**Tasks and Functionalities of Data Mining**

[Data Mining](https://www.geeksforgeeks.org/data-mining/) functions are used to define the trends or correlations contained in data mining activities. In comparison, data mining **activities** can be divided into 2 categories:

**1]Descriptive Data Mining:**

This category of data mining is concerned with finding patterns and relationships in the data that can provide insight into the underlying structure of the data. Descriptive data mining is often used to summarize or explore the data, and it can be used to answer questions such as: What are the most common patterns or relationships in the data? Are there any clusters or groups of data points that share common characteristics? What are the outliers in the data, and what do they represent?  
Some common techniques used in descriptive data mining include:

**Cluster analysis:**

This technique is used to identify groups of data points that share similar characteristics. Clustering can be used for segmentation, anomaly detection, and summarization.

**Association rule mining:**

This technique is used to identify relationships between variables in the data. It can be used to discover co-occurring events or to identify patterns in transaction data.

**Visualization:**

This technique is used to represent the data in a visual format that can help users to identify patterns or trends that may not be apparent in the raw data.

**2]Predictive Data Mining:**This category of data mining is concerned with developing models that can predict future behavior or outcomes based on historical data. Predictive data mining is often used for classification or regression tasks, and it can be used to answer questions such as: What is the likelihood that a customer will churn? What is the expected revenue for a new product launch? What is the probability of a loan defaulting?  
Some common techniques used in predictive data mining include:

**Decision trees:**This technique is used to create a model that can predict the value of a target variable based on the values of several input variables. Decision trees are often used for classification tasks.

**Neural networks:**This technique is used to create a model that can learn to recognize patterns in the data. Neural networks are often used for image recognition, speech recognition, and natural language processing.

**Regression analysis:** This technique is used to create a model that can predict the value of a target variable based on the values of several input variables. Regression analysis is often used for prediction tasks.

**Both descriptive and predictive data mining techniques are important**for gaining insights and making better decisions. Descriptive data mining can be used to explore the data and identify patterns, while predictive data mining can be used to make predictions based on those patterns. Together, these techniques can help organizations to understand their data and make informed decisions based on that understanding.

**Data Mining Functionality:**

**1. Class/Concept Descriptions:** Classes or definitions can be correlated with results. In simplified, descriptive and yet accurate ways, it can be helpful to define individual groups and concepts. These class or concept definitions are referred to as class/concept descriptions.

* **Data Characterization:** This refers to the summary of general characteristics or features of the class that is under the study. The output of the data characterization can be presented in various forms include pie charts, bar charts, curves, multidimensional data cubes.

**Example:** To study the characteristics of software products with sales increased by 10% in the previous years. To summarize the characteristics of the customer who spend more than $5000 a year at AllElectronics, the result is general profile of those customers such as that they are 40-50 years old, employee and have excellent credit rating.

* **Data Discrimination:** It compares common features of class which is under study. It is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

**Example:**we may want to compare two groups of customers those who shop for computer products regular and those who rarely shop for such products(less than 3 times a year), the resulting description provides a general comparative profile of those customers, such as 80% of the customers who frequently purchased computer products are between 20 and 40 years old and have a university degree, and 60% of the customers who infrequently buys such products are either seniors or youth, and have no university degree.

**2. Mining Frequent Patterns, Associations, and Correlations:** Frequent patterns are nothing but things that are found to be most common in the data. There are different kinds of frequencies that can be observed in the dataset.

* **Frequent item set:** This applies to a number of items that can be seen together regularly for eg: milk and sugar.
* **Frequent Subsequence:** This refers to the pattern series that often occurs regularly such as purchasing a phone followed by a back cover.
* **Frequent Substructure:** It refers to the different kinds of data structures such as trees and graphs that may be combined with the itemset or subsequence.

**Association Analysis:** The process involves uncovering the relationship between data and deciding the rules of the association. It is a way of discovering the relationship between various items.

