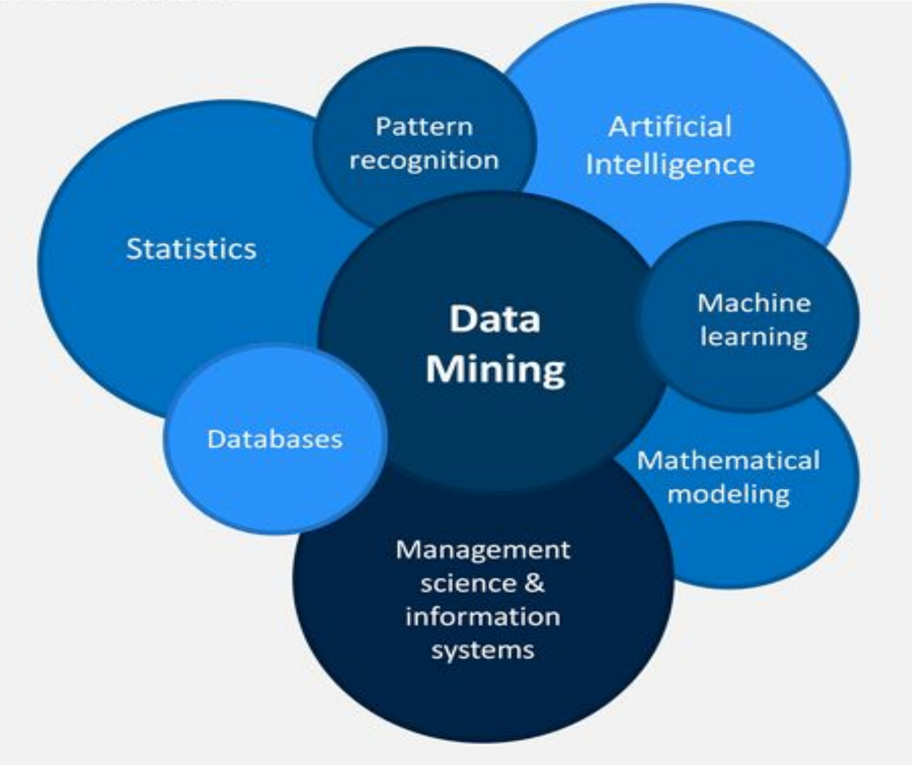
Data Mining & Big Data Analytics

UNIT - I: Chapter 1 – INTRODUCTION TO DATA MINING

Dr.K.YEMUNARANE
Department of Computer Science
KGiSL Institute of Information Management
Yemunarane.k@kgisliim.ac.in

Contents

- ❑ Introduction to Data Mining
- ❑ What is Data Mining?
- ❑ History of Data Mining
- ❑ Why Data Mining?
- ❑ Data Mining and its process
- ❑ Uses of Data Mining



Content 1

Introduction to Data Mining

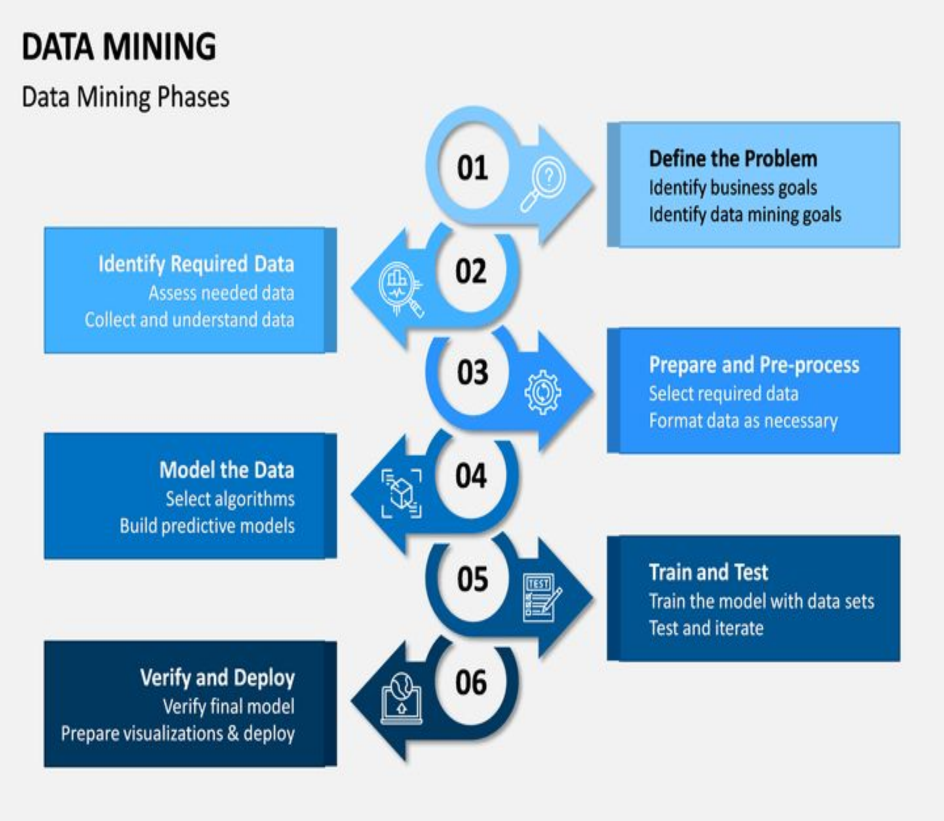
Content 1

Introduction to Data Mining

What is Data Mining?

- **Data Mining** is a set of method that applies to large and complex databases.

- Eliminate the randomness and discover the hidden pattern.



DATA MINING

Data Mining Phases

01 Define the Problem
Identify business goals
Identify data mining goals

02 Identify Required Data
Assess needed data
Collect and understand data

03 Prepare and Pre-process
Select required data
Format data as necessary

04 Model the Data
Select algorithms
Build predictive models

05 Train and Test
Train the model with data sets
Test and iterate

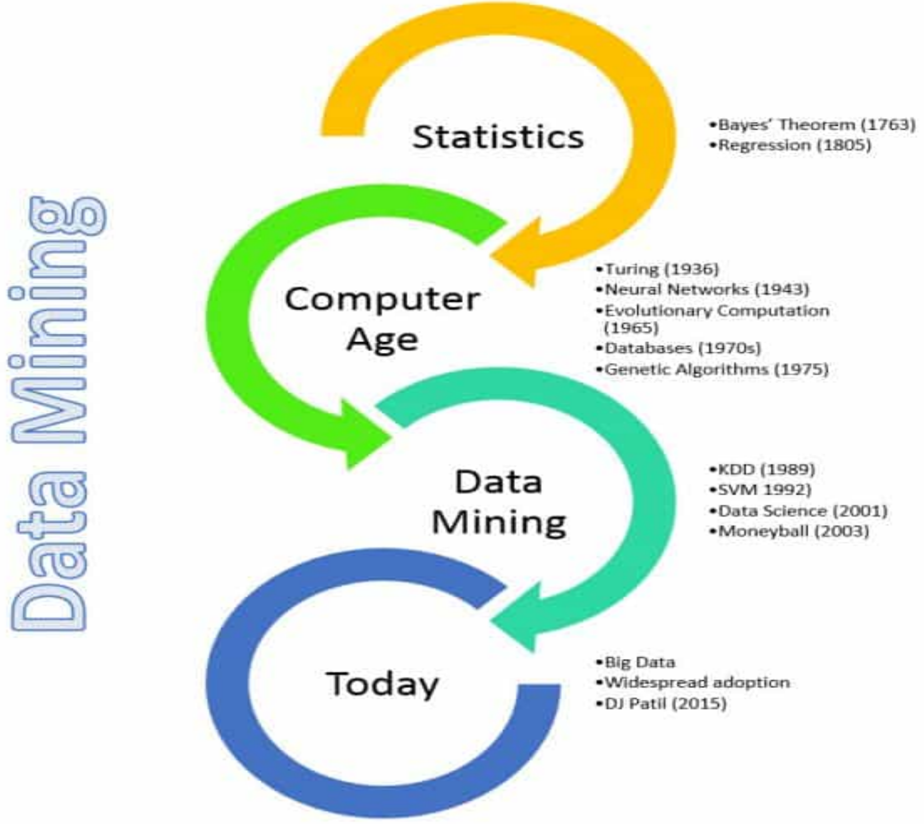
06 Verify and Deploy
Verify final model
Prepare visualizations & deploy

Content 1

Data Mining History

- In 1960s statisticians used the terms “Data Fishing” or “Data Dredging”.

- “Data Mining” appeared around 1990 in the database community.



Data Mining

Statistics

- Bayes’ Theorem (1763)
- Regression (1805)

Computer Age

- Turing (1936)
- Neural Networks (1943)
- Evolutionary Computation (1965)
- Databases (1970s)
- Genetic Algorithms (1975)

Data Mining

- KDD (1989)
- SVM 1992)
- Data Science (2001)
- Moneyball (2003)

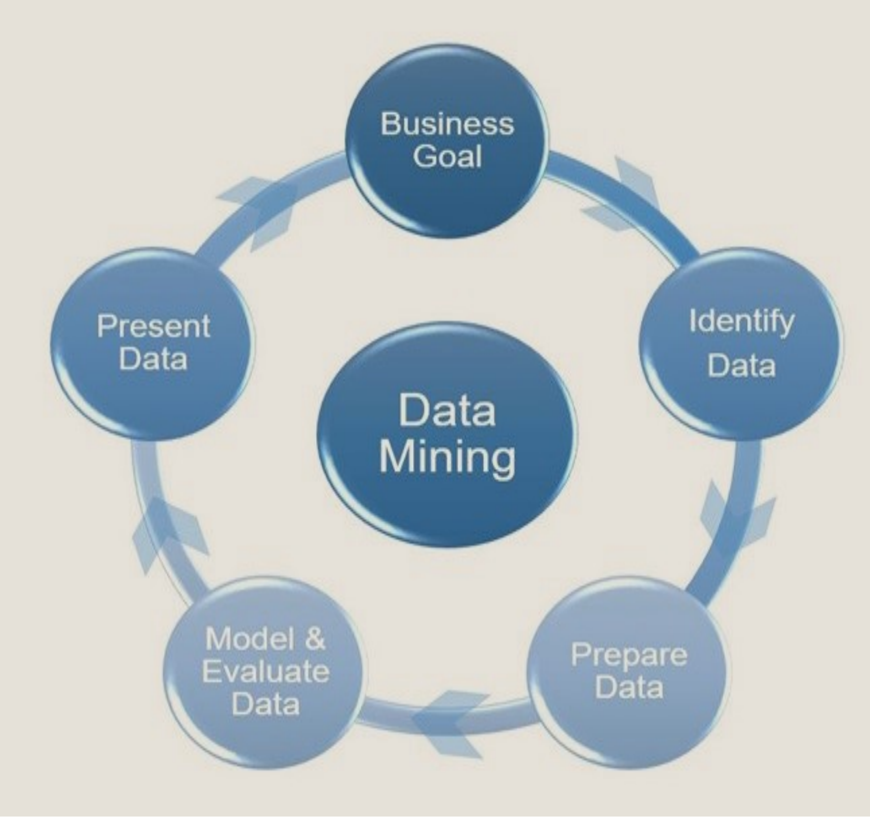
Today

- Big Data
- Widespread adoption
- DJ Patil (2015)

Content 1

Why do we need Data Mining?

