Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called ***Knowledge Discovery in Database (KDD)***. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

What is Data Mining?

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

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

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

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

Data Mining Applications

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

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

Data Mining in Healthcare

Data mining in Education

Data Mining in CRM (Customer Relationship Management)

Data Mining in Fraud detection

Data Mining in Lie Detection

Data Mining Financial Banking

Data Mining in Market Basket Analysis

Types of Data Mining

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

Relational Database:

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the

database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.

Data warehouses:

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

Data Repositories:

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

Object-Relational Database:

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

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

Transactional Database:

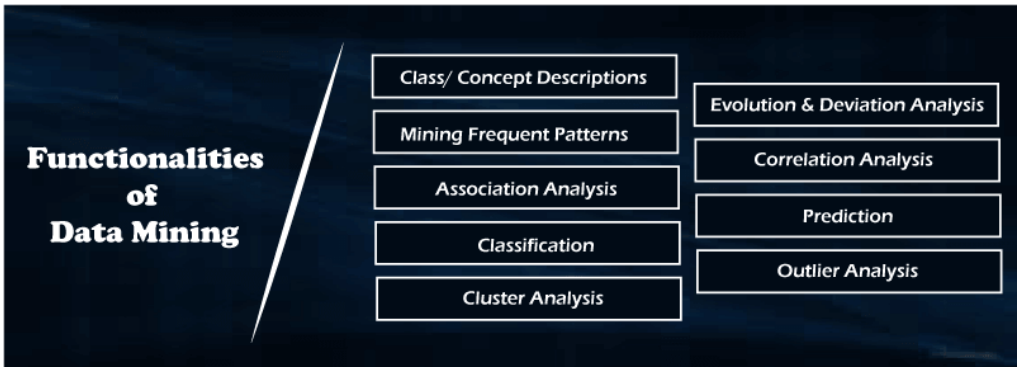
A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database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

Tasks and Functionalities of Data Mining

Functionalities of Data Mining

Data mining functionalities are used to represent the type of patterns that have to be discovered in data mining tasks. Data mining tasks can be classified into two types: descriptive and predictive. Descriptive mining tasks define the common features of the data in the database, and the predictive mining tasks act in inference on the current information to develop predictions.

Data mining is extensively used in many areas or sectors. It is used to predict and characterize data. But the ultimate objective in **Data Mining Functionalities** is to observe the various trends in data mining. There are several data mining functionalities that the organized and scientific methods offer, such as:



1. Class/Concept Descriptions

A class or concept implies there is a data set or set of features that define the class or a concept. A class can be a category of items on a shop floor, and a concept could be the abstract idea on which data may be categorized like products to be put on clearance sale and non-sale products. There are two concepts here, one that helps with grouping and the other that helps in differentiating.

- **Data Characterization:** This refers to the summary of general characteristics or features of the class, resulting in specific rules that define a target class. A data analysis technique called Attribute-oriented Induction is employed on the data set for achieving characterization.
- **Data Discrimination:** Discrimination is used to separate distinct data sets based on the disparity in attribute values. It compares features of a class with features of one or more contrasting classes.g., bar charts, curves and pie charts.

2. Mining Frequent Patterns

One of the functions of data mining is finding data patterns. Frequent patterns are things that are discovered to be most common in data. Various types of frequency can be found in the dataset.

- **Frequent item set:**This term refers to a group of items that are commonly found together, such as milk and sugar.
- **Frequent substructure:** It refers to the various types of data structures that can be combined with an item set or subsequences, such as trees and graphs.
- **Frequent Subsequence:** A regular pattern series, such as buying a phone followed by a cover.

3. Association Analysis

It analyses the set of items that generally occur together in a transactional dataset. It is also known as Market Basket Analysis for its wide use in retail sales. Two parameters are used for determining the association rules:

- t provides which identifies the common item set in the database.
- Confidence is the conditional probability that an item occurs when another item occurs in a transaction.

4. Classification

Classification is a data mining technique that categorizes items in a collection based on some predefined properties. It uses methods like if-then, decision trees or neural networks to predict a class or essentially classify a collection of items. A training set containing items whose properties are known is used to train the system to predict the category of items from an unknown collection of items.

