BUSINESS INTELLIGENCE AND DATA ANALYTICS
– WE PREDICT THE FUTURE –

Institute of Analytics
Recognized Programme

University of Petra

307304
Business Intelligence and Data Mining

Chapter 1 – Introduction to Data Mining

https://www.javatpoint.com/implementation-process-of-data-mining

What is Data Mining?

- **Definition:** Data mining is the process of discovering patterns, correlations, trends, and useful information from large datasets, often involving methods from statistics, machine learning, and database systems.
- **Pattern Recognition:** Identifying meaningful patterns in data, such as associations, clusters, or classifications.

The Importance of Data Mining

- **Application Areas:**
  - **Business:** Customer segmentation, market basket analysis, fraud detection.
  - **Healthcare:** Disease prediction, patient risk analysis.
  - **Finance:** Credit scoring, stock market analysis.
  - **Retail:** Recommendation systems, inventory forecasting.
- **Why it Matters:**
  - Turning raw data into actionable insights.
  - Helping organizations make informed decisions.
  - Enabling predictive and prescriptive analytics.

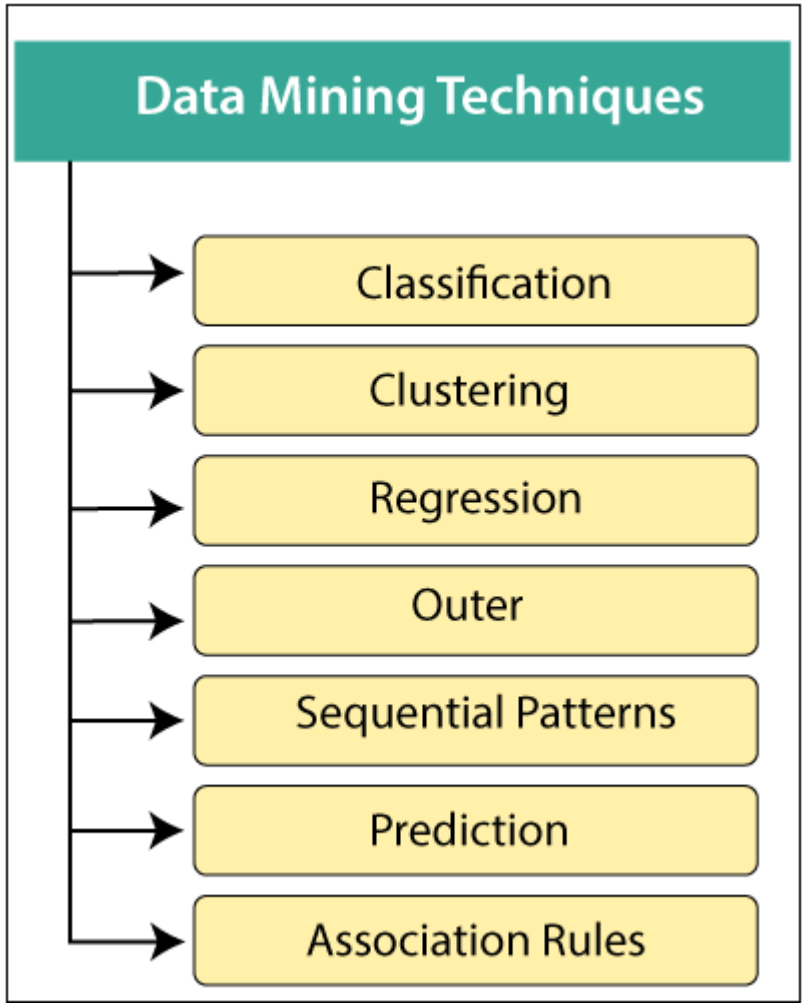
Types of Data Mining Tasks

1. Descriptive Tasks:

- **Clustering:** Grouping data into clusters based on similarity.
  - Example: Segmenting customers into different groups based on purchasing behavior.
- **Association Rule Mining:** Discovering relationships between variables.
  - Example: Market Basket Analysis—finding products often bought together.

2. Predictive Tasks:

- **Classification:** Assigning data to predefined categories.
  - Example: Email classification as spam or not.
- **Regression:** Predicting a continuous value.
  - Example: Predicting house prices based on features like size and location.

