# Methodologies in Data Mining


**Riska Yanu Fa'rifah**

**Information Systems Undergraduate Program**
**School of Industrial Engineering**

**PLO & CLO to be achieved:**

**PLO02 –** Able to design, develop, implement, and evaluate information system-based solutions to meet organizational needs towards becoming a data-driven organization.

**CLO02 –** The students are capable of developing information system-based solutions using the appropriate development methodology.

**Outline:**

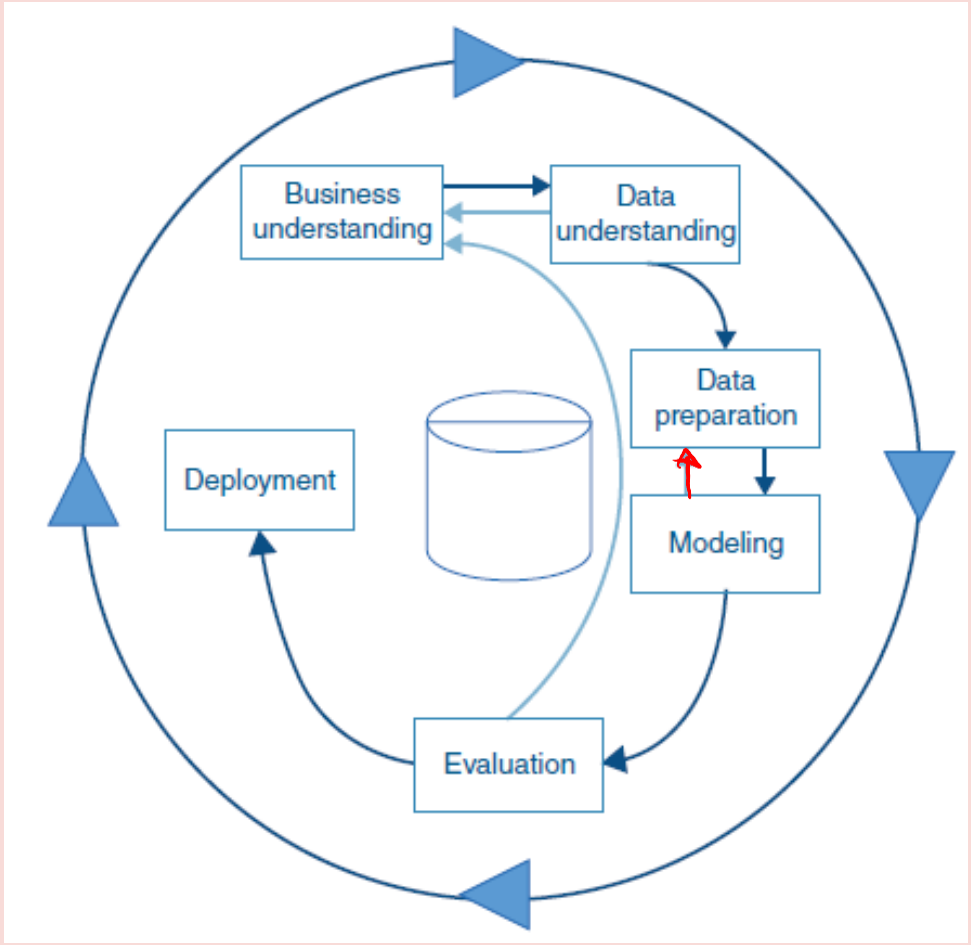
1. **Types of Methodologies in Data Mining**
2. **Business Understanding**
3. **Data Understanding**
4. **Data Preparation**
5. **Modelling**
6. **Evaluation**
7. **Deployment**