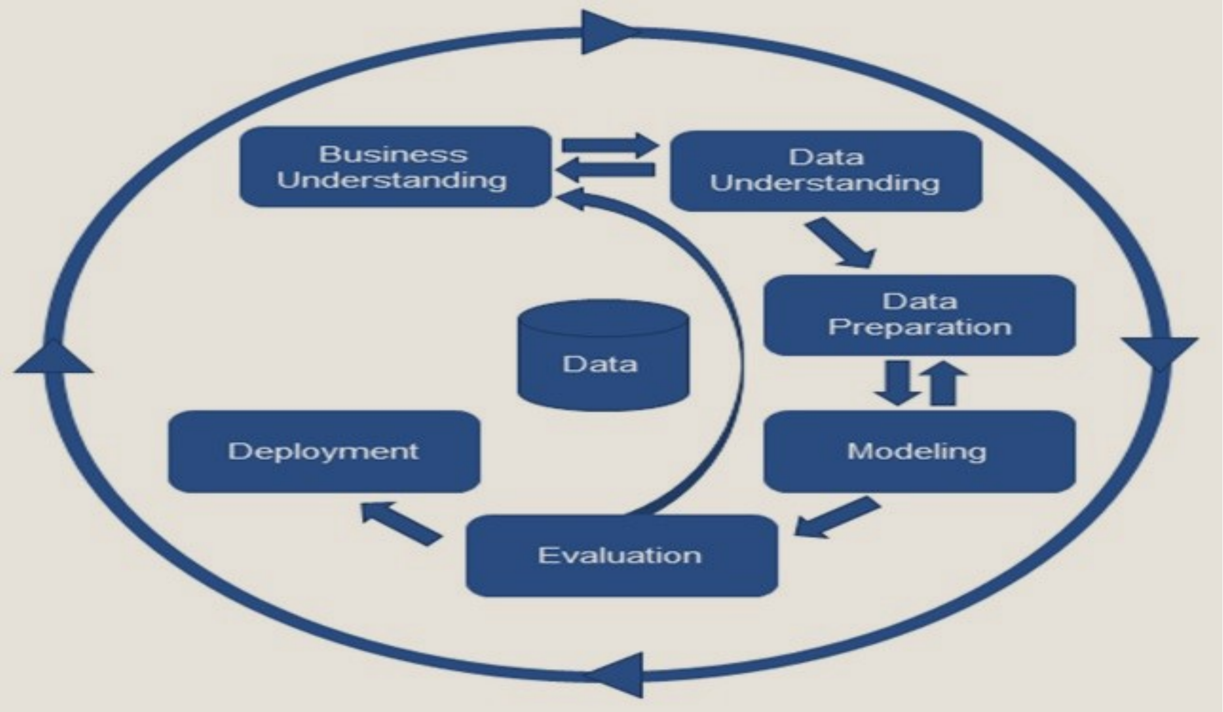
- ❑ Data mining **helps to develop smart market decision, run accurate campaigns, make predictions**, and more.
- ❑ With the help of Data mining, we can **analyze customer behaviors and their insights.** This leads to great success and data-driven business.
- ❑ Data mining is the procedure **of capturing large sets of data in order to identify the insights and visions of that data.**
- ❑ This helps the business to **take accurate and better decisions in an organization.**



Content 1

Data Mining and its process

- ❑ Requirement gathering
- ❑ Data exploration
- ❑ Data preparations
- ❑ Modeling
- ❑ Evaluation
- ❑ Deployment



Content 1

Uses of Data Mining

Data mining services can be used for the following functions

Research and surveys.

 Information collection

 Customer opinions

 Data scanning

 Extraction of information

 Pre-processing of data

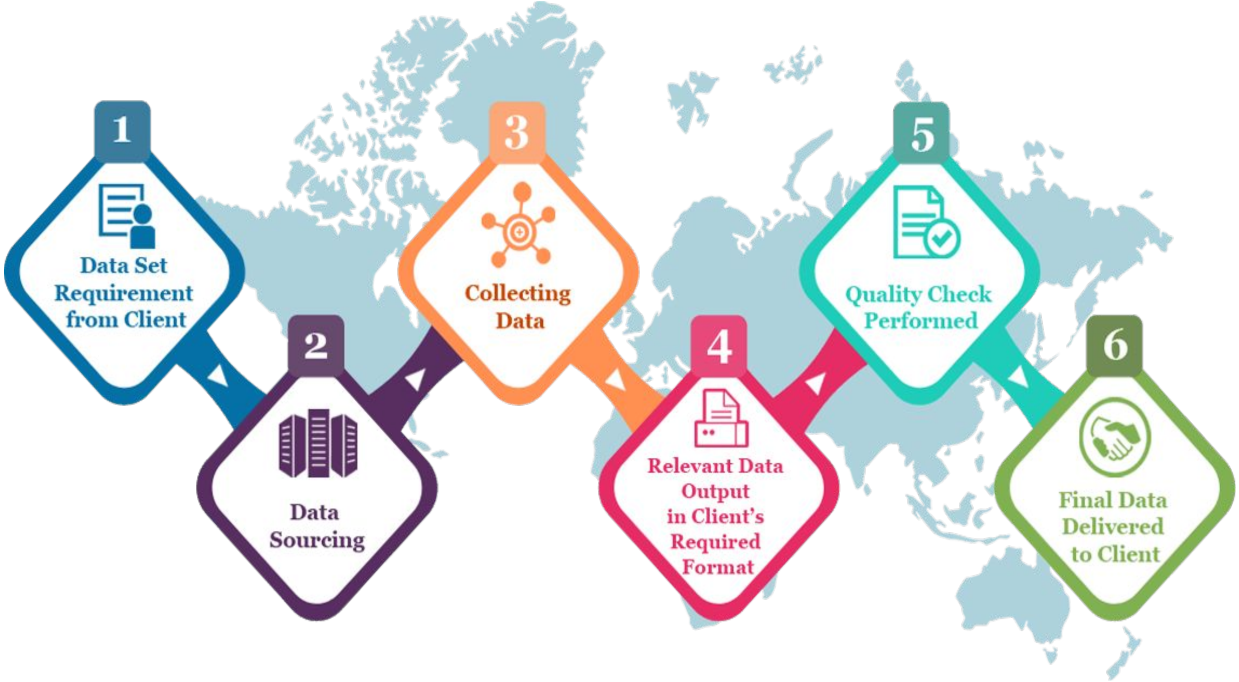
 Web data

 Competitor analysis

 Online research

 News

 Updating data


1
Data Set Requirement from Client
2
Data Sourcing
3
Collecting Data
4
Relevant Data Output in Client's Required Format
5
Quality Check Performed
6
Final Data Delivered to Client

Content 2

TYPES OF DATA MINING

Content 2

Types of Data Mining

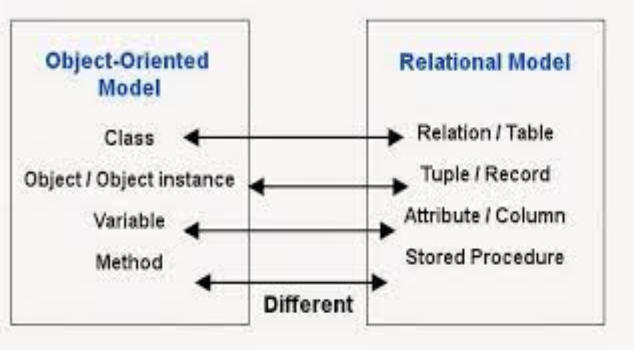
Relational Database

Data warehouses

Data Repositories

Object-Relational Database

Transactional Database



Object-Oriented Model

Relational Model

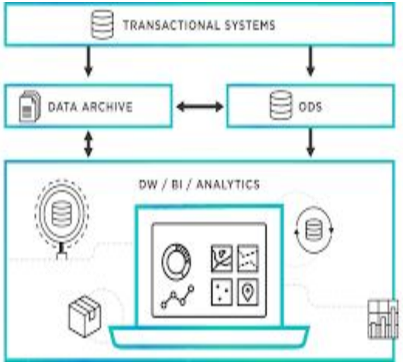
Class ⟷ Relation / Table

Object / Object instance ⟷ Tuple / Record

Variable ⟷ Attribute / Column

Method ⟷ Stored Procedure

Different



TRANSACTIONAL SYSTEMS

DATA ARCHIVE ⟷ ODS

DW / BI / ANALYTICS

Content 2

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 search ability, reporting, and organization.

