**Data Mining**

**Definition:Data Mining** is defined as the procedure of extracting information from huge sets of data.

In other words, data mining is mining knowledge from data.

**Terminologies involved in data mining:**Knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis etc.

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation.

The information or knowledge extracted so can be used for any of the following applications

* Market Analysis
* Fraud Detection
* Customer Retention
* Production Control
* Science Exploration

**DATA WAREHOUSING**

A **data warehouse**, or enterprise data warehouse (EDW), is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis, data mining, artificial intelligence (AI), and machine learning.

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

**Using Data Warehouse Information**

There are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively. They can gather data, analyze it, and take decisions based on the information present in the warehouse. The information gathered in a warehouse can be used in any of the following domains

**Tuning Production Strategies** − The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.

**Customer Analysis** − Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.

**Operations Analysis** − Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

**FUNCTIONS OF DATA WAREHOUSE TOOLS AND UTILITIES:**

* **Data Extraction** − Involves gathering data from multiple heterogeneous sources.
* **Data Cleaning** − Involves finding and correcting the errors in data.
* **Data Transformation** − Involves converting the data from legacy format to warehouse format.
* **Data Loading** − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
* **Refreshing** − Involves updating from data sources to warehouse.

**ETL:**

ETL stands for **E**xtract, **T**ransform, and **L**oad.

It is defined as a Data Integration service and allows companies to [combine data from various sources](https://hevodata.com/integrations/pipeline/) into a single, consistent data store that is loaded into a Data Warehouse or any other target system.

**STEPS INVOLVED IN THE ETL PROCESS:**

* **Extraction**: In this, the structured or unstructured data is extracted from its source and consolidated into a single repository. ETL tools automate the extraction process and create a more efficient and reliable workflow for handling large volumes of data and multiple sources.
* **Transformation**: In order to improve data integrity the data needs to be transformed such as it needs to be sorted, standardized, and redundant data should be removed. This step ensures that raw data which arrives at its new destination is fully compatible and ready to use.
* **Loading**: This is the final step of the [ETL process](https://hevodata.com/learn/etl-process/) which involves loading the data into the final destination(data lake or data warehouse). The data can be loaded all at once(full load) or at scheduled intervals(incremental load).

**ETL Tools** are applications/platforms that enable users to execute ETL processes. In simple terms, these tools help businesses move data from one or many desparate data sources to a destination. These help in making the data both digestible and accessible (and in turn analysis-ready) in the desired location – often a data warehouse.

ETL tools are the first essential step in the data warehousing process that eventually make more informed decisions in less time.

**TYPES OF ETL TOOLS:**

* **Enterprise ETL Tools**

The ETL tools are often bundled as part of a larger platform and appeal to enterprises with older, legacy systems that they need to work with and build on. These ETL tools can handle pipelines efficiently and are highly scalable since they were one of the first to offer ETL tools and mature in the market. These tools support most relational and non-relational databases.

* **Custom ETL Tools**

In this, the custom tools and pipelines are created using scripting languages like SQL or Python. While this gives you an opportunity for customization and higher flexibility, it also requires more administration and maintenance.

* **Cloud-Based ETL Tools**

These tools integrate with proprietary data sources and ingest data from different web apps or on-premises sources. These tools move data between systems and copy, transform, and enrich data before writing it to data warehouses or data lakes.

* **Open-Source ETL Tools**

Many ETL tools today are free and provide easy-to-use user interfaces for designing data exchange processes and monitoring the flow of information. An advantage of open-source solutions is that organizations can access the source code to study the tool infrastructure and extend the functionality.

**Best ETL Tools for 2023**

**HevoData ,Pentaho,Talend,AWS Glue , Azure Data Factory etc…**

**CLASSIFICATION IN DATA MINING**

Classification in data mining is a common technique that separates data points into different classes. It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple ones.

It primarily involves using algorithms that you can easily modify to improve the data quality. The primary goal of classification is to connect a variable of interest with the required variables. The algorithm establishes the link between the variables for prediction. The algorithm used for classification in data mining is called the classifier, and observations are called the instances.

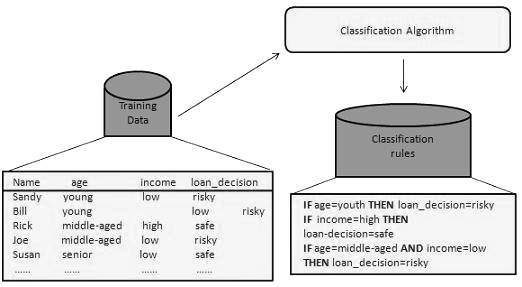
**Example:**A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.

**THE DATA CLASSIFICATION PROCESS INCLUDES TWO STEPS**

1. Building the Classifier or Model
2. Using Classifier for Classification
3. **Building the Classifier or Model**

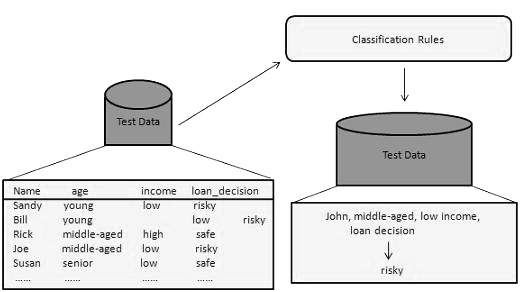
This step is the learning step or the learning phase.In this step the classification algorithms build the classifier.