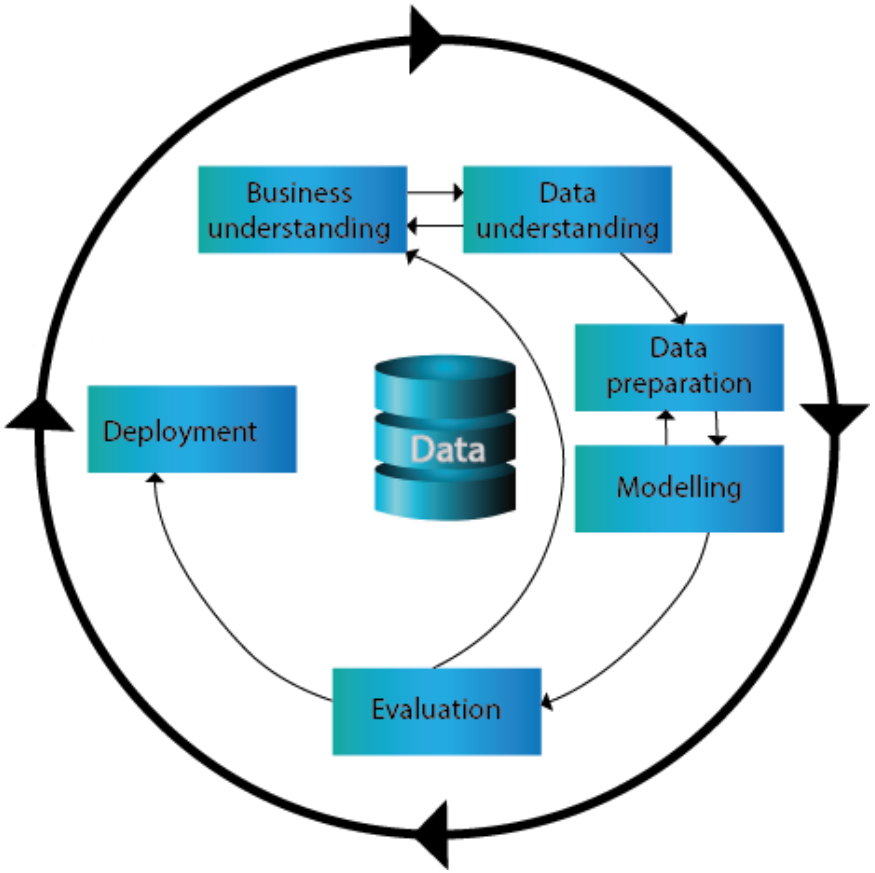


The Data Mining Process

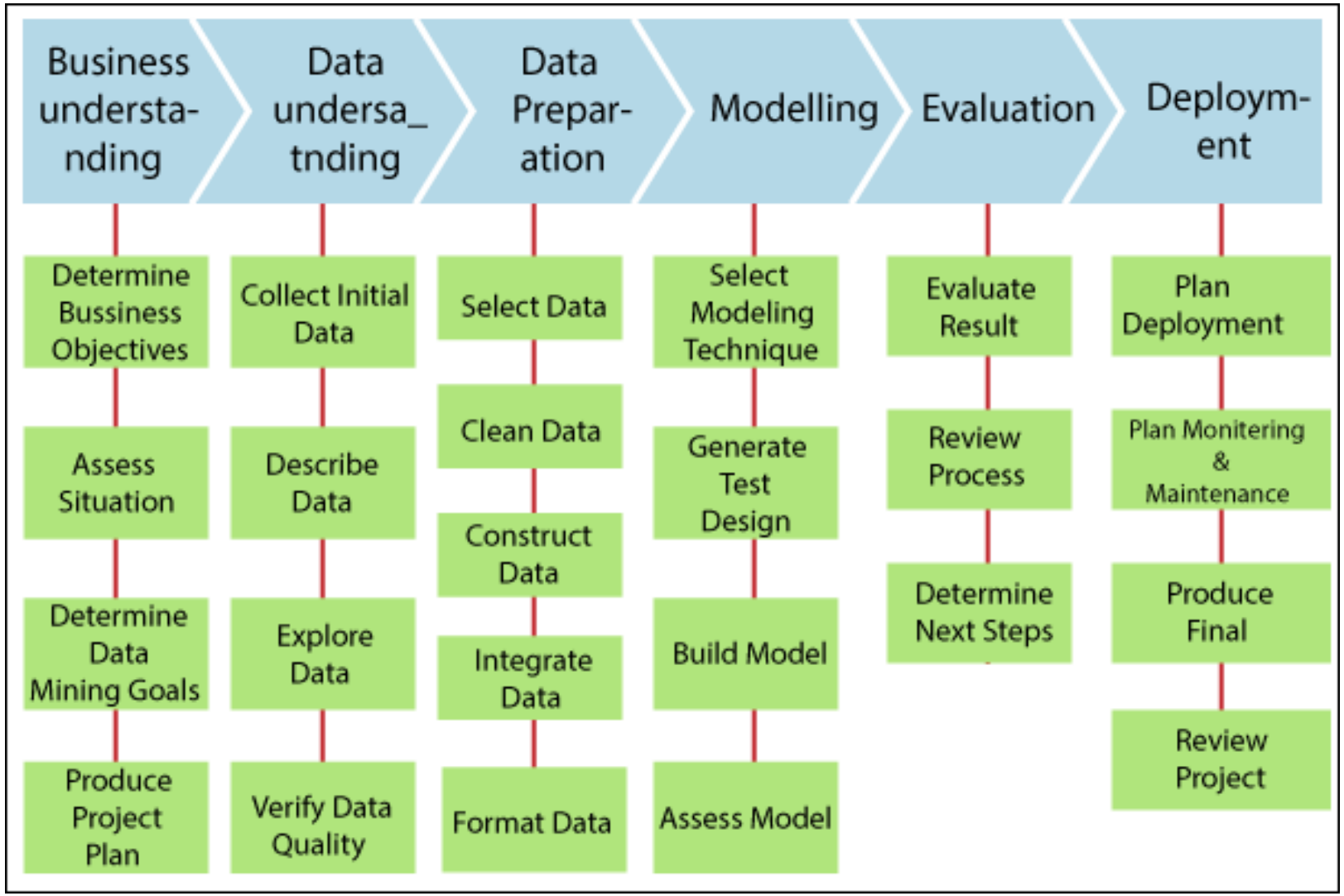
- Many different sectors are taking advantage of data mining to boost their business efficiency, including manufacturing, chemical, marketing, aerospace, etc.
- Therefore, the need for a standardized data mining process was needed.
- Data mining techniques must be reliable, repeatable by company individuals with little or no knowledge of the data mining context.
- As a result, a cross-industry standard process for data mining (CRISP-DM - The Cross-Industry Standard Process for Data Mining) was first introduced in 1990, after going through many workshops, and contribution for more than 300 organizations.

The Cross-Industry Standard Process for Data Mining (CRISP-DM)

- Cross-industry Standard Process of Data Mining (CRISP-DM) is a standard methodology/best practices guideline for Data Mining.
- It is comprised of six phases designed as a cyclical method as the given figure.



The Cross-Industry Standard Process for Data Mining (CRISP-DM)



1. Business understanding:

It focuses on understanding the project goals and requirements form a business point of view, then converting this information into a data mining problem afterward a preliminary plan designed to accomplish the target.

Tasks:

- Determine business objectives
- Access situation
- Determine data mining goals
- Produce a project plan

Determine business objectives:

- It understands the project targets and prerequisites from a business point of view.
- Thoroughly understand what the customer wants to achieve.
- Reveal significant factors, at the starting, it can impact the result of the project.

1. Business understanding:

Access situation:

- It requires a more detailed analysis of facts about all the resources, constraints, assumptions, and others that ought to be considered.

Determine data mining goals:

- A business goal states the target of the business terminology. For example, increase catalog sales to the existing customer.
- A data mining goal describes the project objectives. For example, It assumes how many objects a customer will buy, given their demographics details (Age, Salary, and City) and the price of the item over the past three years.

Produce a project plan:

- It states the targeted plan to accomplish the business and data mining plan.
- The project plan should define the expected set of steps to be performed during the rest of the project, including the latest technique and better selection of tools.

1. Business Understanding Scenario:

An e-commerce company wants to improve its marketing strategies to boost sales on its website. The data analysis team is tasked with a project to understand customer behavior and identify factors that influence purchasing decisions.

