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.

In other words, we can say that 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.

Advantages of Data Mining

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

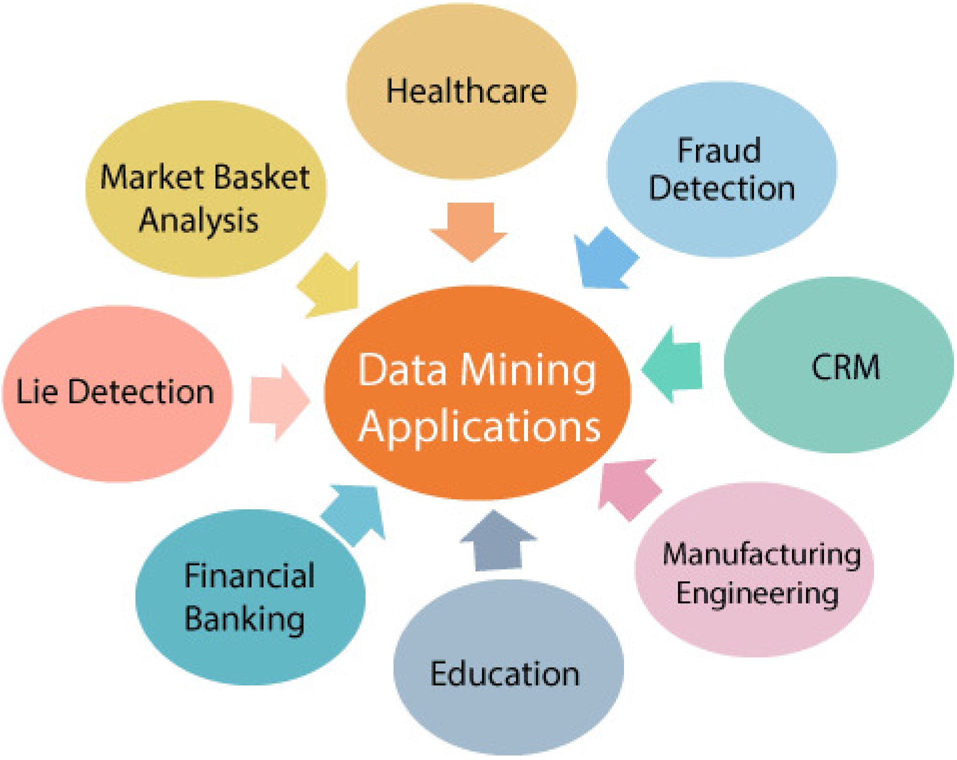
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.
- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Data Mining Applications

