Data Mining Introduction

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data.

Data mining is also called ***Knowledge Discovery in Database (KDD)***.

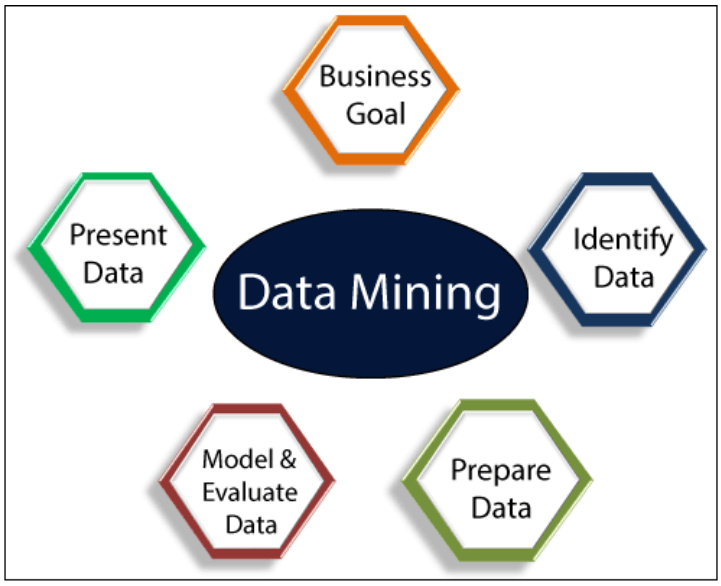
The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

What is Data Mining?

- DM is 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
- DM 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,
    - for efficient analysis,
    - for helping decision making and other data requirement
    - to eventually cost-cutting and generating revenue.
- DM is the act of
    - automatically searching for large stores of information
    - to find trends and patterns
    - that go beyond simple analysis procedures.
- DM utilizes complex mathematical algorithms for data segments and evaluates the probability of future events.
- DM is also called Knowledge Discovery of Data (KDD) refers to
    - the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

DM in Business:

- DM 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.



DM and Data Science

- Data Mining is an activity which is a part of a broader Knowledge Discovery in Databases (KDD) Process while Data Science is a field of study just like Applied Mathematics or Computer Science.
- Often Data Science is looked upon in a broad sense while Data Mining is considered a niche.
- Some activities under Data Mining such as statistical analysis, writing data flows and pattern recognition can intersect with Data Science. Hence, Data Mining becomes a subset of Data Science.
- Machine Learning in Data Mining is used more in pattern recognition while in Data Science it has a more general use.
- Data Science and Data Mining should not be confused with Big Data Analytics and one can have both Miners and Scientists working on big datasets.

Example : Consider a scenario where you are a major retailer in India. You have 50 stores operating in 10 major cities in India and you have been operational for 10 years.

Scenario 1: You want to study the last 8 years' data to find the number of sales of sweets during festive seasons of 3 cities. It is a job of Data Mining expert. A Data Miner would probably go through historical information stored in legacy systems and employ algorithms to extract trends.

Scenario 2: You want to know which sweets have received more positive reviews. In this case, your sources of data may not be limited to databases, they could extend to social websites or customer feedback messages. It is a job of Data Scientist. A person employed as a Data Scientist is more suited to apply algorithms and conduct this socio-computational analysis.

Data Science vs Data Mining Comparison Table

Below is the comparison table between Data Science and Data Mining.

Basis for comparison	Data Mining	Data Science
What is it?	A technique	An area
Focus	Business process	Scientific study
Goal	Make data more usable	Building Data-centric products for an organization
Output	Patterns	Varied
Purpose	Finding trends previously not known	Social analysis, building predictive models, unearthing unknown facts, and more
Vocational Perspective	Someone with a knowledge of navigating across data and statistical understanding can conduct data mining	A person needs to understand Machine Learning, Programming, info-graphic techniques and have the domain knowledge to become a data scientist
Extent	Data mining can be a subset of Data Science as Mining activities are part of the Data Science pipeline	Multidisciplinary – Data Science consists of Data Visualizations, Computational Social Sciences, Statistics, Data Mining, Natural Language Processing, et cetera
Deals with (the type of data)	Mostly structured	All forms of data – structured, semi-structured and unstructured
Other less popular names	Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction	Data-driven Science

Types of Data Mining

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.

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.

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.