**Example:**Suppose we want to know which items are frequently purchased together. An example for such a rule mined from a transactional database is,

**buys (X, “computer”) ⇒ buys (X, “software”) [support = 1%, confidence = 50%],**

where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all the transactions under analysis show that computer and software are purchased together. This association rule involves a single attribute or predicate (i.e., buys) that repeats. Association rules that contain a single predicate are referred to as single-dimensional association rules.

**age (X, “20…29”) ∧ income (X, “40K..49K”) ⇒ buys (X, “laptop”)**

**[support = 2%, confidence = 60%].**

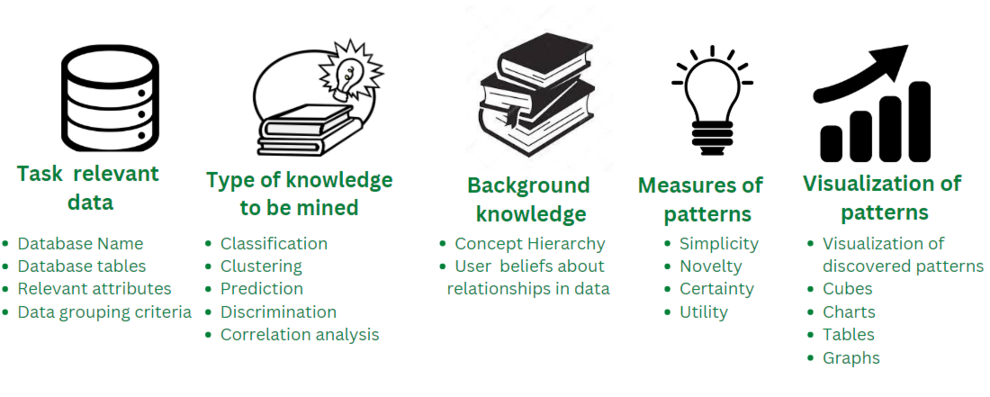
The**rule says that**2% are 20 to 29 years old with an income of $40,000 to $49,000 and have purchased a laptop. There is a 60% probability that a customer in this age and income group will purchase a laptop. The association involving more than one attribute or predicate can be referred to as a multidimensional association rule.

Typically, association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold. Additional analysis can be performed to uncover interesting statistical correlations between associated attribute–value pairs.

**Correlation Analysis:** Correlation is a mathematical technique that can show whether and how strongly the pairs of attributes are related to each other. For example, Highted people tend to have more weight.

**Data Mining Task Primitives**

Data mining task primitives refer to the basic building blocks or components that are used to construct a data mining process. These primitives are used to represent the most common and fundamental tasks that are performed during the data mining process.  The use of data mining task primitives can provide a modular and reusable approach, which can improve the performance, efficiency, and understandability of the data mining process.



The Data Mining Task Primitives are as follows:

1. **The set of task relevant data to be mined:** It refers to the specific data that is relevant and necessary for a particular task or analysis being conducted using data mining techniques. This data may include specific attributes, variables, or characteristics that are relevant to the task at hand, such as customer demographics, sales data, or website usage statistics. The data selected for mining is typically a subset of the overall data available, as not all data may be necessary or relevant for the task.  For **example**:Extracting the database name, database tables, and relevant required attributes from the dataset from the provided input database.
2. **Kind of knowledge to be mined:**It refers to the type of information or insights that are being sought through the use of data mining techniques. This describes the data mining tasks that must be carried out. It includes various tasks such as classification, clustering, discrimination, characterization, association, and evolution analysis. For **example**, It determines the task to be performed on the relevant data in order to mine useful information such as classification, clustering, prediction, discrimination, outlier detection, and correlation analysis.
3. **Background knowledge to be used in the discovery process:** It refers to any prior information or understanding that is used to guide the data mining process. This can include domain-specific knowledge, such as industry-specific terminology, trends, or best practices, as well as knowledge about the data itself. The use of background knowledge can help to improve the accuracy and relevance of the insights obtained from the data mining process. For **example**, The use of background knowledge such as concept hierarchies, and user beliefs about relationships in data in order to evaluate and perform more efficiently.
4. **Interestingness measures and thresholds for pattern evaluation:**It refers to the methods and criteria used to evaluate the quality and relevance of the patterns or insights discovered through data mining. Interestingness measures are used to quantify the degree to which a pattern is considered to be interesting or relevant based on certain criteria, such as its frequency, confidence, or lift. These measures are used to identify patterns that are meaningful or relevant to the task. Thresholds for pattern evaluation, on the other hand, are used to set a minimum level of interestingness that a pattern must meet in order to be considered for further analysis or action.  For **example**: Evaluating the interestingness and interestingness measures such as utility, certainty, and novelty for the data and setting an appropriate threshold value for the pattern evaluation.
5. **Representation for visualizing the discovered pattern:** It refers to the methods used to represent the patterns or insights discovered through data mining in a way that is easy to understand and interpret. Visualization techniques such as charts, graphs, and maps are commonly used to represent the data and can help to highlight important trends, patterns, or relationships within the data. Visualizing the discovered pattern helps to make the insights obtained from the data mining process more accessible and understandable to a wider audience, including non-technical stakeholders. For **example**Presentation and visualization of discovered pattern data using various visualization techniques such as barplot, charts, graphs, tables, etc.

**Advantages of Data Mining Task Primitives**

The use of data mining task primitives has several advantages, including:

1. **Modularity**: Data mining task primitives provide a modular approach to data mining, which allows for flexibility and the ability to easily modify or replace specific steps in the process.
2. **Reusability**: Data mining task primitives can be reused across different data mining projects, which can save time and effort.
3. **Standardization**: Data mining task primitives provide a standardized approach to data mining, which can improve the consistency and quality of the data mining process.
4. **Understandability**: Data mining task primitives are easy to understand and communicate, which can improve collaboration and communication among team members.
5. **Improved Performance**: Data mining task primitives can improve the performance of the data mining process by reducing the amount of data that needs to be processed, and by optimizing the data for specific data mining algorithms.
6. **Flexibility**: Data mining task primitives can be combined and repeated in various ways to achieve the goals of the data mining process, making it more adaptable to the specific needs of the project.
7. **Efficient use of resources**: Data mining task primitives can help to make more efficient use of resources, as they allow to perform specific tasks with the right tools, avoiding unnecessary steps and reducing the time and computational power needed.

**Data Integration in Data Mining**

Last Updated : 01 Feb, 2023

**INTRODUCTION :**

* Data integration in data mining refers to the process of combining data from multiple sources into a single, unified view. This can involve cleaning and transforming the data, as well as resolving any inconsistencies or conflicts that may exist between the different sources. The goal of data integration is to make the data more useful and meaningful for the purposes of analysis and decision making. Techniques used in data integration include data warehousing, ETL (extract, transform, load) processes, and data federation.