5. Prediction

It defines predict some unavailable data values or spending trends. An object can be anticipated based on the attribute values of the object and attribute values of the classes. It can be a prediction of missing numerical values or increase or decrease trends in time-related information. There are primarily two types of predictions in data mining: numeric and class predictions.

- ***Numeric predictions*** are made by creating a linear regression model that is based on historical data. Prediction of numeric values helps businesses ramp up for a future event that might impact the business positively or negatively.
- ***Class predictions*** are used to fill in missing class information for products using a training data set where the class for products is known.

6. Cluster Analysis

In image processing, pattern recognition and bioinformatics, clustering is a popular data mining functionality. It is similar to classification, but the classes are not predefined. Data attributes represent the classes. Similar data are grouped together, with the difference being that a class label is not known. Clustering algorithms group data based on similar features and dissimilarities.

7. Outlier Analysis

Outlier analysis is important to understand the quality of data. If there are too many outliers, you cannot trust the data or draw patterns. An outlier analysis determines if there is something out of turn in the data and whether it indicates a situation that a business needs to consider and take

measures to mitigate. An outlier analysis of the data that cannot be grouped into any classes by the algorithms is pulled up.

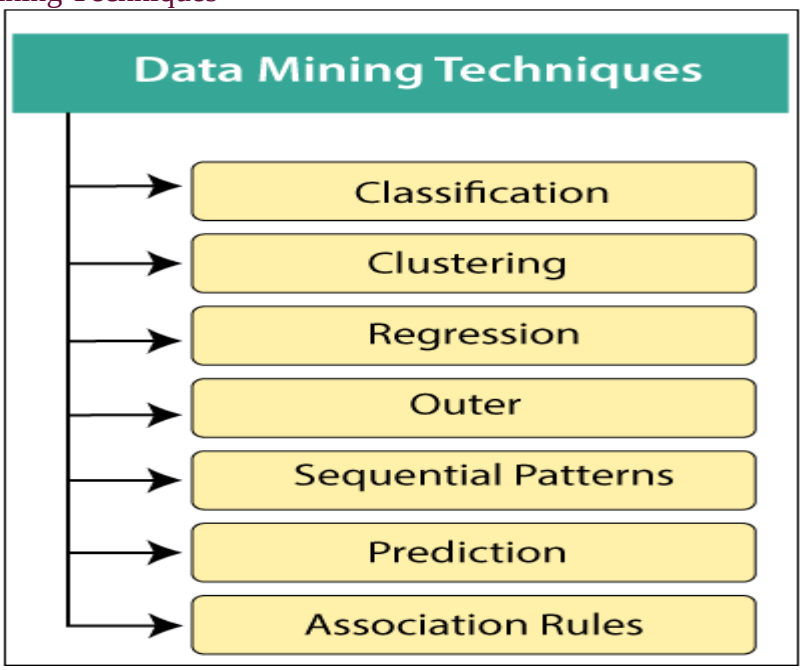
8. Evolution and Deviation Analysis

Evolution Analysis pertains to the study of data sets that change over time. Evolution analysis models are designed to capture evolutionary trends in data helping to characterize, classify, cluster or discriminate time-related data.

9. Correlation Analysis

Correlation is a mathematical technique for determining whether and how strongly two attributes is related to one another. It refers to the various types of data structures, such as trees and graphs, that can be combined with an item set or subsequence. It determines how well two numerically measured continuous variables are linked.

Data Mining Techniques



1. Classification:

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

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

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

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

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

iv. **Classification of data mining frameworks according to data mining techniques used:** This classification is as per the data analysis approach utilized, such as neural networks, machine learning, genetic algorithms, visualization, statistics, data warehouse-oriented or database-oriented, etc.
The classification can also take into account, the level of user interaction involved in the data mining procedure, such as query-driven systems, autonomous systems, or interactive exploratory systems.

2. Clustering:

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

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

3. Regression:

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

demand, and competition. Primarily it gives the exact relationship between two or more variables in the given data set.