Advantages of Data Mining

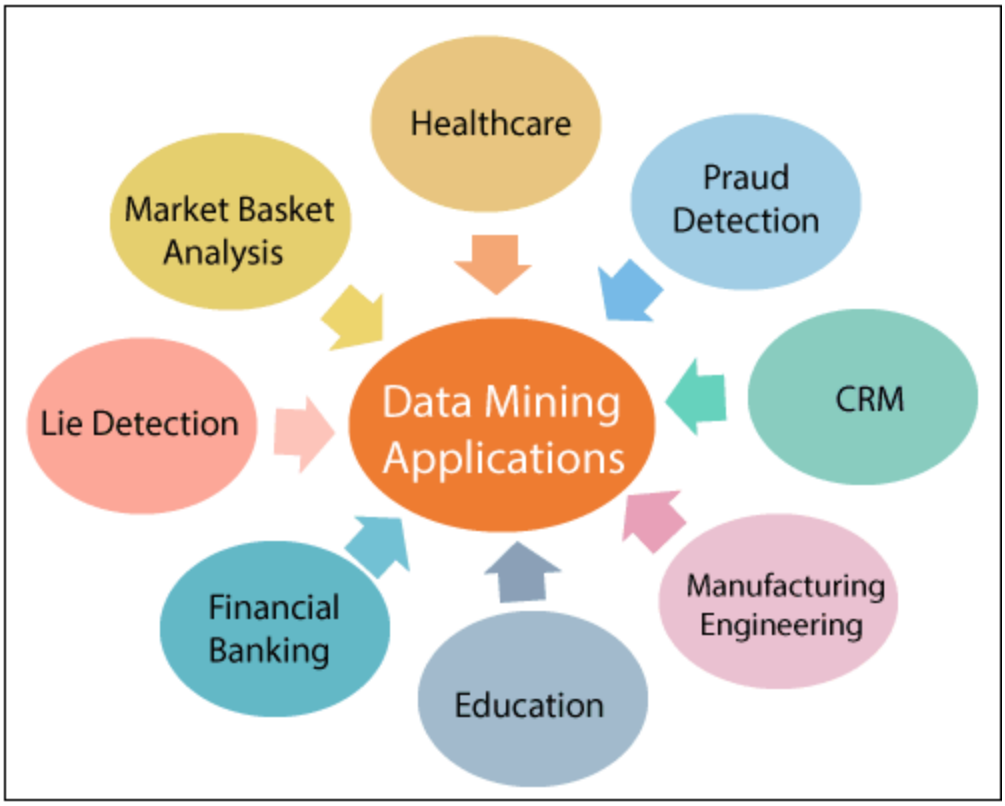
- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making 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:

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.

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.

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.

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.

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.

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.

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.

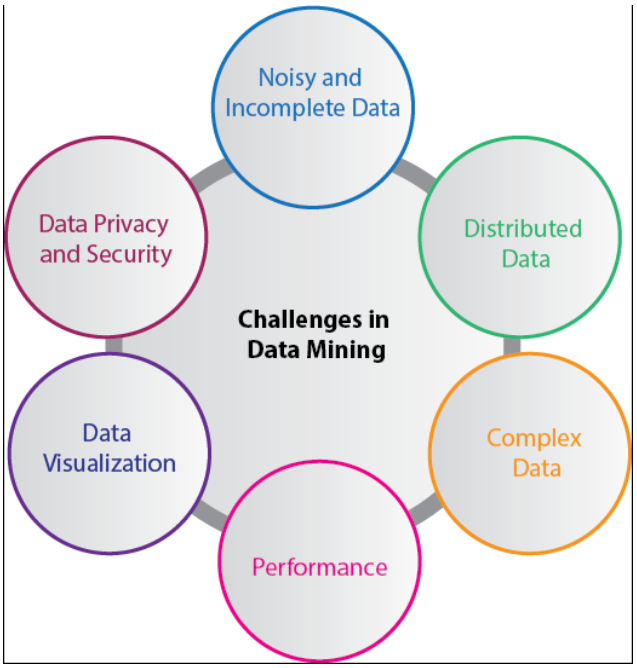
Data Mining Financial Banking:

- 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.



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)makes 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.

There are many more challenges in data mining in addition to the problems above-mentioned. More problems are disclosed as the actual data mining process begins, and the success of data mining relies on getting rid of all these difficulties.