Key Steps:

- **Define the Business Objective**:
  - The primary goal is to increase the website's sales by 15% over the next six months.
  - The focus is on understanding customer behavior, such as the number of items added to the cart, time spent on the website, and engagement with promotional offers.
- **Key Questions**:
  - Which age groups or demographics are more likely to make a purchase?
  - What impact do promotions (discounts, free shipping) have on purchasing decisions?
  - What products are frequently added to the cart but not purchased?
- **Key Performance Indicators (KPIs)**:
  - Number of completed purchases.
  - Conversion rate (percentage of visitors who make a purchase).
  - Average order value.
  - Ratio of products added to the cart vs. products purchased.
- **Risk Assessment**:
  - Some customer behaviors might be missing from the data (e.g., interactions with promotions outside the website).
  - Data privacy regulations (such as GDPR) need to be respected.
- **Resources**:
  - The data analysis team has tools like Python, R, and access to customer databases.
  - There's a budget for acquiring external data or tools if needed.

2- Data Understanding:

Data understanding starts with an original data collection and proceeds with operations to get familiar with the data, to data quality issues, to find better insight in data, or to detect interesting subsets for concealed information hypothesis.

Tasks:

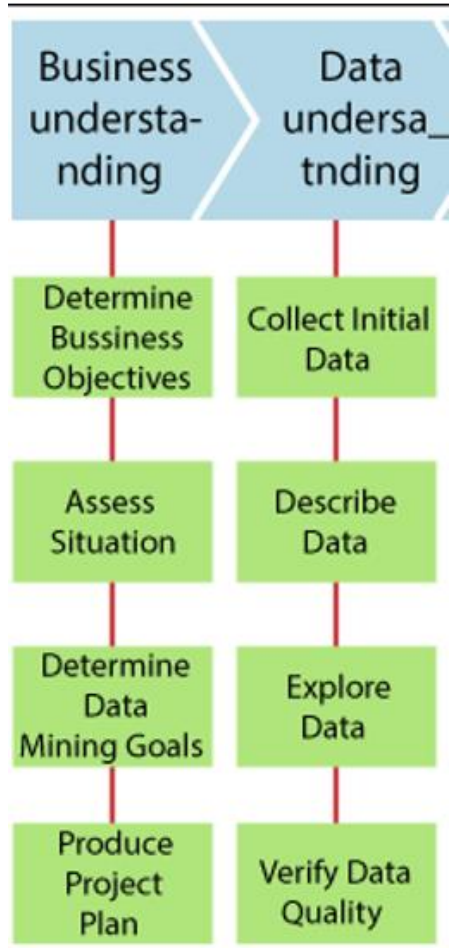
- Collects initial data
- Describe data
- Explore data
- Verify data quality

Collect initial data:

- It acquires the information mentioned in the project resources.
- It includes data loading if needed for data understanding.
- It may lead to original data preparation steps.
- If various information sources are acquired then integration is an extra issue, either here or at the subsequent stage of data preparation.

Describe data:

- It examines the "gross" or "surface" characteristics of the information obtained.
- It reports on the outcomes.



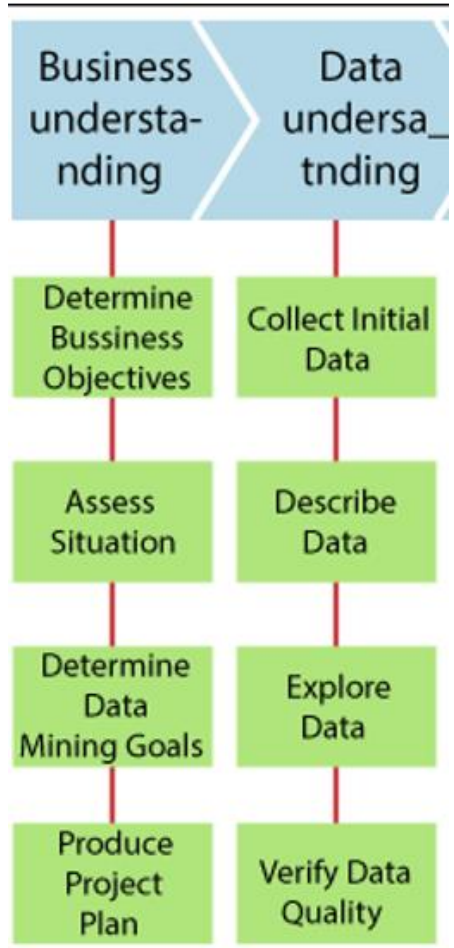
2- Data Understanding:

Explore data:

- Addressing data mining issues that can be resolved by **querying, visualizing,** and **reporting,** including:
  - Distribution of important characteristics, results of simple aggregation.
  - Establish the relationship between the small number of attributes.
  - Characteristics of important sub-populations, simple statistical analysis.
- It may refine the data mining objectives.
- It may contribute or refine the information description, and quality reports.
- It may feed into the transformation and other necessary information preparation.

Verify data quality:

- It examines the data quality and addressing questions.



2. Data Understanding Scenario:

The team gathers available data from the website, including customer interactions and sales records. The data contains:

- Customer browsing behaviors (e.g., number of site visits, time spent on pages).
- Data on promotional interactions (e.g., response to discounts).
- Sales data (purchased products, cart abandonment).
- Demographic information (age, location).

Key Steps:

- **Explore the Data**: The team performs an initial analysis of the data to identify inconsistencies, missing values, and unusual patterns. For example, they find that a large percentage of customers add items to their cart but don't complete the purchase.
- **Identify Data Quality Issues**: Missing values in some fields (e.g., customer age), outliers in transaction amounts, and incomplete logs of promotional responses.

3. Data Preparation

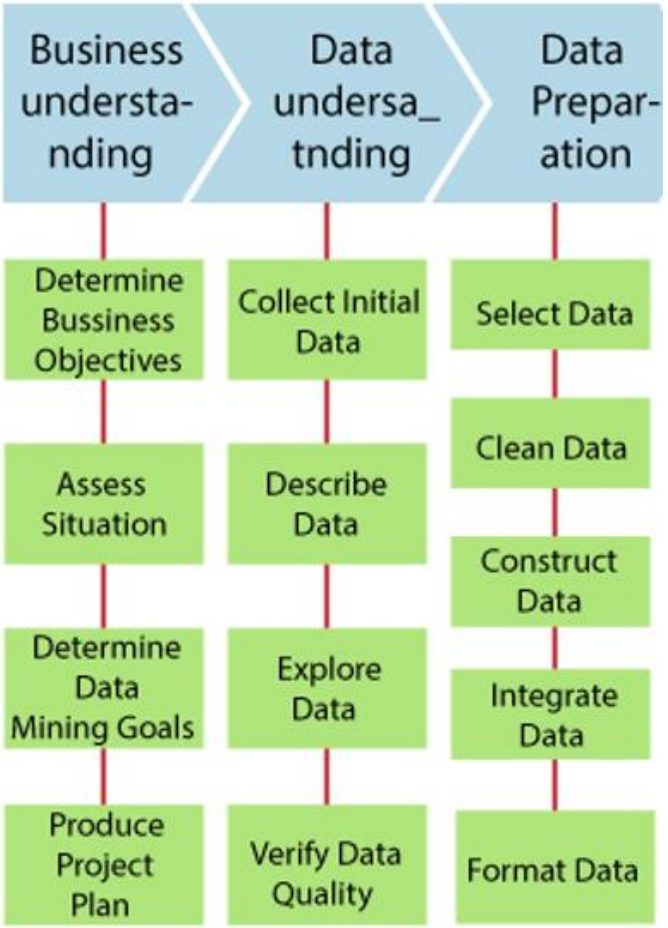
- It usually takes more than 90 percent of the time.
- It covers all operations to build the final data set from the original raw information.
- Data preparation is probable to be done several times and not in any prescribed order.

