Imputation Methods

Data Mining & Neural Networks

Dr. Wilmer Garzón

Director, Master's Program in Data Science
Department of Computer Engineering

Escuela Colombiana de Ingeniería
Universidade da Coruña

2025

Imputation Methods for Missing Data

- In data mining and neural networks, **handling missing data is crucial for building robust and accurate models**.
- This session will explore various imputation techniques, their advantages, limitations, and practical applications.

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

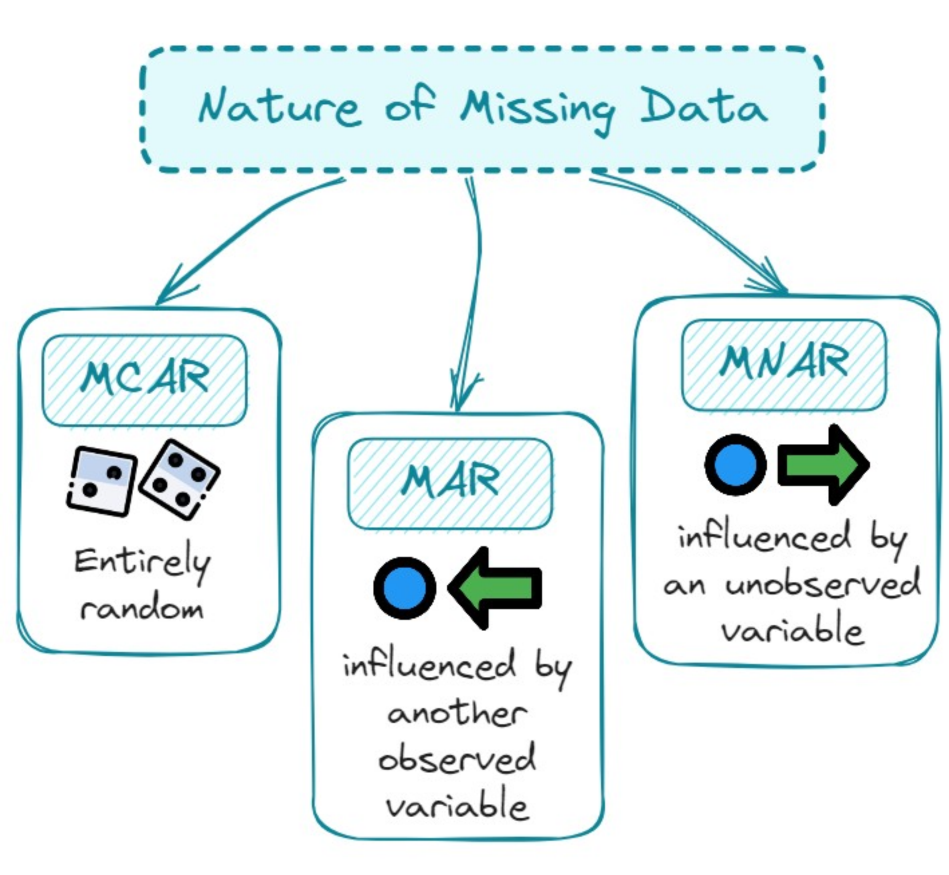
UNIVERSIDAD

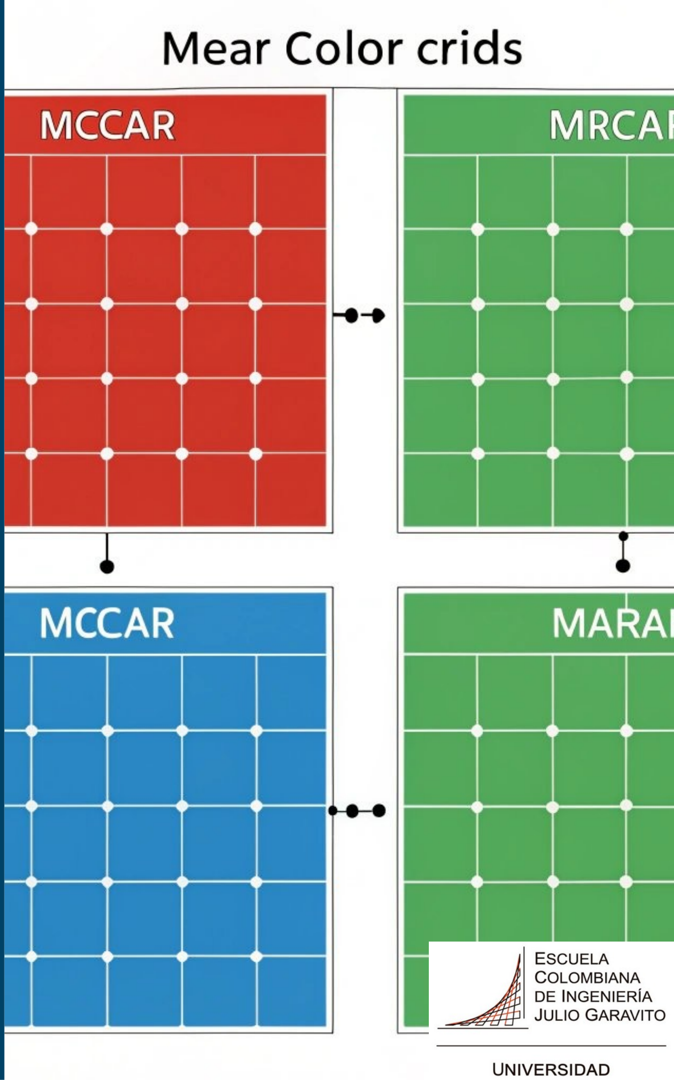
Introduction to Missing Data