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

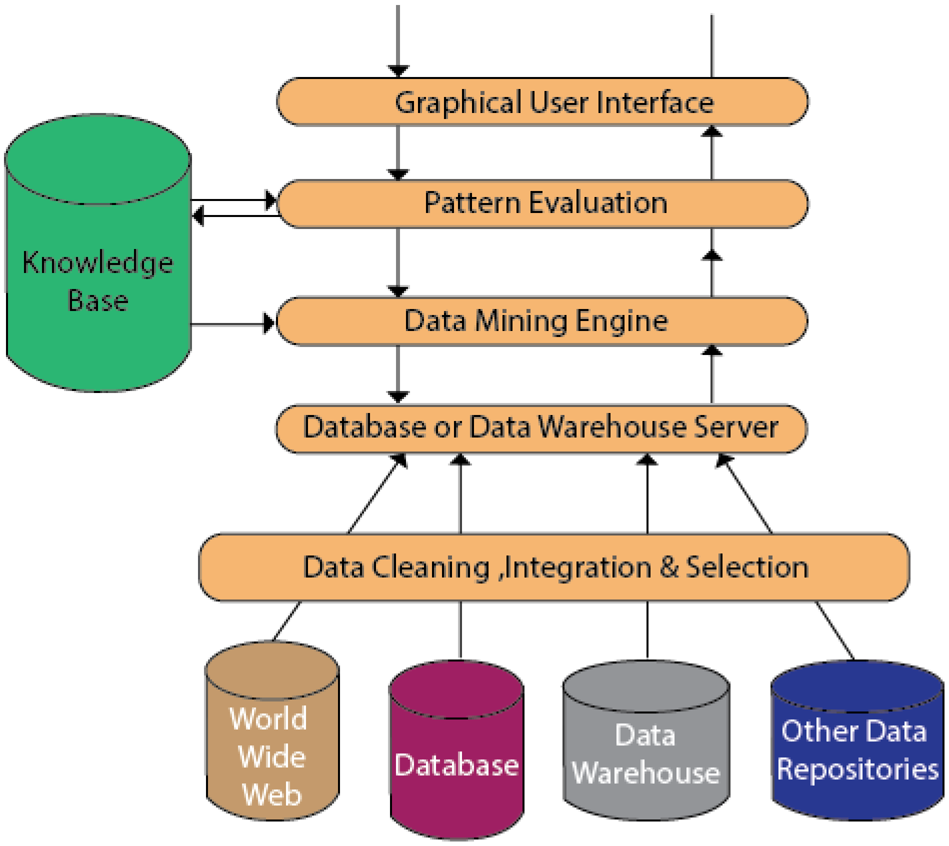


Introduction

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

Data Mining Architecture

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



Data Source:

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

Different processes:

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

Database or Data Warehouse Server:

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

Data Mining Engine:

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

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

Pattern Evaluation Module:

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

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

Graphical User Interface:

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

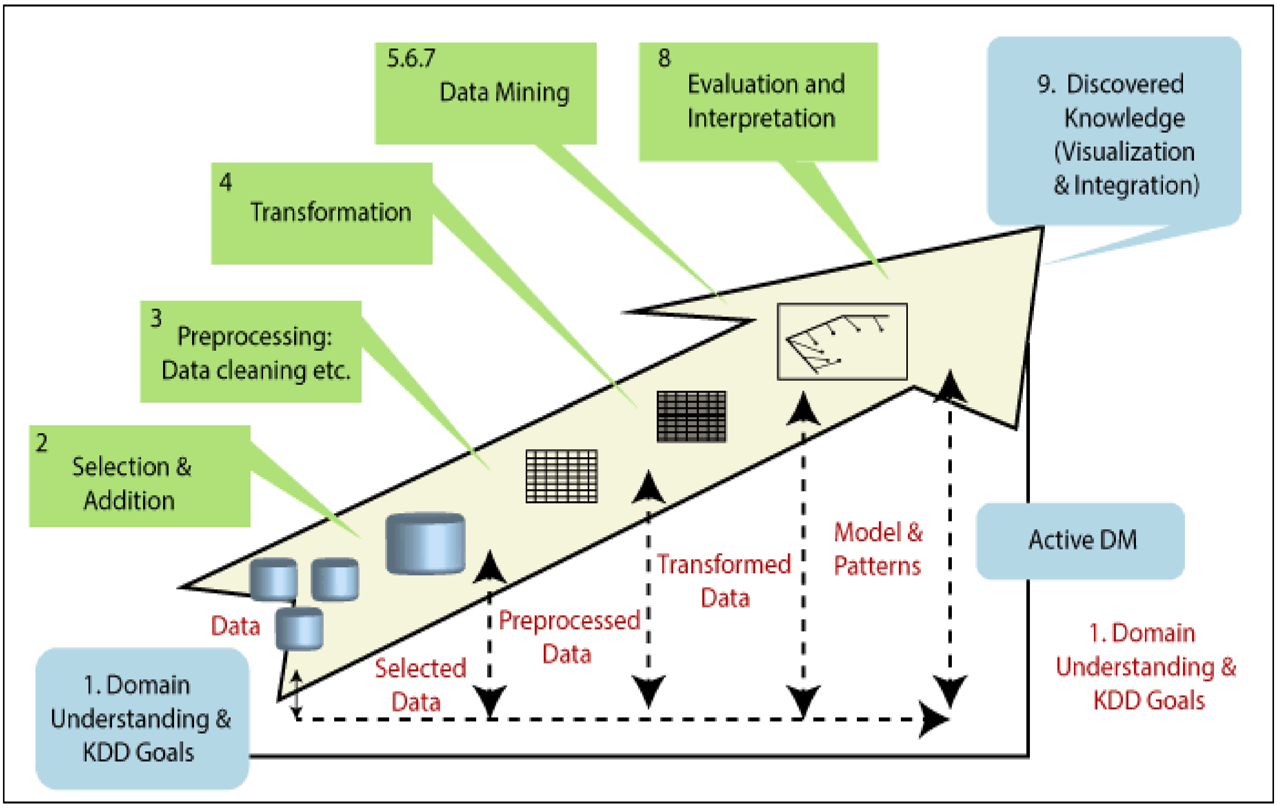
Knowledge Base:

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

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 has many imaginative aspects in the sense that one cant presents one formula or make a complete scientific categorization for the correct decisions for each step and application type. Thus, it is needed to understand the process and the different requirements and possibilities in each stage.

The process begins with determining the KDD objectives and ends with the implementation of the discovered knowledge. At that point, the loop is closed, and the Active Data Mining starts. Subsequently, changes would need to be made in the application domain. For example, offering various features to cell phone users in order to reduce churn. This closes the loop, and the impacts are then measured on the new data repositories, and the KDD process again. Following is a concise description of the nine-step KDD process, Beginning with a managerial step:


5.6.7
Data Mining
8
Evaluation and Interpretation
9. Discovered Knowledge (Visualization & Integration)
4
Transformation
3
Preprocessing: Data cleaning etc.
2
Selection & Addition
Data
Selected Data
Preprocessed Data
Transformed Data
Model & Patterns
Active DM
1. Domain Understanding & KDD Goals
1. Domain Understanding & KDD Goals

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. The expansion to which one pays attention to this level relies upon numerous factors. Regardless, studying the aspects is significant and regularly revealing by itself, to enterprise data frameworks.

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. For example, in medical assessments, the quotient of attributes may often be the most significant factor and not each one by itself. In business, we may need to think about impacts beyond our control as well as efforts and transient issues. For example, studying the impact of advertising accumulation. However, if we do not utilize the right transformation at the starting, then we may acquire an amazing effect that insights to us about the transformation required in the next iteration. Thus, the KDD process follows upon itself and prompts an understanding of the transformation required.

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.

;

Data Mining Techniques/Task/Functionality

Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets. These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees. Thus, data mining incorporates analysis and prediction.

data mining **activities** can be divided into 2 categories:

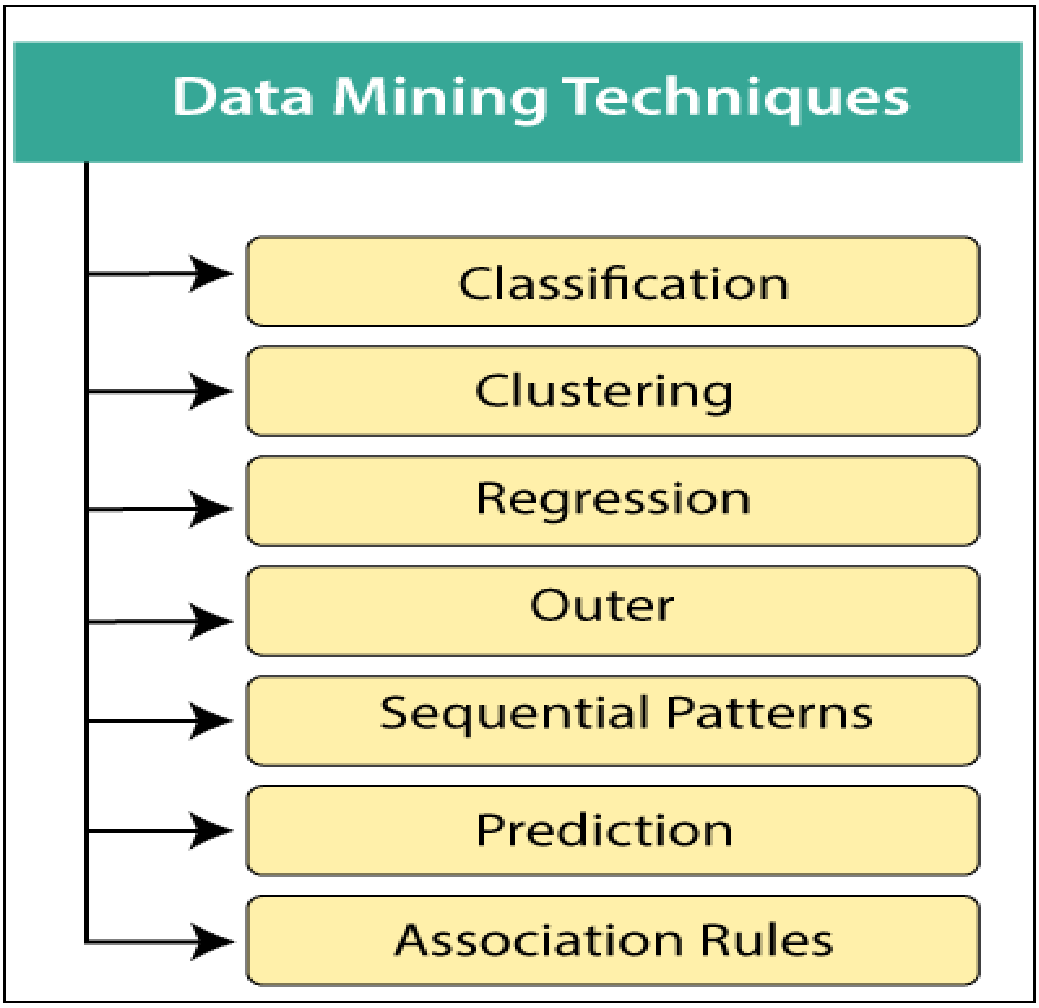
1. **Descriptive Data Mining:**
   It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set.
   For examples: count, average etc.

1. Clustering
2. 2.Summarization
3. 3.Association Rules.
4. Sequence Discovery

2. **Predictive Data Mining:**
   It helps developers to provide unlabeled definitions of

attributes. Based on previous tests, the software estimates the characteristics that are absent.
For example: Judging from the findings of a patient’s medical examinations that is he suffering from any particular disease.