**Reference:**

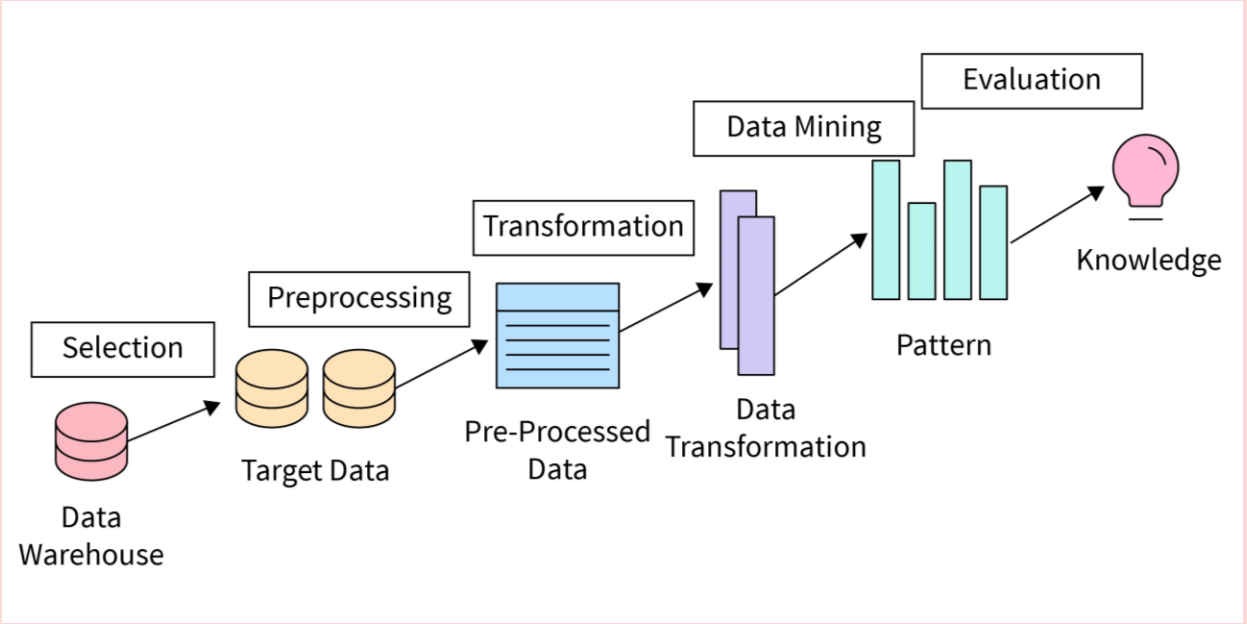
1. Han, Jiawei, Jian Pei, and Hanghang Tong. 2023. Data Mining Concepts and Techniques. 4th ed. ed. Beth LoGiudice. Cambridge: Katey Birtcher.
2. Moreira, J. M., Carvalho, A. C. P. L. F., & Horváth, T. (2018). A General Introduction to Data Analytics. In A General Introduction to Data Analytics. https://doi.org/10.1002/9781119296294
3. Provost & Fawcett. (2013). Data science-what you need to know about analytic-thinking and decision-making. Journal of Chemical Information and Modeling, 53(9), 1689–1699.

# Types of Methodologies in Data Mining

## CRISP - DM


Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Deployment

## KDD


Selection
Preprocessing
Transformation
Data Mining
Evaluation
Data Warehouse
Target Data
Pre-Processed Data
Data Transformation
Pattern
Knowledge

# Business Understanding

**Business Understanding:**

1. Understand the business domain, being able to define the problem from the business domain perspective, and finally being able to translate such business problems into a data analytics problem.
2. This phase focuses on understanding the business objectives and requirements of a data mining project.
3. The Business Understanding lays the foundation for the entire project by ensuring that the analytics efforts align with the organization’s strategic goals..

**Stages in Business Understanding:**

1. Define Business Objectives
2. Assess Current Situation
3. Formulate Data Mining Problem
4. Determine Project Objectives
5. Plan Project

**The Example of Business Understanding (Predicting Hospital Readmissions):**

1. **Define Business Objectives**
   - **Identify goals:** The primary goal is to reduce hospital readmission rates for patients with chronic conditions, thereby improving patient outcomes and reducing costs associated with unnecessary readmissions
   - **Specific objective:** Aim to decrease readmissions by 15% within the next year
2. **Assess Current Situation**
   - **Analyze existing processes:**
     - Review current discharge processes and follow-up protocols.
     - Identify which patients are frequently readmitted and investigate the reasons behind their readmissions
   - **Stakeholder input:** Gather insights from healthcare providers, case managers, and patient to understand the challenges faced in managing post-discharge care.