Data Mining Techniques

- Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets.
- These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees.
- Thus, data mining incorporates analysis and prediction.
- Depending on various methods and technologies from the intersection of machine learning, database management, and statistics, professionals in data mining have devoted their objectives to better understanding how to process and make conclusions from the huge amount of data, but what are the methods they use to make it happen?
- Major major data mining techniques are association, classification, clustering, prediction, sequential patterns, and regression.

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

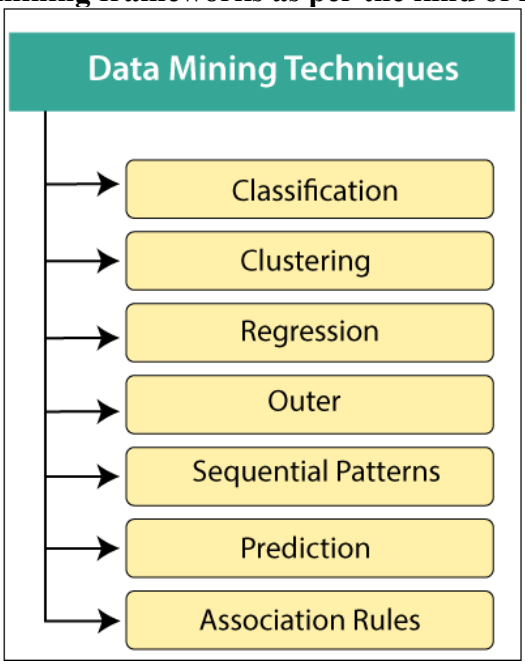
i. **Classification of Data mining frameworks as per the type of data sources mined:**
This classification is as per the type of data handled. For example, multimedia, spatial data, text data, time-series data, World Wide Web, and so on..

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.

Classification of data mining frameworks as per the kind of knowledge discovered:



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.

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 Outilier 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:

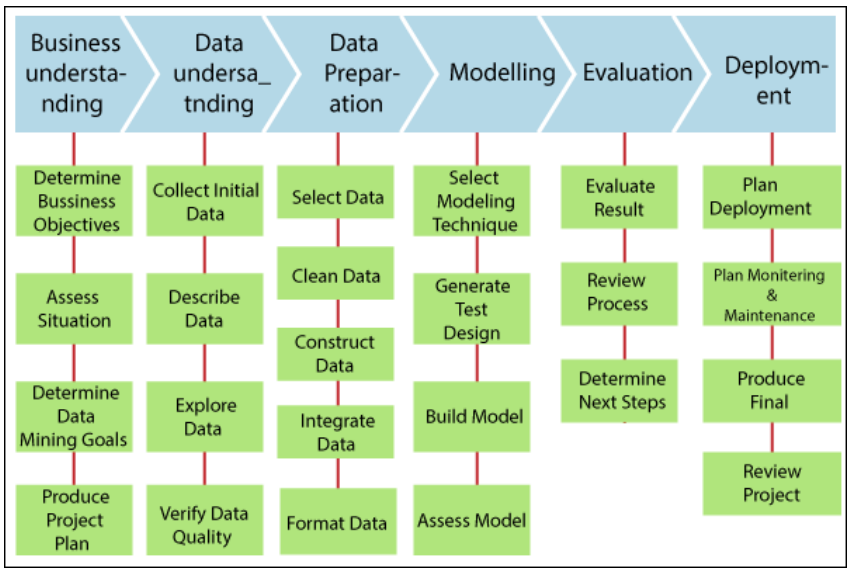
Prediction used a combination of other data mining techniques such as trends, clustering, classification, etc. It analyzes past events or instances in the right sequence to predict a future event.

next →← prev

Data Mining Implementation Process

Many different sectors are taking advantage of data mining to boost their business efficiency, including manufacturing, chemical, marketing, aerospace, etc. Therefore, the need for a conventional data mining process improved effectively. Data mining techniques must be reliable, repeatable by company individuals with little or no knowledge of the data mining context. As a result, a cross-industry standard process for data mining (CRISP-DM) was first introduced in 1990, after going through many workshops, and contribution for more than 300 organizations.

