	Pithapur Rajahs Government College(A) Kakinada **Department of Computer Science**	**Program:** II B.Sc. (Data Science) **Semester :** III
Course 7	Course Name: DATA MINING TECHNIQUES USING R	
	Hours Allocated: 3hrs/week	Credits: 3

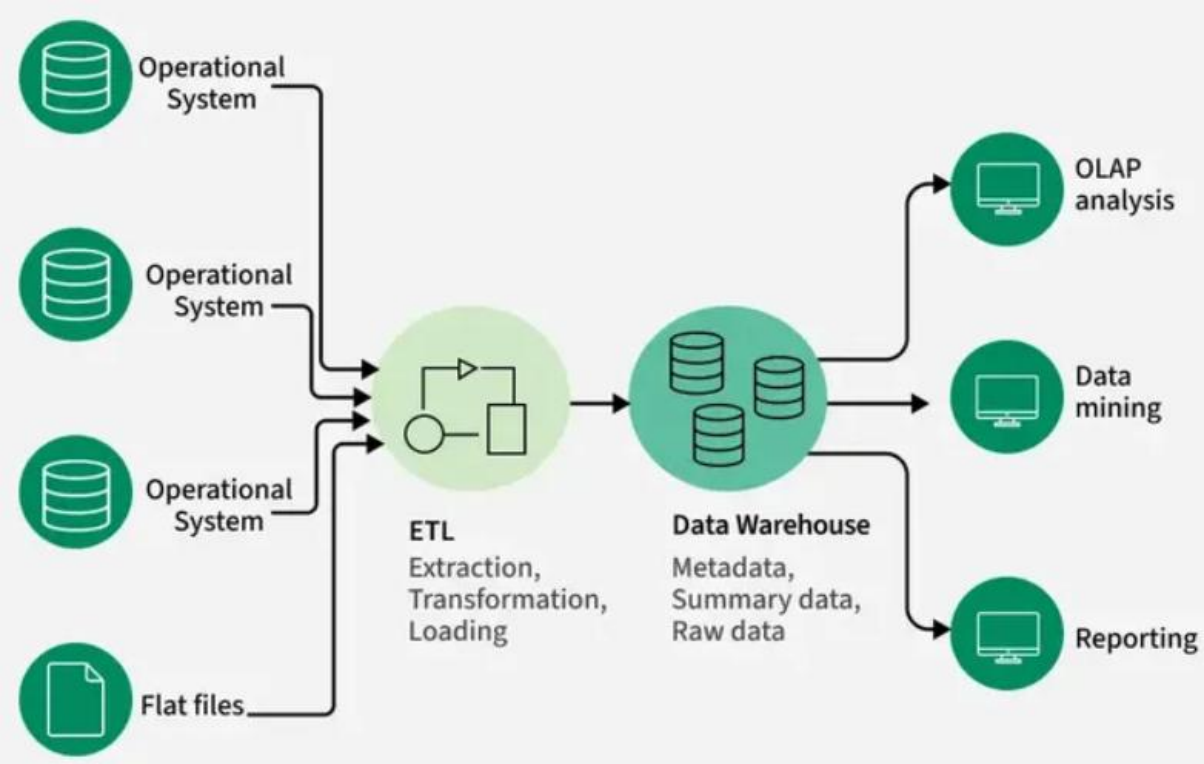
UNIT – I SYLLABUS

An idea on Data Warehouse, Data mining-KDD versus data mining, Stages of the Data Mining Process-Task primitives., Data Mining Techniques – Data mining knowledge representation.

An idea on Data Warehouse

A data warehouse is a centralized system used for storing and managing large volumes of data from various sources. It is designed to help businesses analyze historical data and make informed decisions. Data from different operational systems is collected, cleaned, and stored in a structured way, enabling efficient querying and reporting.

- Goal is to produce statistical results that may help in decision-making.
- Ensures fast data retrieval even with the vast datasets.



Need for Data Warehousing

1. Handling Large Volumes of Data: Traditional databases can only store a limited amount of data (MBs to GBs), whereas a data warehouse is designed to handle much larger datasets (TBs), allowing businesses to store and manage massive amounts of historical data.

2. Enhanced Analytics: Transactional databases are not optimized for analytical purposes. A data warehouse is built specifically for data analysis, enabling businesses to perform complex queries and gain insights from historical data.