Content 2

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 Source
Data Warehouse
Analysis

Content 2

Data Repositories

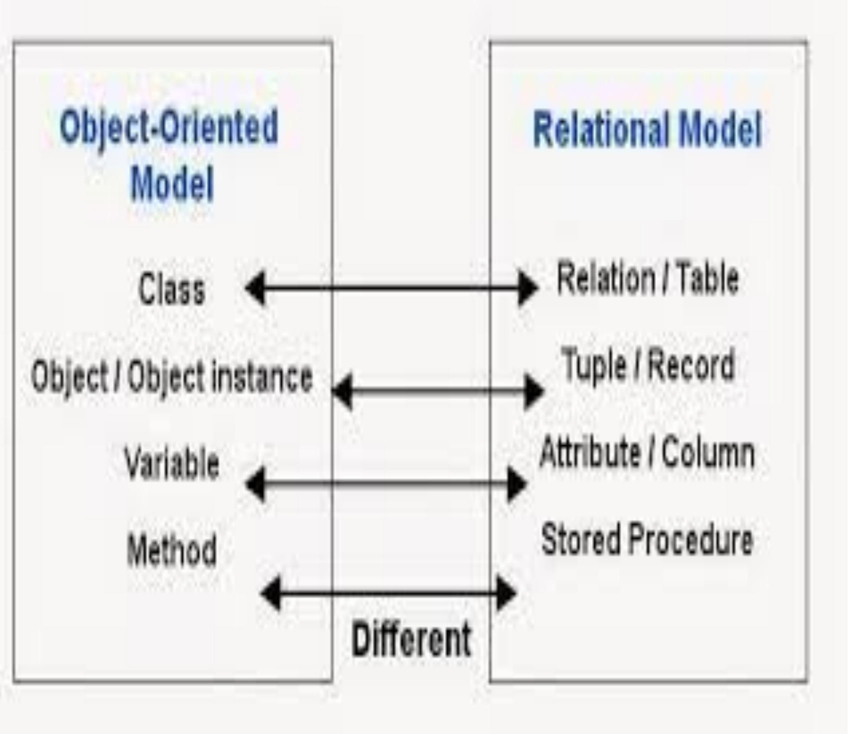
The Data Repository generally refers to a **destination for data storage.** However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. **For example**, a group of databases, where an organization has kept various kinds of information.

Content 2

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.

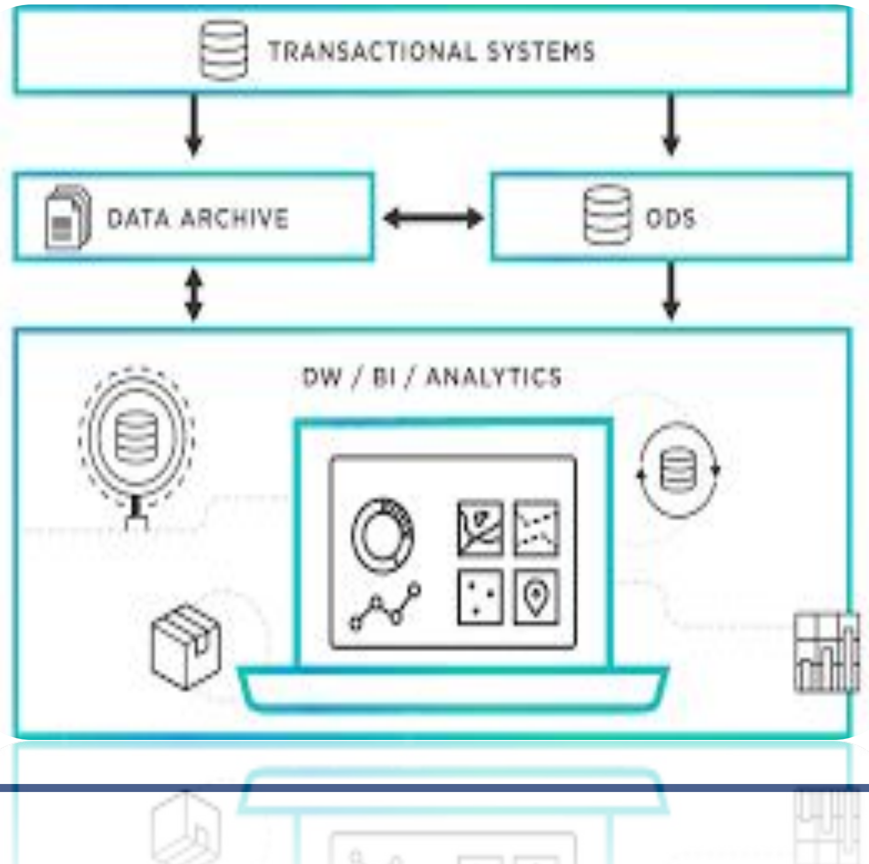
One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.



Content 2

Transactional Database

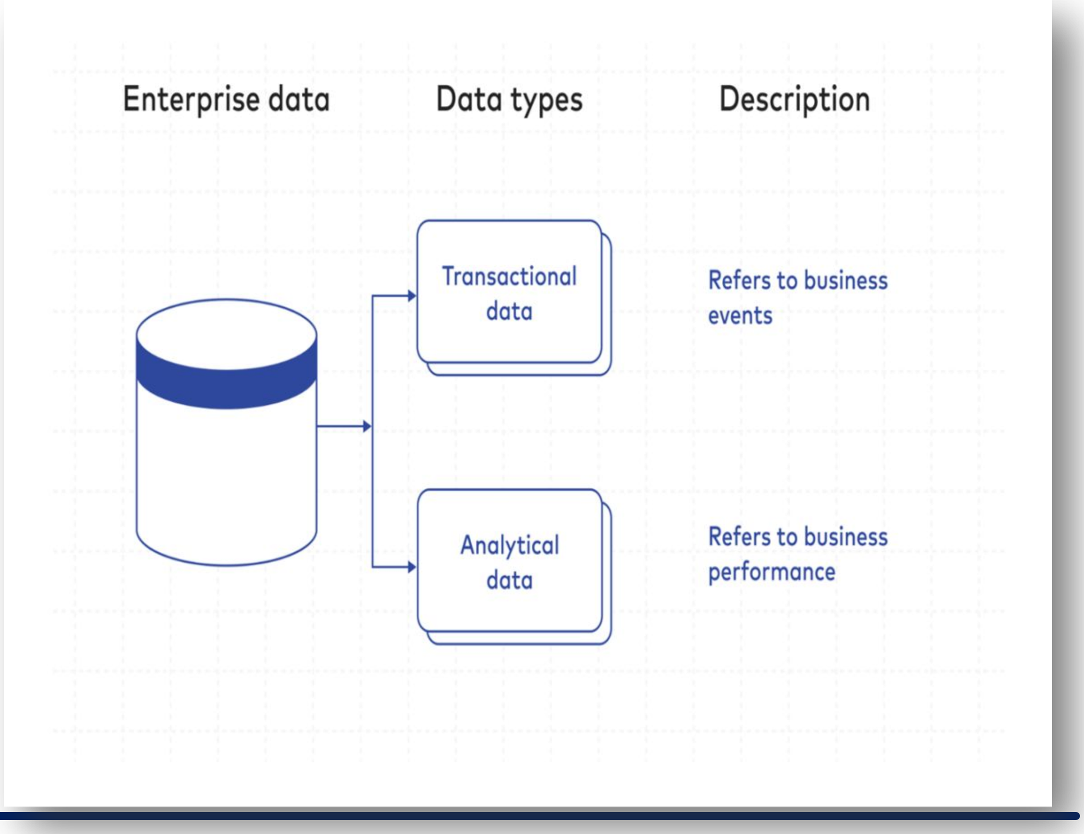
A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database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



Content 2

- **Transaction** refers to the usage of a database. **Relational** refers to the way in which a given database stores data.

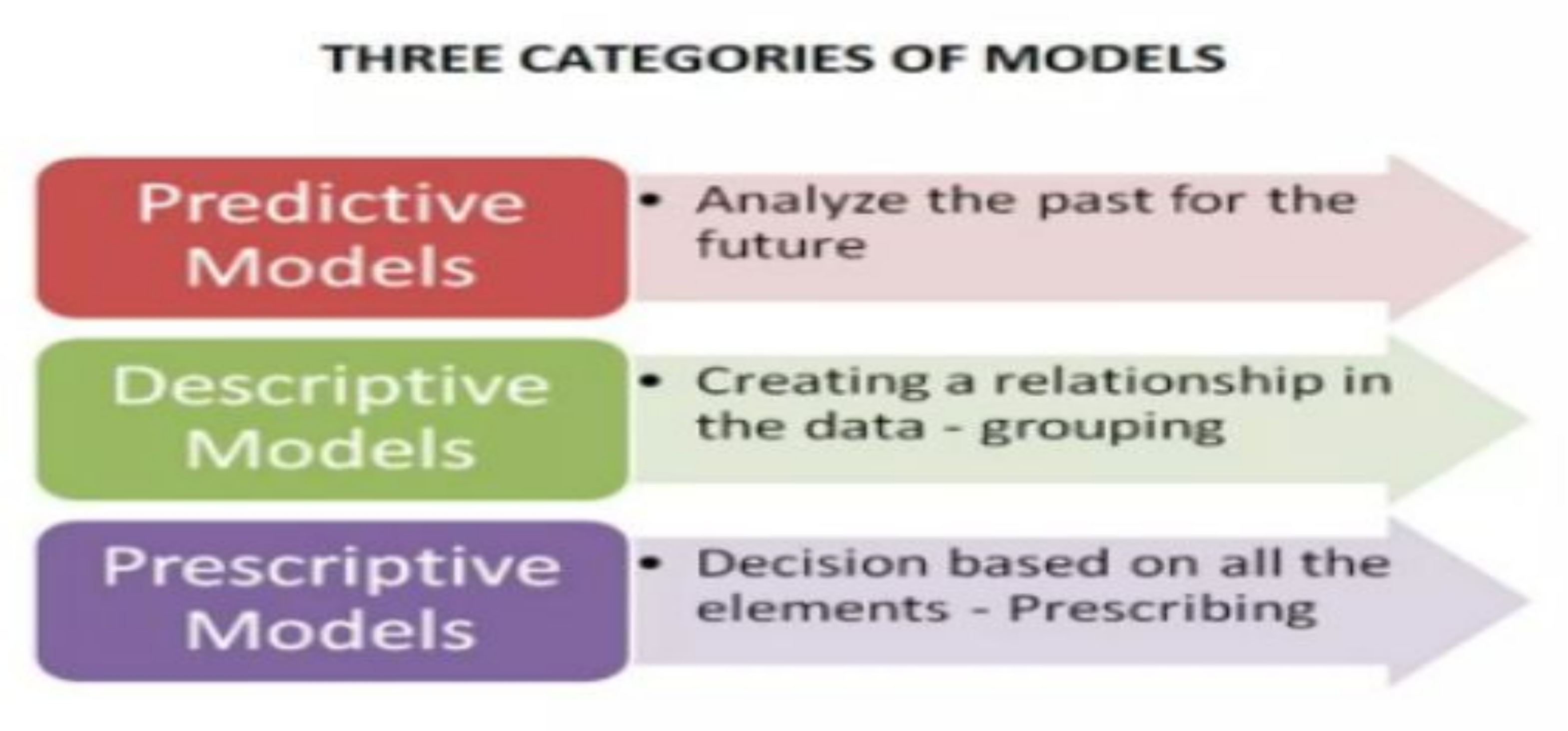
- The three major transactional databases include **CRM (customer relationship management), HRM (human resources management),** and **ERP (enterprise resource planning)**.