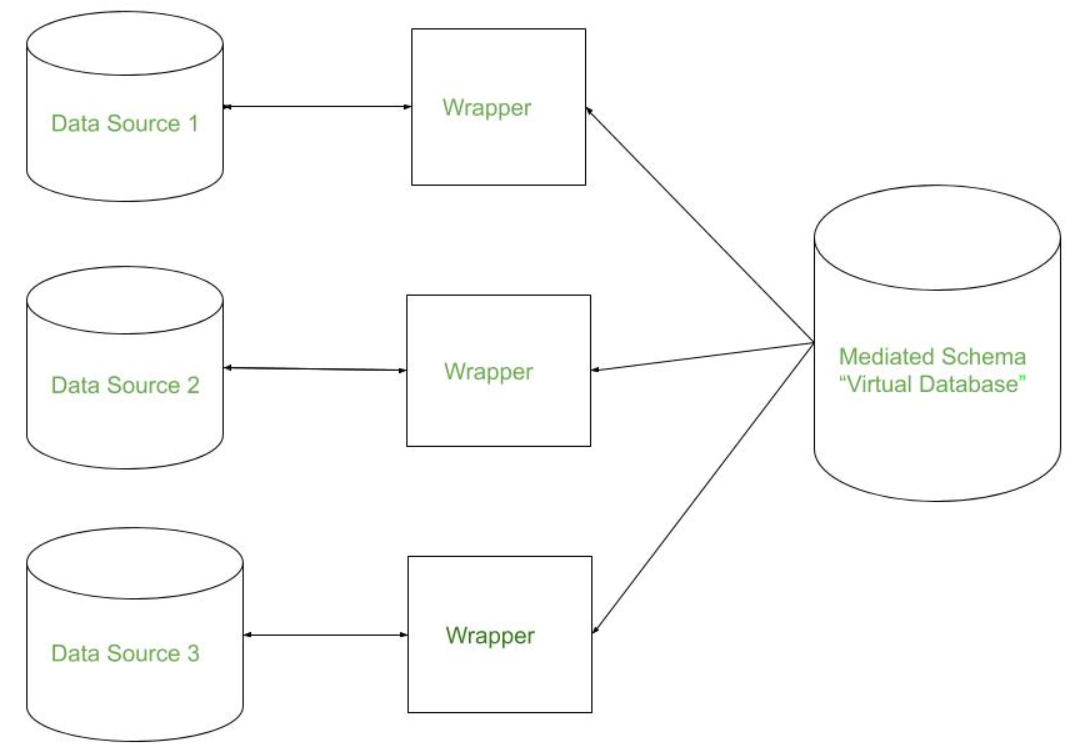
Data Integration is a data preprocessing technique that combines data from multiple heterogeneous data sources into a coherent data store and provides a unified view of the data. These sources may include multiple data cubes, databases, or flat files.

The data integration approaches are formally defined as triple <G, S, M> where,   
G stand for the global schema,   
S stands for the heterogeneous source of schema,   
M stands for mapping between the queries of source and global schema.

**What is data integration :**  
    Data integration is the process of combining data from multiple sources into a cohesive and consistent view. This process involves identifying and accessing the different data sources, mapping the data to a common format, and reconciling any inconsistencies or discrepancies between the sources. The goal of data integration is to make it easier to access and analyze data that is spread across multiple systems or platforms, in order to gain a more complete and accurate understanding of the data.

    Data integration can be challenging due to the variety of data formats, structures, and semantics used by different data sources. Different data sources may use different data types, naming conventions, and schemas, making it difficult to combine the data into a single view. Data integration typically involves a combination of manual and automated processes, including data profiling, data mapping, data transformation, and data reconciliation.

   Data integration is used in a wide range of applications, such as business intelligence, data warehousing, master data management, and analytics. Data integration can be critical to the success of these applications, as it enables organizations to access and analyze data that is spread across different systems, departments, and lines of business, in order to make better decisions, improve operational efficiency, and gain a competitive advantage.



There are mainly 2 major approaches for data integration – one is the “tight coupling approach” and another is the “loose coupling approach”.

**Tight Coupling:**

This approach involves creating a centralized repository or data warehouse to store the integrated data. The data is extracted from various sources, transformed and loaded into a data warehouse. Data is integrated in a tightly coupled manner, meaning that the data is integrated at a high level, such as at the level of the entire dataset or schema. This approach is also known as data warehousing, and it enables data consistency and integrity, but it can be inflexible and difficult to change or update.

* Here, a data warehouse is treated as an information retrieval component.
* In this coupling, data is combined from different sources into a single physical location through the process of ETL – Extraction, Transformation, and Loading.