- Missing data occurs when no **value is stored for a variable** in an observation.
- It can arise from data **entry errors, equipment malfunctions, or non-responses in surveys**.
- Ignoring **missing data can lead to biased results** and reduced model performance.

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD

Types of Missing Data



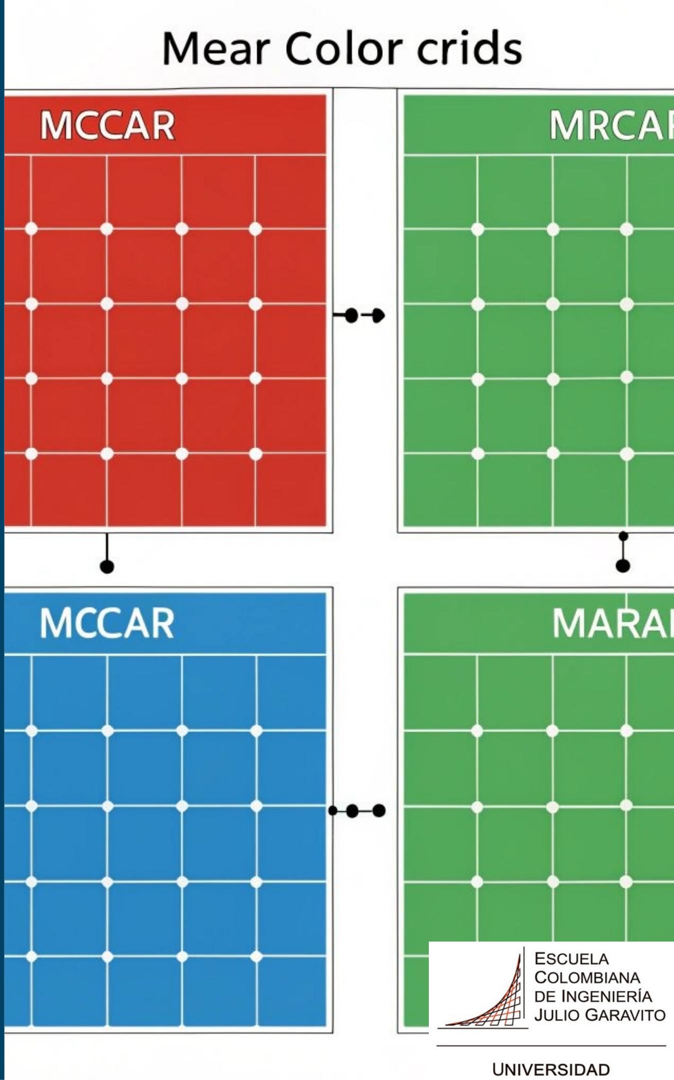


ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD

Types of Missing Data (1)

- **1: MCAR (Missing Completely at Random)**:
    - The missingness occurs **entirely at random** and is **unrelated to any values in the dataset**.
    - The missingness does **not depend on any variable**, neither the ones measured nor the ones missing.