Content 3

DATA MINING FUNCTIONALITIES

Content 3



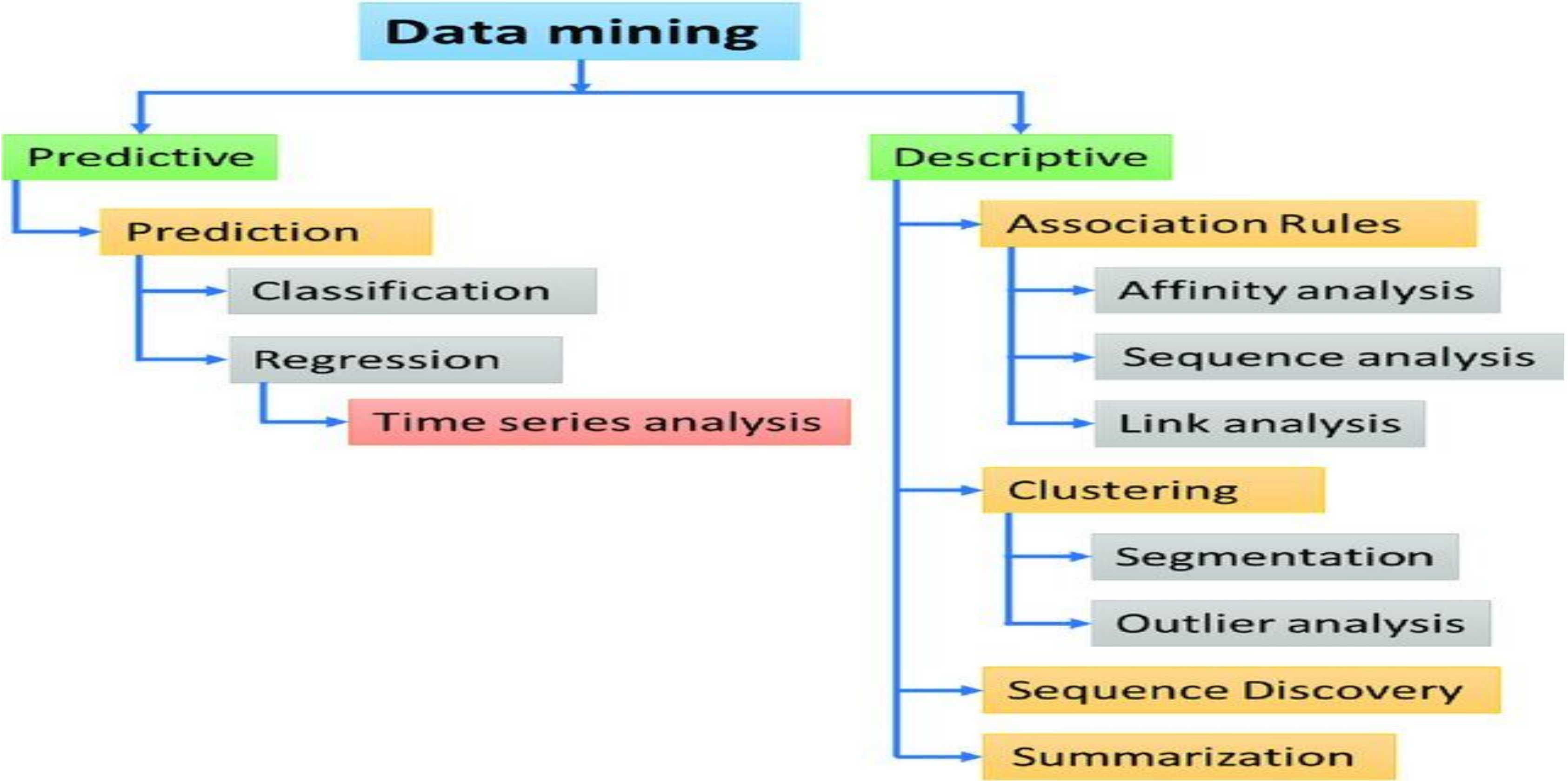
THREE CATEGORIES OF MODELS

Predictive Models
- Analyze the past for the future

Descriptive Models
- Creating a relationship in the data - grouping

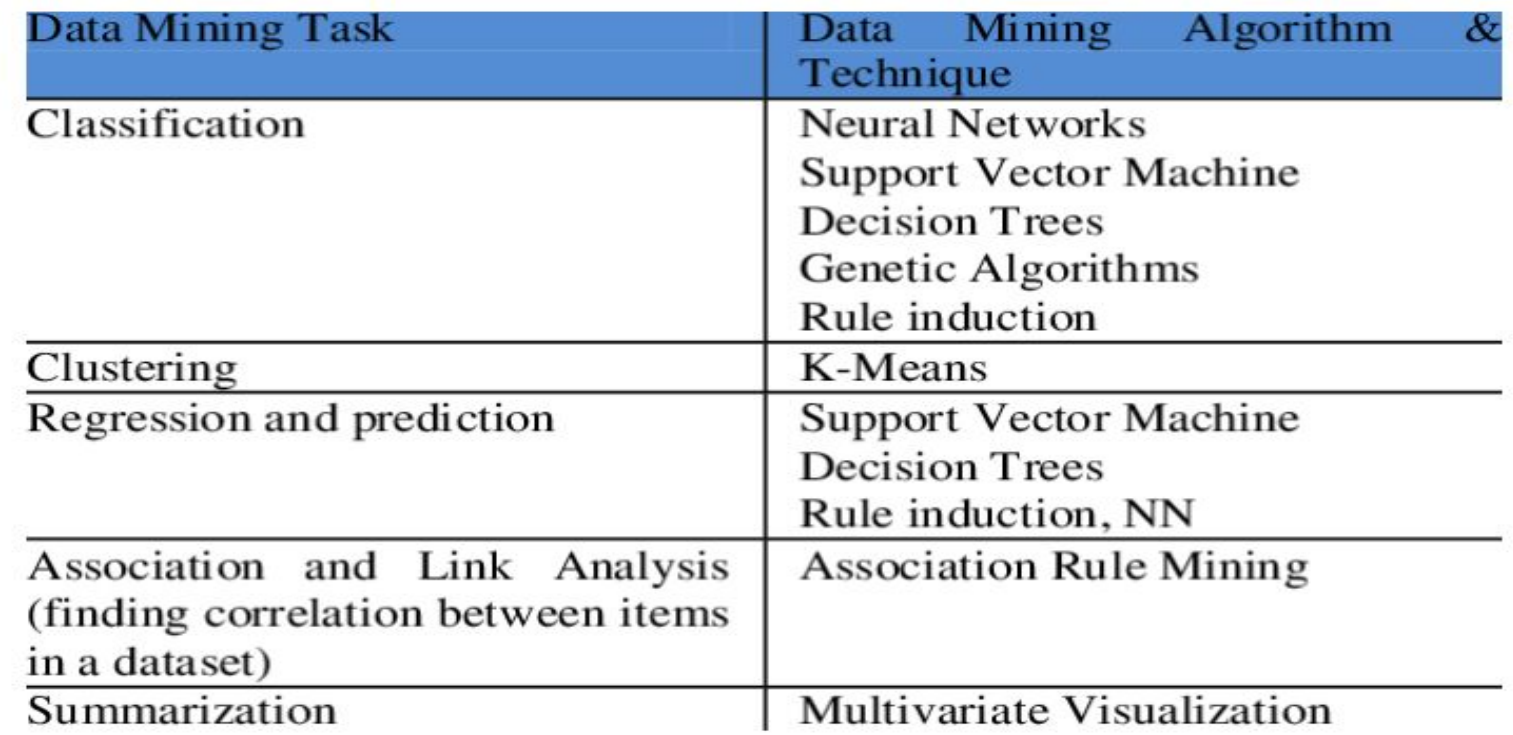
Prescriptive Models
- Decision based on all the elements - Prescribing

Content 3


Data mining
Predictive
Prediction
Classification
Regression
Time series analysis
Descriptive
Association Rules
Affinity analysis
Sequence analysis
Link analysis
Clustering
Segmentation
Outlier analysis
Sequence Discovery
Summarization

Content 3

DATA MINING TASKS & TECHNIQUES



Data Mining Task	Data Mining Algorithm & Technique
Classification	Neural Networks Support Vector Machine Decision Trees Genetic Algorithms Rule induction
Clustering	K-Means
Regression and prediction	Support Vector Machine Decision Trees Rule induction, NN
Association and Link Analysis (finding correlation between items in a dataset)	Association Rule Mining
Summarization	Multivariate Visualization

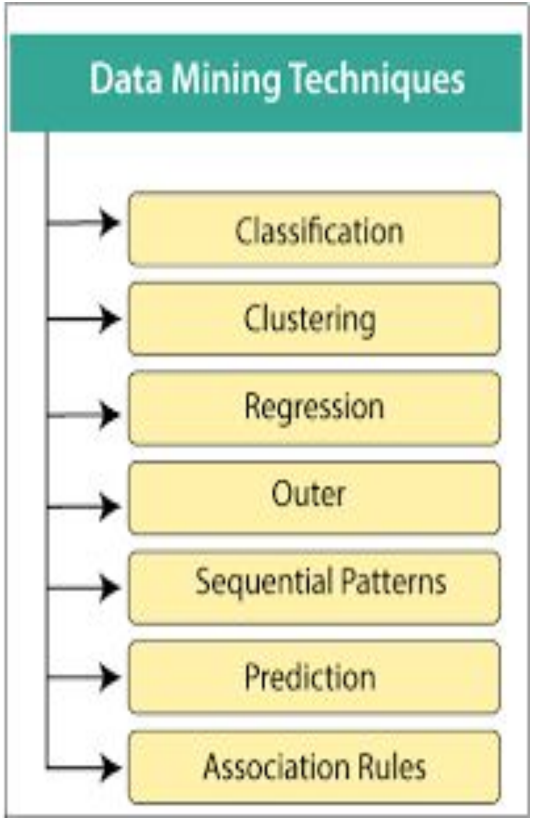
Content 3

DATA MINING FUNCTIONALITIES

There are a number of data mining functionalities that the organized and scientific methods offer.