**Loose Coupling:**

This approach involves integrating data at the lowest level, such as at the level of individual data elements or records. Data is integrated in a loosely coupled manner, meaning that the data is integrated at a low level, and it allows data to be integrated without having to create a central repository or data warehouse. This approach is also known as data federation, and it enables data flexibility and easy updates, but it can be difficult to maintain consistency and integrity across multiple data sources.

* Here, an interface is provided that takes the query from the user, transforms it in a way the source database can understand, and then sends the query directly to the source databases to obtain the result.
* And the data only remains in the actual source databases.

**Issues in Data Integration:**

 There are several issues that can arise when integrating data from multiple sources, including:

1. **Data Quality:** Inconsistencies and errors in the data can make it difficult to combine and analyze.
2. **Data Semantics:** Different sources may use different terms or definitions for the same data, making it difficult to combine and understand the data.
3. **Data Heterogeneity:** Different sources may use different data formats, structures, or schemas, making it difficult to combine and analyze the data.
4. **Data Privacy and Security:** Protecting sensitive information and maintaining security can be difficult when integrating data from multiple sources.
5. **Scalability:** Integrating large amounts of data from multiple sources can be computationally expensive and time-consuming.
6. **Data Governance:** Managing and maintaining the integration of data from multiple sources can be difficult, especially when it comes to ensuring data accuracy, consistency, and timeliness.
7. **Performance:** Integrating data from multiple sources can also affect the performance of the system.
8. **Integration with existing systems:**Integrating new data sources with existing systems can be a complex task, requiring significant effort and resources.
9. **Complexity:** The complexity of integrating data from multiple sources can be high, requiring specialized skills and knowledge.

# Classification of Data Mining Systems

Data mining refers to the process of extracting important data from raw data. It analyses the data patterns in huge sets of data with the help of several software. Ever since the development of data mining, it is being incorporated by researchers in the research and development field.

With [Data mining](https://www.javatpoint.com/data-mining), businesses are found to gain more profit. It has not only helped in understanding customer demand but also in developing effective strategies to enforce overall business turnover. It has helped in determining business objectives for making clear decisions.

Data collection and data warehousing, and computer processing are some of the strongest pillars of data mining. Data mining utilizes the concept of mathematical algorithms to segment the data and assess the possibility of occurrence of future events.

To understand the system and meet the desired requirements, data mining can be classified into the following systems:

* Classification based on the mined Databases
* Classification based on the type of mined knowledge
* Classification based on statistics
* Classification based on Machine Learning
* Classification based on visualization
* Classification based on Information Science
* Classification based on utilized techniques
* Classification based on adapted applications

## Classification Based on the mined Databases

A data mining system can be classified based on the types of databases that have been mined. A database system can be further segmented based on distinct principles, such as data models, types of data, etc., which further assist in classifying a data mining system.

For example, if we want to classify a database based on the data model, we need to select either relational, transactional, object-relational or data warehouse mining systems.

## Classification Based on the type of Knowledge Mined

A data mining system categorized based on the kind of knowledge mind may have the following functionalities:

1. Characterization
2. Discrimination
3. Association and Correlation Analysis
4. Classification
5. Prediction
6. Outlier Analysis
7. Evolution Analysis

## Classification Based on the Techniques Utilized

A data mining system can also be classified based on the type of techniques that are being incorporated. These techniques can be assessed based on the involvement of user interaction involved or the methods of analysis employed.

## Classification Based on the Applications Adapted

Data mining systems classified based on adapted applications adapted are as follows:

1. Finance
2. Telecommunications
3. DNA
4. Stock Markets
5. E-mail

## Examples of Classification Task

Following is some of the main examples of classification tasks:

* Classification helps in determining tumor cells as benign or malignant.
* Classification of credit card transactions as fraudulent or legitimate.
* Classification of secondary structures of protein as alpha-helix, beta-sheet, or random coil.
* Classification of news stories into distinct categories such as finance, weather, entertainment, sports, etc.

## Integration schemes of Database and Data warehouse systems

**No Coupling**

In no coupling schema, the data mining system does not use any database or data warehouse system functions.