- **Example**: Imagine a survey where **some questionnaires were lost in the mail randomly, with no connection to the respondents' answers** or characteristics.

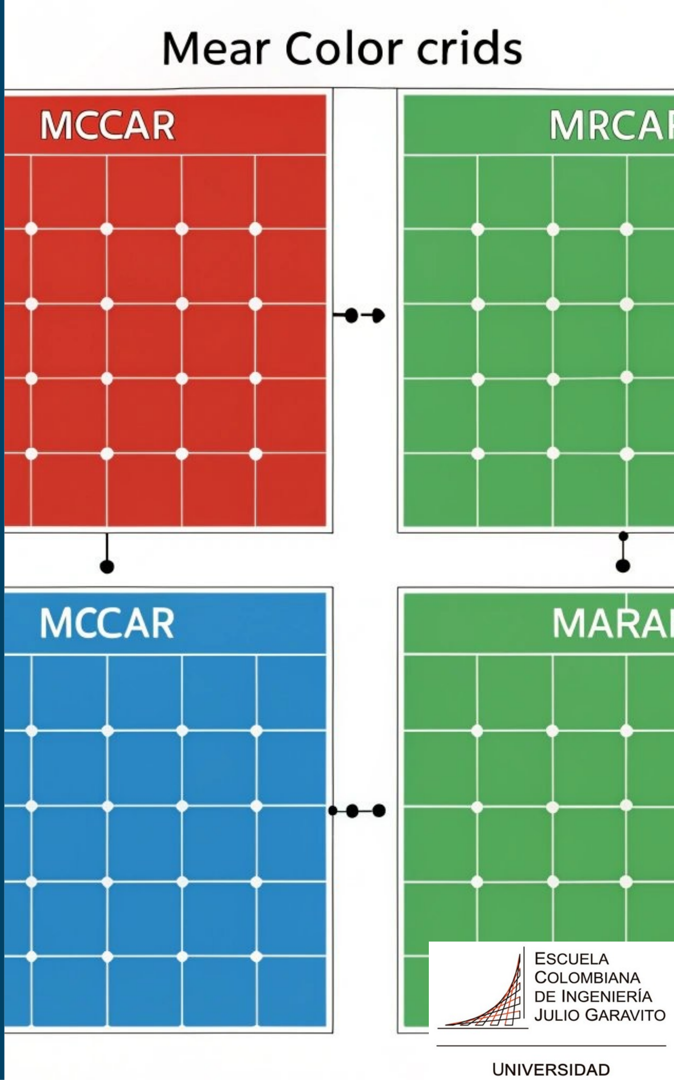


ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD

Types of Missing Data (2)

- **2: MAR (Missing at Random)**:
  - The **missingness is related to observed data**.
  - The probability of **a value being missing depends only on observed data**.
  - The missingness can be related to other variables that are observed in the dataset.

- **Example**: In a medical study, **older patients may be less likely to answer income questions**.
  - Here, **missing income data depend on age (observed)**, but not on the income values themselves.