Tasks:

- Select data
- Clean data
- Construct data
- Integrate data
- Format data

Select data:

- It decides which information to be used for evaluation.
- In the data selection criteria include significance to data mining objectives, quality and technical limitations such as data volume boundaries or data types.
- It covers the selection of characteristics and the choice of the document in the table.



3. Data Preparation

Clean data:

- It may involve the selection of clean subsets of data, inserting appropriate defaults or more ambitious methods, such as estimating missing information by modeling.

Construct data:

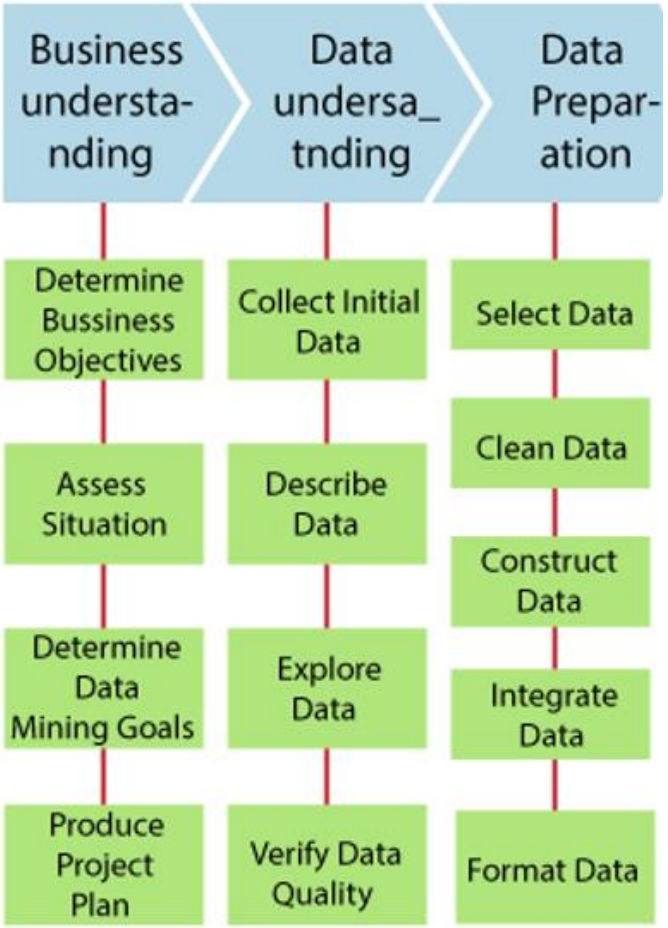
- It comprises of Constructive information preparation, such as generating derived characteristics, complete new documents, or transformed values of current characteristics.

Integrate data:

- Integrate data refers to the methods whereby data is combined from various tables, or documents to create new documents or values.

Format data:

- Formatting data refer mainly to linguistic changes produced to information that does not alter their significance but may require a modeling tool.



3. Data Preparation Scenario:

The team cleans and organizes the data to make it ready for modeling. This includes:

- **Data Cleaning**: Handling missing or incorrect values (e.g., replacing null ages with averages or discarding incomplete records).
- **Feature Selection**: Choosing relevant features like session duration, items added to cart, purchased products, and responses to promotions.
- **Data Transformation**: Formatting the data appropriately (e.g., converting timestamps into meaningful categories like "weekday vs. weekend," calculating average time spent per session).

Key Steps:

Prepare for Modeling: The data is now consistent and structured, with unnecessary information removed and relevant features created. For example, creating a binary feature indicating whether a customer responded to a promotion.

4. Modeling

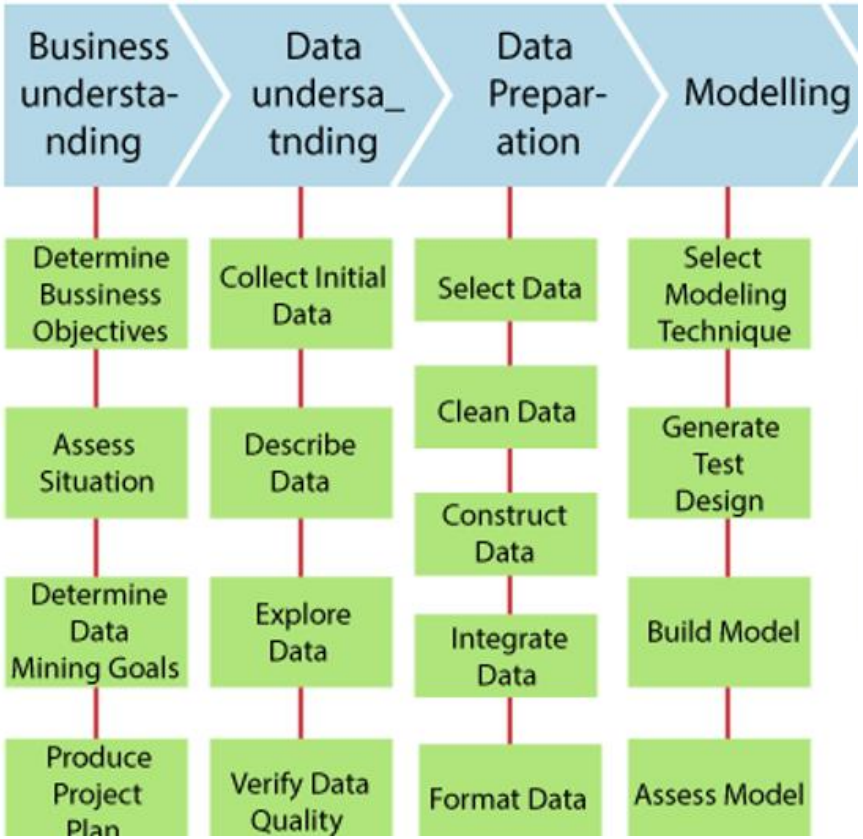
In modeling, various modeling methods are selected and applied, and their parameters are measured to optimum values. Some methods gave particular requirements on the form of data. Therefore, stepping back to the data preparation phase is necessary.