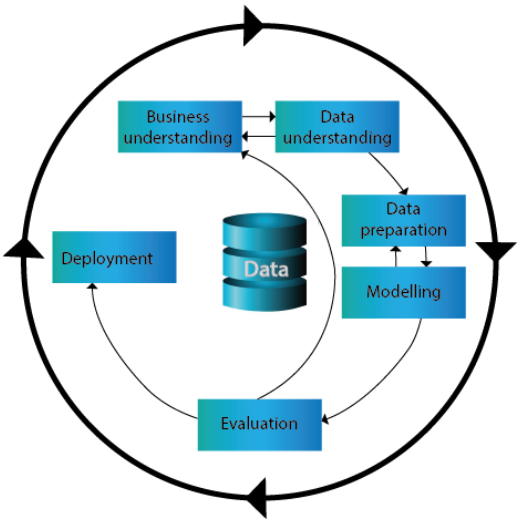
Data mining is described as a process of finding hidden precious data by evaluating the huge quantity of information stored in data warehouses, using multiple data mining techniques such as Artificial Intelligence (AI), Machine learning and statistics.



Let's examine the implementation process for data mining in details:

The Cross-Industry Standard Process for Data Mining (CRISP-DM)

Cross-industry Standard Process of Data Mining (CRISP-DM) comprises of six phases designed as a cyclical method as the given figure:



1. Business understanding:

It focuses on understanding the project goals and requirements form a business point of view, then converting this information into a data mining problem afterward a preliminary plan designed to accomplish the target.

Tasks:

- Determine business objectives
- Access situation
- Determine data mining goals
- Produce a project plan

Determine business objectives:

- It Understands the project targets and prerequisites from a business point of view.
- Thoroughly understand what the customer wants to achieve.
- Reveal significant factors, at the starting, it can impact the result of the project.

Access situation:

- It requires a more detailed analysis of facts about all the resources, constraints, assumptions, and others that ought to be considered.

Determine data mining goals:

- A business goal states the target of the business terminology. For example, increase catalog sales to the existing customer.
- A data mining goal describes the project objectives. For example, It assumes how many objects a customer will buy, given their demographics details (Age, Salary, and City) and the price of the item over the past three years.

Produce a project plan:

- It states the targeted plan to accomplish the business and data mining plan.
- The project plan should define the expected set of steps to be performed during the rest of the project, including the latest technique and better selection of tools.

2. Data Understanding:

Data understanding starts with an original data collection and proceeds with operations to get familiar with the data, to data quality issues, to find better insight in data, or to detect interesting subsets for concealed information hypothesis.

Tasks:

- Collects initial data
- Describe data
- Explore data
- Verify data quality

Collect initial data:

- It acquires the information mentioned in the project resources.
- It includes data loading if needed for data understanding.
- It may lead to original data preparation steps.
- If various information sources are acquired then integration is an extra issue, either here or at the subsequent stage of data preparation.

Describe data:

- It examines the "gross" or "surface" characteristics of the information obtained.
- It reports on the outcomes.

Explore data:

- Addressing data mining issues that can be resolved by **querying, visualizing,** and **reporting,** including:
  - Distribution of important characteristics, results of simple aggregation.
  - Establish the relationship between the small number of attributes.
  - Characteristics of important sub-populations, simple statical analysis.

- It may refine the data mining objectives.
- It may contribute or refine the information description, and quality reports.
- It may feed into the transformation and other necessary information preparation.

Verify data quality:

- It examines the data quality and addressing questions.

3. Data Preparation:

- It usually takes more than 90 percent of the time.
- It covers all operations to build the final data set from the original raw information.
- Data preparation is probable to be done several times and not in any prescribed order.

Tasks:

- Select data
- Clean data
- Construct data
- Integrate data
- Format data

Select data:

- It decides which information to be used for evaluation.
- In the data selection criteria include significance to data mining objectives, quality and technical limitations such as data volume boundaries or data types.
- It covers the selection of characteristics and the choice of the document in the table.

Clean data:

- It may involve the selection of clean subsets of data, inserting appropriate defaults or more ambitious methods, such as estimating missing information by modeling.

Construct data:

- It comprises of Constructive information preparation, such as generating derived characteristics, complete new documents, or transformed values of current characteristics.

Integrate data:

- Integrate data refers to the methods whereby data is combined from various tables, or documents to create new documents or values.

Format data:

- Formatting data refer mainly to linguistic changes produced to information that does not alter their significance but may require a modeling tool.