4. Association Rules:

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

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

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

These are three major measurements technique:

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

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

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

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

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

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

5. Outer detection:

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

6. Sequential Patterns:

The sequential pattern is a data mining technique specialized for **evaluating sequential data** to discover sequential patterns. It comprises of finding interesting subsequences in a set of sequences, where the stake of a sequence can be measured in terms of different criteria like length, occurrence frequency, etc.
In other words, this technique of data mining helps to discover or recognize similar patterns in transaction data over some time.

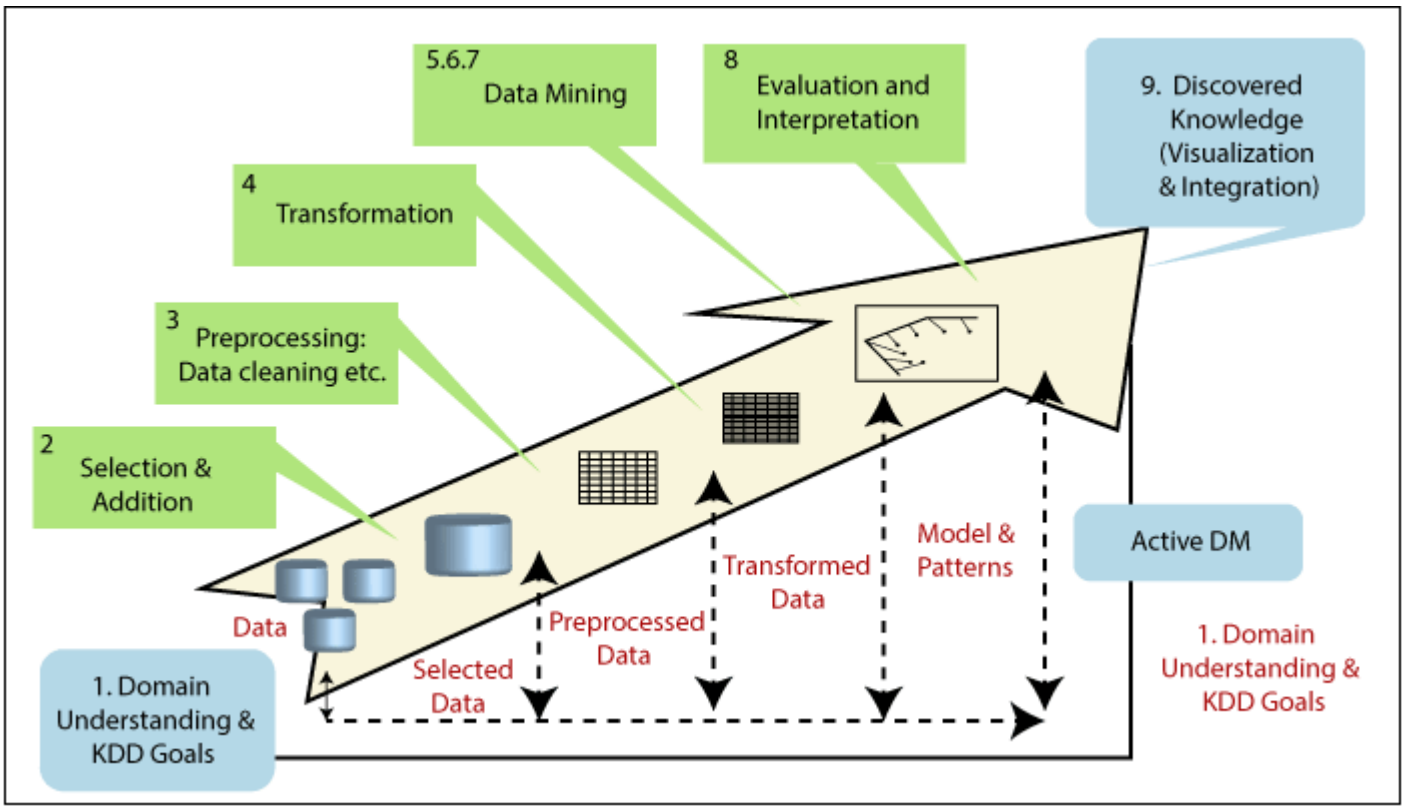
7. Prediction:

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

The KDD Process

The knowledge discovery process(illustrates in the given figure) is iterative and interactive, comprises of nine steps. The process is iterative at each stage, implying that moving back to the previous actions might be required.