Tasks:

- Select modeling technique
- Generate test design
- Build model
- Access model

Select modeling technique:

- It selects the real modeling method that is to be used. For example, decision tree, neural network.
- If various methods are applied, then it performs this task individually for each method.



4. Modeling

Generate test Design:

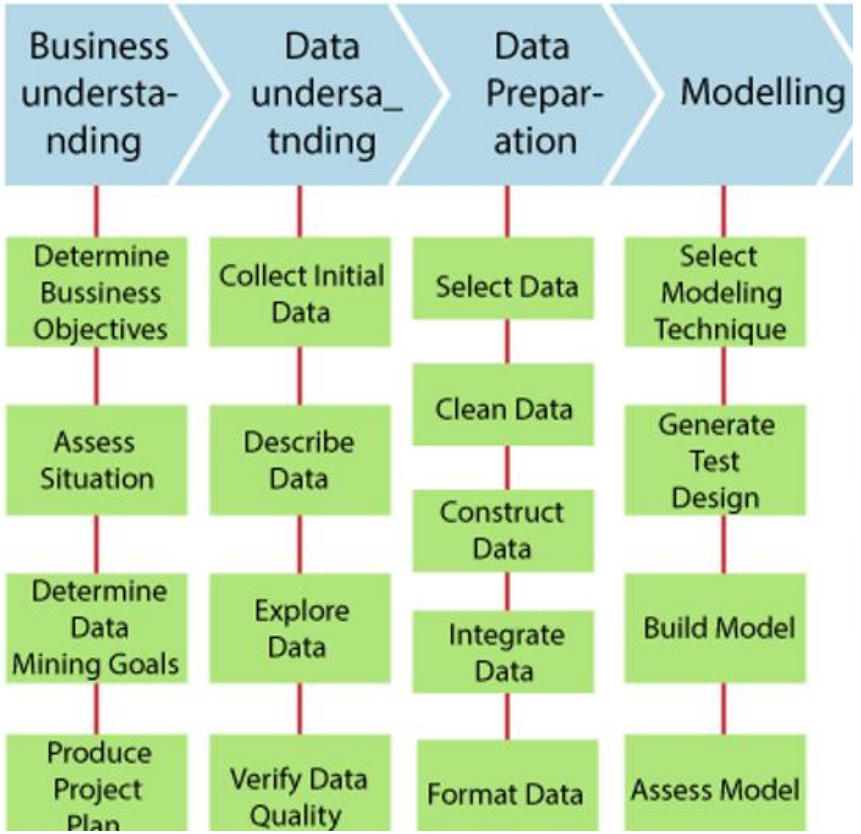
- Generate a procedure or mechanism for testing the validity and quality of the model before constructing a model. For example, in classification, error rates are commonly used as quality measures for data mining models. Therefore, typically separate the data set into train and test set, build the model on the train set and assess its quality on the separate test set.

Build model:

- To create one or more models, we need to run the modeling tool on the prepared data set.

Assess model:

- It interprets the models according to its domain expertise, the data mining success criteria, and the required design.
- It assesses the success of the application of modeling and discovers methods more technically.
- It Contacts business analytics and domain specialists later to discuss the outcomes of data mining in the business context.



4. Modeling Scenario:

The team applies data mining algorithms to extract insights and predictions from the prepared data. Some potential models include:

- **Clustering**: Segmenting customers into groups based on their purchasing behavior. For instance, one group may frequently visit the site but rarely make purchases, while another group may buy consistently but only when promotions are active.
- **Classification**: Building a model to predict which customers are likely to respond to specific promotions, using demographic data and past interactions.
- **Market Basket Analysis**: Identifying which products are often bought together, so they can be marketed as bundles or promoted to relevant customer segments.
- **Predictive Modeling**: Creating models to predict the likelihood of purchase after items are added to the cart, based on past customer behavior.

Key Steps:

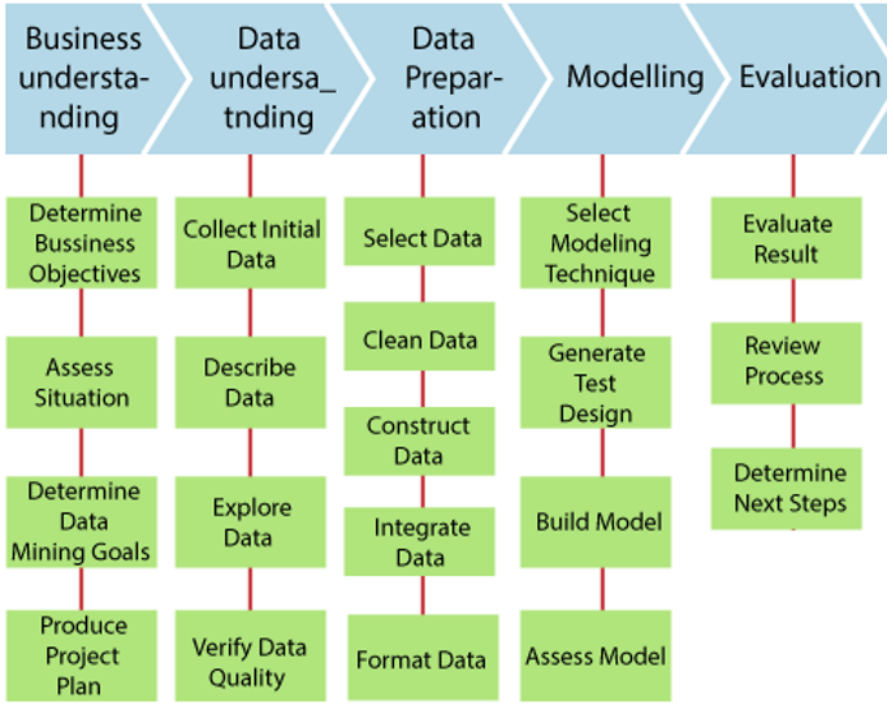
Test Models: The team applies different algorithms (e.g., decision trees, k-means clustering) to the data and tests their performance using training and validation sets.

5. Evaluation

- At the last of this phase, a decision on the use of the data mining results should be reached.
- It evaluates the model efficiently, and review the steps executed to build the model and to ensure that the business objectives are properly achieved.
- The main objective of the evaluation is to determine some significant business issue that has not been regarded adequately.
- At the last of this phase, a decision on the use of the data mining outcomes should be reached.