4. Modeling:

In modeling, various modeling methods are selected and applied, and their parameters are measured to optimum values. Some methods gave particular requirements on the form of data. Therefore, stepping back to the data preparation phase is necessary.

Tasks:

- Select modeling technique
- Generate test design
- Build model
- Access model

Select modeling technique:

- It selects the real modeling method that is to be used. For example, decision tree, neural network.
- If various methods are applied,then it performs this task individually for each method.

Generate test Design:

- Generate a procedure or mechanism for testing the validity and quality of the model before constructing a model. For example, in classification, error rates are commonly used as quality measures for data mining models. Therefore, typically separate the data set into train and test set, build the model on the train set and assess its quality on the separate test set.

Build model:

- To create one or more models, we need to run the modeling tool on the prepared data set.

Assess model:

- It interprets the models according to its domain expertise, the data mining success criteria, and the required design.
- It assesses the success of the application of modeling and discovers methods more technically.
- It Contacts business analytics and domain specialists later to discuss the outcomes of data mining in the business context.

5. Evaluation:

- At the last of this phase, a decision on the use of the data mining results should be reached.
- It evaluates the model efficiently, and review the steps executed to build the model and to ensure that the business objectives are properly achieved.
- The main objective of the evaluation is to determine some significant business issue that has not been regarded adequately.
- At the last of this phase, a decision on the use of the data mining outcomes should be reached.

Tasks:

- Evaluate results
- Review process
- Determine next steps

Evaluate results:

- It assesses the degree to which the model meets the organization's business objectives.
- It tests the model on test apps in the actual implementation when time and budget limitations permit and also assesses other data mining results produced.
- It unveils additional difficulties, suggestions, or information for future instructions.

Review process:

- The review process does a more detailed evaluation of the data mining engagement to determine when there is a significant factor or task that has been somehow ignored.
- It reviews quality assurance problems.

Determine next steps:

- It decides how to proceed at this stage.
- It decides whether to complete the project and move on to deployment when necessary or whether to initiate further iterations or set up new data-mining initiatives.it includes resources analysis and budget that influence the decisions.

6. Deployment:

Determine:

- Deployment refers to how the outcomes need to be utilized.

Deploy data mining results by:

- It includes scoring a database, utilizing results as company guidelines, interactive internet scoring.

- The information acquired will need to be organized and presented in a way that can be used by the client. However, the deployment phase can be as easy as producing. However, depending on the demands, the deployment phase may be as simple as generating a report or as complicated as applying a repeatable data mining method across the organizations.

Tasks:

- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project

Plan deployment:

- To deploy the data mining outcomes into the business, takes the assessment results and concludes a strategy for deployment.
- It refers to documentation of the process for later deployment.

Plan monitoring and maintenance:

- It is important when the data mining results become part of the day-to-day business and its environment.
- It helps to avoid unnecessarily long periods of misuse of data mining results.
- It needs a detailed analysis of the monitoring process.

Produce final report:

- A final report can be drawn up by the project leader and his team.
- It may only be a summary of the project and its experience.
- It may be a final and comprehensive presentation of data mining.

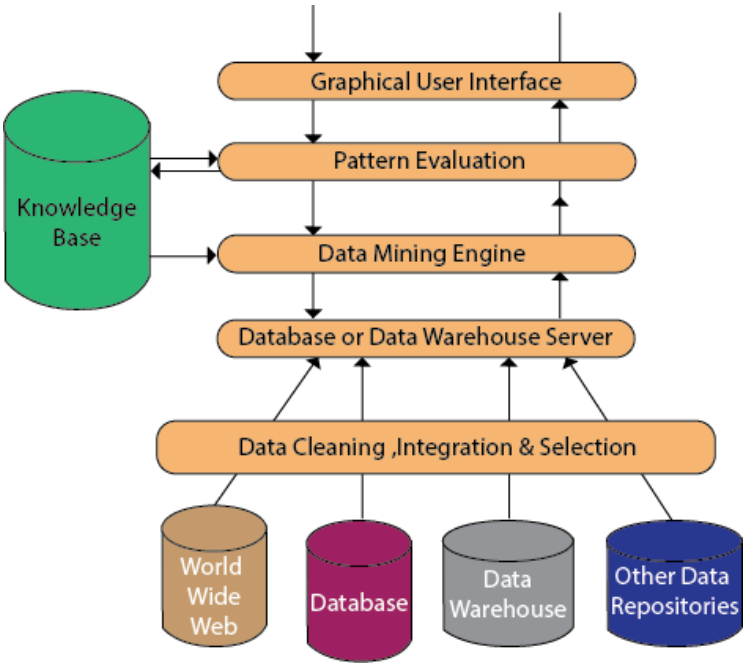
Review project:

- Review projects evaluate what went right and what went wrong, what was done wrong, and what needs to be improved.