**Loose Coupling**

In loose coupling, data mining utilizes some of the database or data warehouse system functionalities. It mainly fetches the data from the data repository managed by these systems and then performs data mining. The results are kept either in the file or any designated place in the database or data warehouse.

**Semi-Tight Coupling**

In semi-tight coupling, data mining is linked to either the DB or DW system and provides an efficient implementation of data mining primitives within the database.

**Tight Coupling**

A data mining system can be effortlessly combined with a database or data warehouse system in tight coupling.

# Data Mining Task Primitives

A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery to direct the mining process or examine the findings from different angles or depths. The data mining primitives specify the following,

1. Set of task-relevant data to be mined.
2. Kind of knowledge to be mined.
3. Background knowledge to be used in the discovery process.
4. Interestingness measures and thresholds for pattern evaluation.
5. Representation for visualizing the discovered patterns.

A data mining query language can be designed to incorporate these primitives, allowing users to interact with data mining systems flexibly. Having a data mining query language provides a foundation on which user-friendly graphical interfaces can be built.

Designing a comprehensive data mining language is challenging because data mining covers a wide spectrum of tasks, from data characterization to evolution analysis. Each task has different requirements. The design of an effective data mining query language requires a deep understanding of the power, limitation, and underlying mechanisms of the various kinds of data mining tasks. This facilitates a data mining system's communication with other information systems and integrates with the overall information processing environment.

## List of Data Mining Task Primitives

A data mining query is defined in terms of the following primitives, such as:

**1. The set of task-relevant data to be mined**

This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (the relevant attributes or dimensions).

In a relational database, the set of task-relevant data can be collected via a relational query involving operations like selection, projection, join, and aggregation.

The data collection process results in a new data relational called the **initial data relation**. The initial data relation can be ordered or grouped according to the conditions specified in the query. This data retrieval can be thought of as a subtask of the data mining task.

This initial relation may or may not correspond to physical relation in the database. Since virtual relations are called Views in the field of databases, the set of task-relevant data for data mining is called a minable view.

**2. The kind of knowledge to be mined**

This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

**3. The background knowledge to be used in the discovery process**

This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and evaluating the patterns found. Concept hierarchies are a popular form of background knowledge, which allows data to be mined at multiple levels of abstraction.

Concept hierarchy defines a sequence of mappings from low-level concepts to higher-level, more general concepts.

* **Rolling Up - Generalization of data:** Allow to view data at more meaningful and explicit abstractions and makes it easier to understand. It compresses the data, and it would require fewer input/output operations.
* **Drilling Down - Specialization of data:** Concept values replaced by lower-level concepts. Based on different user viewpoints, there may be more than one concept hierarchy for a given attribute or dimension.

An example of a concept hierarchy for the attribute (or dimension) age is shown below. User beliefs regarding relationships in the data are another form of background knowledge.

**4. The interestingness measures and thresholds for pattern evaluation**

Different kinds of knowledge may have different interesting measures. They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. For example, interesting measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

* **Simplicity:** A factor contributing to the interestingness of a pattern is the pattern's overall simplicity for human comprehension. For example, the more complex the structure of a rule is, the more difficult it is to interpret, and hence, the less interesting it is likely to be. Objective measures of pattern simplicity can be viewed as functions of the pattern structure, defined in terms of the pattern size in bits or the number of attributes or operators appearing in the pattern.
* **Certainty (Confidence):** Each discovered pattern should have a measure of certainty associated with it that assesses the validity or "trustworthiness" of the pattern. A certainty measure for association rules of the form "A =>B" where A and B are sets of items is confidence. Confidence is a certainty measure. Given a set of task-relevant data tuples, the confidence of "A => B" is defined as  
  Confidence (A=>B) = # tuples containing both A and B /# tuples containing A
* **Utility (Support):** The potential usefulness of a pattern is a factor defining its interestingness. It can be estimated by a utility function, such as support. The support of an association pattern refers to the percentage of task-relevant data tuples (or transactions) for which the pattern is true.  
  Utility (support): usefulness of a pattern  
  Support (A=>B) = # tuples containing both A and B / total #of tuples
* **Novelty:** Novel patterns are those that contribute new information or increased performance to the given pattern set. For example -> A data exception. Another strategy for detecting novelty is to remove redundant patterns.