Tasks:

- Evaluate results
- Review process
- Determine next steps



5. Evaluation

Evaluate results:

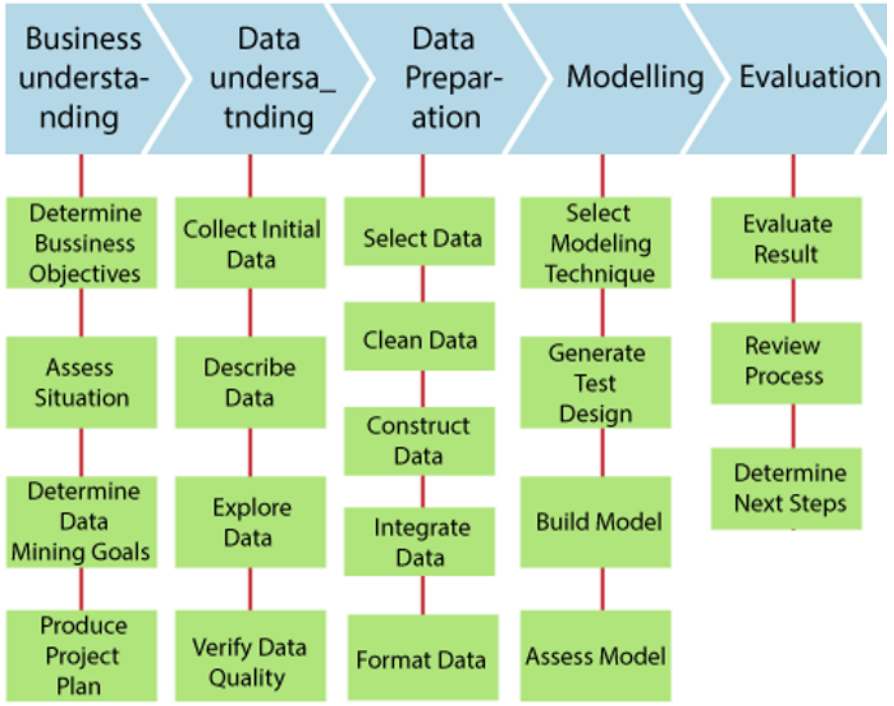
- It assesses the degree to which the model meets the organization's business objectives.
- It tests the model on test apps in the actual implementation when time and budget limitations permit and also assesses other data mining results produced.
- It unveils additional difficulties, suggestions, or information for future instructions.

Review process:

- The review process does a more detailed evaluation of the data mining engagement to determine when there is a significant factor or task that has been somehow ignored.
- It reviews quality assurance problems.

Determine next steps:

- It decides how to proceed at this stage.
- It decides whether to complete the project and move on to deployment when necessary or whether to initiate further iterations or set up new data-mining initiatives.it includes resources analysis and budget that influence the decisions.



5. Evaluation Scenario:

The models are evaluated to ensure they meet the business goals. For example:

- **Accuracy Testing**: Comparing the model's predictions with actual customer behavior to assess accuracy.
- **Business Relevance**: Ensuring that the insights provided by the models align with the business objectives, such as increasing sales and optimizing marketing campaigns.

Example: The classification model successfully identifies customers with a 90% accuracy who are likely to make a purchase after receiving a specific promotional offer.

Key Steps:

Iterate or Refine Models: If the models don't meet the required performance, adjustments are made (e.g., tweaking parameters, selecting different features).

6. Deployment

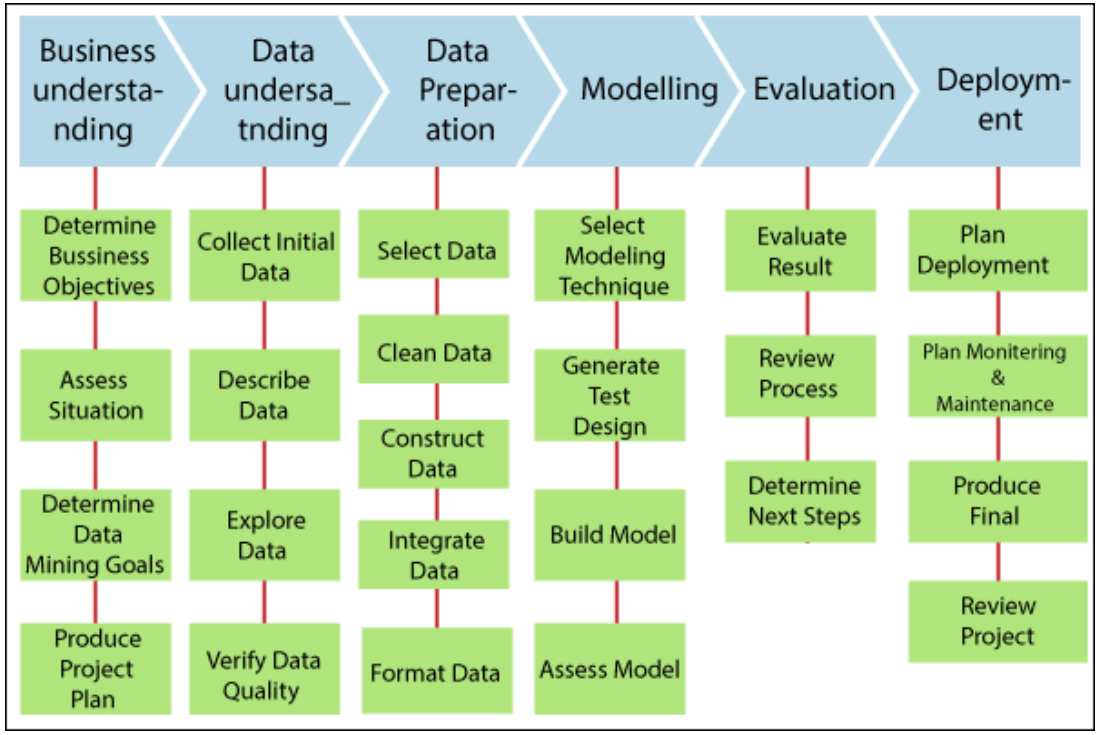
Deployment refers to how the outcomes need to be utilized.

Deploy data mining results by:

- It includes scoring a database, utilizing results as company guidelines, interactive internet scoring.
- The information acquired will need to be organized and presented in a way that can be used by the client. However, the deployment phase can be as easy as producing. However, depending on the demands, the deployment phase may be as simple as generating a report or as complicated as applying a repeatable data mining method across the organizations.