3. **Formulate Data Mining Problem**
   - **Translate business goals:**
     - The data mining problem is to develop a predictive model that identifies patients at high risk of readmission within 30 days of discharge.
     - Define key factors influencing readmissions, such as previous admission history, treatment plans, and demographic information.
4. **Determine Project Objectives**
   - **Set success criteria:**
     - Success will be measured by the model's accuracy in identifying at-risk patients, with a target of at least 75% precision and recall.
     - Additional success criteria may include improved patient follow-up rates and a decrease in overall readmission costs.

5. **Plan Project**
   - **Outline the project scope:**
     - Define the project timeline, including data collection, analysis, model development, and evaluation phases.
     - Identify key personnel, including data analysts, healthcare professionals, and IT support.
   - **Resource requirements:** Determine necessary data sources, tools for analysis (example: machine learning software), and budget considerations.

# Data Understanding

**Data Understanding:**

1. This involves the collection of necessary data and its initial visualization and summarization to obtain the first insights, particularly, but not exclusively, regarding data quality issues such as missing data or outliers.
2. This phase focuses on collecting and analyzing the data relevant to the problem identified in the Business Understanding stage.
3. The goal is to gain insights into the data's characteristics, quality, and potential issues that may affect the analysis.

**Stages in Data Understanding:**

1. Data Collection
2. Data Description
3. Data Exploration
4. Data Quality Assessment

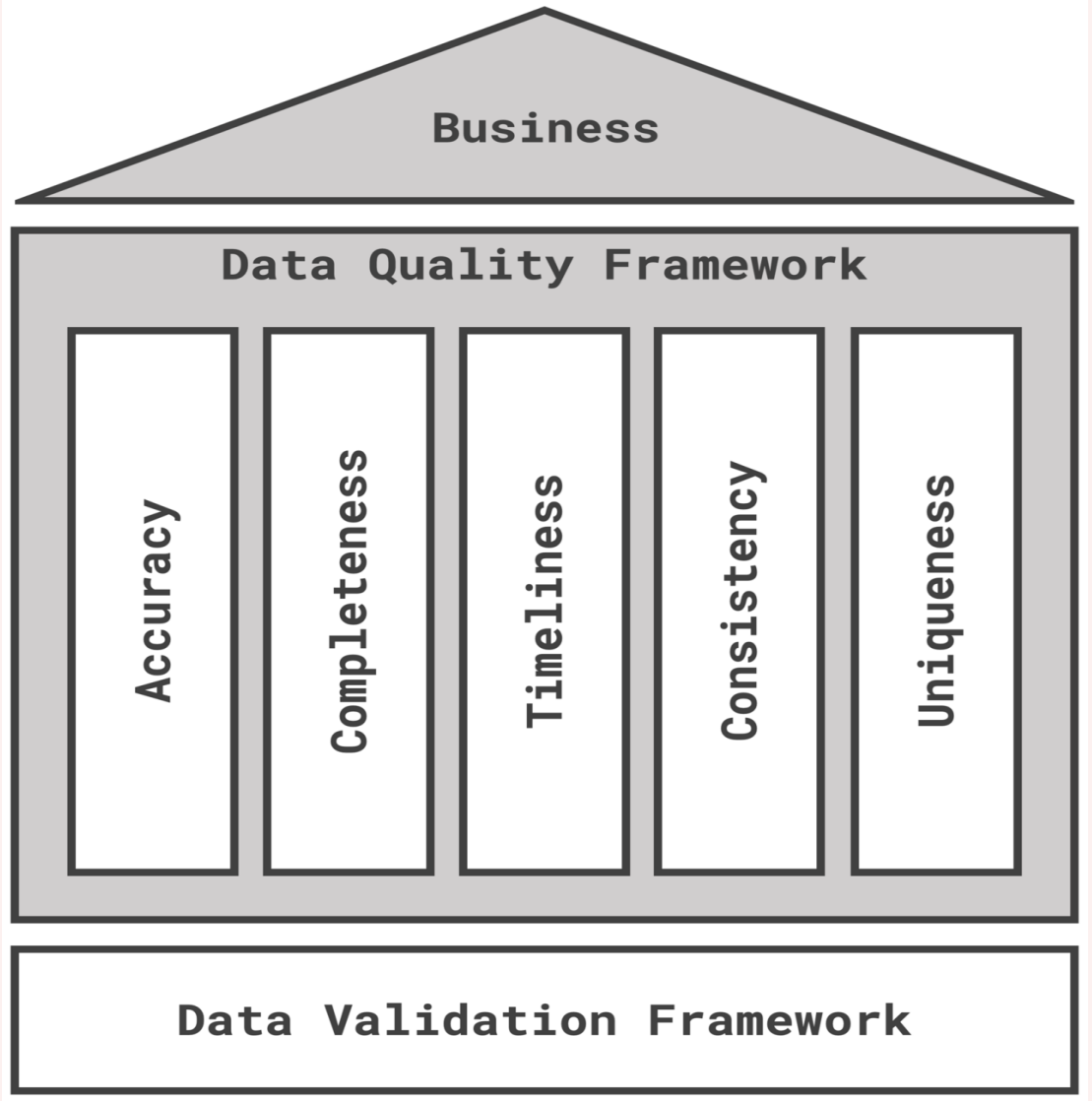
**The Example of Data Understanding (Predicting Hospital Readmissions):**

