictions. It helps predict customer behavior, develops customer profiles, identifies cross-selling opportunities.

## **Benefits of Data Mining:**

- Data mining technique helps companies to get knowledge-based information.
- Data mining helps organizations to make the profitable adjustments in operation and production.
- The data mining is a cost-effective and efficient solution compared to other statistical data applications.
- Data mining helps with the decision-making process.



- Facilitates automated prediction of trends and behaviors as well as automated discovery of hidden patterns.
- It can be implemented in new systems as well as existing platforms
- It is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

## **Disadvantages of Data Mining**

- There are chances of companies may sell useful information of their customers to other companies for money. For example, American Express has sold credit card purchases of their customers to the other companies.
- Many data mining analytics software is difficult to operate and requires advance training to work on.
- Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of correct data mining tool is a very difficult task.
- The data mining techniques are not accurate, and so it can cause serious consequences in certain conditions.

Applications	Usage
Communications	Data mining techniques are used in communication sector to predict customer behavior to offer highly targeted and relevant campaigns.
Insurance	Data mining helps insurance companies to price their products profitable and promote new offers to their new or existing customers.
Education	Data mining benefits educators to access student data, predict achievement levels and find students or groups of students which need extra attention. For example, students who are weak in maths subject.
Manufacturing	With the help of Data Mining Manufacturers can predict wear and tear of production assets. They can anticipate maintenance which helps them reduce them to minimize downtime.
Banking	Data mining helps finance sector to get a view of market risks and manage regulatory compliance. It helps banks to identify probable defaulters to decide whether to issue credit cards, loans, etc.
Retail	Data Mining techniques help retail malls and grocery stores identify and arrange most sellable items in the most attentive positions. It helps store owners to comes up with the offer which encourages customers to increase their spending.
Service Providers	Service providers like mobile phone and utility industries use Data Mining to predict the reasons when a customer leaves their company. They analyze billing details, customer service interactions, complaints made to the company to assign each customer a probability score and offers incentives.
E-Commerce	E-commerce websites use Data Mining to offer cross-sells and up-sells through their websites. One of the most famous names is Amazon, who use Data mining techniques to get more customers into their eCommerce store.

Super Markets	Data Mining allows supermarket's develop rules to predict if their shoppers were likely to be expecting. By evaluating their buying pattern, they could find woman customers who are most likely pregnant. They can start targeting products like baby powder, baby shop, diapers and so on.
Crime Investigation	Data Mining helps crime investigation agencies to deploy police workforce (where is a crime most likely to happen and when?), who to search at a border crossing etc.
Bioinformatics	Data Mining helps to mine biological data from massive datasets gathered in biology and medicine.

## Where is Data Mining used?

### Summary:

- Data Mining is all about explaining the past and predicting the future for analysis.
- Data mining helps to extract information from huge sets of data. It is the procedure of mining knowledge from data.
- Data mining process includes business understanding, Data Understanding, Data Preparation, Modelling, Evolution, Deployment.
- Important Data mining techniques are Classification, clustering, Regression, Association rules, Outer detection, Sequential Patterns, and prediction
- R-language and Oracle Data mining are prominent data mining tools.
- Data mining technique helps companies to get knowledge-based information.
- The main drawback of data mining is that many analytics software is difficult to operate and requires advance training to work on.
- Data mining is used in diverse industries such as Communications, Insurance, Education, Manufacturing, Banking, Retail, Service providers, eCommerce, Supermarkets Bioinformatics.