Tasks:

- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project



6. Deployment

Plan deployment:

- To deploy the data mining outcomes into the business, takes the assessment results and concludes a strategy for deployment.
- It refers to documentation of the process for later deployment.

Plan monitoring and maintenance:

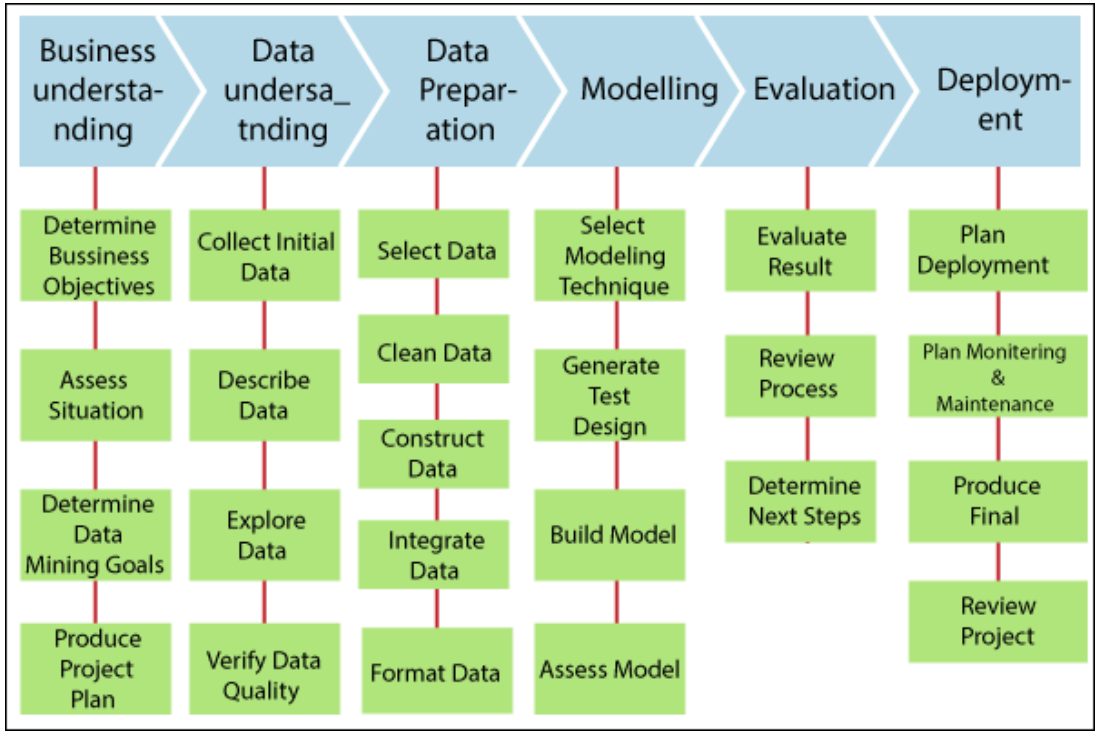
- It is important when the data mining results become part of the day-to-day business and its environment.
- It helps to avoid unnecessarily long periods of misuse of data mining results.
- It needs a detailed analysis of the monitoring process.

Produce final report:

- A final report can be drawn up by the project leader and his team.
- It may only be a summary of the project and its experience.
- It may be a final and comprehensive presentation of data mining.

Review project:

- Review projects evaluate what went right and what went wrong, what was done wrong, and what needs to be improved.



6. Deployment Scenario:

After the models have been evaluated and refined, the team deploys the results into the company's marketing strategy. This could include:

- **Targeted Marketing Campaigns**: Use the classification model to send personalized promotional offers to customers identified as likely buyers.
- **Product Recommendations**: Based on market basket analysis, the website can display product recommendations that are likely to appeal to individual customers.
- **Dynamic Promotions**: Implementing the predictive model to offer discounts or other incentives to customers who are predicted to abandon their cart without purchasing.

Key Steps:

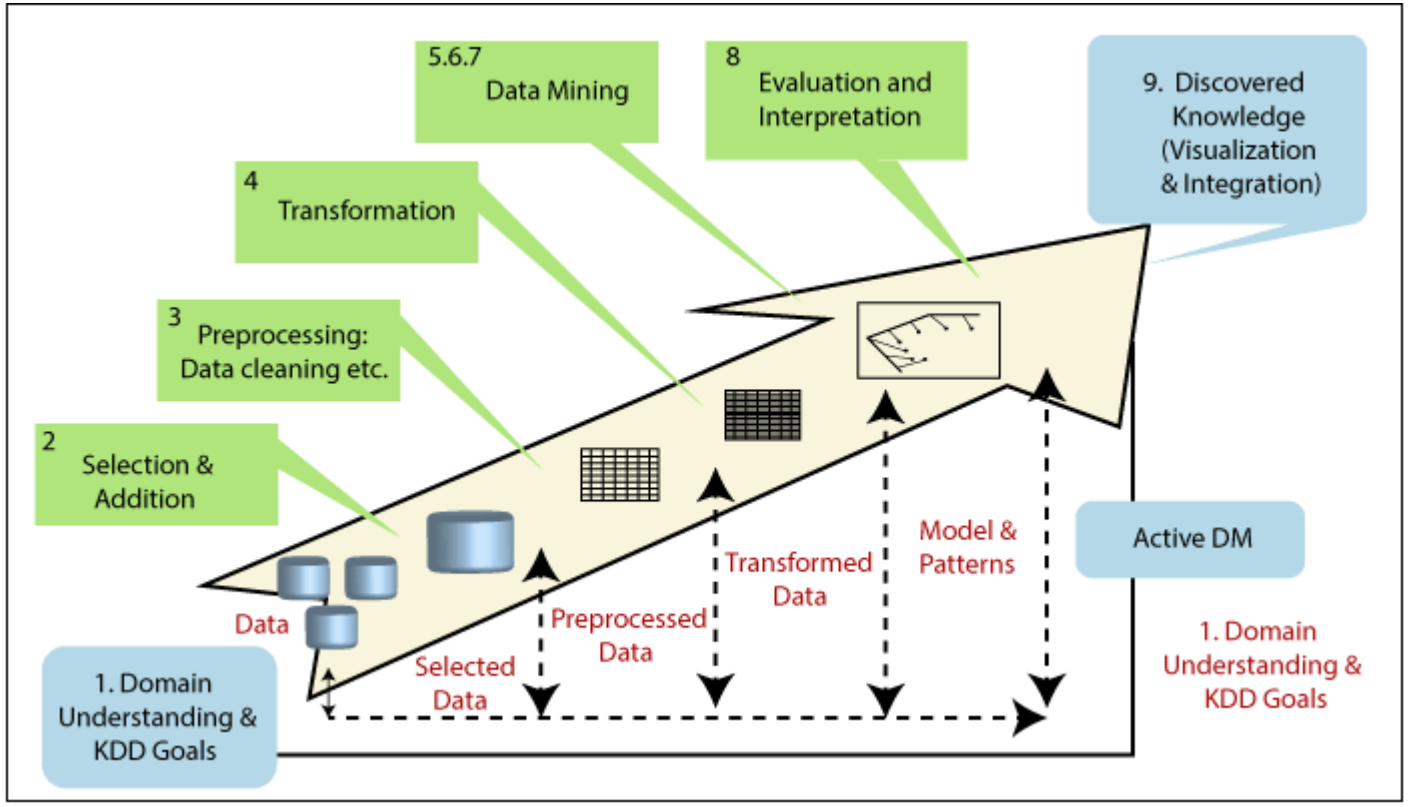
- **Monitor and Optimize**: The company tracks the performance of these initiatives to see how they affect conversion rates and sales, and further adjusts the models if necessary.
- **Training and Documentation**: The sales and marketing teams are trained on how to use the insights generated from the models, and automated reports are generated to continuously monitor key metrics (e.g., conversion rates, average order values).