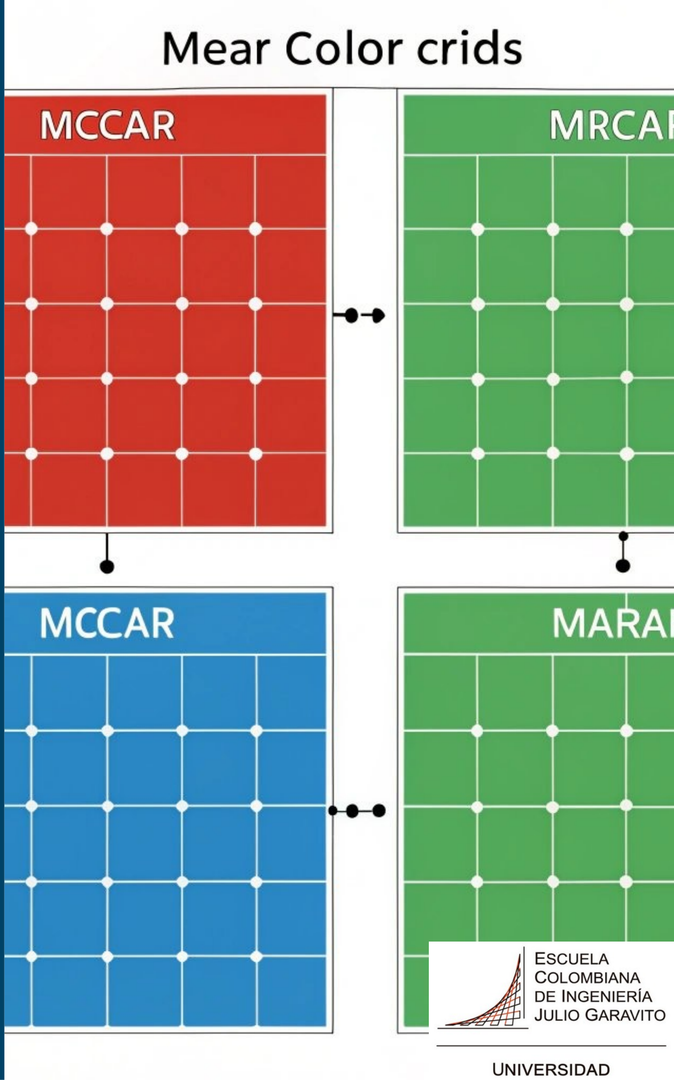


ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD

Types of Missing Data (3)

- **3: MNAR (Missing Not at Random)**:
    - **The missingness is related to unobserved data.**
    - Missingness is related to the value of the missing data.

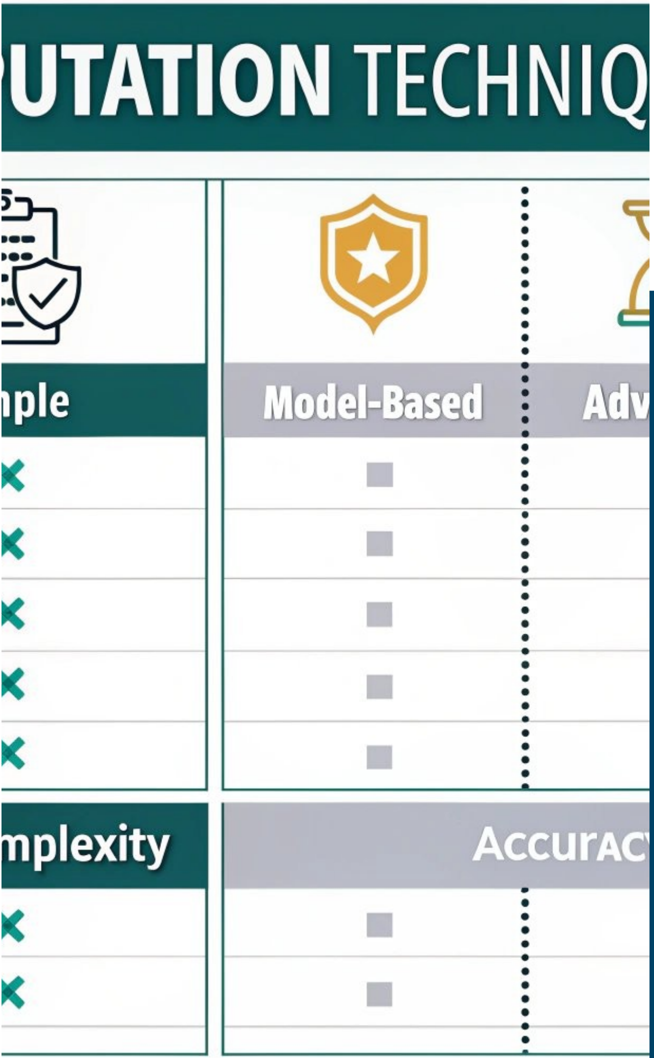
- **Example**: In a survey about income, **people with very high or very low incomes may choose not to report their income**
    - The missingness depends on the income value itself.



ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD

Why Imputation Matters

- Imputation allows us to retain all observations by **filling in missing values**, preserving statistical power, and reducing bias.



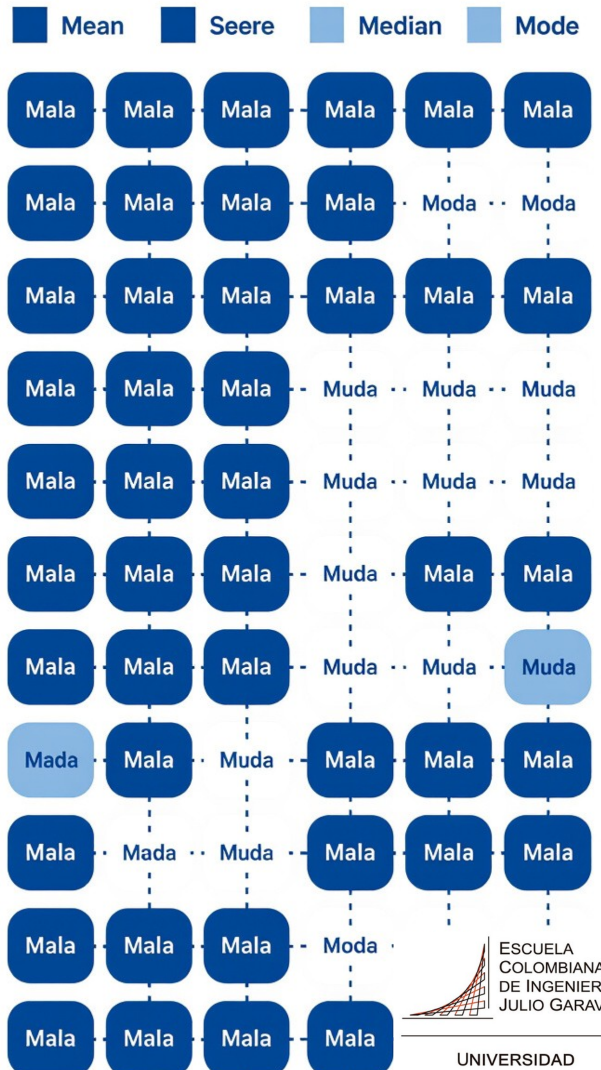
Overview of Imputation Techniques

- Imputation methods can be broadly categorized into:
  - **Simple Imputation** (mean, median, mode)
  - **Model-based Imputation** (regression, k-NN, MICE)
  - **Advanced Techniques** (deep learning, GANs) Each method has trade-offs in terms of complexity, accuracy, and computational cost.

Mean/Median/Mode Imputation

This is the simplest method where missing values are replaced with the **mean (for continuous)**, **median (for skewed)**, or **mode (for categorical)** of the variable.

- **Pros**: Easy to implement, fast.
- **Cons**: Ignores feature relationships, underestimates variance.
- **Example**: Replacing missing age values with the average age.