3. Centralized Data Storage: A data warehouse acts as a central repository for all organizational data, helping businesses to integrate data from multiple sources and have a unified view of their operations for better decision-making.

4. Trend Analysis: By storing historical data, a data warehouse allows businesses to analyze trends over time, enabling them to make strategic decisions based on past performance and predict future outcomes.

5. Support for Business Intelligence: Data warehouses support business intelligence tools and reporting systems, providing decision-makers with easy access to critical information, which enhances operational efficiency and supports data-driven strategies.

Data Mining

Data mining is the process of extracting insights from large datasets using statistical and computational techniques. It can involve structured, semi-structured or unstructured data stored in databases, data warehouses or data lakes. The goal is to uncover hidden patterns and relationships to support informed decision-making and predictions using methods like clustering, classification, regression and anomaly detection.

Data mining is widely used in industries such as marketing, finance, healthcare and telecommunications. For example, it helps identify customer segments in marketing or detect disease risk factors in healthcare. However, it also raises ethical concerns particularly regarding privacy and the misuse of personal data, requiring careful safeguards.

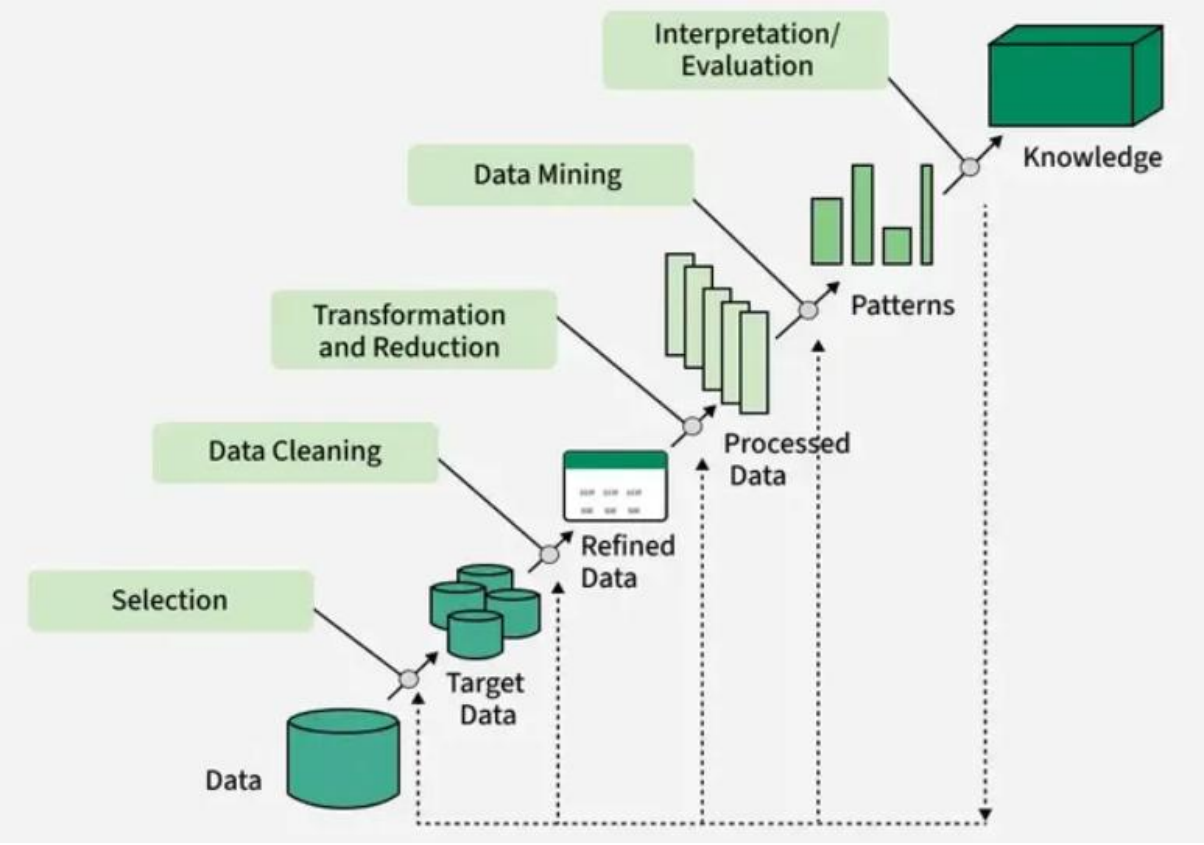
KDD versus data mining

Knowledge Discovery in Databases (KDD) refers to the complete process of uncovering valuable knowledge from large datasets. It starts with the selection of relevant data, followed by preprocessing to clean and organize it, transformation to prepare it for analysis, data mining to uncover patterns and relationships, and concludes with the evaluation and interpretation of results, ultimately producing valuable knowledge or insights. KDD is widely utilized in fields like machine learning, pattern recognition, statistics, artificial intelligence, and data visualization.