KDD (Knowledge Discovery in Databases)

- **Developed**: In the late 1980s to early 1990s, as a more **academic** framework for extracting knowledge from large datasets.
- **Focus**: The entire process of transforming raw data into useful knowledge.
- **Primary Goal**: To develop a framework for researchers and data analysts to systematically extract knowledge, focusing on the data-driven aspect of discovery.

Key Steps in the KDD Process:

1. Domain Understanding & KDD Goals: Define the problem and objectives.
2. Selection & Addition: Choose relevant data sources.
3. Preprocessing (Data Cleaning): Clean and prepare the data.
4. Transformation: Convert data into a suitable format for analysis. 5-7. Data Mining: Apply algorithms to discover patterns.
5. Evaluation and Interpretation: Assess the discovered patterns.
6. Discovered Knowledge: Visualize and integrate the knowledge for practical use.



Aspect	KDD (Knowledge Discovery in Databases)	CRISP-DM (Cross Industry Standard Process for Data Mining)
Primary Focus	Discovery of new patterns or knowledge from data	Solving specific business problems using data mining
Purpose	**Research-oriented**; theoretical exploration of data	Practical, structured approach for **business-driven** data mining projects
Audience	Data scientists, researchers, and academics	Industry professionals, business analysts, data practitioners
Origin	Developed in late 1980s to early 1990s in academic contexts	Developed in 1996 by a consortium of companies (SPSS, Teradata, Daimler-Benz)
Process Goal	Understanding and discovering unknown patterns in data	Aligning data mining efforts with business goals and deploying actionable insights
Stages	5 Stages: Data Selection, Preprocessing, Transformation, Data Mining, Interpretation & Evaluation	6 Stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment
Emphasis	Data-centric (focus on selection, preparation, and algorithm application)	Business-centric (focus on understanding business needs and deploying insights)
Application Context	Academic research, experimental data analysis	Real-world business applications, operational problem-solving
Flexibility	More exploratory, flexible in terms of data exploration and discovery	Structured and repeatable, tailored to industry needs
Outcome	New knowledge, insights, and understanding	Practical solutions and actionable insights for business improvement
Commonly Used In	Research institutions, universities, and data science labs	Businesses across industries like finance, retail, marketing, etc.
Data Mining Stage	A core part of the overall process, but with focus on knowledge extraction	One of the stages; focuses on ali

Business Cases - Examples

Key Techniques Mentioned:

- **Clustering**: Used in segmentation (e.g., customer or patient segmentation).
- **Classification**: Applied for binary or multi-class decisions (e.g., fraud detection, churn prediction).
- **Regression**: Used for continuous prediction (e.g., price optimization, demand forecasting).
- **Association Rule Mining**: Discovering relationships between variables (e.g., market basket analysis).
- **Outlier Detection**: Detecting anomalies (e.g., fraud detection, anomaly detection in production).
- **Dimension Reduction**: Reducing the number of variables while retaining important patterns (e.g., risk management, portfolio optimization).

Data Mining Technique	Industry	Problem Domain	Business Example
Clustering	Retail	Customer Segmentation	Grouping customers based on purchasing behavior to target marketing campaigns
	Healthcare	Patient Segmentation	Grouping patients with similar treatment responses for personalized medicine
	Pharmaceuticals	Drug Effectiveness	Identifying patterns in drug response among patient groups
Classification	Finance	Credit Scoring	Predicting credit risk and determining if a customer is likely to default on a loan
	Healthcare	Disease Diagnosis	Predicting the likelihood of heart disease based on patient attributes
	Telecom	Customer Churn Prediction	Predicting which customers are likely to cancel their subscription
	Insurance	Customer Retention	Identifying factors leading to customer cancellations to improve retention strategies
	Banking	Loan Approval	Assessing if a loan applicant is eligible for approval based on past behavior
	Education	Student Dropout Prediction	Predicting students likely to drop out based on academic performance and other factors
	Retail	Personalized Marketing	Using purchase history to create personalized offers for customers
	Marketing	Campaign Effectiveness	Predicting the likelihood of a customer responding to a marketing campaign

Data Mining Technique	Industry	Problem Domain	Business Example
Regression	E-Commerce	Price Optimization	Predicting the best price point for maximizing sales
	Real Estate	Property Value Prediction	Predicting house prices based on location, size, and features
	Retail	Inventory Management	Optimizing inventory levels to avoid stockouts or overstocking
	Logistics	Demand Forecasting	Predicting future demand for products to adjust supply chain operations
	Manufacturing	Product Failure Prediction	Predicting when a machine is likely to fail based on sensor data
Outlier Detection	Finance	Fraud Detection	Detecting unusual spending patterns on credit cards to flag fraudulent transactions
	Manufacturing	Anomaly Detection in Production	Identifying faulty products in a manufacturing process
	Telecom	Network Optimization	Detecting unusual network traffic patterns to prevent failures
Association Rule Mining	Retail	Market Basket Analysis	Finding combinations of items frequently bought together in a supermarket
	E-Commerce	Customer Purchase Path Analysis	Analyzing the sequence of customer actions leading to a purchase
Dimension Reduction	Banking	Risk Management	Predicting financial risks based on market and economic conditions
	Banking	Portfolio Optimization	Reducing the dimensionality of variables to optimize investment portfolios