1. **Data Collection**
   - **Gather Data**: Collect relevant datasets needed for the analysis. In this case, data may include:
     - Patient Demographics: Age, gender, socioeconomic status.
     - Medical History: Previous admissions, chronic conditions, and comorbidities.
     - Treatment Data: Details of treatment administered during the hospital stay.
     - Discharge Information: Date of discharge, discharge instructions, follow-up appointments.
     - Readmission Data: Whether the patient was readmitted within 30 days and the reason for readmission.
   - **Sources:** Data may be sourced from electronic health records (EHR), patient management systems, and hospital databases.

2. **Data Description**
   - **Summarize Data:** Describe the structure and content of the dataset, including data types and variable counts
3. **Data Exploration**
   - **Exploratory Data Analysis (EDA):** Conduct EDA to identify patterns and insights, including:
     - Distribution analysis: Analyze the distribution of key variables, such as the age of patients or the frequency of different diagnoses.
     - Correlation analysis: Examine correlations between variables. For example, check if there's a correlation between the length of stay and readmission rates.
     - Visualization: use histograms, box plots, or scatter plots to understand relationships and distributions.

## 4. Data Quality Assessment

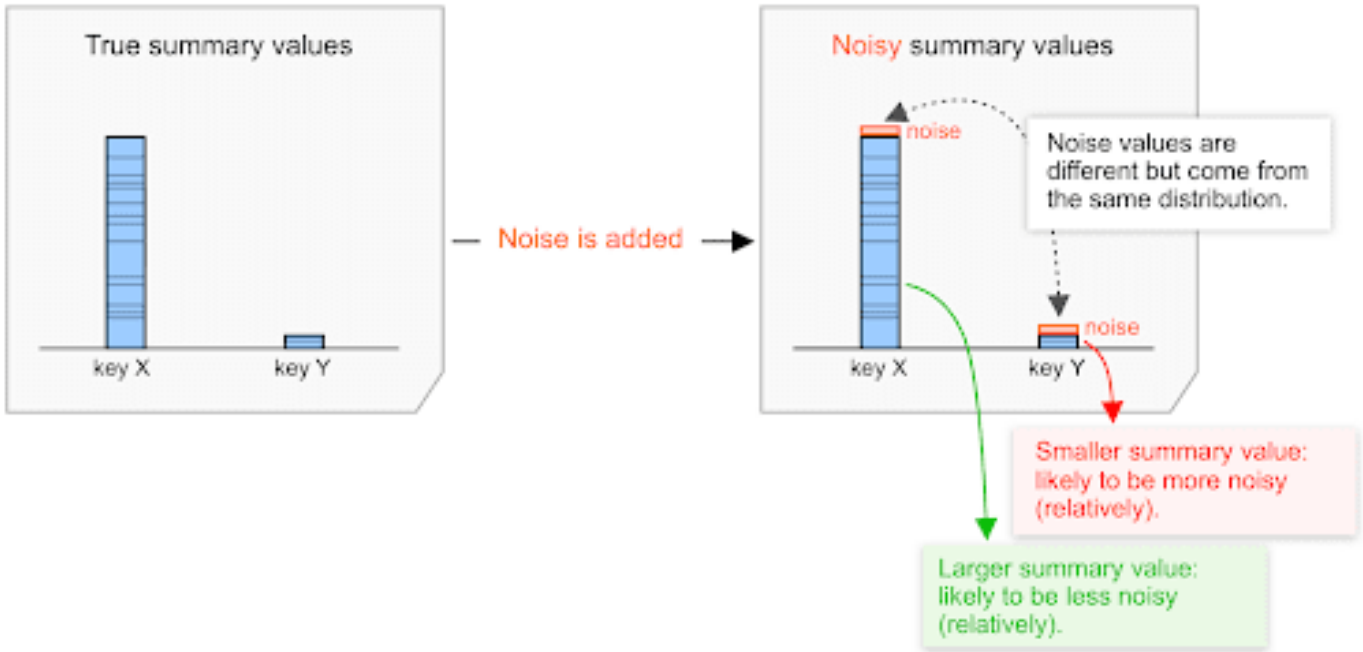
- Data have quality if they satisfy the requirements of the intended use.
- There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, uniqueness.



