•**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

•**goal of the data mining** process is to extract information from a data set and transform it into an understandable structure for further use.

**•The key properties of data mining are**

→Automatic discovery of patterns →Prediction of likely outcomes →Creation of actionable information →Focus on large datasets and databases

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

→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.

→A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights

→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.

→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

→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

**•Advantages of Data Mining**

→The Data Mining technique enables organizations to obtain knowledge-based data.

→Data mining enables organizations to make lucrative modifications in operation and production.

→Compared with other statistical data applications, data mining is a cost-efficient.

→Data Mining helps the decision-making process of an organization.

→It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.

→It can be induced in the new system as well as the existing platforms.

**•Disadvantages of Data Mining**

→Many data mining analytics software is difficult to operate and needs advance training to work on.

→Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.

→The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

**•Data Mining Applications**

→healthcare →Market basket analysis →education →manufacturing engineering →customer relationship management →fraud detection →lie detection → financial banking

**•Challenges of Implementation in Data mining**

→incomplete and noisy data →data distribution →complex data →performance →Data privacy and security →Data visualization.

**•Task of data mining**

→Anomaly detection (Outlier/change/deviation detection) :The identification of unusual data records, that might be interesting or data errors that require further investigation.

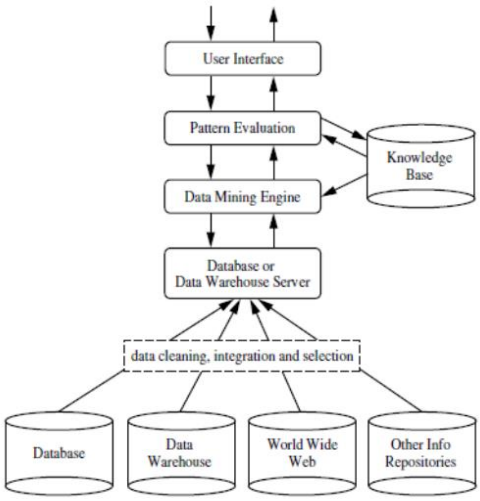
→Association rule learning (Dependency modelling) : Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

→Clustering : is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

→Regression : attempts to find a function which models the data with the least error.

→Summarization: providing a more compact representation of the data set, including visualization and report generation.

**•Architecture of data mining**



**•Explanation**

→Knowledge Base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.

→Data Mining Engine: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

→Pattern Evaluation Module: This component typically employs interestingness measures interacts with the data mining modules so as to focus the search toward interesting patterns.

→User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory datamining based on the intermediate data mining results

**•Data mining process/steps**

1.State the problem and formulate the hypothesis :

2. Collect the data
3.Preprocessing the data : includes at least two common tasks:

→Outlier detection (and removal) : Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values.

There are two strategies for dealing with outliers:

a. Detect and eventually remove outliers as a part of the preprocessing phase, or

b. Develop robust modeling methods that are insensitive to outliers.

→Scaling, encoding, and selecting features – Data preprocessing includes several steps such as variable scaling and different types of encoding. For example, one feature with the range [0, 1] and the other with the range [−100, 1000] will not have the same weights in the applied technique; they will also influence the final data-mining results differently.

4. Estimate the model : main task is to select and implement of the appropriate data-mining technique.

5. Interpret the model and draw conclusions:

**•Data mining implementation process**

1.Business understanding: In this phase, business and data-mining goals are established.

2.Data understanding: In this phase, sanity check on data is performed to check whether its appropriate for the data mining goals.

3.Data preparation: In this phase, data is made production ready where by The data from different sources should be selected, cleaned, transformed, formatted, anonymized, and constructed (if required)

4.Data transformation: Data transformation operations would contribute toward the success of the mining process. It includes

→Smoothing: It helps to remove noise from the data.

→ Aggregation: Summary or aggregation operations are applied to the data. I.e., the weekly sales data is aggregated to calculate the monthly and yearly total.

→ Generalization: In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.

→ Normalization: Normalization performed when the attribute data are scaled up o scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.

→ Attribute construction: these attributes are constructed and included the given set of attributes helpful for data mining

5.Modelling: In this phase, mathematical models are used to determine data patterns.

6. Evaluation: In this phase, patterns identified are evaluated against the business objectives

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

**•Classification of Data mining Systems:**

→Database Technology →Statistics →Machine Learning →Information Science →Visualization

**•Data mining technique**

1.**Classification**: This technique is used to obtain important and relevant information about data and metadata. Data mining techniques can be classified by different criteria, as follows:

→Classification of Data mining frameworks as per the type of data sources mined: 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: For example. Object-oriented database, transactional database, relational database, and so on.

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

**3.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.

**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

**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. Also may be used in various domains like intrusion, detection, fraud detection, etc

**6.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.

**7.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

**•Some Other Classification Criteria:**

→Classification according to kind of databases mined

Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example if we classify the database according to data model then we may have a relational, transactional, object- relational, or data warehouse mining system.

→Classification according to kind of knowledge mined

It is means data mining system are classified on the basis of functionalities such as:

−Characterization −Discrimination −Association and Correlation Analysis −Classification −Prediction −Clustering −Outlier Analysis −Evolution Analysis

→Classification according to kinds of techniques utilized

We can describes these techniques according to degree of user interaction involved or the methods of analysis employed.

→Classification according to applications adapted

We can classify the data mining system according to application adapted. These applications are as follows:

-Finance -Telecommunications -DNA -Stock Markets -E-mail

**•Major Issues In Data Mining:**

→Skilled Experts are needed to formulate the data mining queries.

→Overfitting: Due to small size training database, a model may not fit future states.

→Data mining needs large databases which sometimes are difficult to manage

→Business practices may need to be modified to determine to use the information uncovered.

→If the data set is not diverse, data mining results may not be accurate.

→Integration information needed from heterogeneous databases and global information systems could be complex

•**Text data mining** can be described as the process of extracting essential data from standard language text. All the data that we generate via text messages, documents, emails, files are written in common language text.

**•Areas of text mining in data mining:**

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

→**Natural Language Processing:** concerned with giving computers the ability to understand text and spoken words in much the same way human beings can

→**Data mining** refers to the extraction of useful data, hidden patterns from large data sets. Data mining tools can predict behaviors and future trends that allow businesses to make a better data-driven decision.

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

**•Text Mining Process**

→Text transformation: is a technique that is used to control the capitalization of the text.

→Text Pre-processing: a significant task and a critical step in Text Mining, Natural Language Processing (NLP), and information retrieval (IR).

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

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

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

**•Applications of Text Mining:**

→**Risk Management** is a systematic and logical procedure of analyzing, identifying, treating, and monitoring the risks involved in any action or process in organizations. Insufficient risk analysis is usually a leading cause of disappointment

→**Customer Care Service:** Text mining methods, particularly NLP, are finding increasing significance in the field of customer care. Organizations are spending in text analytics programming to improve their overall experience by accessing the textual data from different sources such as customer feedback, surveys, customer calls, etc

→**Business Intelligence:** Companies and business firms have started to use text mining strategies as a major aspect of their business intelligence

→**Social Media Analysis**: Social media analysis helps to track the online data, and there are numerous text mining tools designed particularly for performance analysis of social media sites

**•Text Mining Approaches in Data Mining:**

**1. Keyword-based Association Analysis**: It collects sets of keywords or terms that often happen together and afterward discover the association relationship among them.

**2. Document Classification Analysis:** Automatic document classification,This analysis is used for the automatic classification of the huge number of online text documents like web pages, emails, etc

**•Web mining** is the process of using data mining techniques and algorithms to extract information directly from the Web by extracting it from Web documents and services, Web content, hyperlinks and server logs.

**•Categories of Web mining:**

→**Web content mining** — This is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files.

→**Web structure mining** — This is the process of analyzing the nodes and connection structure of a website through the use of graph theory.

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

**•Comparison Between Data mining and Web mining:**

→Data Mining is the process that attempts to discover pattern and hidden knowledge in large data sets in any system. **WHILE** Web Mining is the process of data mining techniques to automatically discover and extract information from web documents.

→Data Mining is very useful for web page analysis. **WHILE** Web Mining is very useful for a particular website and e-service.

→Data mining Used by Data scientist and data engineers. **WHILE** Used by Data scientists along with data analysts.

→Data Mining is access data privately. **WHILE** Web Mining is access data publicly.

→In Data Mining get the information from explicit structure. **WHILE** In Web Mining get the information from structured, unstructured and semi structured web pages.

→Data mining includes approaches for data cleansing, machine learning algorithms. Statistics and probability **WHILE** Web Mining includes application level knowledge, data engineering with mathematical modules like statistics and probability.