Let us look at a few major ones:

- Classification
- Association Analysis
- Cluster Analysis
- Prediction
- Outlier Analysis
- Evolution & Deviation Analysis
- Mining Frequent patterns
- Correlation analysis
- Concept/Class Description: Characterization and Discrimination



Content 3

Association Analysis

It analyses the set of items that generally occur together in a transactional dataset.

There are two parameters that are used for determining the association rules –

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

Content 3

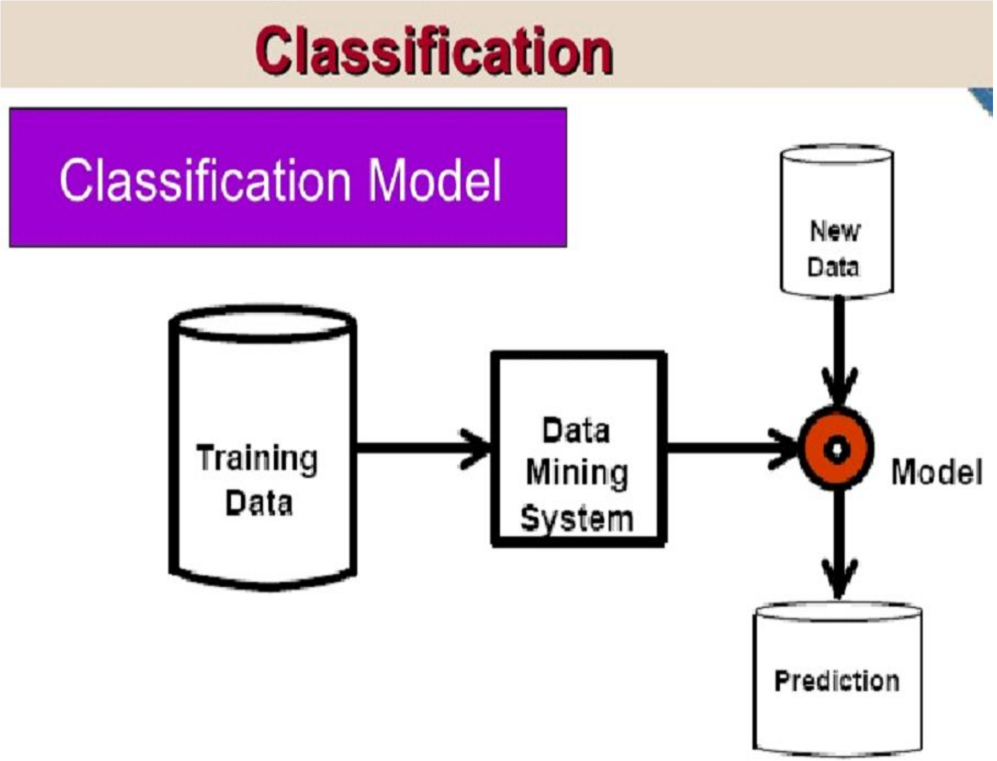
Clustering

It is similar to classification but the classes are not predefined. The classes are represented by data attributes. It is unsupervised learning. The objects are clustered or grouped, depends on the principle of maximizing the intraclass similarity and minimizing the intraclass similarity.

Content 3

Classification

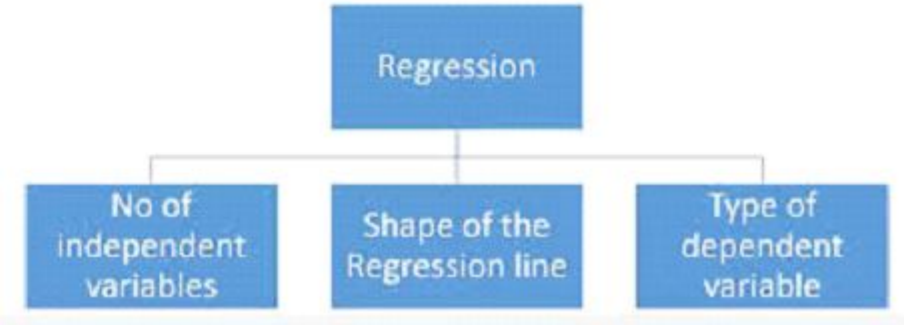
Classification is the procedure of discovering a model that represents and distinguishes data classes or concepts, for the objective of being able to use the model to predict the class of objects whose class label is anonymous. The derived model is established on the analysis of a set of training data (i.e., data objects whose class label is common).



Content 3

Regression

Regression is a technique of data mining which analyze the the relationship between variables. It creates predictive models. Regression technique can analyze and predict the results based on previously known data by applying formulas. Regression is very useful for finding the information on the basis of existing known information. Algorithms used for regression are Multivariate, Multiple Regression Algorithm,etc.



Regression

No of independent variables

Shape of the Regression line

Type of dependent variable

Content 3

Types of Regression

Regression is divided into five different types:

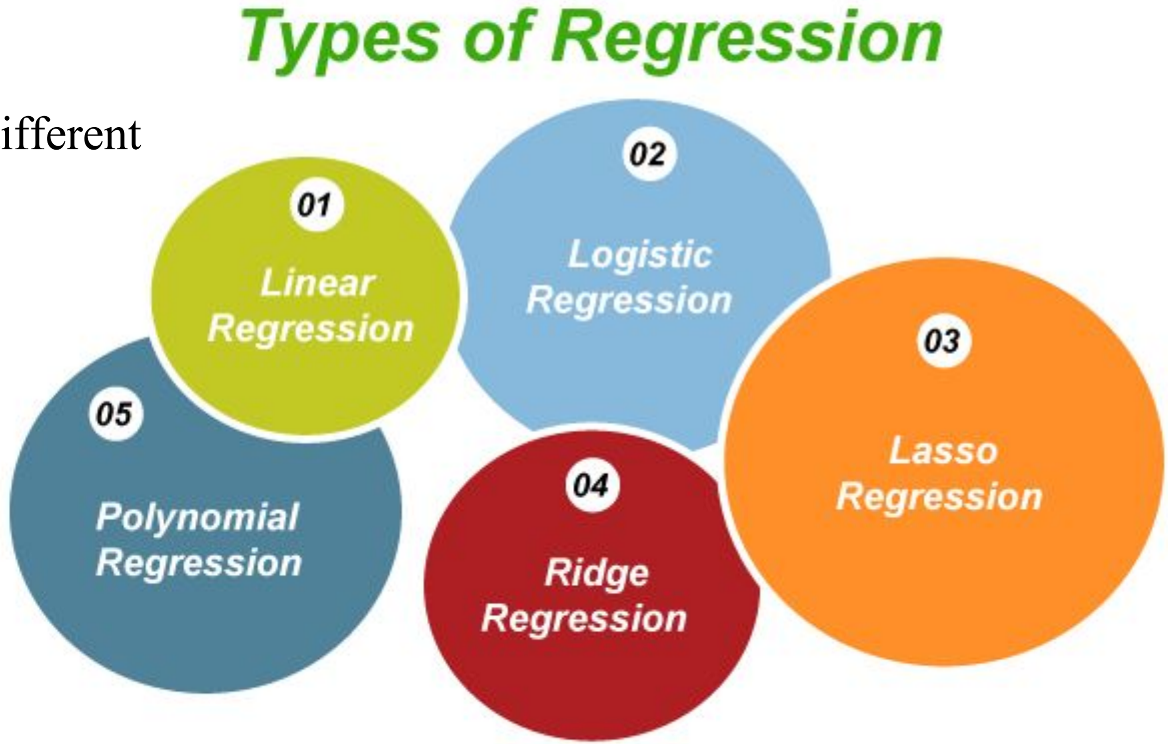
1.Linear Regression

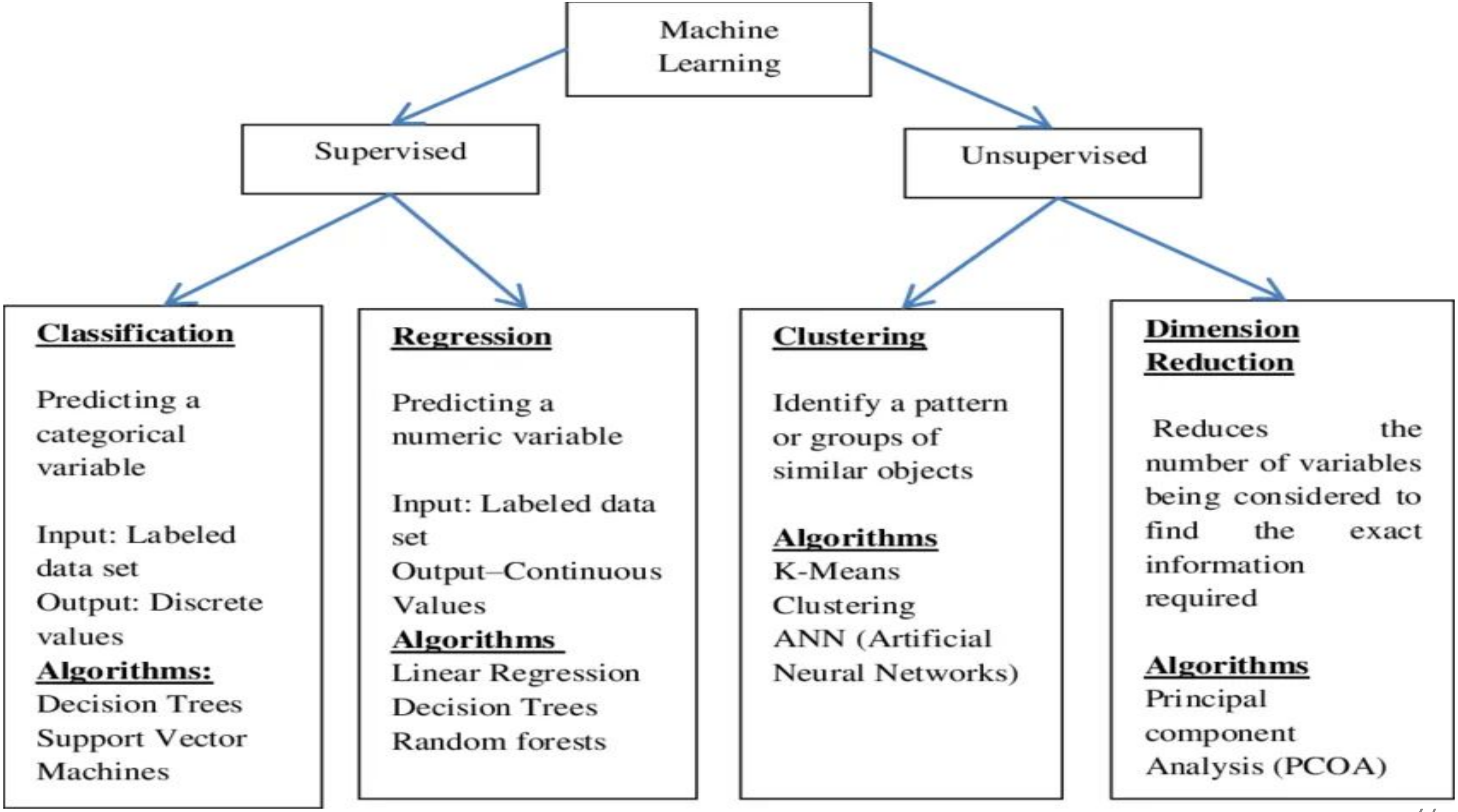
2.Logistic Regression

3.Lasso Regression

4.Ridge Regression

5.Polynomial Regression





Machine Learning

Supervised

Unsupervised

Classification

Predicting a categorical variable

Input: Labeled data set

Output: Discrete values

Algorithms:

Decision Trees

Support Vector Machines

Regression

Predicting a numeric variable

Input: Labeled data set

Output–Continuous Values

Algorithms

Linear Regression

Decision Trees

Random forests

Clustering

Identify a pattern or groups of similar objects

Algorithms

K-Means Clustering

ANN (Artificial Neural Networks)

Dimension Reduction

Reduces the number of variables being considered to find the exact information required

Algorithms

Principal component Analysis (PCOA)

Regression Model	Advantages	Disadvantages
Linear Regression	1. Works well irrespective of the dataset size. 2. Gives information about the relevance of features.	1. The assumptions of Linear Regression.
Polynomial Regression	1. Works on any size of the dataset. 2. Works very well on non-linear problems.	1. We need to choose the right polynomial degree for good bias/ variance tradeoff.
Support Vector Regression	1. Easily adaptable. 2. Works very well on non-linear problems. 3. Not biased by outliers (object that deviates significantly from the rest).	1. Compulsory to apply feature scaling. 2. Not well known. 3. Difficult to understand.
Decision Tree Regression	1. Interpretability. 2. Works well on both linear and non-linear problems. 3. No need to apply feature scaling.	1. Poor results on small datasets. 2. Overfitting can easily occur.
Random Forest Regression	1. Powerful. 2. Accurate. 3. Good performance on many problems including non-linear.	1. No interpretability. 2. Overfitting can easily occur. 3. We need to choose the number of trees.

Content 3

Linear Regression

Linear regression is the type of regression that forms a **relationship between the target variable** and one or more independent variables utilizing a straight line.

Logistic Regression

When the dependent **variable is binary in nature**, i.e., 0 and 1, true or false, success or failure, the **logistic regression technique comes into existence**. Here, the target value (Y) ranges from 0 to 1, and it is primarily used for classification-based problems. Unlike linear regression, **it does not need any independent and dependent variables** to have a linear relationship.

Content 3

Lasso Regression

The term LASSO stands for **Least Absolute Shrinkage and Selection Operator**. Lasso regression is a **linear type of regression that utilizes shrinkage**. In Lasso regression, all the data points are shrunk **towards a central point**, also known as the mean. The lasso process is most fitted for **simple and sparse models with fewer parameters** than other regression. This type of regression is well fitted for models that suffer from multicollinearity.

Ridge Regression

Ridge regression refers to a process that is used to analyze various regression data that have the issue of multicollinearity. Multicollinearity is the existence of a linear correlation between two independent variables.

Content 3

Polynomial Regression

If the power of the **independent variable is more than 1 in the regression equation**, it is termed a polynomial equation. With the help of the example given below, we will understand the concept of polynomial regression.

$$\mathbf{Y = a + b * x^2}$$

Content 4

Classification of Data Mining systems

Content 4

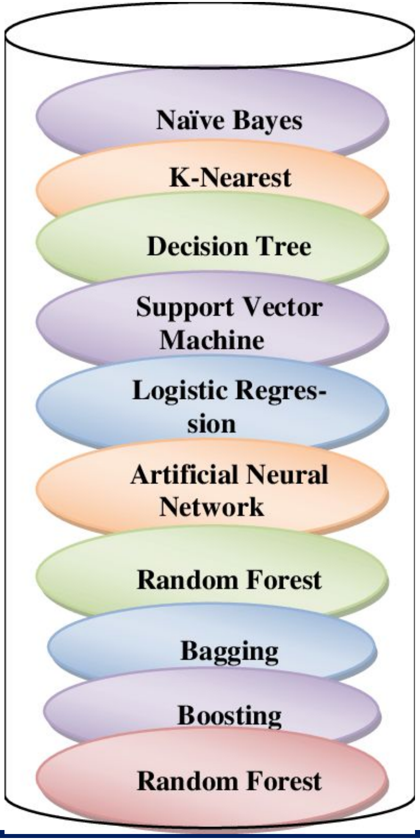
What is Classification in Data Mining?

Types of Classification Techniques in Data Mining

- Generative
- Discriminative

Classifiers in Machine Learning

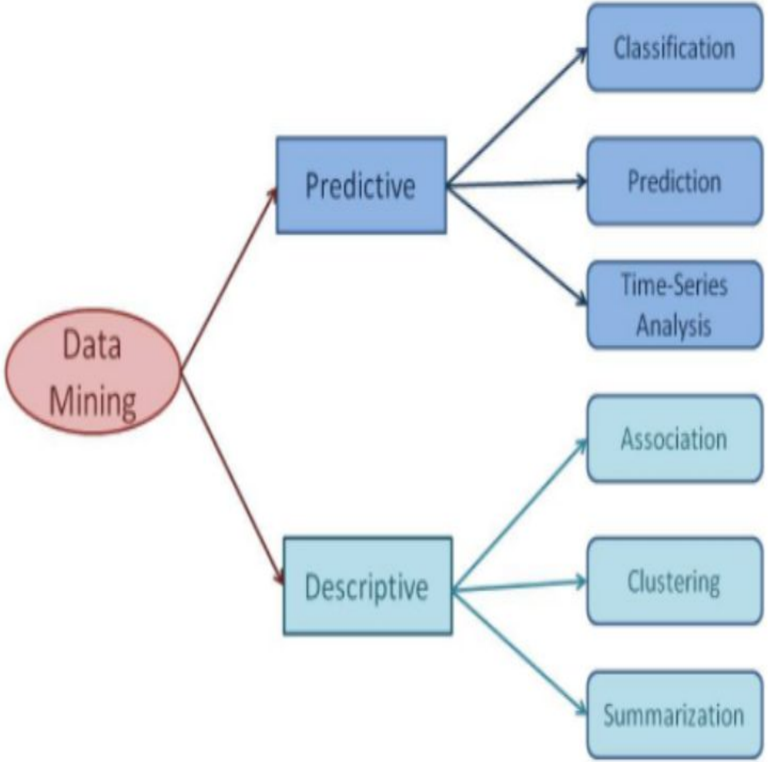
1. Logistic Regression
2. Linear Regression
3. Decision Trees
4. Random forest
5. Naive Bayes
6. Support Vector Machine
7. K-Nearest Neighbours



Content 4

What is Classification in Data Mining?

- Classification in data mining is a common technique that separates data points into different classes.
- It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple ones.
- It primarily involves using algorithms that you can easily modify to improve the data quality.



Content 4 Types of Classification Techniques in Data Mining

- Generative
- Discriminative

Generative

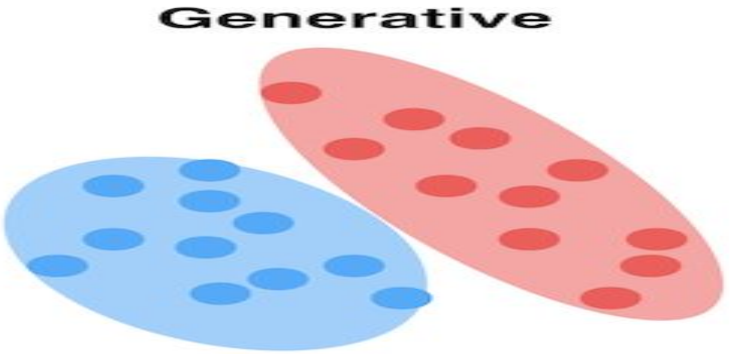
Generative models are models where the focus is **the distribution of individual classes in a dataset and the learning algorithms** tend to model the underlying patterns/distribution of the data points. These models use the intuition of joint probability in theory, creating instances where a given feature (*x*)/input and the desired output/label (*y*) exist at the same time.

Examples of machine learning generative models

Naive Bayes (and generally Bayesian networks)

Hidden Markov model

Linear discriminant analysis (LDA), a dimensionality reduction technique



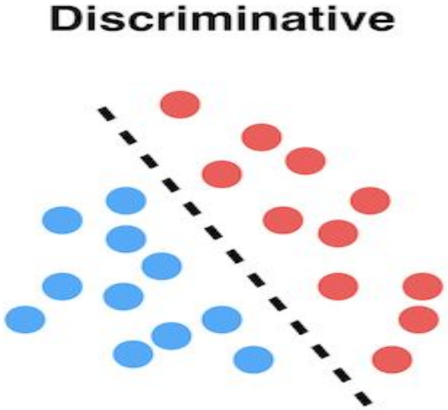
Content 4

Discriminative Models

Discriminative models, also called *conditional models*, tend to learn the boundary between classes/labels in a dataset. Unlike generative models, the goal here is to find the *decision boundary* separating one class from another.

"Another key difference between these two types of models is that while a generative model focuses on explaining how the data was generated, a discriminative model focuses on predicting labels of the data."