The classifier is built from the training set made up of database tuples and their associated class labels. A model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data



1. **Using Classifier for Classification**

In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



**PREDICTION:**

To find a numerical output, prediction is used. The training dataset contains the inputs and numerical output values. According to the training dataset, the algorithm generates a model or predictor. When fresh data is provided, the model should find a numerical output. This approach, unlike classification, does not have a class label. A continuous-valued function or ordered value is predicted by the model.

Example: 1. Predicting the worth of a home based on facts like the number of rooms, total area, and so on.

1. A marketing manager needs to forecast how much a specific consumer will spend during a sale. In this scenario, we are bothered to forecast a numerical value. In this situation, a model or predictor that forecasts a continuous or ordered value function will be built.

**REGRESSION IN DATA MINING:**

Regression refers to a data mining technique that is used to predict the numeric values in a given data set. Regression involves the technique of fitting a straight line or a curve on numerous data points.

For example, regression might be used to predict the product or service cost or other variables. It is also used in various industries for business and marketing behavior, trend analysis, and financial forecast.

**Regression is divided into five different types**

1. Linear Regression
2. Logistic Regression
3. Lasso Regression
4. Ridge Regression
5. Polynomial Regression

The most popular types of regressions are linear and logistic regressions.

**LINEAR REGRESSION**

Linear regression is the type of regression that forms a relationship between the target variable and one or more independent variables utilizing a straight line. The given equation represents the equation of linear regression.

It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression. The linear regression model provides a sloped straight line representing the relationship between the variables.



The values for x and y variables are training datasets (data points) for Linear Regression model representation.

Y = a + b\*X + e.

Where,

a represents the intercept of the line (The point where the line or curve crosses the axis of the graph is called intercept. If a point crosses the x-axis, then it is called the x-intercept. If a point crosses the y-axis, then it is called the y-intercept.)

b represents the slope of the regression line

e represents the random error

X represent the predictor variable (independent)

Y represent the target variable (dependent).

**Applications:**

1. Medical researchers can use this regression model to determine the relationship between independent characteristics, such as age and body weight, and dependent ones, such as blood pressure. This can help reveal the risk factors associated with diseases. They can use this information to identify high-risk patients and promote healthy lifestyles.
2. Financial analysts use linear models to evaluate a company's operational performance and forecast returns on investment. They also use it in the capital asset pricing model, which studies the relationship between the expected investment returns and the associated market risks. It shows companies if an investment has a fair price and contributes to decisions on whether or not to invest in the asset

**LOGISTIC REGRESSION**

When the dependent variable is binary in nature, i.e., 0 and 1, true or false, success or failure, the logistic regression technique comes into existence. Here, the target value (Y) ranges from 0 to 1, and it is primarily used for classification-based problems. Unlike linear regression, it does not need any independent and dependent variables to have a linear relationship. **Example**: Acceptance into university based on student grades.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a either Yes or No, 0 or 1, true or False, etc. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).



**Logistic regression applications in business**

1. An [e-commerce](https://www.techtarget.com/searchcio/definition/e-commerce) company that mails expensive promotional offers to customers, would like to know whether a particular customer is likely to respond to the offers or not i.e., whether that consumer will be a "responder" or a "non-responder."

2. Likewise, a credit card company will develop a model to help it predict if a customer is going to default on its credit card based on such characteristics as annual income, monthly credit card payments and the number of defaults.

3.A medical researcher may want to know the impact of a new drug on treatment outcomes across different age groups. This involves a lot of nested multiplication and division for comparing the outcomes of young and older people who never received a treatment, younger people who received the treatment, older people who received the treatment, and then the whole spontaneous healing rate of the entire group.

4. Logistic regression has become particularly popular in online advertising, enabling [marketers to predict the likelihood of specific website users](https://www.techtarget.com/searchbusinessanalytics/tip/Predictive-analytics-in-marketing-Achieving-success) who will click on particular advertisements as a yes or no percentage.

5.In [healthcare to identify risk factors for diseases](https://www.techtarget.com/searchbusinessanalytics/tip/Predictive-analytics-in-healthcare-12-valuable-use-cases) and plan preventive measures;

6. In drug research to learn the effectiveness of medicines on health outcomes across age, genderetc

7. In weather forecasting apps to predict snowfall and weather conditions;

8. in political polls to determine if voters will vote for a particular candidate;

9. in banking to predict the chances that a loan applicant will default on a loan or not, based on annual income, past defaults and past debts.

**Note: The main difference between logistic and**[**linear regression**](https://www.techtarget.com/searchenterpriseai/definition/linear-regression)**is that logistic regression provides a constant output, while linear regression provides a continuous output.**

**TIME SERIES ANALYSIS:**

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

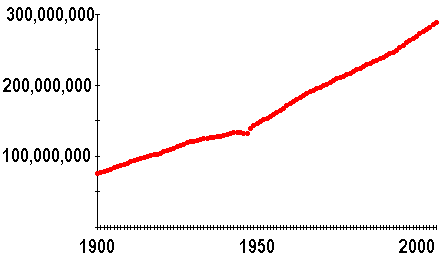
time series analysis can show how variables change over time. In other words, time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data.