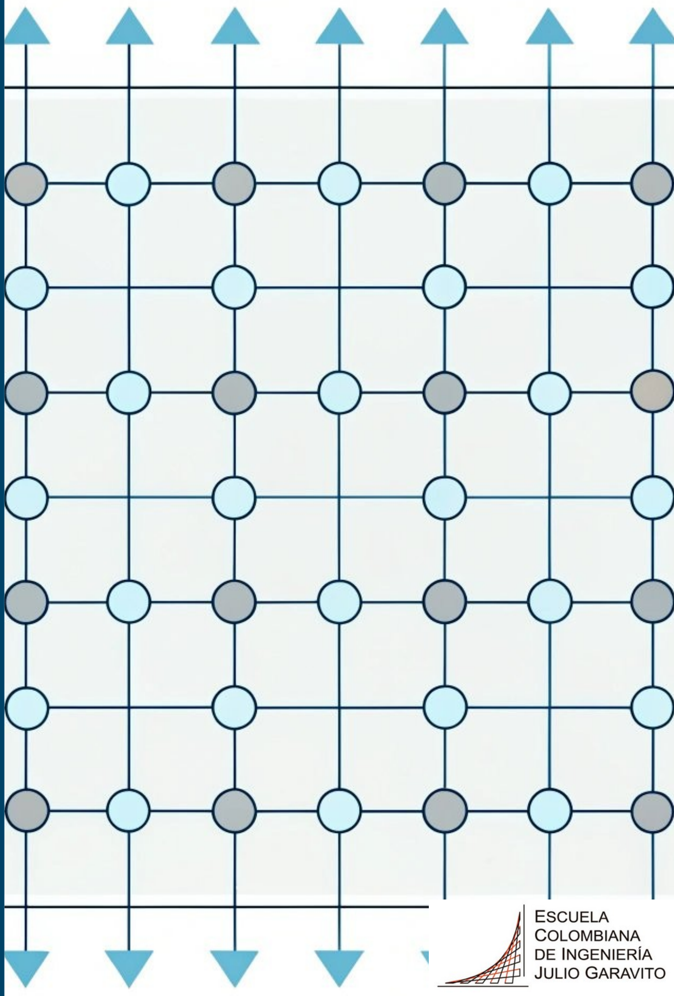
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

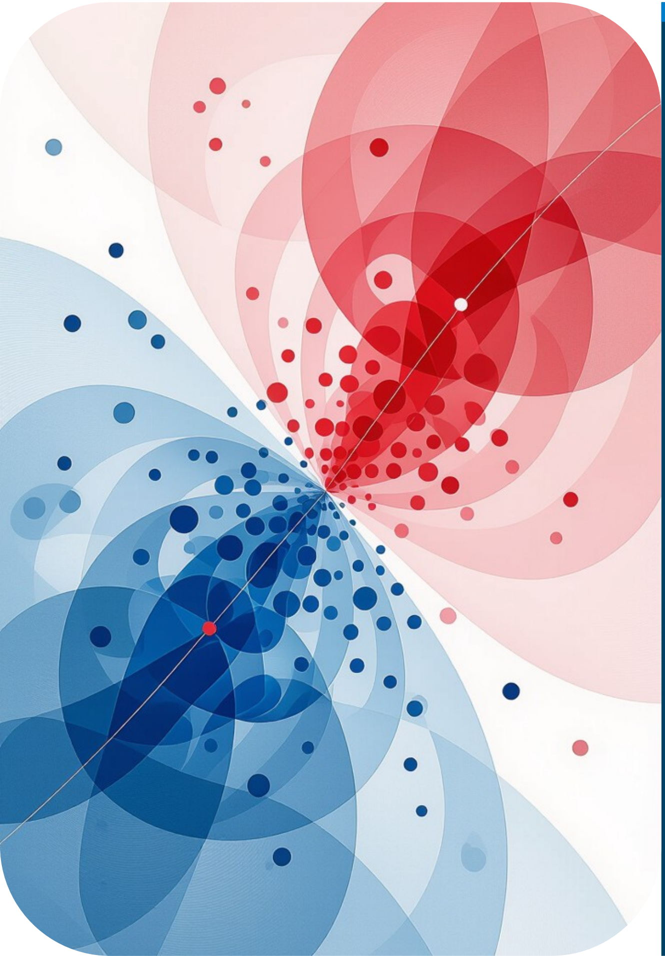
UNIVERSIDAD

Constant Value Imputation

A fixed value (e.g., -999 or “Unknown”) is used to fill missing entries.

- **Pros**: Useful for flagging missingness explicitly.
- **Cons**: Can introduce bias or mislead models if not handled properly. Often used in tree-based models that can treat such values separately.