The KDD process is iterative, involving repeated refinements to ensure the accuracy and reliability of the knowledge extracted. The whole process consists of the following steps:

1. **Data Selection**
2. **Data Cleaning and Preprocessing**

3. **Data Transformation and Reduction**
4. **Data Mining**
5. **Evaluation and Interpretation of Results**



Data Selection

Data Selection is the initial step in the Knowledge Discovery in Databases (KDD) process, where relevant data is identified and chosen for analysis. It involves selecting a dataset or focusing on specific variables, samples, or subsets of data that will be used to extract meaningful insights.

- It ensures that only the most relevant data is used for analysis, improving efficiency and accuracy.
- It involves selecting the entire dataset or narrowing it down to particular features or subsets based on the task's goals.
- Data is selected after thoroughly understanding the application domain.

By carefully selecting data, we ensure that the KDD process delivers accurate, relevant, and actionable insights.

Data Cleaning

In the KDD process, Data Cleaning is essential for ensuring that the dataset is accurate and reliable by correcting errors, handling missing values, removing duplicates, and addressing noisy or outlier data.

- **Missing Values:** Gaps in data are filled with the mean or most probable value to maintain dataset completeness.
- **Noisy Data:** Noise is reduced using techniques like binning, regression, or clustering to smooth or group the data.
- **Removing Duplicates:** Duplicate records are removed to maintain consistency and avoid errors in analysis.

Data cleaning is crucial in KDD to enhance the quality of the data and improve the effectiveness of data mining.

Data Transformation and Reduction

Data Transformation in KDD involves converting data into a format that is more suitable for analysis.

- **Normalization**: Scaling data to a common range for consistency across variables.
- **Discretization**: Converting continuous data into discrete categories for simpler analysis.
- **Data Aggregation**: Summarizing multiple data points (e.g., averages or totals) to simplify analysis.
- **Concept Hierarchy Generation**: Organizing data into hierarchies for a clearer, higher-level view.

Data Reduction helps simplify the dataset while preserving key information.

- **Dimensionality Reduction** (e.g., PCA): Reducing the number of variables while keeping essential data.
- **Numerosity Reduction**: Reducing data points using methods like sampling to maintain critical patterns.
- **Data Compression**: Compacting data for easier storage and processing.

Together, these techniques ensure that the data is ready for deeper analysis and mining.

Data Mining

Data Mining is the process of discovering valuable, previously unknown patterns from large datasets through automatic or semi-automatic means. It involves exploring vast amounts of data to extract useful information that can drive decision-making.

Key characteristics of data mining patterns include:

- **Validity**: Patterns that hold true even with new data.
- **Novelty**: Insights that are non-obvious and surprising.
- **Usefulness**: Information that can be acted upon for practical outcomes.
- **Understandability**: Patterns that are interpretable and meaningful to humans.

In the KDD process, choosing the data mining task is critical. Depending on the objective, the task could involve classification, regression, clustering, or association rule mining. After determining the task, selecting the appropriate data mining algorithms is essential. These algorithms are chosen based on their ability to efficiently and accurately identify patterns that align with the goals of the analysis.

Difference between KDD and Data Mining

Parameter	KDD	Data Mining
Definition	KDD is the overall process of discovering valid, novel, potentially useful, and ultimately understandable patterns and relationships in large datasets.	Data Mining is a subset of KDD, focused on the extraction of useful patterns and insights from large datasets.
Objective	To extract valuable knowledge and insights from data to support decision-making and understanding.	To identify patterns, relationships, and trends within data to generate useful insights.
Techniques Used	Involves multiple steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation.	Includes techniques like association rules, classification, clustering, regression, decision trees, neural networks, and dimensionality reduction.
Output	Generates structured knowledge in the form of rules, models, and insights that can aid in decision-making or predictions.	Results in patterns, relationships, or associations that can improve understanding or decision-making.
Focus	Focuses on the discovery of useful knowledge, with an emphasis on interpreting and validating the findings.	Focuses on discovering patterns, relationships, and trends within data without necessarily considering the broader context.
Role of Domain Expertise	Domain expertise is important in KDD, as it helps in defining the goals of the process, choosing	Domain expertise is less critical in data mining, as the focus is on using algorithms to detect