Data Mining Architecture

Data mining is a significant method where previously unknown and potentially useful information is extracted from the vast amount of data. The data mining process involves several components, and these components constitute a data mining system architecture.

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.



Data Source:

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses. Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Different processes:

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server. These procedures are not as easy as we think. Several methods may be performed on the data as part of selection, integration, and cleaning.

Database or Data Warehouse Server:

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine:

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module:

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

This segment commonly employs stake measures that cooperate with the data mining modules to focus the search towards fascinating patterns. It might utilize a stake threshold to filter out discovered patterns. On the other hand, the pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used. For efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

Graphical User Interface:

The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process. This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

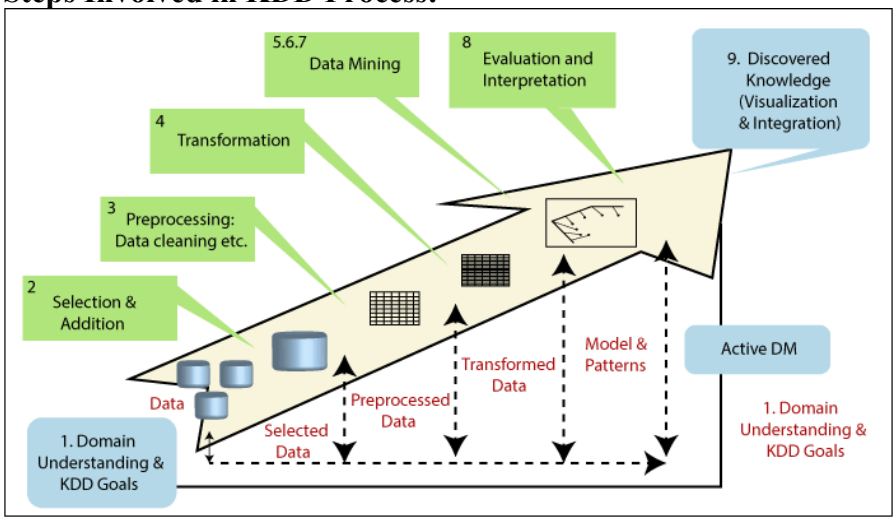
Knowledge Base:

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process. The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable. The pattern assessment module regularly interacts with the knowledge base to get inputs, and also update it.

KDD and DM

Data Mining also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

Steps Involved in KDD Process:



1.a Define the objective:

- This is the initial preliminary step.
- It develops the scene for understanding what should be done with the various decisions like transformation, algorithms, representation, etc.
- The individuals who are in charge of a KDD venture need to understand and characterize the objectives of the end-user and the environment in which the knowledge discovery process will occur (involves relevant prior knowledge).

1.b Preparing DataSet

- Discovering data accessiblity, obtaining important data,
- Collecting and integrating all the data for knowledge discovery
- This process is important because of DM learns and discovers from the accessible data.
- If some significant attributes are missing, at that point, then the entire study may be unsuccessful from this respect, the more attributes are considered.
- On the other hand, to organize, collect, and operate advanced data repositories is expensive. The interactive and iterative aspect of the KDD is taking place where process will begin with best available data sets and later expands and observes the impact in terms of knowledge discovery and modeling.

2. **Data Cleaning**: Data cleaning is defined as removal of incorrect, incomplete, irrelevant, duplicate or irregularly formatted information.

- Cleaning in case of **Missing values**.
- Cleaning **noisy** data, where noise is a random or variance error.

- Cleaning with **Data discrepancy detection** and **Data transformation tools**.

3. **Data Integration**: Data integration is defined as heterogeneous data from multiple sources combined in a common source (DataWarehouse).
   - Data integration using **Data Migration tools**.
   - Data integration using **Data Synchronization tools**.
   - Data integration using **ETL**(Extract-Load-Transformation) process.
4. **Data Selection**: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   - Data selection using **Neural network**.
   - Data selection using **Decision Trees**.
   - Data selection using **Naive bayes**.
   - Data selection using **Clustering**, **Regression**, etc.