1.Classification
2.Regression
3.Time Series Analysis
4.Prediction



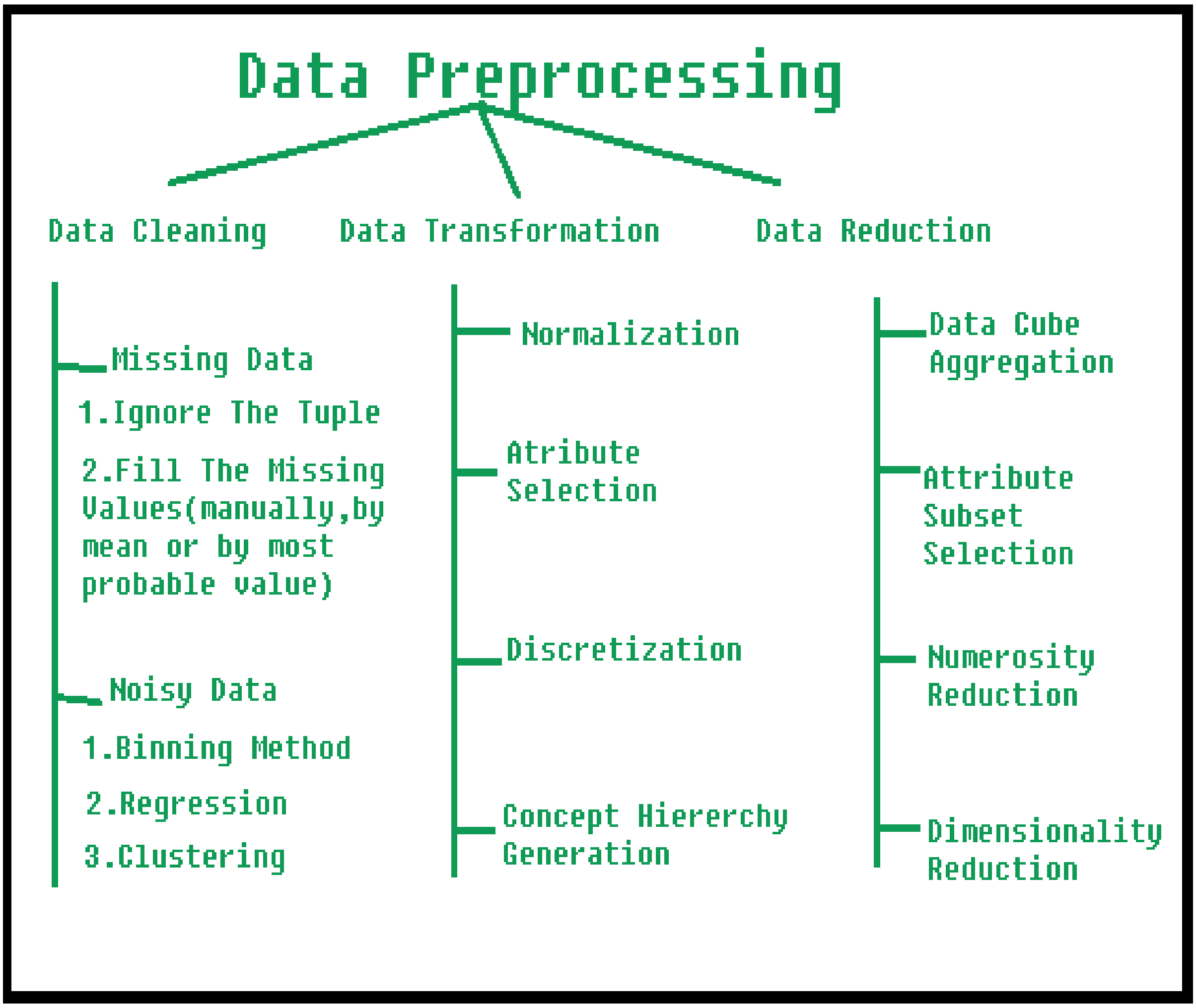
1. Classification:

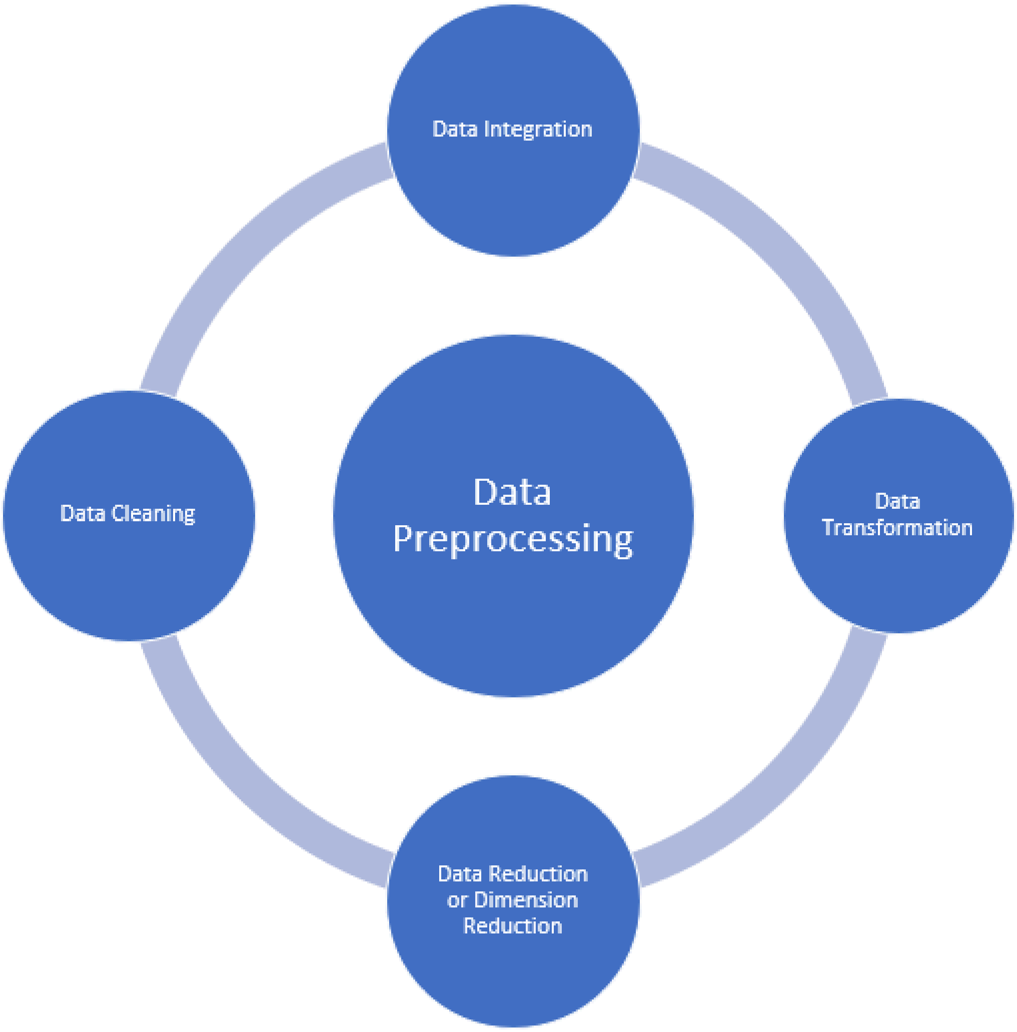
This technique is used to obtain important and relevant information about data and metadata. This data mining technique helps to classify data in different classes.

Data Preprocessing in Data Mining

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.


Data Preprocessing
Data Cleaning
Data Transformation
Data Reduction
Missing Data
1.Ignore The Tuple
2.Fill The Missing Values(manually,by mean or by most probable value)
Noisy Data
1.Binning Method
2.Regression
3.Clustering
Normalization
Atribute Selection
Discretization
Concept Hiererchy Generation
Data Cube Aggregation
Attribute Subset Selection
Numerosity Reduction
Dimensionality Reduction



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”.

5.

3. Data Reduction:
Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**
   Aggregation operation is applied to data for the construction of the data cube.

2. **Attribute Subset Selection:**
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than

significance level can be discarded.

3. **Numerosity Reduction:**
   This enable to store the model of data instead of whole data, for example: Regression Models.

4. **Dimensionality Reduction:**
   This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Component Analysis).

Data Transformation in Data Mining

The data are transformed in ways that are ideal for mining the data. The data transformation involves steps that are:

1. Smoothing:

It is a process that is used to remove noise from the dataset using some algorithms It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.

2. Aggregation:

Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.

The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies.

For **example**, Sales, data may be aggregated to compute monthly& annual total amounts.

3. Discretization:
It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.

Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.

For **example**, (1-10, 11-20) (age:- young, middle age, senior).

4. Attribute Construction:
Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.

5. Generalization:
It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).

For **example**, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

6. Normalization: Data normalization involves converting all data variable into a given range.