Parameter	KDD	Data Mining
	appropriate data, and interpreting the results.	patterns, often without prior domain-specific knowledge

Stages of the Data Mining Process

INTRODUCTION:

The data mining process typically involves the following steps:

Business Understanding: This step involves understanding the problem that needs to be solved and defining the objectives of the data mining project. This includes identifying the business problem, understanding the goals and objectives of the project, and defining the KPIs that will be used to measure success. This step is important because it helps ensure that the data mining project is aligned with business goals and objectives.

Data Understanding: This step involves collecting and exploring the data to gain a better understanding of its structure, quality, and content. This includes understanding the sources of the data, identifying any data quality issues, and exploring the data to identify patterns and relationships. This step is important because it helps ensure that the data is suitable for analysis.

Data Preparation: This step involves preparing the data for analysis. This includes cleaning the data to remove any errors or inconsistencies, transforming the data to make it suitable for analysis, and integrating the data from different sources to create a single dataset. This step is important because it ensures that the data is in a format that can be used for modeling.

Modeling: This step involves building a predictive model using machine learning algorithms. This includes selecting an appropriate algorithm, training the model on the data, and evaluating its performance. This step is important because it is the heart of the data mining process and involves developing a model that can accurately predict outcomes on new data.

Evaluation: This step involves evaluating the performance of the model. This includes using statistical measures to assess how well the model is able to predict outcomes on new data. This step is important because it helps ensure that the model is accurate and can be used in the real world.

Deployment: This step involves deploying the model into the production environment. This includes integrating the model into existing systems and processes to make predictions in real-time. This step is important because it allows the model to be used in a practical setting and to generate value for the organization.

Alternative names for Data Mining :

1. Knowledge discovery (mining) in databases (KDD)

2. Knowledge extraction

3. Data/pattern analysis

4. Data archaeology

5. Data dredging

6. Information harvesting

7. Business intelligence

Task primitives

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.

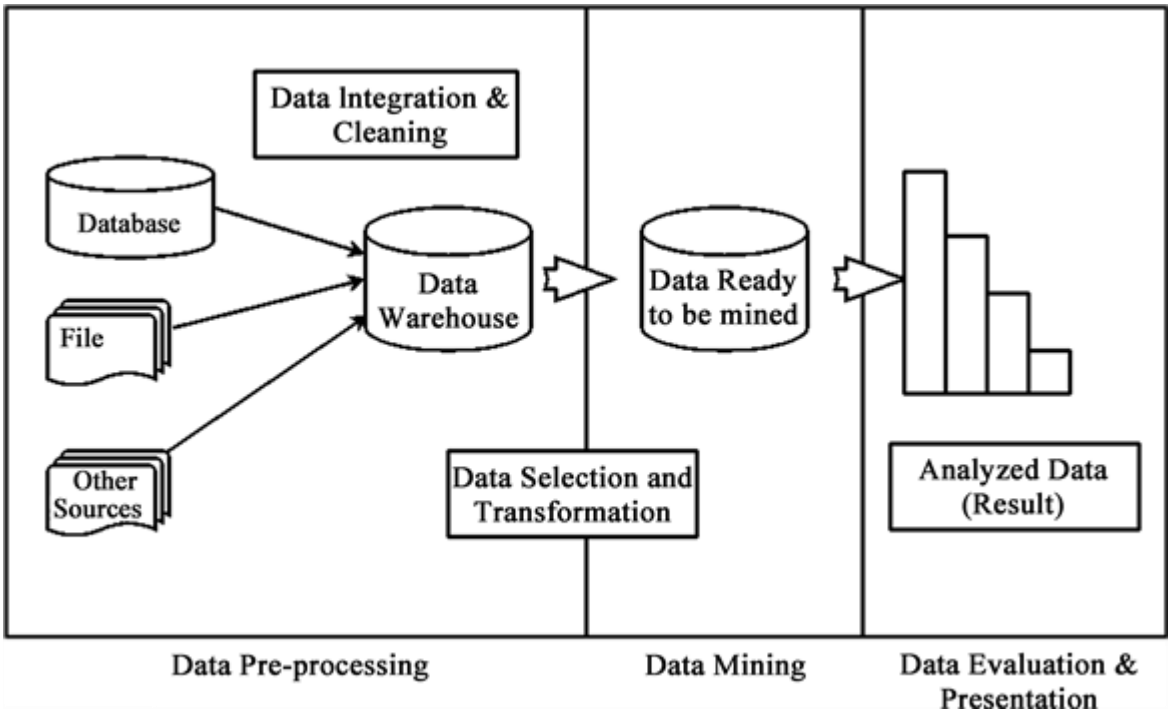
Regression analysis: This technique is used to create a model that can predict the value of a target variable based on the values of several input variables. Regression analysis is often used for prediction tasks.

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 Techniques

Data Mining is the process of discovering useful patterns and insights from large amounts of data. Data science, information technology, and artisanal practices put together to reassemble the collected information into something valuable. Researchers and professionals are working to develop newer, faster, cheaper, and more accurate ways to accomplish this process. Various other terms are attached to data mining, like "**knowledge mining from data**," "**knowledge extraction**," "**data analysis**," and "**data dredging**," which all simply refer to the same idea.

Data mining is often a synonym for **Knowledge Discovery from Data (KDD)**. Some people see data mining as a key part of KDD, where smart methods are used to find patterns in the data. The term "Knowledge Discovery in Databases" (KDD) was first coined by Gregory Piatetsky-Shapiro in 1989. However, "data mining" became more widely used in business and media. Today, both terms are often used interchangeably.



Data Mining Techniques

1. Association

Association analysis looks for patterns where certain items or conditions tend to appear together in a dataset. It's commonly used in market basket analysis to see which products are often bought together. One method, called associative classification, generates rules from the data and uses them to build a model for predictions.

2. Classification

Classification builds models to sort data into different categories. The model is trained on data with known labels and is then used to predict labels for unknown data. Some examples of classification models are:

- **Decision Tree**
- **SVM(Support Vector Machine)**
- **Generalized Linear Models**
- **Bayesian classification**
- **Classification by Backpropagation**
- **K-NN Classifier**
- **Rule-Based Classification**
- **Frequent-Pattern Based Classification**
- **Rough Set Theory**
- **Fuzzy Logic**

3. Prediction

Prediction is similar to classification, but instead of predicting categories, it predicts continuous values (like numbers). The goal is to build a model that can estimate the value of a specific attribute for new data.

4. Clustering

Clustering groups similar data points together without using predefined categories. It helps discover hidden patterns in the data by organizing objects into clusters where items in each cluster are more similar to each other than to those in other clusters.

5. Regression

Regression is used to predict continuous values, like prices or temperatures, based on past data. There are two main types: linear regression, which looks for a straight-line relationship, and multiple linear regression, which uses more variables to make predictions.

6. Artificial Neural Network (ANN) Classifier

An artificial neural network (ANN) is a model inspired by how the human brain works. It learns from data by adjusting connections between artificial neurons. Neural networks are

great for recognizing complex patterns but require a lot of training and can be hard to interpret.

7. Outlier Detection

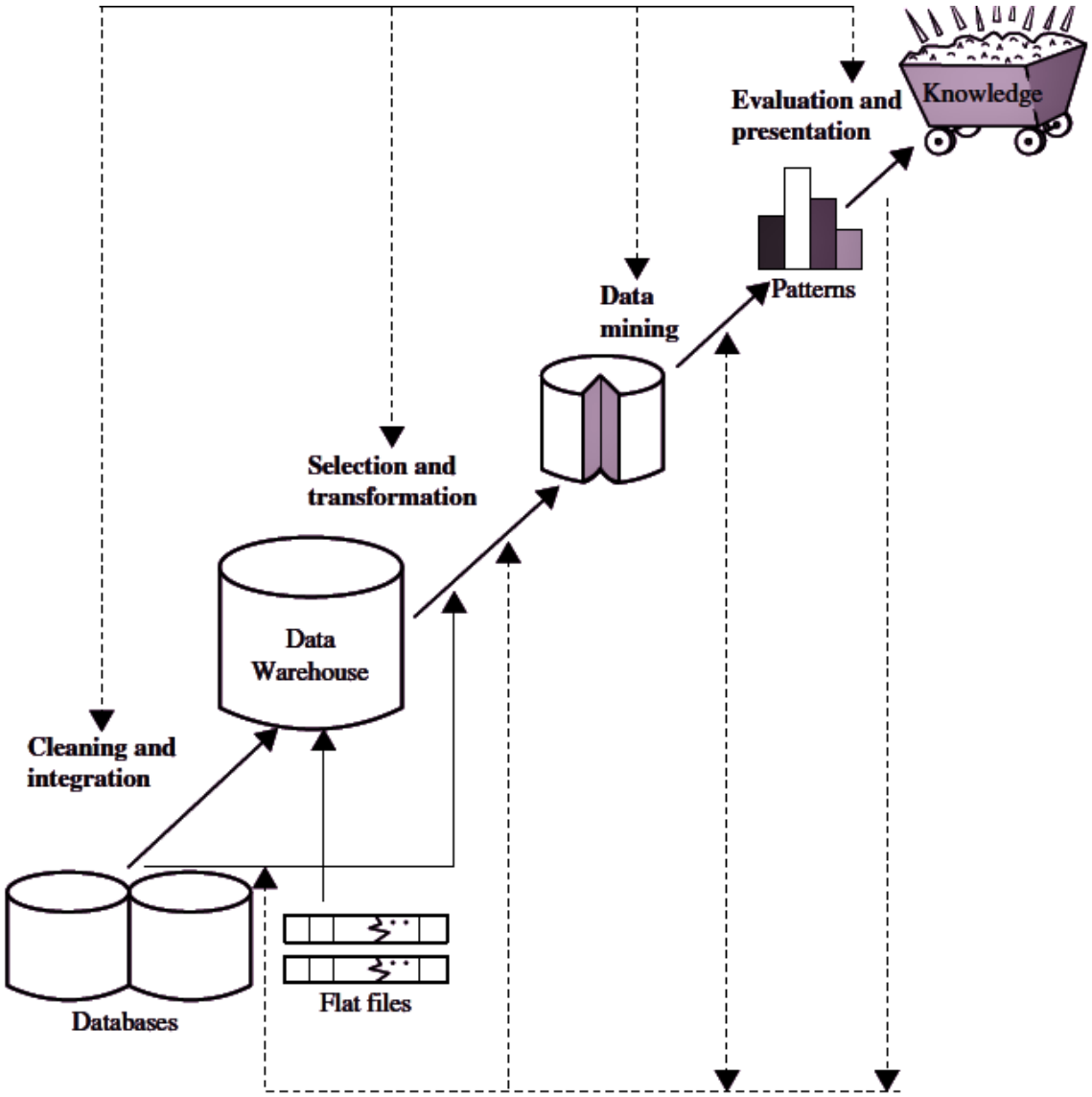
Outlier detection identifies data points that are very different from the rest of the data. These unusual points, called outliers, can be spotted using statistical methods or by checking if they are far away from other data points.

8. Genetic Algorithm

Genetic algorithms are inspired by natural selection. They solve problems by evolving solutions over several generations. Each solution is like a "species," and the fittest solutions are kept and improved over time, simulating "survival of the fittest" to find the best solution to a problem.

Data Mining Knowledge Representation

Knowledge representation in data mining is the process of structuring and organizing extracted patterns and insights from data in a way that is understandable and usable by both humans and machines. It involves transforming raw data into a meaningful format, enabling data mining algorithms to identify patterns, relationships, and trends. Effective knowledge representation is crucial for transforming raw data into actionable insights and facilitating decision-making.



Key aspects of knowledge representation in data mining:

Structured Data: Knowledge representation aims to structure data in a way that is easily understood and utilized by data mining algorithms.

Types of Representations: Various methods are employed, including decision trees, association rules, classification models, and clustering results.

Decision Trees: These graphical representations help classify data by splitting it into branches based on attribute values.

Association Rules: These rules identify relationships between items in a dataset, such as "If a customer buys milk, they are also likely to buy bread".

Classification Models: These models categorize data into predefined classes based on learned patterns.

Clustering: This technique groups similar data points together, revealing hidden structures in the data.

Visualization: Visual representations like parallel coordinates, Chernoff faces, and other icon-based techniques help in understanding data distributions and patterns.

Why is knowledge representation important?

Actionable Insights: It transforms raw data into meaningful information that can be used for decision-making.

Improved Understanding: It facilitates the discovery of patterns, relationships, and trends within large datasets.

Effective Decision Making: It enables users to make informed decisions based on the extracted knowledge.

Facilitates Communication: It provides a clear and concise way to communicate findings to both technical and non-technical audiences.