The process begins with determining the KDD objectives and ends with the implementation of the discovered knowledge. Following is a concise description of the nine-step KDD process, Beginning with a managerial step:



1. Building up an understanding of the application domain

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

2. Choosing and creating a data set on which discovery will be performed

Once defined the objectives, the data that will be utilized for the knowledge discovery process should be determined. This incorporates discovering what data is accessible, obtaining important data, and afterward integrating all the data for knowledge discovery onto one set involves the qualities that will be considered for the process. This process is important because of Data Mining learns and discovers from the accessible data. This is the evidence base for building the models. If some significant attributes are missing, at that point, then the entire study may be unsuccessful from this respect, the more attributes are considered. On the other hand, to organize, collect, and operate advanced data repositories is expensive, and there is an arrangement with the opportunity for best understanding the phenomena. This arrangement refers to an aspect where the interactive and iterative aspect of the KDD is taking place. This begins with the best available data sets and later expands and observes the impact in terms of knowledge discovery and modeling.

3. Preprocessing and cleansing

In this step, data reliability is improved. It incorporates data clearing, for example, Handling the missing quantities and removal of noise or outliers. It might include complex statistical techniques or use a Data Mining algorithm in this context. For example, when one suspects that a specific attribute of lacking reliability or has many missing data, at this point, this attribute could turn into the objective of the Data Mining supervised algorithm. A prediction model for these attributes will be created, and after that, missing data can be predicted.

4. Data Transformation

In this stage, the creation of appropriate data for Data Mining is prepared and developed. Techniques here incorporate dimension reduction(for example, feature selection and extraction and record sampling), also attribute transformation(for example, discretization of numerical attributes and functional transformation). This step can be essential for the success of the entire KDD project, and it is typically very project-specific.

5. Prediction and description

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

6. Selecting the Data Mining algorithm

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

7. Utilizing the Data Mining algorithm

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

8. Evaluation

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

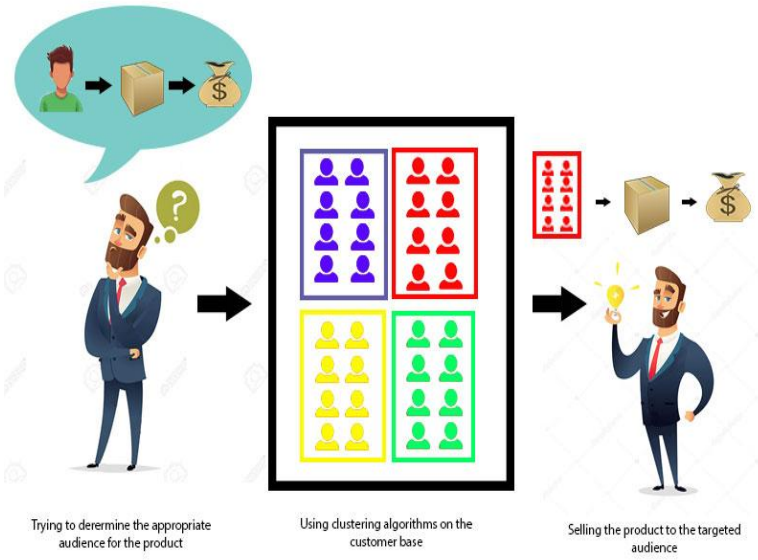
9. Using the discovered knowledge

Now, we are prepared to include the knowledge into another system for further activity. The knowledge becomes effective in the sense that we may make changes to the system and measure the impacts. The accomplishment of this step decides the effectiveness of the whole KDD process. For example, the knowledge was discovered from a certain static depiction, it is usually a set of data, but now the data becomes dynamic. Data structures may change certain quantities that become unavailable, and the data domain might be modified, such as an attribute that may have a value that was not expected previously.

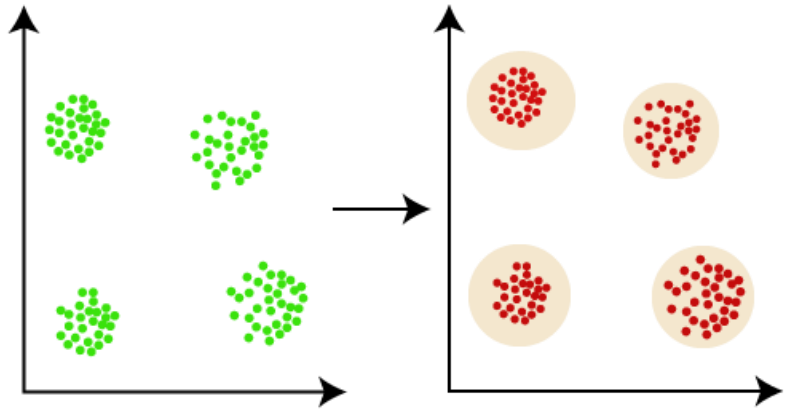
Clustering in Data Mining

Clustering is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group.

Clustering helps to splits data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters. Now that the data from our customer base is divided into clusters, we can make an informed decision about who we think is best suited for this product.



Let's understand this with an example, suppose we are a market manager, and we have a new tempting product to sell. We are sure that the product would bring enormous profit, as long as it is sold to the right people. So, how can we tell who is best suited for the product from our company's huge customer base?



Showing four clusters formed from the set of unlabeled data

Clustering, falling under the category of **unsupervised machine learning**, is one of the problems that machine learning algorithms solve.

Clustering only utilizes input data, to determine patterns, anomalies, or similarities in its input data.

A good clustering algorithm aims to obtain clusters whose:

- The intra-cluster similarities are high, It implies that the data present inside the cluster is similar to one another.
- The inter-cluster similarity is low, and it means each cluster holds data that is not similar to other data.

What is a Cluster?

- A cluster is a subset of similar objects
- A subset of objects such that the distance between any of the two objects in the cluster is less than the distance between any object in the cluster and any object that is not located inside it.
- A connected region of a multidimensional space with a comparatively high density of objects.

What is clustering in Data Mining?

- 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

Important points:

- Data objects of a cluster can be considered as one group.
- We 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.
- The over-classification main advantage is that it is adaptable to modifications, and it helps single out important characteristics that differentiate between distinct groups.

Applications of cluster analysis in data mining:

- In many applications, clustering analysis is widely used, such as data analysis, market research, pattern recognition, and image processing.
- It assists marketers to find different groups in their client base and based on the purchasing patterns. They can characterize their customer groups.
- It helps in allocating documents on the internet for data discovery.
- Clustering is also used in tracking applications such as detection of credit card fraud.

- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to analyze the characteristics of each cluster.
- 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.
- It helps in the identification of areas of similar land that are used in an earth observation database and the identification of house groups in a city according to house type, value, and geographical location.

Comparison between Data Mining and Data Warehousing:

S. No.	Basis of Comparison	Data Warehousing	Data Mining
1.	**Definition**	A data warehouse is a database system that is designed for analytical analysis instead of transactional work.	Data mining is the process of analyzing data patterns.
2.	**Process**	Data is stored periodically.	Data is analyzed regularly.
3.	**Purpose**	Data warehousing is the process of extracting and storing data to allow easier reporting.	Data mining is the use of pattern recognition logic to identify patterns.
4.	**Managing Authorities**	Data warehousing is solely carried out by engineers.	Data mining is carried out by business users with the help of engineers.
5.	**Data**	Data warehousing is the	Data mining is considered as a process of