   NOTE: After this step, data reliability is improved. It incorporates data clearing, for example, Handling the missing quantities and removal of noise or outliers. It might include complex statistical techniques or use a Data Mining algorithm in this context. For example, when one suspects that a specific attribute of lacking reliability or has many missing data, at this point, this attribute could turn into the objective of the Data Mining supervised algorithm. A prediction model for these attributes will be created, and after that, missing data can be predicted. The expansion to which one pays attention to this level relies upon numerous factors. Regardless, studying the aspects is significant and regularly revealing by itself, to enterprise data frameworks.

5. **Data Transformation**: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

   Data Transformation is a two step process:
   - **Data Mapping**: Assigning elements from source base to destination to capture transformations.
   - **Code generation**: Creation of the actual transformation program.

   In this stage, the creation of appropriate data for Data Mining is prepared and developed. Techniques here incorporate dimension reduction(for example, feature selection and extraction and record sampling), also attribute transformation(for example, discretization of numerical attributes and functional transformation). This step can be essential for the success of the entire KDD project, and it is typically very project-specific. For example, in medical assessments, the quotient of attributes may often be the most significant factor and not each one by itself. In business, we may need to think about impacts beyond our control as well as efforts and transient issues. For example, studying the impact of advertising accumulation. However, if we do not utilize the right transformation at the starting, then we may acquire an amazing effect that insights to us about the transformation required in the next iteration. Thus, the KDD process follows upon itself and prompts an understanding of the transformation required.

6. **Data Mining**: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   - Transforms task relevant data into **patterns**.
   - Decides purpose of model using **classification** or **characterization**.

   NOTE:

   - **Prediction and description**

     We are now prepared to decide on which kind of Data Mining to use, for example, classification, regression, clustering, etc. This mainly relies on the KDD objectives, and also on the previous steps. There are two significant objectives in Data Mining, the first one is a prediction, and the second one is the description. Prediction is usually referred to as supervised Data Mining, while descriptive Data Mining incorporates the unsupervised and visualization aspects of Data Mining. Most Data Mining techniques depend on inductive learning, where a model is built explicitly or implicitly by generalizing from an adequate number of preparing models. The fundamental assumption of the inductive approach is that the prepared model applies to future cases. The technique also takes into account the level of meta-learning for the specific set of accessible data.

   - **Selecting the Data Mining algorithm**

     Having the technique, we now decide on the strategies. This stage incorporates choosing a particular technique to be used for searching patterns that include multiple inducers. For example, considering precision versus understandability, the previous is better with neural networks, while the latter is better with decision trees. For each system of meta-learning, there are several possibilities of how it can be succeeded. Meta-learning focuses on clarifying what causes a Data Mining algorithm to be fruitful or not in a specific issue. Thus, this methodology attempts to understand the situation under which a Data Mining algorithm is most suitable. Each algorithm has parameters and strategies of leaning, such as ten folds cross-validation or another division for training and testing.

   - **Utilizing the Data Mining algorithm**

     At last, the implementation of the Data Mining algorithm is reached. In this stage, we may need to utilize the algorithm several times until a satisfying outcome is obtained. For example, by turning the algorithms control parameters, such as the minimum number of instances in a single leaf of a decision tree.

7. **Pattern Evaluation**: Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.

- Find **interestingness score** of each pattern.
- Uses **summarization** and **Visualization** to make data understandable by user.

Note :

In this step, we assess and interpret the mined patterns, rules, and reliability to the objective characterized in the first step. Here we consider the preprocessing steps as for their impact on the Data Mining algorithm results. For example, including a feature in step 4, and repeat from there. This step focuses on the comprehensibility and utility of the induced model. In this step, the identified knowledge is also recorded for further use. The last step is the use, and overall feedback and discovery results acquire by Data Mining.