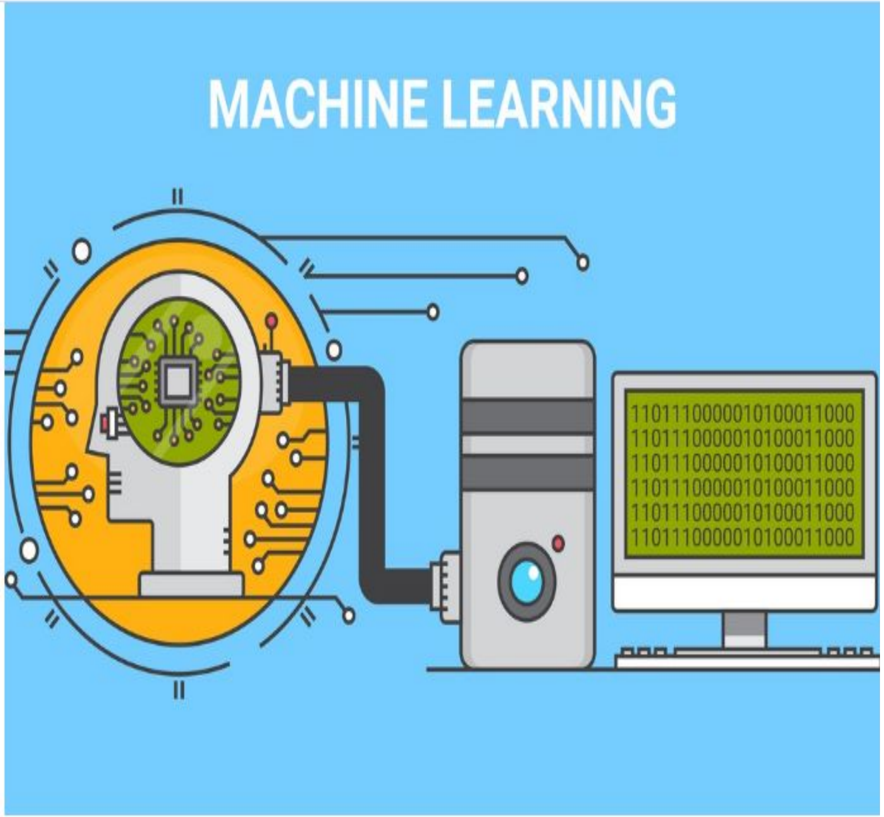
Examples of discriminative models in machine learning are:
Logistic regression
Support vector machine
Decision tree
Random forest



Content 4

Classifiers in Machine Learning

1. Logistic Regression
2. Linear Regression
3. Decision Trees
4. Random forest
5. Naive Bayes
6. Support Vector Machine
7. K-Nearest Neighbours

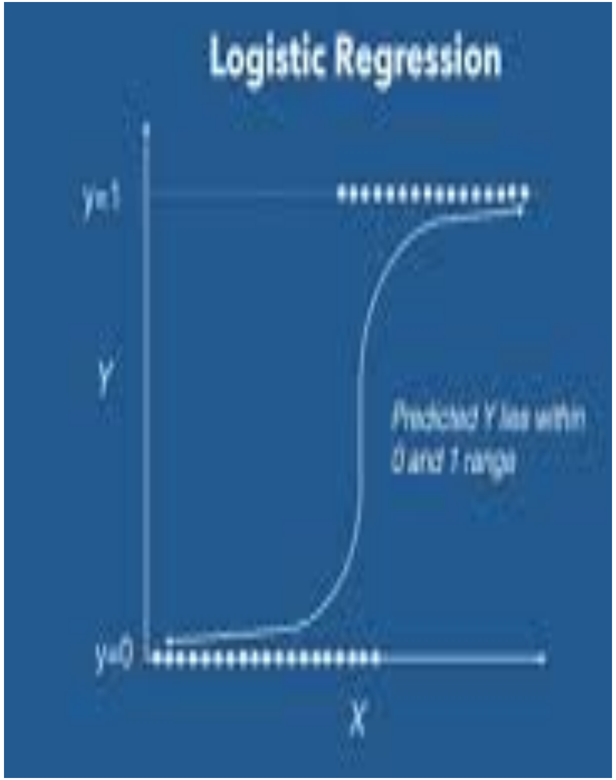


Content 4

Logistic Regression

Logistic regression is a **statistical analysis method to predict a binary outcome**, such as yes or no, based on prior observations of a data set.

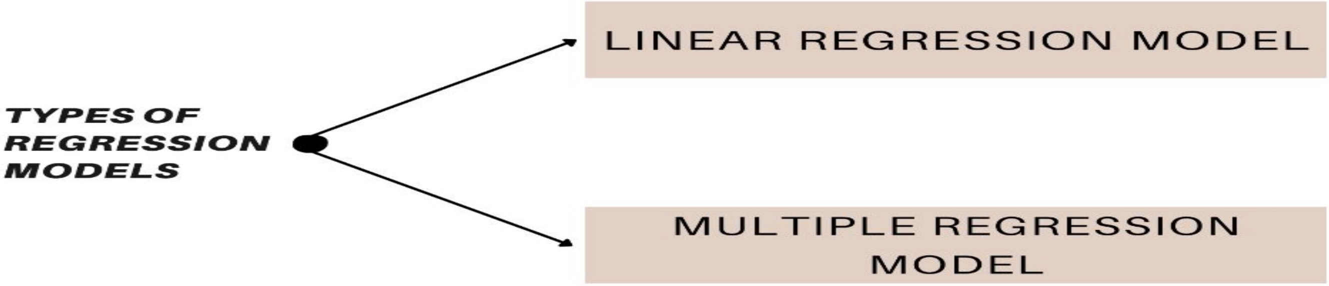
A logistic regression model predicts a **dependent data variable by analyzing the relationship between one or more existing independent variables.** For example, a logistic regression could be used to predict whether a political candidate will win or lose an election



Content 4

Two types of Regression can be observed in data mining. Those two types are given below:

- Linear Regression Model
- Multiple Regression Model



TYPES OF REGRESSION MODELS

LINEAR REGRESSION MODEL

MULTIPLE REGRESSION MODEL

Content 4

Linear regression

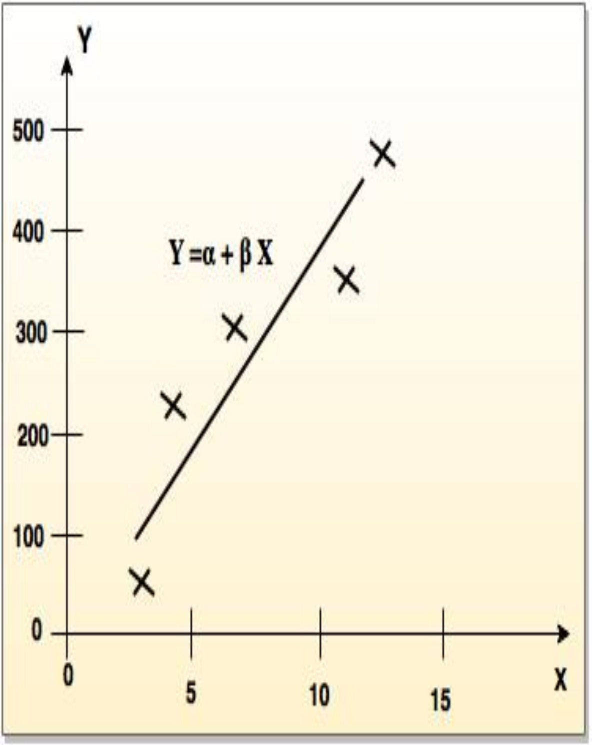
- It is simplest form of regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observe the data.
- Linear regression attempts to find the mathematical relationship between variables.
- If outcome is straight line then it is considered as linear model and if it is curved line, then it is a non linear model.

The relationship between dependent variable is given by straight line and it has only one independent variable.

$$\mathbf{Y = \alpha + B\ X}$$

Model **'Y'**, is a linear function of **'X'**.

The value of 'Y' increases or decreases in linear manner according to which the value of 'X' also changes.



Linear Regression

Content 4

Multiple Regression

Multiple Regression Model is generally used to explain the relationship between multiple independent or multiple predictor variables.

It can be considered as one of the most popular models for predictions in data mining.

In general, it uses two or more than two independent variables to predict an outcome for the users.

The formula that is used in the multiple regression model is given below:

$$\mathbf{Y = a_0 + a_1x1 + a_2x2 + a_3x3 + a_4x4 + \ldots\ldots. + a_kx_k + e}$$

Where, Y is the response variable (the values that have to be predicted).

X1+X2+X3+X4+Xk are the independent predictors.

e is the random error in the above formula.

A0, A1, A3, A4, Ak are the regression coefficients.

Content 4

Content 4

Regression	Classification
Regression refers to a type of **supervised machine** learning technique that is used to predict any continuous-valued attribute.	Classification refers to a process of assigning predefined class labels to instances based on their attributes**.** **(supervised machine learning technique)**
In regression, the nature of the predicted data is ordered.	In classification, the nature of the predicated data is unordered.
The regression can be further divided into linear regression and non-linear regression.	Classification is divided into two categories: binary classifier and multi-class classifier.
In the regression process, the calculations are basically done by utilizing the **root mean square error**.	In the classification process, the calculations are basically done by **measuring the efficiency**.
Examples of regressions are regression tree, linear regression, etc.	The examples of classifications are the decision tree.

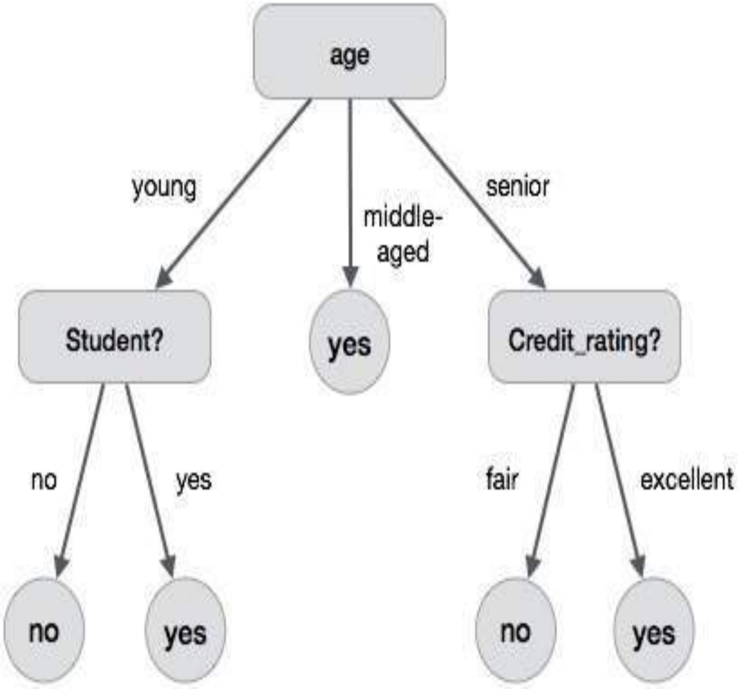
Content 4

Decision Tree

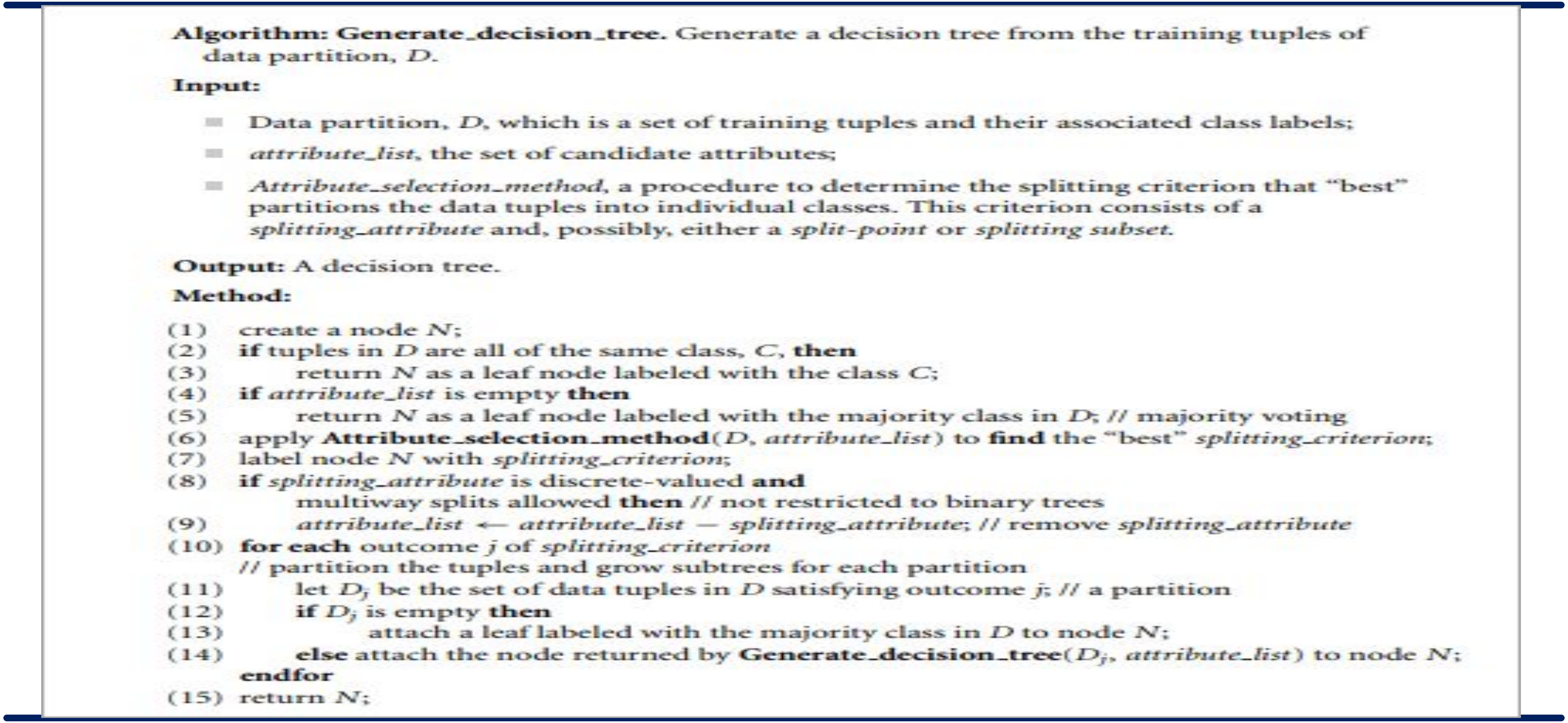
A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

The benefits of having a decision tree are as follows :

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.



Content 4



Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition, *D*.

Input:

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split-point* or *splitting subset*.

Output: A decision tree.

Method:

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C, then
(3)        return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)        return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
           multiway splits allowed then // not restricted to binary trees
(9)        attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)       let D_j be the set of data tuples in D satisfying outcome j; // a partition
(12)       if D_j is empty then
(13)             attach a leaf labeled with the majority class in D to node N;
(14)       else attach the node returned by Generate_decision_tree(D_j, attribute_list) to node N;
     endfor
(15) return N;
```

Content 4

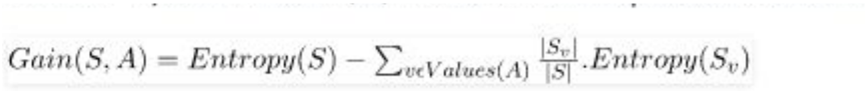
In Decision Tree the major challenge is to identification of the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:

Information Gain
Gini Index

Information Gain

we use a node in a decision tree to partition the training instances into smaller subsets the entropy changes. Information gain is a measure of this change in entropy.

Definition: Suppose S is a set of instances, A is an attribute, S_v is the subset of S with A = v, and Values (A) is the set of all possible values of A, then



$$Gain(S,A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|}.Entropy(S_v)$$

Content 4

Entropy

Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy more the information content.

Definition: Suppose S is a set of instances, A is an attribute, S_v is the subset of S with A = v, and Values (A) is the set of all possible values of A, then

Gini Index

- Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.
- It means an attribute with lower Gini index should be preferred.
- Sk learn supports “Gini” criteria for Gini Index and by default, it takes “gini” value.
- The Formula for the calculation of the of the Gini Index is given below.

Content 4

Tree Pruning

Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Tree Pruning Approaches

There are two approaches to prune a tree

Pre-pruning − The tree is pruned by halting its construction early.

Post-pruning - This approach removes a sub-tree from a fully grown tree.

Cost Complexity

The cost complexity is measured by the following two parameters

- Number of leaves in the tree, and
- Error rate of the tree.

Content 5

Major Issues in Data Mining.

Major Issues in Data Mining.

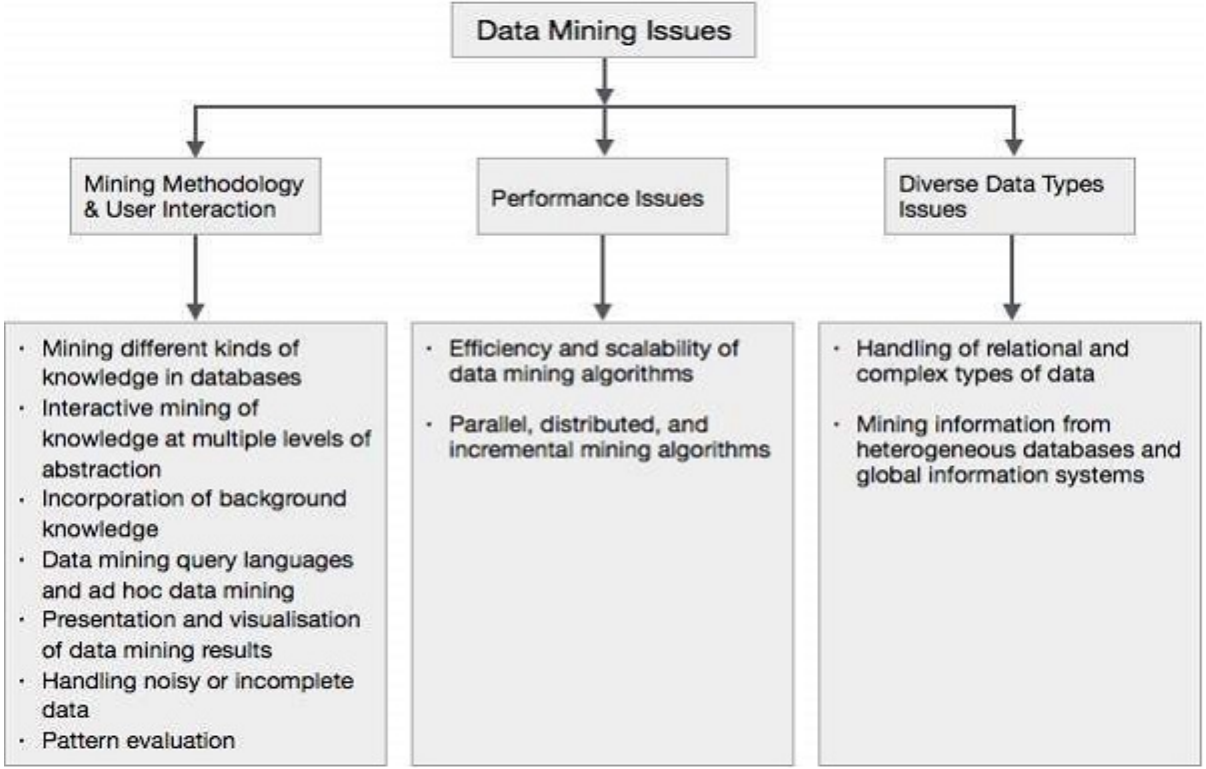
Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding :

Mining Methodology and User Interaction

Performance Issues
Diverse Data Types Issues

Content 5

The following diagram describes the major issues.



Data Mining Issues

Mining Methodology & User Interaction

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Presentation and visualisation of data mining results
- Handling noisy or incomplete data
- Pattern evaluation

Performance Issues

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, and incremental mining algorithms

Diverse Data Types Issues

- Handling of relational and complex types of data
- Mining information from heterogeneous databases and global information systems

Content 5

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.

Content 5

Mining Methodology and User Interaction Issues

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.

Content 5

Mining Methodology and User Interaction Issues

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.

Content 5

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.

Content 5

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.

End

Chapter 1 to 5

- Introduction
- Relational Databases
- Data Warehouses Transactional databases
- Data Mining functionalities
- Classification of Data Mining systems
- Major Issues in Data Mining.