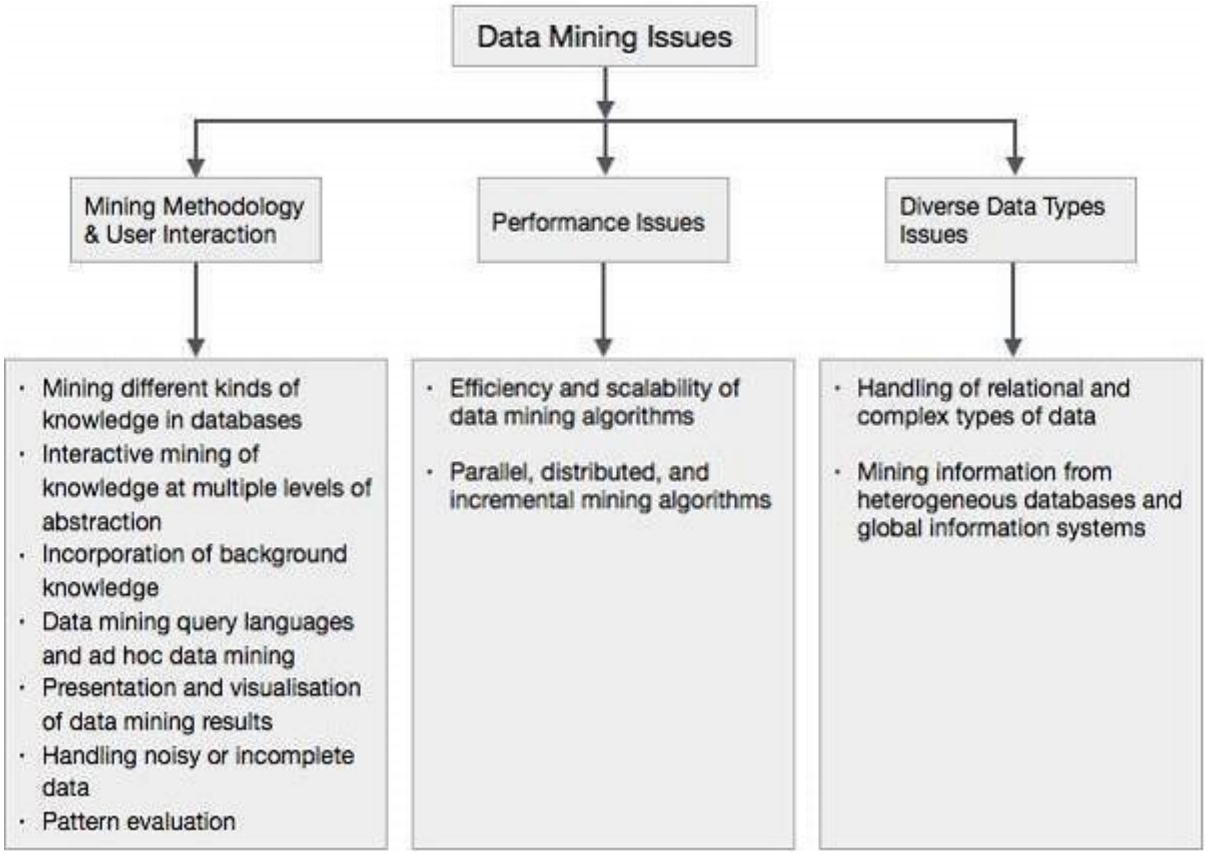
S. No.	Basis of Comparison	Data Warehousing	Data Mining
	Handling	process of pooling all relevant data together.	extracting data from large data sets.
6.	**Functionality**	Subject-oriented, integrated, time-varying and non-volatile constitute data warehouses.	AI, statistics, databases, and machine learning systems are all used in data mining technologies.
7.	**Task**	Data warehousing is the process of extracting and storing data in order to make reporting more efficient.	Pattern recognition logic is used in data mining to find patterns.
8.	**Uses**	It extracts data and stores it in an orderly format, making reporting easier and faster.	This procedure employs pattern recognition tools to aid in the identification of access patterns.
9.	**Examples**	When a data warehouse is connected with operational business systems like CRM (Customer Relationship Management) systems, it adds value.	Data mining aids in the creation of suggestive patterns of key parameters. Customer purchasing behavior, items, and sales are examples. As a result, businesses will be able to make the required adjustments to their operations and production.

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 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.

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.