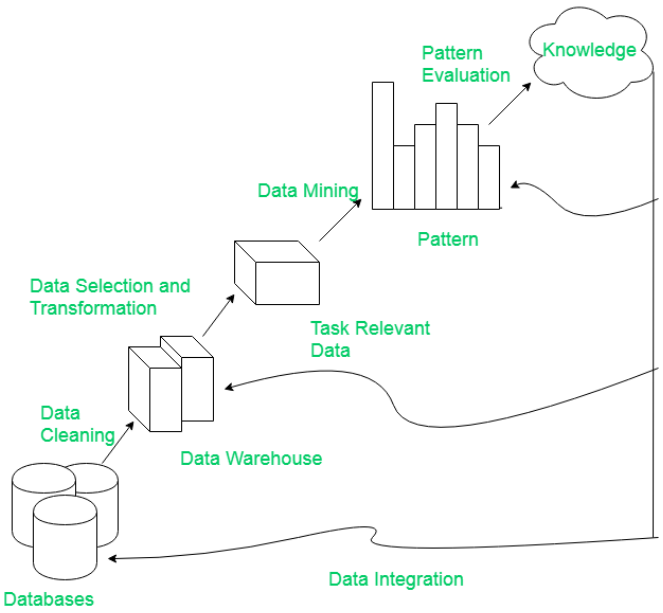
8. **Knowledge representation**: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
   - Generate **reports**.
   - Generate **tables**.
   - Generate **discriminant rules**, **classification rules**, **characterization rules**, etc.

Note:

- KDD is an **iterative process** where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.
- **Preprocessing of databases** consists of **Data cleaning** and **Data Integration**
- **Using the discovered knowledge**
    - Now, we are prepared to include the knowledge into another system for further activity. The knowledge becomes effective in the sense that we may make changes to the system and measure the impacts. The accomplishment of this step decides the effectiveness of the whole KDD process. There are numerous challenges in this step, such as losing the "laboratory conditions" under which we have worked. For example, the knowledge was discovered from a certain static depiction, it is usually a set of data, but now the data becomes dynamic. Data structures may change certain quantities that become unavailable, and the data domain might be modified, such as an attribute that may have a value that was not expected previously.

Data Mining: Data Warehouse Process

KDD incorporating Data Warehouse



Data Warehouses are information gathered from multiple sources and saved under a schema that is living on the identical site. It is made with the aid of diverse techniques inclusive of the following processes :

1. Data Cleanup:

2. Data Integration:

Data integration is the process of integrating data from different assets right into a unified view. The integration manner starts with a startup and includes steps which include refinement, ETL mapping, and conversion. Data integration ultimately permits analytics tools to create powerful and cheap enterprise intelligence.

In a typical data integration procedure, the client sends a request for information to the master server. The master server prepares the vital records from internal and external assets. Extracts facts from sources and then integrates them into a single information set. It is then returned again to the client for use.

3. Data Transformation:

Data transformation is the manner of converting information from one layout or shape to another layout or structure. Data Transformation is critical for features which include data integration and information management. Data transformation has different capabilities: you could alternate the records types relying on the desires of your project, enrich or aggregate the records through casting off invalid or duplicate data.

Generally, the technique consists of two stages.

In the **first step**, you should:

- Perform an information search that identifies assets and data types.
- Determine the structure and information changes that occur.

- Mapping data to discover how character fields are mapped, edited, inserted, filtered, and stored.

In the **second step**, you must:

- Extract data from the original source. The size of the supply can range from a connected tool to a dependable useful resource along with a database or streaming resources, including telemetry or logging files from clients who use your web application.
- Send data to the target site.
- The target may be a database or a data warehouse that manages structured and unstructured records.

4. Loading Data:

Data loading is the manner of copying and loading data from a report, folder or application to a database or similar utility. This is usually done via copying digital data from the source and pasting or loading the records into a data warehouse or processing tools.

Data-loading is used in data extraction and loading methods. Typically, such information is loaded in a different format than the original location of the source.

5. Data Refreshing:

In this process, the data stored in the warehouse is periodically refreshed so that they maintain its integrity.

A data warehouse is a model of Multidimensional data structures that are known as "Data Cube" in which every dimension represents an attribute or different set of attributes in the schema of the data and each cell is used to store the value. Data is gathered from various sources such as Hospitals, Banks, Organizations and many more and goes through a process called ETL(Extract, Transform, Load).

1. **Extract:** This process reads the data from the database of various sources.
2. **Transform:** It transforms the data stored inside the databases into data cubes so that it can be loaded inside the warehouse.
3. **Load:** It is a process of writing the transformed data into the data warehouse.

This process can be seen in the illustration below: