Contents

Data Mining

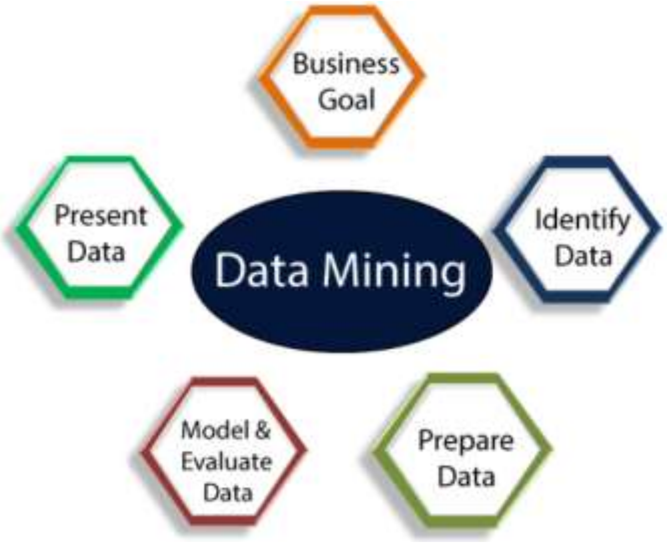
The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

In other words, we can say that Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD).

Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

There are tonnes of information available on various platforms, but very little knowledge is accessible. The biggest challenge is to analyze the data to extract important information that can be used to solve a problem or for company development. There are many powerful instruments and techniques available to mine data and find better insight from it.



Data mining can be performed on the following types of data:

Relational Database:

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.

Data warehouses:

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

Data Repositories:

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

Transactional Database:

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

Advantages of Data Mining

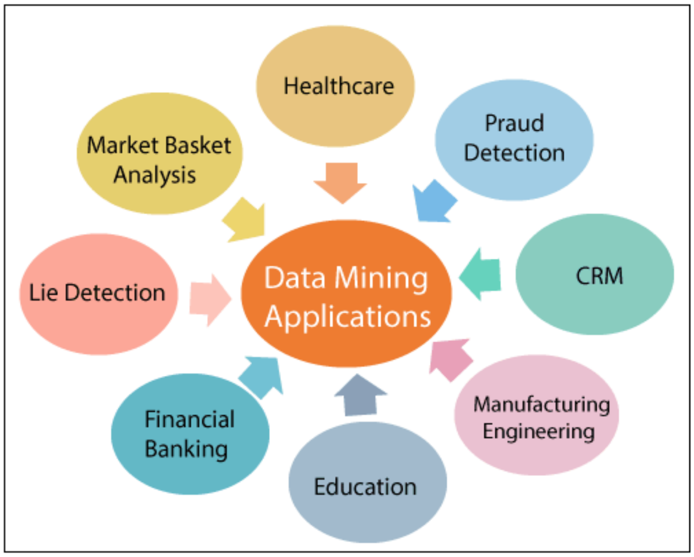
- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.
- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Data Mining Applications

Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction, and corporate profits. Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.



These are the following areas where data mining is widely used:

1. **Data Mining in Healthcare:**

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

2. **Data Mining in Market Basket Analysis:**

Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

3. **Data mining in Education:**

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments. EDM objectives are recognized as affirming student's future learning behavior, studying the impact of educational support, and promoting learning science. An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

4. Data Mining in Manufacturing Engineering:

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

5. Data Mining in CRM (Customer Relationship Management):

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

6. Data Mining in Fraud detection:

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

7. Data Mining in Lie Detection:

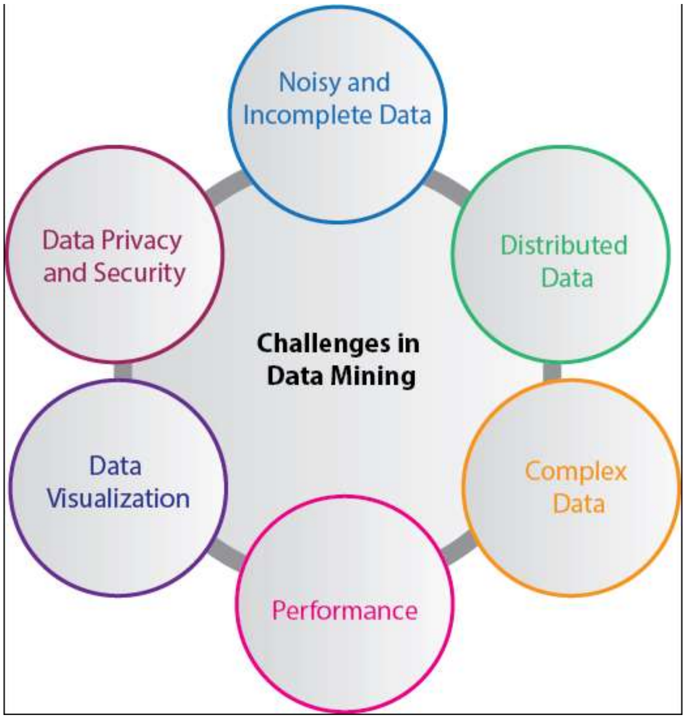
Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task. Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc. This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text. The information collected from the previous investigations is compared, and a model for lie detection is constructed.

8. Data Mining Financial Banking:

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts. The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.

Challenges of Implementation in Data mining

Although data mining is very powerful, it faces many challenges during its execution. Various challenges could be related to performance, data, methods, and techniques, etc. The process of data mining becomes effective when the challenges or problems are correctly recognized and adequately resolved.


Noisy and Incomplete Data
Distributed Data
Complex Data
Performance
Data Visualization
Data Privacy and Security
Challenges in Data Mining

Incomplete and noisy data:

The process of extracting useful data from large volumes of data is data mining. The data in the real-world is heterogeneous, incomplete, and noisy. Data in huge quantities will usually be inaccurate or unreliable. These problems may occur due to data measuring instrument or because of human errors. Suppose a retail chain collects phone numbers of customers who spend more than $ 500, and the accounting employees put the information into their system. The person may make a digit mistake when entering the phone number, which results in incorrect data. Even some customers may not be willing to disclose their phone numbers, which results in incomplete data. The data could get changed due to human or system error. All these consequences (noisy and incomplete data) make data mining challenging.

Data Distribution:

Real-worlds data is usually stored on various platforms in a distributed computing environment. It might be in a database, individual systems, or even on the internet. Practically, It is a quite tough task to make all the data to a centralized data repository mainly due to organizational and technical concerns. For example, various regional offices may have their servers to store their data. It is not feasible to store, all the data from all the offices on a central server. Therefore, data mining requires the development of tools and algorithms that allow the mining of distributed data.

Complex Data:
Real-world data is heterogeneous, and it could be multimedia data, including audio and video, images, complex data, spatial data, time series, and so on. Managing these various types of data and extracting useful information is a tough task. Most of the time, new technologies, new tools, and methodologies would have to be refined to obtain specific information.

Performance:
The data mining system's performance relies primarily on the efficiency of algorithms and techniques used. If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.

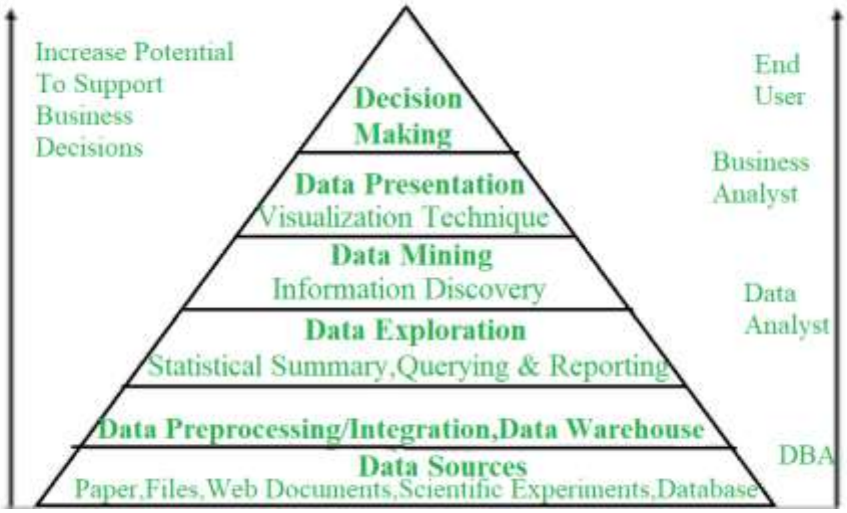
Data Privacy and Security:
Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a retailer analyzes the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.

Data Visualization:
In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it successful.

Prerequisites

Before learning the concepts of Data Mining, you should have a basic understanding of Statistics, Database Knowledge, and Basic programming language.

Data Mining and Business Intelligence:



Data Mining Process

Data Mining is a process of discovering various models, summaries, and derived values from a given collection of data.

The general experimental procedure adapted to data-mining problem involves following steps:

1. State problem and formulate hypothesis –

In this step, a modeler usually specifies a group of variables for unknown dependency and, if possible, a general sort of this dependency as an initial hypothesis. There could also be several hypotheses formulated for one problem at this stage. The primary step requires combined expertise of an application domain and a data-mining model. In practice, it always means an in-depth interaction between data-mining expert and application expert. In successful data-mining applications, this cooperation does not stop within initial phase. It continues during whole data-mining process.

2. Collect data –

This step cares about how information is generated and picked up. Generally, there are two distinct possibilities. The primary is when data-generation process is under control of an expert (modeler). This approach is understood as a designed experiment. The second possibility is when expert cannot influence data generation process. This is often referred to as observational approach.

An observational setting, namely, random data generation, is assumed in most data-mining applications. Typically, sampling distribution is totally unknown after data are collected, or it is partially and implicitly given within data-collection procedure. It is vital, however, to know how data collection affects its theoretical distribution since such a piece of prior knowledge is often useful for modeling and, later, for ultimate interpretation of results. Also, it is important to form sure that information used for estimating a model and therefore data used later for testing and applying a model come from an equivalent, unknown, sampling distribution. If this is often not case, estimated model cannot be successfully utilized in a final application of results.

3. Data Preprocessing –

In the observational setting, data is usually "collected" from prevailing databases, data warehouses, and data marts.

Data preprocessing usually includes a minimum of two common tasks:

- **(i) Outlier Detection (and removal):**

  Outliers are unusual data values that are not according to most observations. Commonly, outliers result from measurement errors, coding, and recording errors, and, sometimes, are natural, abnormal values. Such non-representative samples can seriously affect model produced later.

  There are two strategies for handling outliers: Detect and eventually remove outliers as a neighborhood of preprocessing phase. And Develop robust modeling methods that are insensitive to outliers.

- **(ii) Scaling, encoding, and selecting features:**

  Data preprocessing includes several steps like variable scaling and differing types of encoding. For instance, one feature with range [0, 1] and other with range [100, 1000] will not have an equivalent weight within applied technique. They are going to also influence ultimate data-mining results differently. Therefore, it is recommended to scale them and convey both features to an equivalent weight for further analysis. Also, application-specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modeling.

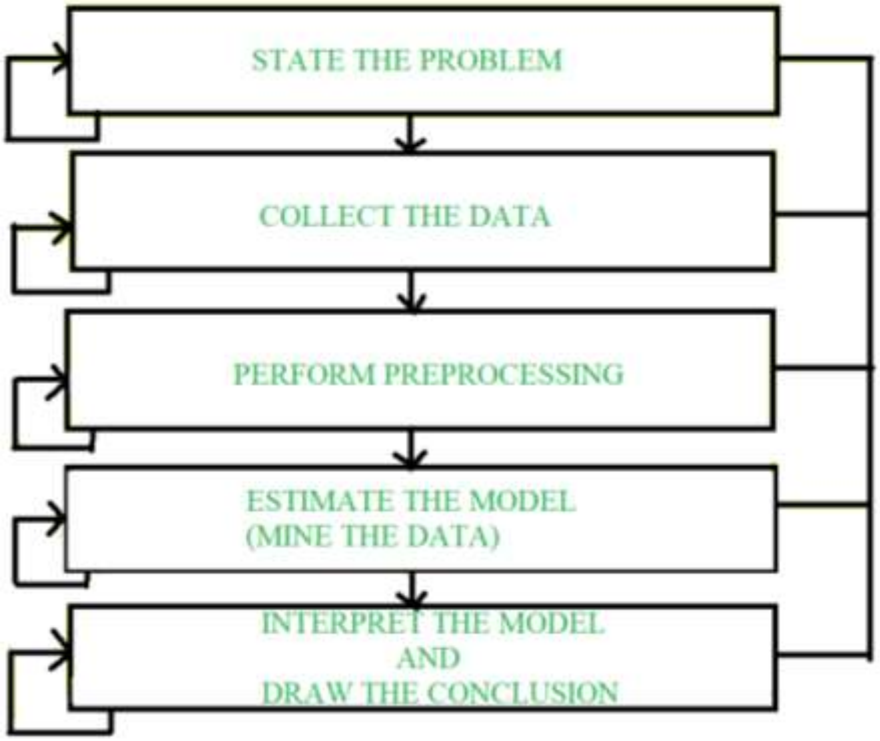
These two classes of preprocessing tasks are only illustrative samples of an outsized spectrum of preprocessing activities during a data-mining process. Data-preprocessing steps should not be considered completely independent from other data-mining phases. In every iteration of data-mining process, all activities, together, could define new and improved data sets for subsequent iterations. Generally, an honest preprocessing method provides an optimal representation for a data-mining technique by incorporating a prior knowledge within sort of application-specific scaling and encoding.

4. Estimate model –

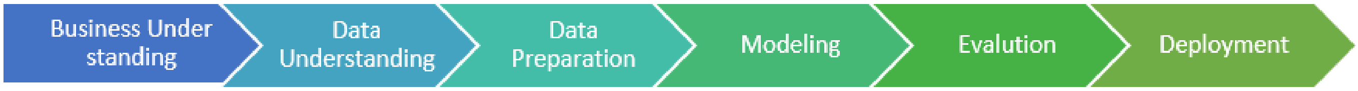
The selection and implementation of acceptable data-mining technique is that main task during this phase. This process is not straightforward. Usually, in practice, implementation is predicated on several models, and selecting simplest one is a further task.

5. **Interpret model and draw conclusions –**

In most cases, data-mining models should help in deciding. Hence, such models got to be interpretable so as to be useful because humans are not likely to base their decisions on complex "black-box" models. Note that goals of accuracy of model and accuracy of its interpretation are somewhat contradictory. Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using high dimensional models. The matter of interpreting these models, also vital, is taken into account a separate task, with specific techniques to validate results.


STATE THE PROBLEM
COLLECT THE DATA
PERFORM PREPROCESSING
ESTIMATE THE MODEL (MINE THE DATA)
INTERPRET THE MODEL AND DRAW THE CONCLUSION

Data Mining Implementation Process



Data Mining Implementation Process

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 o 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:

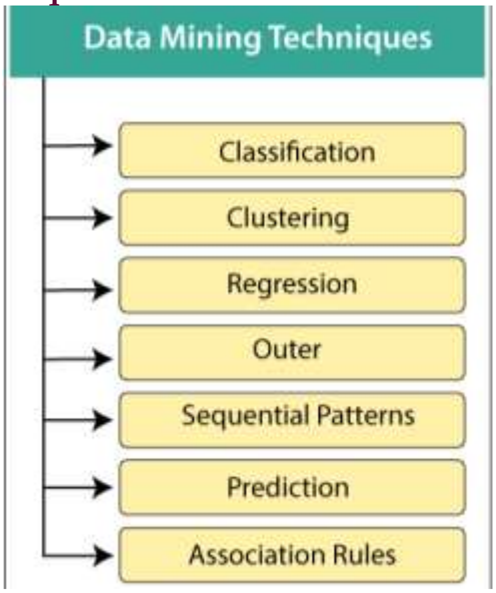
In the deployment phase, you ship your data mining discoveries to everyday business operations.

- 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

Classification of Data Mining Systems:

1. Database Technology
2. Statistics
3. Machine Learning
4. Information Science
5. Visualization

Data Mining Techniques



1. Classification:

This technique is used to obtain important and relevant information about data and metadata. This data mining technique helps to classify data in different classes.

Data mining techniques can be classified by different criteria, as follows:

i. **Classification of Data mining frameworks as per the type of data sources mined:** This classification is as per the type of data handled. For example, multimedia, spatial data, text data, time-series data, World Wide Web, and so on.

ii. **Classification of data mining frameworks as per the database involved:** This classification based on the data model involved. For example. Object-oriented database, transactional database, relational database, and so on.

iii. **Classification of data mining frameworks as per the kind of knowledge discovered:** This classification depends on the types of knowledge discovered or data mining functionalities. For example, discrimination, classification, clustering, characterization, etc. some frameworks tend to be extensive frameworks offering a few data mining functionalities together.

iv. **Classification of data mining frameworks according to data mining techniques used:** This classification is as per the data analysis approach utilized, such as neural networks, machine learning, genetic algorithms, visualization, statistics, data warehouse-oriented or database-oriented, etc.

The classification can also take into account, the level of user interaction involved in the data mining procedure, such as query-driven systems, autonomous systems, or interactive exploratory systems.

2. Clustering:

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. Data modeling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis. From a machine learning point of view, clusters relate to hidden patterns, the search for clusters is unsupervised learning, and the subsequent framework represents a data concept. From a practical point of view, clustering plays an extraordinary job in data mining applications. For example, scientific data exploration, text mining, information retrieval, spatial database applications, CRM, Web analysis, computational biology, medical diagnostics, and much more.

In other words, we can say that Clustering analysis is a data mining technique to identify similar data. This technique helps to recognize the differences and similarities between the data. Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.

3. Regression:

Regression analysis is the data mining process is used to identify and analyze the relationship between variables because of the presence of the other factor. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modeling. For example, we might use it to project certain costs, depending on other factors such as availability, consumer demand, and competition. Primarily it gives the exact relationship between two or more variables in the given data set.

4. Association Rules:

This data mining technique helps to discover a link between two or more items. It finds a hidden pattern in the data set.

Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases. Association rule mining has several applications and is commonly used to help sales correlations in data or medical data sets.

The way the algorithm works is that you have various data, For example, a list of grocery items that you have been buying for the last six months. It calculates a percentage of items being purchased together.

These are three major measurements technique:

- **Lift:**

  This measurement technique measures the accuracy of the confidence over how often item B is purchased.

  (Confidence) / (item B)/ (Entire dataset)

- **Support:**

  This measurement technique measures how often multiple items are purchased and compared it to the overall dataset.

  (Item A + Item B) / (Entire dataset)

- **Confidence:**

  This measurement technique measures how often item B is purchased when item A is purchased as well.

  (Item A + Item B)/ (Item A)

5. Outer detection:

This type of data mining technique relates to the observation of data items in the data set, which do not match an expected pattern or expected behavior. This technique may be used in various domains like intrusion, detection, fraud detection, etc. It is also known as Outlier Analysis or Outlier mining. The outlier is a data point that diverges too much from the rest of the dataset. The majority of the real-world datasets have an outlier. Outlier detection plays a significant role in the data mining field. Outlier detection is valuable in numerous fields like network interruption identification, credit or debit card fraud detection, detecting outlying in wireless sensor network data, etc.

6. Sequential Patterns:

The sequential pattern is a data mining technique specialized for **evaluating sequential data** to discover sequential patterns. It comprises of finding interesting subsequences in a set of sequences, where the stake of a sequence can be measured in terms of different criteria like length, occurrence frequency, etc.

In other words, this technique of data mining helps to discover or recognize similar patterns in transaction data over some time.

7. Prediction:

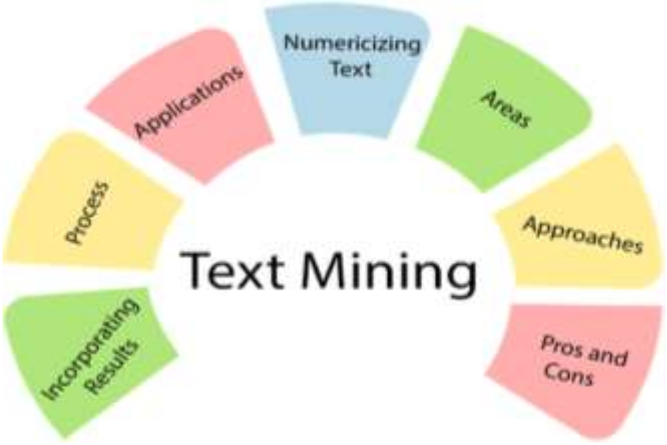
Prediction used a combination of other data mining techniques such as trends, clustering, classification, etc. It analyzes past events or instances in the right sequence to predict a future event.

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

Text Data Mining

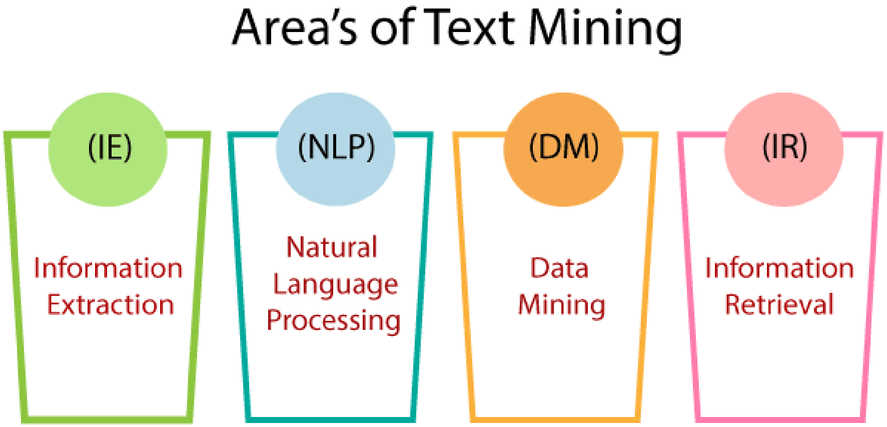
Text data mining can be described as the process of extracting essential data from standard language text. All the data that we generate via text messages, documents, emails, files are written in common language text. Text mining is primarily used to draw useful insights or patterns from such data.



The text mining market has experienced exponential growth and adoption over the last few years and also expected to gain significant growth and adoption in the coming future. One of the primary reasons behind the adoption of text mining is higher competition in the business market, many organizations seeking value-added solutions to compete with other organizations. With increasing completion in business and changing customer perspectives, organizations are making huge investments to find a solution that is capable of analyzing customer and competitor data to improve competitiveness. The primary source of data is e-commerce websites, social media platforms, published articles, survey, and many more. The larger part of the generated data is unstructured, which makes it challenging and expensive for the organizations to analyze with the help of the people. This challenge integrates with the exponential growth in data generation has led to the growth of analytical tools. It is not only able to handle large volumes of text data but also helps in decision-making purposes. Text mining software empowers a user to draw useful information from a huge set of data available sources.

Areas of text mining in data mining:

These are the following area of text mining:



i. **Information Extraction:**

The automatic extraction of structured data such as entities, entities relationships, and attributes describing entities from an unstructured source is called information extraction.

ii. **Natural Language Processing:**

NLP stands for Natural language processing. Computer software can understand human language as same as it is spoken. NLP is primarily a component of artificial intelligence (AI). The development of the NLP application is difficult because computers generally expect humans to "Speak" to them in a programming language that is accurate, clear, and exceptionally structured. Human speech is usually not authentic so that it can depend on many complex variables, including slang, social context, and regional dialects.

iii. **Data Mining:**

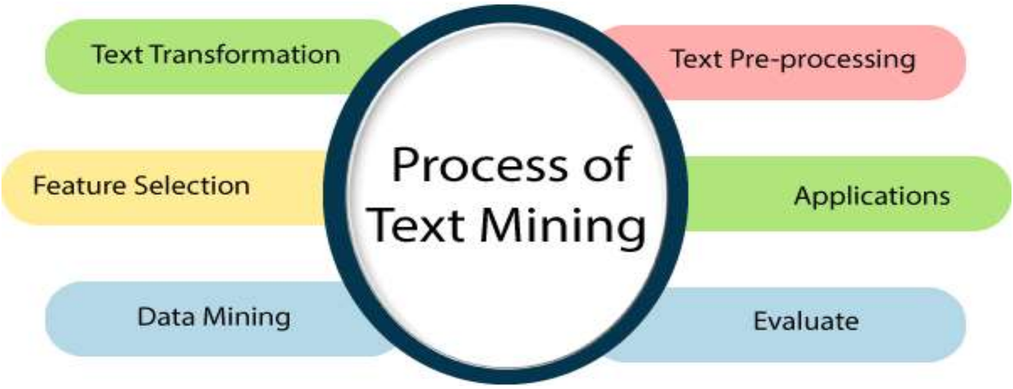
Data mining refers to the extraction of useful data, hidden patterns from large data sets. Data mining tools can predict behaviors and future trends that allow businesses to make a better data-driven decision. Data mining tools can be used to resolve many business problems that have traditionally been too time-consuming.

iv. **Information Retrieval:**

Information retrieval deals with retrieving useful data from data that is stored in our systems. Alternately, as an analogy, we can view search engines that happen on websites such as e-commerce sites or any other sites as part of information retrieval.

Text Mining Process:

The text mining process incorporates the following steps to extract the data from the document.



i. **Text transformation**

A text transformation is a technique that is used to control the capitalization of the text. Here the two major way of document representation is given.

a. Bag of words
b. Vector Space

ii. **Text Pre-processing**

Pre-processing is a significant task and a critical step in Text Mining, Natural Language Processing (NLP), and information retrieval (IR). In the field of text mining, data pre-processing is used for extracting useful information and knowledge from unstructured text data. Information Retrieval (IR) is a matter of choosing which documents in a collection should be retrieved to fulfill the user's need.

iii. **Feature selection:**

Feature selection is a significant part of data mining. Feature selection can be defined as the process of reducing the input of processing or finding the essential information sources. The feature selection is also called **variable selection**.

iv. **Data Mining:**

Now, in this step, the text mining procedure merges with the conventional process. Classic Data Mining procedures are used in the structural database.

v. **Evaluate:**

Afterward, it evaluates the results. Once the result is evaluated, the result abandon.

Applications of Text Mining:

These are the following text mining applications:

i. **Risk Management:**

Risk Management is a systematic and logical procedure of analyzing, identifying, treating, and monitoring the risks involved in any action or process in organizations. Insufficient risk analysis is usually a leading cause of disappointment. It is particularly true in the financial organizations where adoption of Risk Management Software based on text mining technology can effectively enhance the ability to diminish risk. It enables the administration of millions of sources and petabytes of text documents, and giving the ability to connect the data. It helps to access the appropriate data at the right time.

ii. **Customer Care Service:**

Text mining methods, particularly NLP, are finding increasing significance in the field of customer care. Organizations are spending in text analytics programming to improve their overall experience by accessing the textual data from different sources such as customer feedback, surveys, customer calls, etc. The primary objective of text analysis is to reduce the response time of the organizations and help to address the complaints of the customer rapidly and productively.

iii. **Business Intelligence:**

Companies and business firms have started to use text mining strategies as a major aspect of their business intelligence. Besides providing significant insights into customer behavior and trends, text mining strategies also support organizations to analyze the qualities and weaknesses of their opponent's so, giving them a competitive advantage in the market.

iv. Social Media Analysis:

Social media analysis helps to track the online data, and there are numerous text mining tools designed particularly for performance analysis of social media sites. These tools help to monitor and interpret the text generated via the internet from the news, emails, blogs, etc. Text mining tools can precisely analyze the total no of posts, followers, and total no of likes of your brand on a social media platform that enables you to understand the response of the individuals who are interacting with your brand and content.

Text Mining Approaches in Data Mining:

These are the following text mining approaches that are used in data mining.

1. Keyword-based Association Analysis:

It collects sets of keywords or terms that often happen together and afterward discover the association relationship among them. First, it preprocesses the text data by parsing, stemming, removing stop words, etc. Once it pre-processed the data, then it induces association mining algorithms. Here, human effort is not required, so the number of unwanted results and the execution time is reduced.

2. Document Classification Analysis:

Automatic document classification:

This analysis is used for the automatic classification of the huge number of online text documents like web pages, emails, etc. Text document classification varies with the classification of relational data as document databases are not organized according to attribute values pairs.

Numericizing text:

i. Stemming algorithms

A significant pre-processing step before ordering of input documents starts with the stemming of words. The terms "stemming" can be defined as a reduction of words to their roots. For example, different grammatical forms of words and ordered are the same. The primary purpose of stemming is to ensure a similar word by text mining program.

ii. **Support for different languages:**
There are some highly language-dependent operations such as stemming, synonyms, the letters that are allowed in words. Therefore, support for various languages is important.

iii. **Exclude certain character:**
Excluding numbers, specific characters, or series of characters, or words that are shorter or longer than a specific number of letters can be done before the ordering of the input documents.

iv. **Include lists, exclude lists (stop-words):**
A particular list of words to be listed can be characterized, and it is useful when we want to search for a specific word. It also classifies the input documents based on the frequencies with which those words occur. Additionally, "stop words," which means terms that are to be rejected from the ordering can be characterized. Normally, a default list of English stop words incorporates "the," "a," "since," etc. These words are used in the respective language very often but communicate very little data in the document.

What does Web Mining mean?

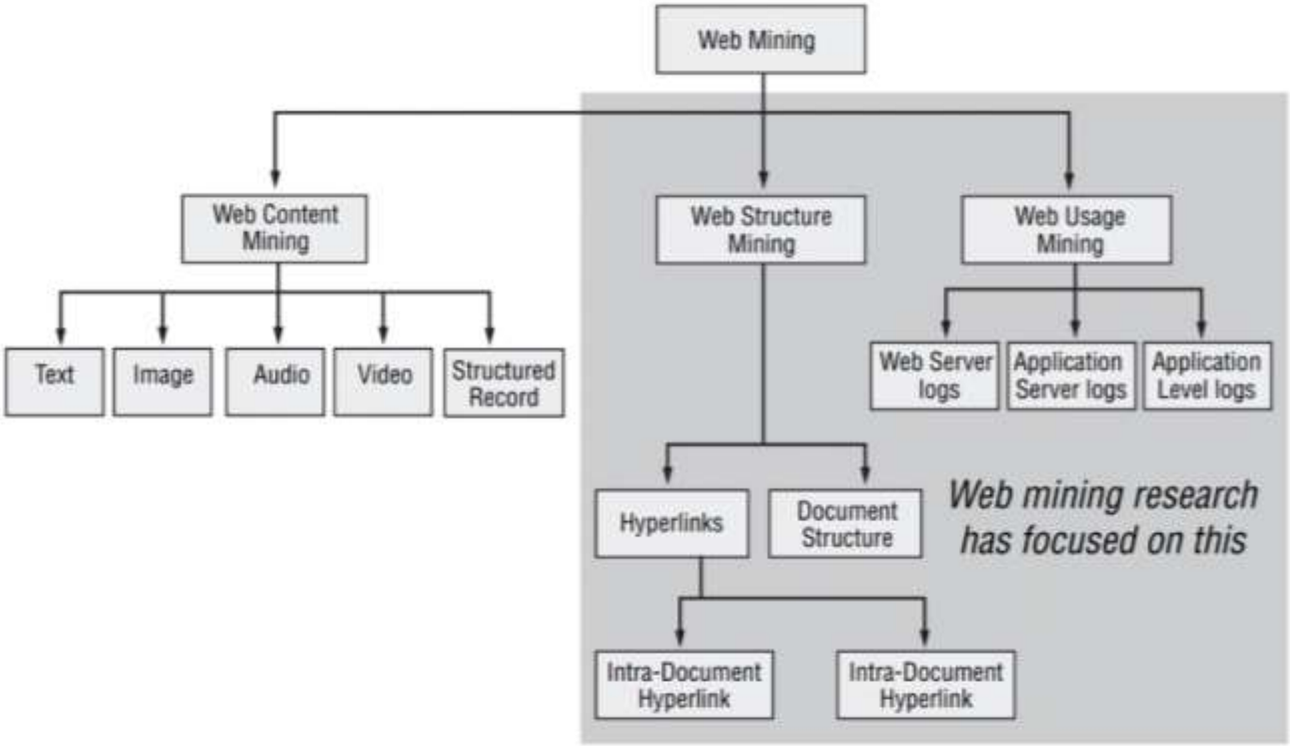
Web mining is the process of using data mining techniques and algorithms to extract information directly from the Web by extracting it from Web documents and services, Web content, hyperlinks and server logs. The goal of Web mining is to look for patterns in Web data by collecting and analyzing information in order to gain insight into trends, the industry and users in general.

Categories of Web mining:

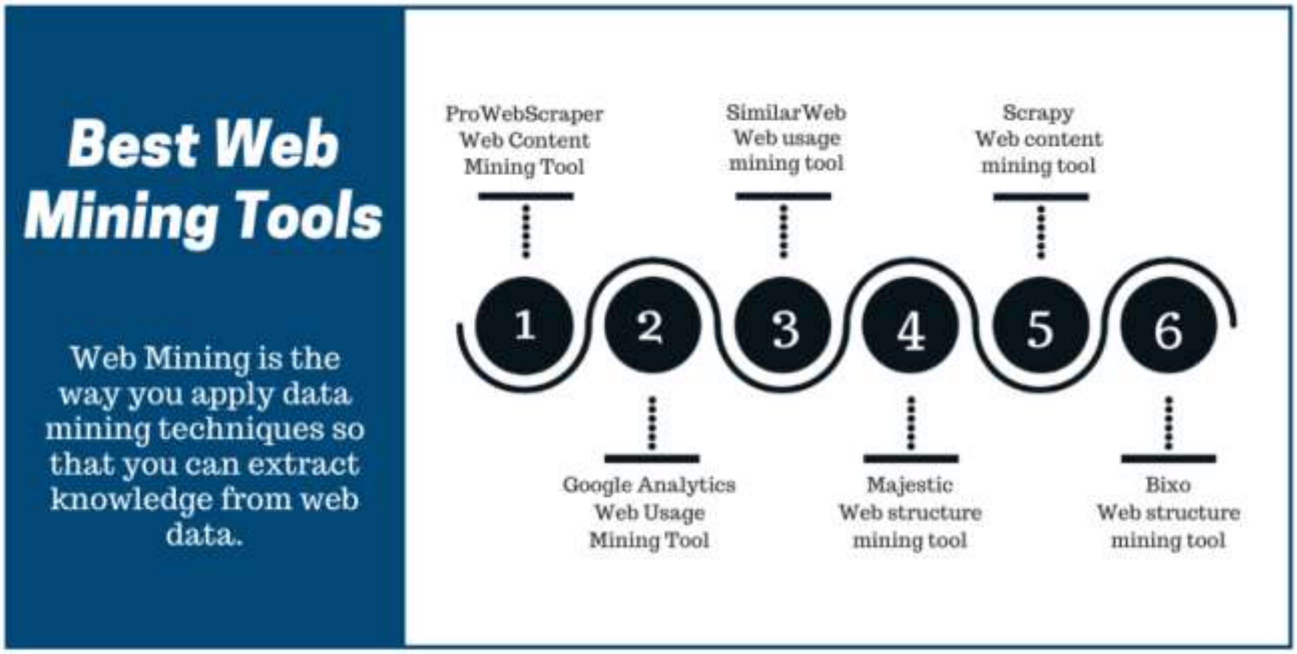
i. **Web content mining** — This is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. Techniques used in this discipline have been heavily drawn from natural language processing (NLP) and information retrieval.

ii. **Web structure mining** — This is the process of analyzing the nodes and connection structure of a website through the use of graph theory. There are two things that can be obtained from this: the structure of a website in terms of how it is connected to other sites and the document structure of the website itself, as to how each page is connected.

iii. **Web usage mining** — This is the process of extracting patterns and information from server logs to gain insight on user activity including where the users are from, how many clicked what items on the site and the types of activities being done on the site.



Web Mining Tools


Best Web Mining Tools
1
ProWebScraper Web Content Mining Tool
2
Google Analytics Web Usage Mining Tool
3
SimilarWeb Web usage mining tool
4
Majestic Web structure mining tool
5
Scrapy Web content mining tool
6
Bixo Web structure mining tool

Web Mining is the way you apply data mining techniques so that you can extract knowledge from web data.

Comparison Between Data mining and Web mining:

Points	Data Mining	Web Mining
Definition	Data Mining is the process that attempts to discover pattern and hidden knowledge in large data sets in any system.	Web Mining is the process of data mining techniques to automatically discover and extract information from web documents.
Application	Data Mining is very useful for web page analysis.	Web Mining is very useful for a particular website and e-service.
Target Users	Data scientist and data engineers.	Data scientists along with data analysts.
Access	Data Mining is access data privately.	Web Mining is access data publicly.
Structure	In Data Mining get the information from explicit structure.	In Web Mining get the information from structured, unstructured and semi-structured web pages.
Problem Type	Clustering, classification, regression, prediction, optimization and control.	Web content mining, Web structure mining.
Tools	It includes tools like machine learning algorithms.	Special tools for web mining are Scrapy, PageRank and Apache logs.
Skills	It includes approaches for data cleansing, machine learning algorithms. Statistics and probability.	It includes application level knowledge, data engineering with mathematical modules like statistics and probability.