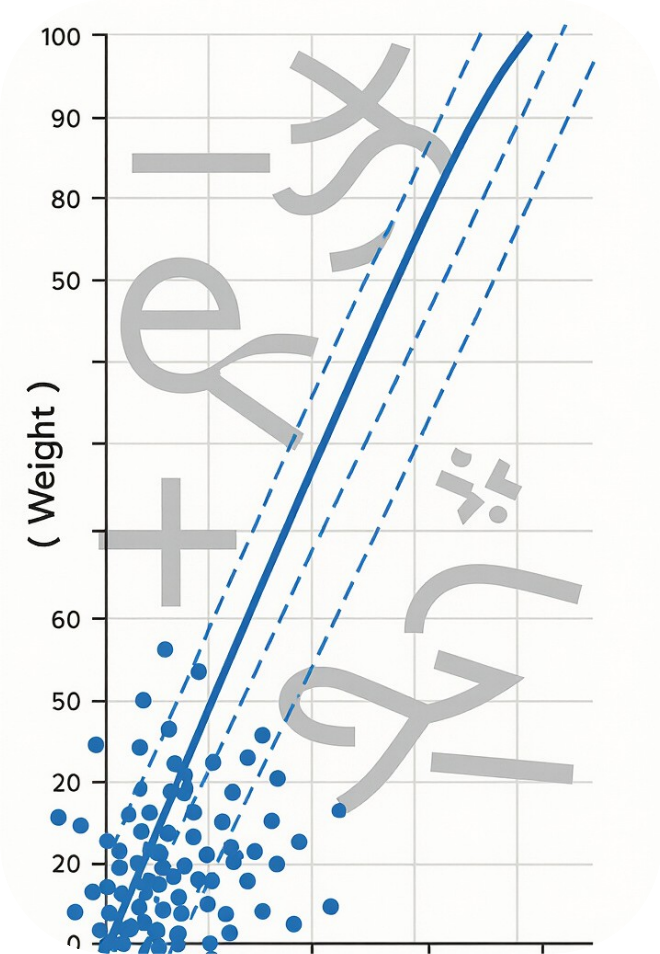


k-Nearest Neighbors (k-NN) Imputation

Imputes missing values using the average of the k most similar instances:

- **Pros**: Captures local structure and relationships.
- **Cons**: Computationally expensive, sensitive to distance metric.
- **Example**: Filling missing income based on similar individuals' profiles.

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD



Regression Imputation

Predicts missing values using a regression model trained on observed data:

- **Pros**: Utilizes relationships between variables.
- **Cons**: Can lead to overfitting, underestimates variability.
- **Example**: Predicting missing weight using height and age.

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD

Multiple Imputation by Chained Equations (MICE)

Each variable with missing values is imputed using a model based on other variables:

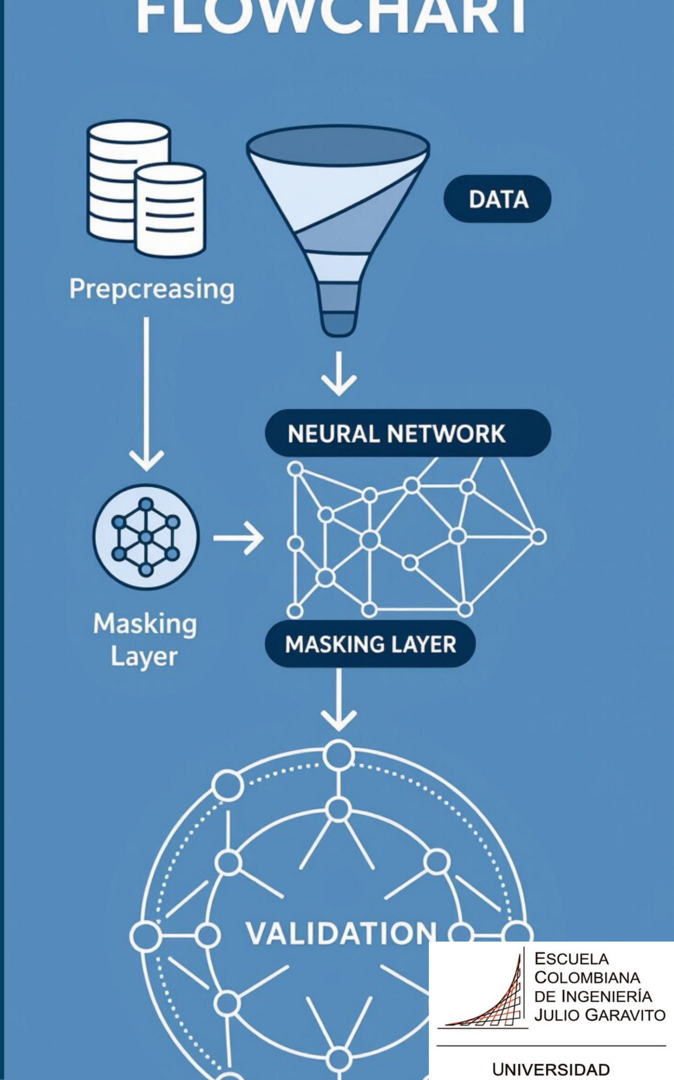
- **Pros**: Accounts for uncertainty, flexible with variable types.
- **Cons**: Computationally intensive, requires careful tuning. Widely used in medical and social sciences.