| Diagnose | Hemoglobin | Hematokrit | Lekosit | Eritrosit | Trombosit | HbA1c | RBG |
|---|---|---|---|---|---|---|---|
| DM TYPE 2 | 120.6 | 37.1 | 7130 | 4.09 | 241000 | 10.6 | 150 |
| DM TYPE 2 + comorbidity | | | | | | 5 | 33 |
| DM TYPE 2 + comorbidity | 13.3 | 39.5 | 8330 | 4.47 | 340000 | 9.2 | 161 |
| DM TYPE 2 + comorbidity | 14.2 | 41.6 | 15.68 | 4.95 | 364000 | 10.8 | 425 |
| DM TYPE 2 | 12.1 | 36.2 | 20590 | 5.33 | 333000 | 11 | 247 |
| DM TYPE 2 | 12.2 | 35.2 | 10020 | 4.26 | 359000 | | |
| DM TYPE 2 + comorbidity | 12.4 | 36.2 | 13570 | 3.95 | 432000 | 11.1 | 398 |

**Main Problem in Data Quality:**

a. Redundant vales
b. Missing values
c. Inconsistencies
d. Outliers
e. Noisy values



**Stages in Data Quality Assessment**

- Checking the redundant values
- Checking the missing values
- Checking the inconsistencies
- Checking the outliers
- Checking the noisy values

# Data Preparation

**Data Preparation:**

1. This involves preparing the data set for the modeling tool, and includes data transformation, feature construction, outlier removal, missing data fulfillment and incomplete instances removal.
2. Data preparation is a crucial stage in the CRISP-DM process.

**Stages in Data Preparation:**

1. Data cleaning ✓
2. Data integration ✓
3. Data reduction ✓
4. Data transformation ✓

**The Example of Data Preparation (Predicting Hospital Readmissions):**

1. **Data Cleaning**
   - **Handle Missing Values:** Identify any missing data in critical fields (e.g., missing age, diagnosis, or follow-up appointment details)
     **Techniques:**
     - Imputation: Fill in missing values using statistical methods (like mean, median, or mode).
     - Removal: Exclude records with significant missing values if they impact the dataset's integrity.
   - **Remove Duplicates:** Check for duplicate patient records and ensure that each patient is represented only once.
   - **Correct Inconsistencies:** Standardize variable entries for uniformity (like ensuring all diagnosis codes follow the same format).

2. **Data Integration**
   - **Merge Datasets:** Combine data from multiple sources into a single dataset. **Example:** Integrate hospital admission records with post-discharge survey data to provide a comprehensive view of patient outcomes.
   - **Resolve Redundancies:** Eliminate duplicate entries that may arise from combining datasets.
3. **Data Reduction**
   - **Dimensionality Reduction (Reduce the Number of Features)**
     - Use techniques to reduce the number of features while retaining the essential information.
     - Methods:
       Principal Component Analysis (PCA): Transform the dataset into a lower dimensional space while preserving variance.
       Feature Selection Techniques: Apply methods like Recursive Feature Elimination (RFE) or using statistical tests to select the most significant variables.
   - **Feature Selection:** Identify and retain only the most important features for the analysis, removing redundant or irrelevant variables.