Time series analysis typically requires a large number of data points to ensure consistency and reliability. An extensive data set ensures you have a representative sample size and that analysis can cut through noisy data.

**Examples:**

* Weather forecast
* Rainfall measurements
* Temperature readings
* Heart rate monitoring (ECG)
* Brain monitoring
* Quarterly sales
* Stock market analysis
* Automated stock trading
* Industry forecasts
* Interest rates

Time series is also used in several nonfinancial contexts, such as measuring the change in population over time. The figure below depicts such a time series for the growth of the U.S. population over the century from 1900 to 2000.



Dia:A time series graph of the population of the United States from 1900 to 2000.

**Summarization**

The term **Data Summarization** can be defined as the presentation of a summary/report of generated data in a comprehensible and informative manner. To relay information about the dataset, summarization is obtained from the entire dataset. It is a carefully performed summary that will convey trends and patterns from the dataset in a simplified manner.

Data has become more complex hence, there is a need to summarize the data to gain useful information. Data summarization has great importance in data mining as it can also help in deciding appropriate statistical tests to use depending on the general trends revealed from the summarization.

**Areas in which Data Summarization is implemented:**

1. [Centrality](https://hevodata.com/learn/data-summarization-in-data-mining/#ds1)
2. [Dispersion](https://hevodata.com/learn/data-summarization-in-data-mining/#ds2)
3. [Distribution of a Sample of Data](https://hevodata.com/learn/data-summarization-in-data-mining/#ds3)

**1) Centrality**

The principle of Centrality is used to describe the center or middle value of the data.Measures used to show the centrality

**Mean:** This is used to calculate the numerical average of the set of values.

**Mode:** This shows the most frequently repeated value in a dataset.

**Median:** This identifies the value in the middle of all the values in the dataset when values are ranked in order.

The most appropriate measure to use will depend largely on the shape of the dataset.

**2) Dispersion**

The dispersion of a sample refers to spreading out the values around the average (center).  It shows the amount of variation or diversity within the data. When the values are close to the center, the sample has low dispersion while high dispersion occurs when they are widely scattered about the center.

**Different measures of dispersion can be used based on the dataset**

1 **Standard deviation:** This provides a standard way of knowing what is normalorextra large or extra small and helps to understand the spread of the variable from the mean. It shows how close all the values are to the mean.

1. **Variance:** This is similar to standard deviation but it measures how tightly or loosely values are spread around the average.

**3. Range:** The range indicates the difference between the largest and the smallest values thereby showing the distance between the extremes.

**3) Distribution of a Sample of Data**

The distribution of sample data values has to do with the shape which refers to how data values are distributed across the range of values in the sample. In simple terms, it means if the values are clustered around the average to show how they are symmetrically arranged around it or if there are more values to one side than the order.

**Two ways to explore the distribution of the sample data are**

1. Graphically
2. through shape statistics.

**Clustering**

Clustering is the method of converting a group of abstract objects into classes of similar objects.Clustering is a method of partitioning a set of data or objects into a set of significant subclasses called clusters.

It helps users to understand the structure or natural grouping in a data set and used either as a stand-alone instrument to get a better insight into data distribution or as a pre-processing step for other algorithms.Data objects of a cluster can be considered as one group. First partition the information set into groups while doing cluster analysis. It is based on data similarities and then assigns the levels to the groups.

**Applications of cluster analysis in data mining:**

1. data analysis, market research, pattern recognition, and image processing.
2. It assists marketers to find different groups in their client base and based on the purchasing patterns. They can characterize their customer groups.
3. It helps in allocating documents on the internet for data discovery.
4. used in tracking applications such as detection of credit card fraud.

5. In terms of biology, It can be used to determine plant and animal taxonomies, categorization of genes with the same functionalities and gain insight into structure inherent to populations.

**REQUIREMENTS OF CLUSTERING IN DATA MINING:**

**1. Scalability:**

Scalability (to increase Style)in clustering implies that as we boost the amount of data objects, the time to perform clustering should approximately scale to the complexity order of the algorithm. If we raise the number of data objects 10 folds, then the time taken to cluster them should also approximately increase 10 times. It means there should be a linear relationship. If that is not the case, then there is some error with our implementation process.

**2. Interpretability:**

The outcomes of clustering should be interpretable, comprehensible, and usable.

**3. Discovery of clusters with attribute shape:**

The clustering algorithm should be able to find arbitrary shape clusters. They should not be limited to only distance measurements that tend to discover a spherical cluster of small sizes.

**4. Ability to deal with different types of attributes:**

Algorithms should be capable of being applied to any data such as data based on intervals (numeric), binary data, and categorical data.

**5. Ability to deal with noisy data:**

Databases contain data that is noisy, missing, or incorrect. Few algorithms are sensitive to such data and may result in poor quality clusters.

**6. High dimensionality:**

The clustering tools should not only be able to handle high dimensional data space but also the low-dimensional space.

**Association Rule:**

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

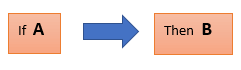
The association rule learning is one of the very important concepts of [machine learning](https://www.javatpoint.com/machine-learning), and it is employed in **Market Basket analysis, Web usage mining, continuous production, etc.** Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby. Consider the below diagram:



How does Association Rule Learning work?