**5. The expected representation for visualizing the discovered patterns**

This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, cross tabs, charts, graphs, decision trees, cubes, or other visual representations.

Users must be able to specify the forms of presentation to be used for displaying the discovered patterns. Some representation forms may be better suited than others for particular kinds of knowledge.

For example, generalized relations and their corresponding cross tabs or pie/bar charts are good for presenting characteristic descriptions, whereas decision trees are common for classification.

## Example of Data Mining Task Primitives

Suppose, as a marketing manager of **AllElectronics**, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than $40,000 and who have bought more than $1,000 worth of items, each of which is priced at no less than $100.

In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules. This data mining query is expressed in DMQL3 as follows, where each line of the query has been enumerated to aid in our discussion.

1. use database AllElectronics\_db
2. use hierarchy location\_hierarchy for T.branch, age\_hierarchy for C.age
3. mine classification as promising\_customers
4. in relevance to C.age, C.income, I.type, I.place\_made, T.branch
5. from customer C, an item I, transaction T
6. where I.item\_ID = T.item\_ID and C.cust\_ID = T.cust\_ID and C.income ≥ 40,000 and I.price ≥ 100
7. group by T.cust\_ID

Data 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. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding −

* Mining Methodology and User Interaction
* Performance Issues
* Diverse Data Types Issues

The following diagram describes the major issues.

## Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

* **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.
* **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.
* **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.
* [**Data mining query languages**](https://www.tutorialspoint.com/data_mining/dm_query_language.htm)**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.
* **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.
* **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.
* **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

## Performance Issues

There can be performance-related issues such as follows −

* **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.
* **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.

[Explore our](https://www.tutorialspoint.com/latest/courses?utm_source=tutorialspoint&utm_medium=tutorials_3p&utm_campaign=internal" \t "_blank)**[latest online courses](https://www.tutorialspoint.com/latest/courses?utm_source=tutorialspoint&utm_medium=tutorials_3p&utm_campaign=internal" \t "_blank)**[and learn new skills at your own pace. Enroll and become a certified expert to boost your career.](https://www.tutorialspoint.com/latest/courses?utm_source=tutorialspoint&utm_medium=tutorials_3p&utm_campaign=internal" \t "_blank)

## Diverse Data Types Issues

* **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
* **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 Preprocessing in Data Mining**

Last Updated : 06 May, 2023

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

**Some common steps in data preprocessing include:**

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

**Data Cleaning:**This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

**Data Integration:**This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

**Data Transformation:**This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

**Data Reduction:**This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

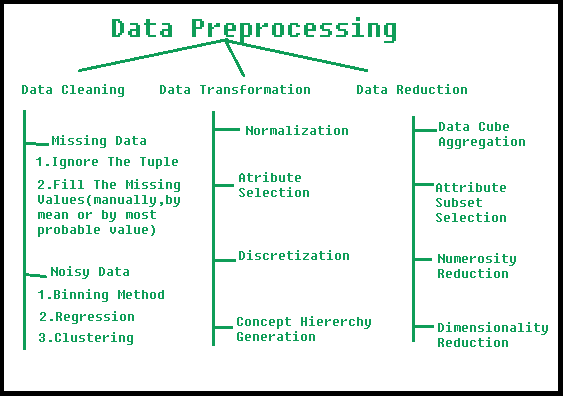
**Data Discretization:**This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

**Data Normalization:**This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

Data preprocessing plays a crucial role in ensuring the quality of data and the accuracy of the analysis results. The specific steps involved in data preprocessing may vary depending on the nature of the data and the analysis goals.