4. **Data Transformation:** is a critical stage in data preparation that involves converting raw data into a format suitable for analysis. This stage aims to convert data into appropriate formats for analysis.
   - **Encoding categorical variables/ attributes:** Convert categorical data into a numerical format for analysis.
     **Method:**
     - One-Hot Encoding: Create binary (0 or 1) columns for each category.
     - Label Encoding: Assign unique integers to each category.
   - **Feature scaling:** Normalize or standardize numerical features to ensure they are on a similar scale, which is particularly important for distance-based algorithms.
     **Methods:**
     - Min-Max Scaling: Scale values to a range between 0 and 1.
     - Z-Score Normalization: Scale values based on their mean and standard deviation.

Continued...

- **Data Aggregation**
  - Summarize and combine data from multiple records to create a single record per patient, reducing the overall volume of data.
  - Example: Aggregate patient visits by calculating the total number of visits or average length of stay for each patient over a specific period.
- **Discretization**
  - Convert continuous variables into discrete categories to simplify analysis.
  - Group continuous values into bins or intervals.
  - Example: Transforming ages into categories like "18-30," "31-50," and "51+" to reduce complexity.

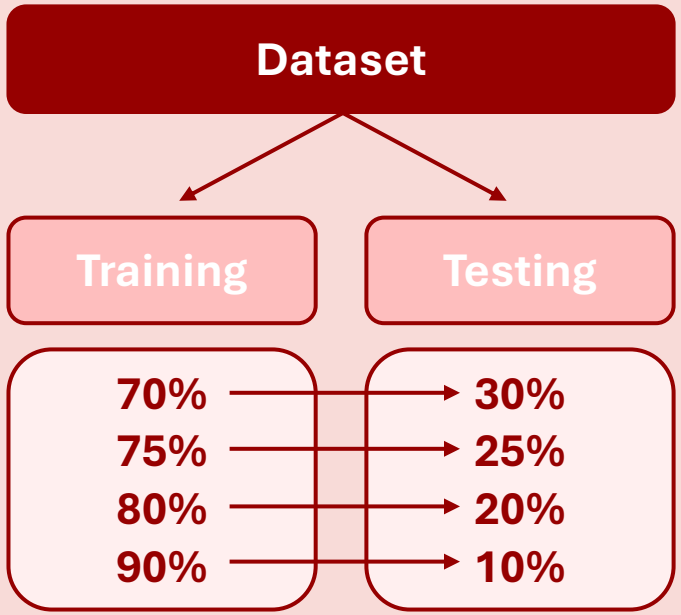
# Modelling

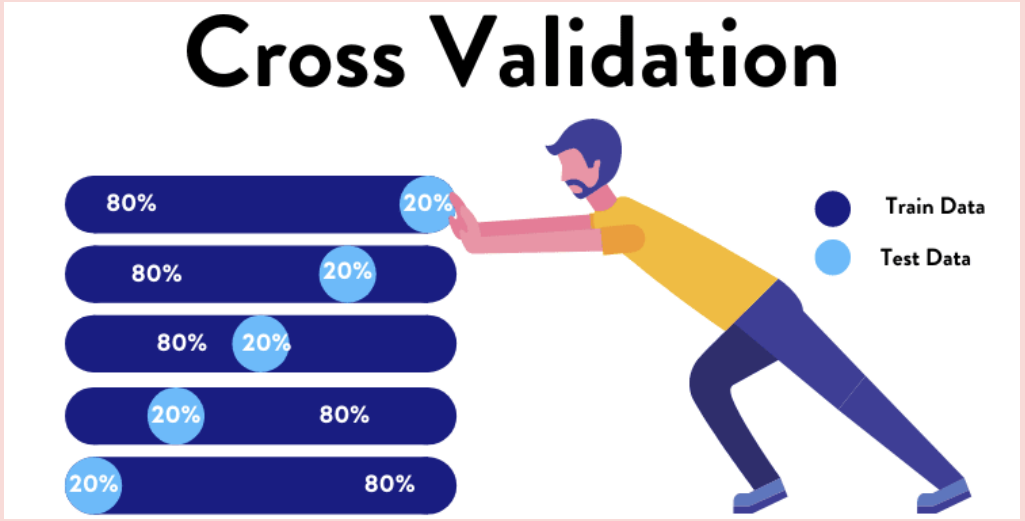
**Modeling:**

1. Refers to the process of applying statistical or machine learning techniques to the data to uncover patterns, relationships, or predictive insights that can help solve the business problem identified at the beginning of the project.
2. Several methods can be used to solve the same problem in analytics, often with specific data requirements.