Association rule learning works on the concept of If and Else Statement, such as if A then B.



Here, If element is called **antecedent**, and then statement is called as **Consequent**. These types of relationships where we can find out some association or relation between two items is known *as single cardinality*. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics. They are,

1. **Support**
2. **Confidence**
3. **Lift**

**Support**

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

Association Rule Learning

**Confidence**

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

Association Rule Learning

**Lift**

It is the strength of any rule, which can be defined as below formula:

Association Rule Learning

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

If **Lift= 1**: The probability of occurrence of antecedent and consequent is independent of each other.

**Lift>1**: It determines the degree to which the two itemsets are dependent to each other.

**Lift<1**: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

**Association rule learning can be divided into three types of algorithms:**

1. **Apriori**
2. **Eclat**
3. **F-P Growth Algorithm**

**Apriori Algorithm**

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

**Eclat Algorithm**stands for **Equivalence Class Transformation**. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.

**F-P Growth Algorithm**

The F-P growth algorithm stands for **Frequent Pattern**, and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern or tree. The purpose of this frequent tree is to extract the most frequent patterns.

**APPLICATIONS OF ASSOCIATION RULE LEARNING**

1. **Market Basket Analysis:**  This technique is commonly used by big retailers to determine the association between items.
2. **Medical Diagnosis:**  it helps in identifying the probability of illness for a particular disease.
3. **Protein Sequence:** The association rules help in determining the synthesis of artificial Proteins.
4. **Catalog Design**

**SEQUENCE DISCOVERY OR SEQUENTIAL PATTERN IN DATA MINING :**

  Sequential pattern mining, also known as GSP (Generalized Sequential Pattern) mining, is a technique used to identify patterns in sequential data. The goal of GSP mining is to discover patterns in data that occur over time, such as customer buying habits, website navigation patterns, or sensor data.

**Sequence:** A sequence is formally defined as the ordered set of items {s1, s2, s3, …,sn}. As the name suggests, it is the sequence of items occurring together. It can be considered as a transaction or purchased items together in a basket.

**Subsequence:** The subset of the sequence is called a subsequence. Suppose {a, b, g, q, y, e, c} is a sequence. The subsequence of this can be {a, b, c} or {y, e}. Observe that the subsequence is not necessarily consecutive items of the sequence. From the sequences of databases, subsequences are found from which the generalized sequence patterns are found at the end.

**Sequence pattern:**A sub-sequence is called a pattern when it is found in multiple sequences. The goal of the GSP algorithm is to mine the sequence patterns from the large database. The database consists of the sequences. When a subsequence has a frequency equal to more than the “support” value.  For example: the pattern <a, b> is a sequence pattern mined from sequences {b, x, c, a}, {a, b, q}, and {a, u, b}.

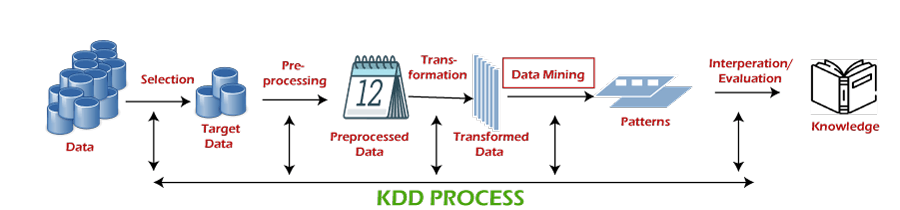
**Uses of GSP(Generalized sequence pattern) mining:**

1. **Market basket analysis:** GSP mining can be used to analyze customer buying habits and identify products that are frequently purchased together. This can help businesses to optimize their product placement and marketing strategies.
2. **Fraud detection:** GSP mining can be used to identify patterns of behavior that are indicative of fraud, such as unusual patterns of transactions or access to sensitive data.
3. **Website navigation:** GSP mining can be used to analyze website navigation patterns, such as the sequence of pages visited by users, and identify areas of the website that are frequently accessed or ignored.
4. **Sensor data analysis**: GSP mining can be used to analyze sensor data, such as data from IoT devices, and identify patterns in the data that are indicative of certain conditions or states.
5. **Social media analysis**:   GSP mining can be used to analyze social media data, such as posts and comments, and identify patterns in the data that indicate trends, sentiment, or other insights.
6. **Medical data analysis**: GSP mining can be used to analyze medical data, such as patient records, and identify patterns in the data that are indicative of certain health conditions or trends.

**KNOWLEDGE DISCOVERY IN DATABASES(KDD):**

The KDD method is a complex and iterative approach to knowledge extraction from[big data](https://www.codingninjas.com/studio/library/big-data). Extraction of knowledge from massive data is accomplished through the intricate and iterative KDD process.

It is an extensive method that includes data mining as one of its steps. It includes utilising a variety of algorithms and statistical methods to sort through large amounts of data and identify relevant and valuable data.



**KDD PROCESS STEPS**

Knowledge discovery in the database process includes the following steps, such as:

1. **Goal identification:** Develop and understand the application domain and the relevant prior knowledge and identify the KDD process's goal from the customer perspective.
2. **Creating a target data set:** Selecting the data set or focusing on a set of variables or data samples on which the discovery was made.
3. **Data cleaning and preprocessing:** Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields
4. **Data reduction and projection:** Finding useful features to represent the data depending on the purpose of the task. The effective number of variables under consideration may be reduced through dimensionality reduction methods or conversion.
5. **Matching process objectives:** KDD with step 1 a method of mining particular. For example, summarization, classification, regression, clustering, and others.
6. **Modeling and exploratory analysis and hypothesis selection:** Choosing the algorithms or data mining and selecting the method or methods to search for data patterns. This process includes deciding which model and parameters may be appropriate and the matching of data mining methods, particularly with the general approach of the KDD process
7. **Data Mining:** The search for patterns of interest in a particular representational form or a set of these representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data mining method to carry out the preceding steps properly.
8. **Presentation and evaluation:** Interpreting mined patterns, possibly returning to some of the steps between steps 1 and 7 for additional iterations. This step may also involve the visualization of the extracted patterns and models or visualization of the data given the models drawn.
9. **Taking action on the discovered knowledge:** Using the knowledge directly, incorporating the knowledge in another system for further action, or simply documenting and reporting to stakeholders. This process also includes checking and resolving potential conflicts with previously believed knowledge (or extracted).

**APPLICATIONS OF KDD**

1. **Business and Marketing:** User analysis, market prediction, Segmenting clients, and focused marketing are all examples of business and marketing databases.
2. **Manufacturing:** Predictive system analysis, process improvement, and quality control.
3. **Finance:** Fraud investigation, evaluation of credit risk, and stock market research in the finance sector can be analysed using the KDD method.
4. **Healthcare:** Drug progress, patient monitoring, and disease diagnosis from a large set of patient data.
5. **Scientific research:** Identifying patterns in massive scientific databases, such as genetics, astronomy, and climate.

**DATA MINING VS KDD.**

| **Key Features** | **Data Mining** | **KDD** |
| --- | --- | --- |
| Basic Definition | Data mining is the process of identifying patterns and extracting details about big data sets using intelligent methods. | The KDD method is a complex and iterative approach to knowledge extraction from big data. |
| Goal | To extract patterns from datasets. | To discover knowledge from datasets. |
| Scope | In the KDD method, the fourth phase is called "data mining." | KDD is a broad method that includes data mining as one of its steps. |
| Used Techniques | Classification  Clustering  Decision Trees  Dimensionality Reduction  Neural Networks  Regression | Data cleaning  Data Integration  Data selection  Data transformation  Data mining  Pattern evaluation  Knowledge Presentation |
| Example | Clustering groups of data elements based on how similar they are. | Data analysis to find patterns and links. |

**Mining Issues :**

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources.

**FACTORS THAT CREATE SOME ISSUES.**

1. Mining Methodology and User Interaction issues
2. Performance Issues
3. Diverse Data Types Issues

**MINING METHODOLOGY AND USER INTERACTION ISSUES:**

1. **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
2. **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
3. **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
4. **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
5. **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
6. **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
7. **Pattern evaluation** − The patterns discovered should be interesting or relevant.

**PERFORMANCE ISSUES :**

1. **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
2. **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

**DIVERSE DATA TYPES ISSUES :**

1. **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data(data related to a specific location on the Earth's surface )etc. It is not possible for one system to mine all these kind of data.
2. **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

**DATA MINING METRICS**

Data mining metrics may be defined as a set of measurements which can help in determining the efficacy of a Data mining Method or Algorithm. They are important to help take the right decision as like as choosing the right data mining technique or algorithm.

Evaluation metrics can be divided into two main categories:

1. Predictive
2. Descriptive.

**PREDICTIVE METRICS** : evaluate how well the model can predict the outcome of new or unseen data, based on the data it was trained on. They are often used for supervised data mining tasks, such as classification or regression.

Examples : accuracy, precision, recall, F1-score, mean squared error, and ROC curve.

**DESCRIPTIVE METRICS** : evaluate how well the model can capture the structure, patterns, or relationships in the data, without a predefined outcome. They are often used for unsupervised data mining tasks, such as clustering or association rule mining.

Examples : silhouette coefficient, Davies-Bouldin index, support, confidence, and lift.

**How to choose the right evaluation metric?**

Selecting the most suitable evaluation metric for data mining is a complex task,since it depends on various elements and compromises. Nevertheless, by adhering to certain general rules, you can make a more educated and reasonable decision, and consequently enhance the quality and usefulness of your data mining outcomes.

1. Define your objective – what are you trying to achieve or optimize? This will help guide the choice of evaluation metric, as different metrics may reflect different aspects of the model performance.
2. Be aware of the data – what are the features, variables, and distributions? How are they associated with the outcome or target variable?
3. compare and validate the results – comparison and validation methods should influence the selection of evaluation metric, as different metrics may have different scales, ranges, or interpretations.

**Social implications on data mining**

There are various social implications of data mining which are as follows −

**Privacy** :In current years privacy concerns have taken on a more important role as merchants, insurance companies, and government agencies amass warehouses including personal records.

The concerns that people have over the group of this data will generally extend to some analytic capabilities used to the data. Users of data mining should start thinking about how their use of this technology will be impacted by legal problems associated with privacy.

**Profiling** − Data Mining and profiling is a developing field that attempts to organize, understand, analyze, reason, and use the explosion of data in this information age. The process contains using algorithms and experience to extract design or anomalies that are very complex, difficult, or time-consuming to recognize.

**Unauthorized Used** − Trends obtain through data mining designed to be used for marketing goals or some other ethical goals, can be misused. Unethical businesses or people can use the data obtained through data mining to take benefit of vulnerable people or discriminate against a specific group of people. Furthermore, the data mining technique is not 100 percent accurate; thus mistakes do appear which can have serious results.

**Data Mining from a Database Perspective.**

A data mining system can be classified according to the kinds of databases on which the data mining is performed. For example, a system is a relational data miner if it discovers knowledge from relational data, or an object-oriented one if it mines knowledge from object-oriented databases.

**Statistical Methods in Data Mining**

Data mining refers to extracting or mining knowledge from large amounts of data. In other words, data mining is the science, art, and technology of discovering large and complex bodies of data in order to discover useful patterns. Theoreticians and practitioners are continually seeking improved techniques to make the process more efficient, cost-effective, and accurate. Any situation can be analyzed in two ways in data mining:

1. **Non-statistical Analysis:** This analysis provides generalized information and includes sound, still images, and moving images.
2. **Statistical Analysis:** In statistics, data is collected, analyzed, explored, and presented to identify patterns and trends. Alternatively, it is referred to as quantitative analysis. It is the analysis of raw data using mathematical formulas, models, and techniques. Through the use of statistical methods, information is extracted from research data, and different ways are available to judge the robustness of research outputs. It is created for the effective handling of large amounts of data that are generally multidimensional and possibly of several complex types.

Statistical analysis is a scientific tool in [AI and ML](https://www.simplilearn.com/artificial-intelligence-masters-program-training-course?source=GhPreviewCoursepages) that helps to collect and analyze large amounts of [data](https://www.simplilearn.com/what-is-data-article) to identify common patterns and trends to convert them into meaningful information. In simple words, statistical analysis is a [data analysis tool](https://www.simplilearn.com/top-data-analysis-tools-article) that helps draw meaningful conclusions from raw and unstructured data. The conclusions are drawn using statistical analysis facilitating decision-making and helping businesses make future predictions on the basis of past trends. It can be defined as a science of collecting and analyzing data to identify trends and patterns and presenting them. Statistical analysis involves working with numbers and is used by businesses and other institutions to make use of data to derive meaningful information.

**Types of Statistical Analysis**

1. [**Descriptive statistical analysis**](https://www.simplilearn.com/what-is-descriptive-statistics-article)involves collecting, interpreting, analyzing, and summarizing data to present them in the form of charts, graphs, and tables. Rather than drawing conclusions, it simply makes the complex data easy to read and understand.
2. **The**[**inferential statistical analysis**](https://www.simplilearn.com/inferential-statistics-article) focuses on drawing meaningful conclusions on the basis of the data analyzed. It studies the relationship between different variables or makes predictions for the whole population.
3. **Predictive statistical analysis** is a type of statistical analysis that analyzes data to derive past trends and predict future events on the basis of them. It uses [machine learning](https://www.simplilearn.com/tutorials/machine-learning-tutorial/statistics-for-machine-learning) algorithms, [data mining](https://www.simplilearn.com/what-is-data-mining-article), [data modelling](https://www.simplilearn.com/what-is-data-modeling-article), and [artificial intelligence](https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/what-is-artificial-intelligence) to conduct the statistical analysis of data.
4. **The prescriptive analysis** conducts the analysis of data and prescribes the best course of action based on the results. It is a type of statistical analysis that helps you make an informed decision.
5. [**Exploratory analysis**](https://www.simplilearn.com/exploratory-data-analysis-article) is similar to inferential analysis, but the difference is that it involves exploring the unknown data associations. It analyzes the potential relationships within the data.
6. **The causal statistical analysis** focuses on determining the cause and effect relationship between different variables within the raw data. In simple words, it determines why something happens and its effect on other variables. This methodology can be used by businesses to determine the reason for failure.

**Importance of Statistical Analysis**

Statistical analysis eliminates unnecessary information and catalogs important data in an uncomplicated manner, making the monumental work of organizing inputs appear so serene. Once the data has been collected, statistical analysis may be utilized for a variety of purposes.

* The statistical analysis aids in summarizing enormous amounts of data into clearly digestible chunks.
* The statistical analysis aids in the effective design of laboratory, field, and survey investigations.
* Statistical analysis may help with solid and efficient planning in any subject of study.
* Statistical analysis aid in establishing broad generalizations and forecasting how much of something will occur under particular conditions.
* Statistical methods, which are effective tools for interpreting numerical data, are applied in practically every field of study. Statistical approaches have been created and are increasingly applied in physical and biological sciences, such as genetics.
* Statistical approaches are used in the job of a businessman, a manufacturer, and a researcher. Statistics departments can be found in banks, insurance businesses, and government agencies.
* A modern administrator, whether in the public or commercial sector, relies on statistical data to make correct decisions.
* Politicians can utilize statistics to support and validate their claims while also explaining the issues they address.

**Benefits of Statistical Analysis**

Statistical analysis can be called a boon to mankind and has many benefits for both individuals and organizations. Given below are some of the reasons why you should consider investing in statistical analysis:

* It can help you determine the monthly, quarterly, yearly figures of sales profits, and costs making it easier to make your decisions.
* It can help you make informed and correct decisions.
* It can help you identify the problem or cause of the failure and make corrections. For example, it can identify the reason for an increase in total costs and help you cut the wasteful expenses.
* It can help you conduct market analysis and make an effective marketing and sales strategy.
* It helps improve the efficiency of different processes.

**Statistical Analysis Methods**

Although there are various methods used to perform data analysis, given below are the 5 most used and popular methods of statistical analysis:

* **Mean:**[Mean or average mean](https://www.simplilearn.com/tutorials/data-analytics-tutorial/measures-of-central-tendency" \t "_blank" \o "Mean or average mean) is one of the most popular methods of statistical analysis. Mean determines the overall trend of the data and is very simple to calculate. Mean is calculated by summing the numbers in the data set together and then dividing it by the number of data points. Despite the ease of calculation and its benefits, it is not advisable to resort to mean as the only statistical indicator as it can result in inaccurate decision making.
* **Standard Deviation**

Standard deviation is another very widely used statistical tool or method. It analyzes the deviation of different data points from the mean of the entire data set. It determines how data of the data set is spread around the mean. You can use it to decide whether the research outcomes can be generalized or not.

* **Regression**

Regression is a statistical tool that helps determine the cause and effect relationship between the variables. It determines the relationship between a dependent and an independent variable. It is generally used to predict future trends and events.

* **Hypothesis Testing**

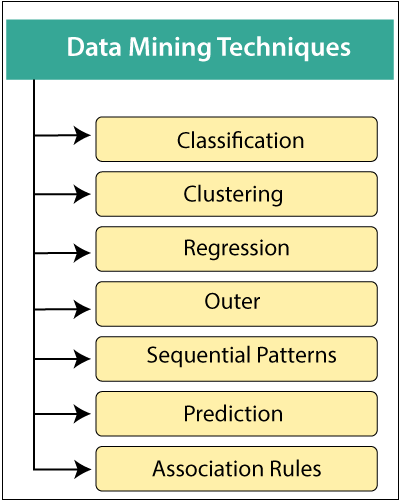
[Hypothesis testing](https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics) can be used to test the validity or trueness of a conclusion or argument against a data set. The hypothesis is an assumption made at the beginning of the research and can hold or be false based on the analysis results.

**Sample Size Determination**

Sample size determination or [data sampling](https://www.simplilearn.com/types-of-sampling-techniques-article) is a technique used to derive a sample from the entire population, which is representative of the population. This method is used when the size of the population is very large. You can choose from among the various data sampling techniques such as snowball sampling, convenience sampling, and random sampling.

**Data Mining Techniques:**

In recent data mining projects, various major data mining techniques have been developed and used.



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:

**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..

**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..

**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..

**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.

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. 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 ,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., 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,Support,Confidence.

5. Outlier 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 uses 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.

**Similarity Measure**

* Similarity measures are mathematical functions used to determine the degree of similarity between two data points or objects. These measures produce a score that indicates how similar or alike the two data points are.
* It takes two data points as input and produces a similarity score as output, typically ranging from 0 (completely dissimilar) to 1 (identical or perfectly similar).
* Similarity measures also have some well-known properties –
* *sim*(*A*,*B*)=1 (or maximum similarity) only if *A*=*B*
* Typical range - (0≤*sim*≤1)
* Symmetry - *sim*(*A*,*B*)=*sim*(*B*,*A*) for all *A* and *B*

Most commonly used similarity measures in data mining.

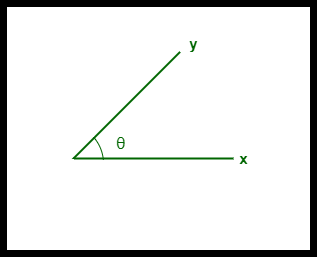
**Cosine Similarity**

Cosine similarity is a widely used similarity measure in data mining and information retrieval. It measures the cosine of the angle between two non-zero vectors in a multi-dimensional space. In the context of data mining, these vectors represent the feature vectors of two data points. The cosine similarity score ranges from 0 to 1, with 0 indicating no similarity and 1 indicating perfect similarity.

The cosine similarity between two vectors is calculated as the dot product of the vectors divided by the product of their magnitudes. This calculation can be represented mathematically as follows –

cos(*θ*)=**A.B** / ∥**A** ∥∥**B**∥where **A** and **B** are the feature vectors of two data points, "." denotes the dot product, and

"||" denotes the magnitude of the vector

. 

**Jaccard Similarity**

The Jaccard similarity is another widely used similarity measure in data mining, particularly in text analysis and clustering. It measures the similarity between two sets of data by calculating the ratio of the intersection of the sets to their union. The Jaccard similarity score ranges from 0 to 1, with 0 indicating no similarity and 1 indicating perfect similarity.

The Jaccard similarity between two sets A and B is calculated as follows -

*J*(*A*,*B*)=∣**A**∩**B**∣ / ∣**A**∪**B**∣

where ∣**A**∩**B**∣ is the size of the intersection of sets **A** and **B**, and ∣**A**∪**B**∣ is the size of the union of sets **A** and **B**.

**Pearson Correlation Coefficient**

The Pearson correlation coefficient is a widely used similarity measure in data mining and statistical analysis.

It measures the linear correlation between two continuous variables, X and Y. The Pearson correlation coefficient ranges from -1 to +1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and +1 indicating a perfect positive correlation. The Pearson correlation coefficient is commonly used in data mining applications such as feature selection and regression analysis. It can help identify variables that are highly correlated with each other, which can be useful for reducing the dimensionality of a dataset. In regression analysis, it can also be used to predict the value of one variable based on the value of another variable.

The Pearson correlation coefficient between two variables, X and Y, is calculated as follows –

*Ρ(X*,*Y)*​= ​cov(*X*,*Y*)​ / *σX*​*σY*  
where  cov(*X*,*Y*) is the covariance between variables *X* and *Y*, and *σX*​ and *σY*​ are the standard deviations of variables *X* and *Y*, respectively.

**Sørensen-Dice Coefficient**

The Sørensen-Dice coefficient, also known as the Dice similarity index or Dice coefficient, is a similarity measure used to compare the similarity between two sets of data, typically used in the context of text or image analysis. The coefficient ranges from 0 to 1, with 0 indicating no similarity and 1 indicating perfect similarity. The Sørensen-Dice coefficient is commonly used in text analysis to compare the similarity between two documents based on the set of words or terms they contain. It is also used in image analysis to compare the similarity between two images based on the set of pixels they contain.

The Sørensen-Dice coefficient between two sets, A and B, is calculated as follows –

*S*(*A*,*B*)= 2 ∣*A*∩*B*∣​ / ∣*A*∣+∣*B*∣

where∣*A*∩*B*∣ is the size of the intersection of sets *A* and *B*, and ∣*A*∣ and ∣*B*∣ are the sizes of sets *A* and *B*, respectively.

**Choosing The Appropriate Similarity Measure**

Choosing an appropriate similarity measure depends on the nature of the data and the specific task at hand. some factors to consider when choosing a similarity measure -

* Different similarity measures are suitable for different data types, such as continuous or categorical data, text or image data, etc. For example, the Pearson correlation coefficient, which is only suitable for continuous variables.
* The choice of similarity measure also depends on the specific task at hand. For example, cosine similarity is often used in information retrieval and text mining, while Jaccard similarity is commonly used in clustering and recommendation systems.
* Some similarity measures are more robust to noise and outliers in the data than others. For example, the Sørensen-Dice coefficient is less sensitive to noise.

**DECISION TREE:**

A decision tree is a type of supervised learning algorithm that is commonly used in machine learning to model and predict outcomes based on input data. It is a tree-like structure where each internal node tests on  attribute, each branch corresponds to attribute value and each leaf node represents the final decision or prediction.

Example of a decision tree

Suppose we want to build a decision tree to predict whether a person is likely to buy a new car based on their demographic and behavior data. The decision tree starts with the root node, which represents the entire dataset. The root node splits the dataset based on the “income” attribute. If the person’s income is less than or equal to $50,000, the decision tree follows the left branch, and if the income is greater than $50,000, the decision tree follows the right branch.

The left branch leads to a node that represents the “age” attribute. If the person’s age is less than or equal to 30, the decision tree follows the left branch, and if the age is greater than 30, the decision tree follows the right branch. The right branch leads to a leaf node that predicts that the person is unlikely to buy a new car.

The left branch leads to another node that represents the “education” attribute. If the person’s education level is less than or equal to high school, the decision tree follows the left branch, and if the education level is greater than high school, the decision tree follows the right branch. The left branch leads to a leaf node that predicts that the person is unlikely to buy a new car. The right branch leads to another node that represents the “credit score” attribute. If the person’s credit score is less than or equal to 650, the decision tree follows the left branch, and if the credit score is greater than 650, the decision tree follows the right branch. The left branch leads to a leaf node that predicts that the person is unlikely to buy a new car. The right branch leads to a leaf node that predicts that the person is likely to buy a new car.

In summary, a decision tree is a graphical representation of all the possible outcomes of a decision based on the input data. It is a powerful tool for modeling and predicting outcomes in a wide range of domains, including business, finance, healthcare, and more.

**Decision tree algorithm:**

* 1. Begin with the entire dataset as the root node of the decision tree.
  2. Determine the best attribute to split the dataset based on a given criterion,
  3. Create a new internal node that corresponds to the best attribute and connects it to the root node.
  4. Partition the dataset into subsets based on the values of the best attribute.
  5. Recursively repeat steps 1-4 for each subset until all instances in a given subset belong to the same class or no further splitting is possible.
  6. Assign a leaf node to each subset that contains instances that belong to the same class.
  7. Make predictions based on the decision tree by traversing it from the root node to a leaf node that corresponds to the instance being classified.

The benefits of having a decision tree are

* It does not require any domain knowledge.
* It is easy to comprehend.
* The learning and classification steps of a decision tree are simple and fast.

The following decision tree is for the concept to buy computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.