By performing these steps, the data mining process becomes more efficient and the results become more accurate.

**Preprocessing in Data Mining:**   
Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



**Steps Involved in Data Preprocessing:**

**1. Data Cleaning:**   
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc. 

* **(a). Missing Data:**   
  This situation arises when some data is missing in the data. It can be handled in various ways.   
  Some of them are:
  1. **Ignore the tuples:**   
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**   
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
* **(b). Noisy Data:**   
  Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**   
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**   
     Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**   
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

**2. Data Transformation:**   
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**   
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**   
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**   
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**   
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. Data Reduction:**   
Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

**Feature Selection:** This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).

**Feature Extraction:**This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

**Sampling:**This involves selecting a subset of data points from the dataset. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

**Clustering:**This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.

**Compression:** This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gzip compression.

# Discretization in data mining

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization. Supervised discretization refers to a method in which the class data is used. Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Now, we can understand this concept with the help of an example

Suppose we have an attribute of Age with the given values

|  |  |
| --- | --- |
| Age | 1,5,9,4,7,11,14,17,13,18, 19,31,33,36,42,44,46,70,74,78,77 |

Table before Discretization

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attribute | Age | Age | Age | Age |
|  | 1,5,4,9,7 | 11,14,17,13,18,19 | 31,33,36,42,44,46 | 70,74,77,78 |
| After Discretization | Child | Young | Mature | Old |

Another example is analytics, where we gather the static data of website visitors. For example, all visitors who visit the site with the IP address of India are shown under country level.

## Some Famous techniques of data discretization

**Histogram analysis**

Histogram refers to a plot used to represent the underlying frequency distribution of a continuous data set. Histogram assists the data inspection for data distribution. For example, Outliers, skewness representation, normal distribution representation, etc.

**Binning**

Binning refers to a data smoothing technique that helps to group a huge number of continuous values into smaller values. For data discretization and the development of idea hierarchy, this technique can also be used.

**Cluster Analysis**

Cluster analysis is a form of data discretization. A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.

**Data discretization using decision tree analysis**

Data discretization refers to a decision tree analysis in which a top-down slicing technique is used. It is done through a supervised procedure. In a numeric attribute discretization, first, you need to select the attribute that has the least entropy, and then you need to run it with the help of a recursive process. The recursive process divides it into various discretized disjoint intervals, from top to bottom, using the same splitting criterion.

**Data discretization using correlation analysis**

Discretizing data by linear regression technique, you can get the best neighboring interval, and then the large intervals are combined to develop a larger overlap to form the final 20 overlapping intervals. It is a supervised procedure.

## Data discretization and concept hierarchy generation

The term hierarchy represents an organizational structure or mapping in which items are ranked according to their levels of importance. In other words, we can say that a hierarchy concept refers to a sequence of mappings with a set of more general concepts to complex concepts. It means mapping is done from low-level concepts to high-level concepts. For example, in computer science, there are different types of hierarchical systems. A document is placed in a folder in windows at a specific place in the tree structure is the best example of a computer hierarchical tree model. There are two types of hierarchy: top-down mapping and the second one is bottom-up mapping.

Let's understand this concept hierarchy for the dimension location with the help of an example.

A particular city can map with the belonging country. For example, New Delhi can be mapped to India, and India can be mapped to Asia.

**Top-down mapping**

Top-down mapping generally starts with the top with some general information and ends with the bottom to the specialized information.

**Bottom-up mapping**

Bottom-up mapping generally starts with the bottom with some specialized information and ends with the top to the generalized information.

### Data discretization and binarization in data mining

Data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. In contrast, data binarization is used to transform the continuous and discrete attributes into binary attributes.

### Why is Discretization important?

As we know, an infinite of degrees of freedom mathematical problem poses with the continuous data. For many purposes, data scientists need the implementation of discretization. It is also used to improve signal noise ratio.