**In modeling we needs:**

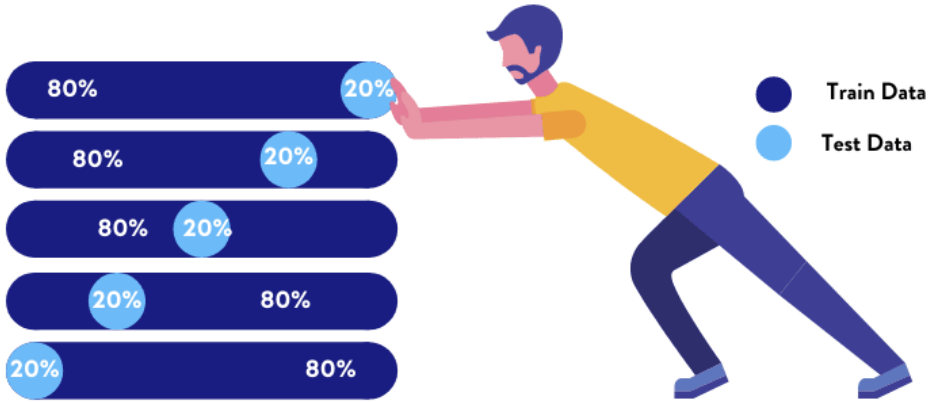
1. Data readiness -> the data needs to be pre-processed like handling missing value, outlier, normalizing or scaling data, and encoding data
2. Splitting data -> in modeling, the data is typically split into a training set (used to build the model) and a testing or validation set (used to evaluate the model's performance). In splitting data, cross validation technique is also often employed.




Resource: dataaspiran.com

# Cross Validation



**In modeling we needs:**

3. Select a suitable algorithm -> depending on the business problem and the type of data, suitable modeling algorithms must be chosen. This could be regression (for predicting continuous values), classification (for predicting categorizing data), clustering (for finding groups within data).
   Types of algorithm:
   - Classification -> Decision Tree, Naïve Bayes, Support Vector Machine, Random Forest, K-Nearest Neighbors, Neural Network
   - Regression -> Linear Regression, Logistic Regression
   - Clustering -> K-mean, K-Medoid
   - Association -> Apriori Algorithm, Frequent Pattern Growth

**The Example of Modeling (Predicting Hospital Readmissions):**

1. Logistic regression -> predicts the probability of a patient being readmitted (binary outcome: readmitted or not).
2. Decision Trees or Random Forests -> to handle complex interactions between variables/ attributes and can automatically identify the most important features.
3. Neural network -> to solve a cases with a large amount of data and complex relationships

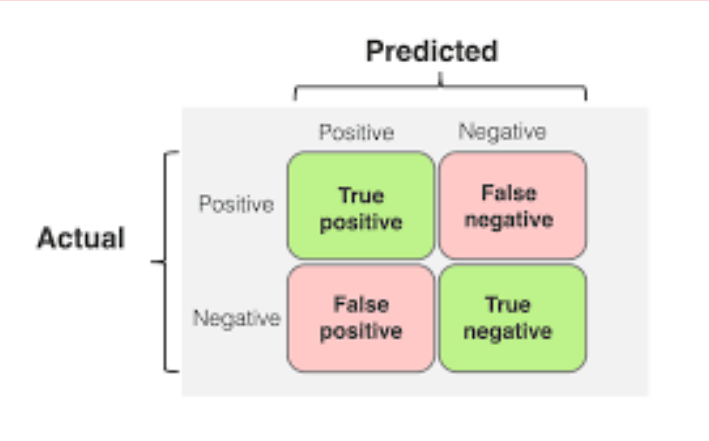
# Evaluation

**Evaluation:**

1. The models built in the modeling phase are thoroughly assessed to ensure they meet the business objectives and deliver actionable insights.
2. The goals of the evaluation are to determine how well the model performs, whether it meets the business objectives, and if it is ready for deployment or needs further refinement.
3. Types of evaluation
   - For Classification -> Confusion matrix (accuracy, precision, recall, f1-score), AUC & ROC curves
   - For Regression -> RMSE (root mean square error), MAE (mean absolute error), R-square
   - For clustering -> Silhouette score, Elbow score

# Evaluation Method for Classification Models:

## Confusion Matrix




$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 - score = 2\frac{precision \times recall}{precision + recall}$$

## AUC & ROC Curve

1. **Receiver Operating Characteristic (ROC):** curve is a graphical representation of a classifier's performance by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold values.

$$TPR = \frac{TP}{TP + FN}$$

$$FTR = \frac{FP}{FP + TN}$$

2. **Area Under the Curve (AUC)**

$$AUC = \sum_{i=1}^{n-1} \left( \frac{TPR_{i+1} + TPR_i}{2} \right) (FPR_{i+1} - FPR_i)$$

AUC is a scalar value between 0 and 1, where:

AUC = 1: Perfect classifier

AUC = 0.5: A model no better than random guessing

AUC < 0.5: A classifier worse than random guessing

# Evaluation Method for Regression Models:

## R-Square (for linear regression)

Also known as the coefficient determinant is a measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s) and commonly used to evaluate the goodness-of-fit of a regression model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$R^2 = 1$ -> Perfect fit, meaning the model explains 100% of the variance in the dependent variable.

$R^2 = 0$ -> The model does not explain any of the variance (the predictions are no better than simply using the mean of the dependent variable).

$0 < R^2 < 1$ -> Indicates the proportion of variance explained by the model.

# Evaluation Method for Regression Models:

## MAE

MAE is a common metric used to measure the accuracy of a regression model. It represents the average of the absolute differences between the actual values and the predicted values. It gives an idea of how far the predictions are from the actual values, on average.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

## RMSE

RMSE is is a metric used to evaluate the accuracy of a regression model by measuring the average magnitude of the errors between the predicted and actual values. It penalizes larger errors more significantly due to the squaring of differences.

$$RMSE = \sqrt{\frac{i}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

A lower RMSE indicates that the model's predictions are closer to the actual values (better performance).

A higher RMSE suggests larger discrepancies between the predicted and actual values, meaning the model performs poorly.

# Evaluation Method for Clustering Models:

**Silhouette score**

The Silhouette score is a metric used to evaluate the quality of clusters in clustering algorithms like K-means or hierarchical clustering. It measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). A higher Silhouette score indicates better-defined clusters.

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

$a_i$: the average distance in the same cluster (cohesion)

$b_i$: the average distance in the nearest neighboring cluster (separation)

$S_i \approx 1$: The point is well-clustered, and its distance to points in its own cluster is much smaller than to points in other clusters.

$S_i \approx 0$: The point is on or very close to the boundary between two clusters.

$S_i \approx -1$: The point is likely misclassified, being closer to another cluster than its own.

# Evaluation Method for Clustering Models:

**Elbow Method (using WCSS Score)**

is used to determine the optimal number of clusters in a clustering algorithm, typically for K-means clustering, k-medoid clustering.

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where elbow point decrease slows down significantly, this point suggests the optimal number of clusters.

# Deployment

**Deployment:**

1. It is the final phase, that focuses on implementing the results of the data mining project into real-world operations
2. This step involves making the model or insights accessible to stakeholders, integrating them into business processes, and ensuring that the outcomes continue to provide value over time.

3. Key aspects:
   - Plan deployment:
     - ✓ Decide how the model or insights will be used.
     - ✓ Determine the technical and business steps necessary for implementation
     - ✓ Choose between various deployment methods
   - Prepare Documentation and Reports:
     - ✓ Create detailed documentation about the model, data, methodology, and results
     - ✓ Provide reports for stakeholders explaining the project outcomes and how to interpret.
   - Implement Model or System:
     - ✓ Integrate the model into production systems, applications, or business processes
     - ✓ Ensure the model is stable, scalable, and properly integrated with existing workflows
   - Review:
     - ✓ Evaluate the success of the project in terms of meeting the business objectives
     - ✓ Gather feedback from stakeholders and users to assess the real-world impact of the deployment

# Methodologies in Data Mining

# Thankyou