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD

Imputation in Neural Networks

Neural networks require complete data. Imputation can be done:

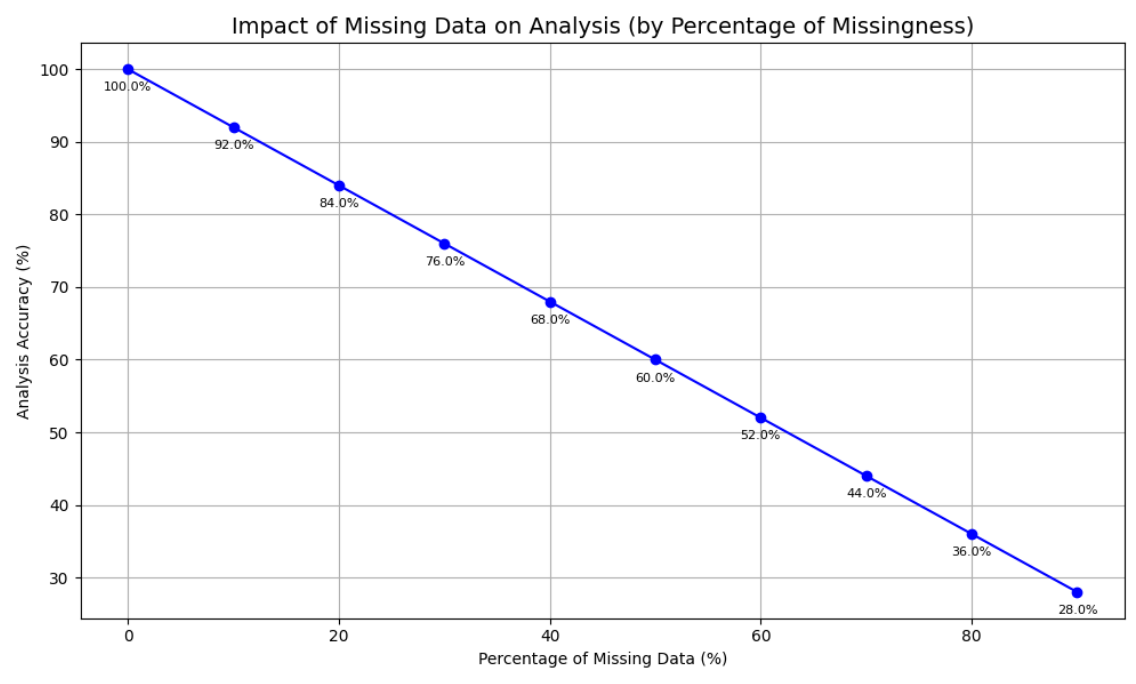
- **Preprocessing step** (before training)
- **Within the model** (e.g., masking layers)
- **Best Practices**: Normalize after imputation, avoid data leakage, validate with cross-validation.



ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD

Impact of Missing Data:

- **0-5%:** minimal impact.
- **5-15%:** Impact becomes noticeable; imputation methods are recommended to reduce bias and loss of power.
- **15-30%:** Missing data can seriously affect results; advanced imputation or modeling techniques should be applied.
- **>30%:** High risk of bias and unreliable conclusions.



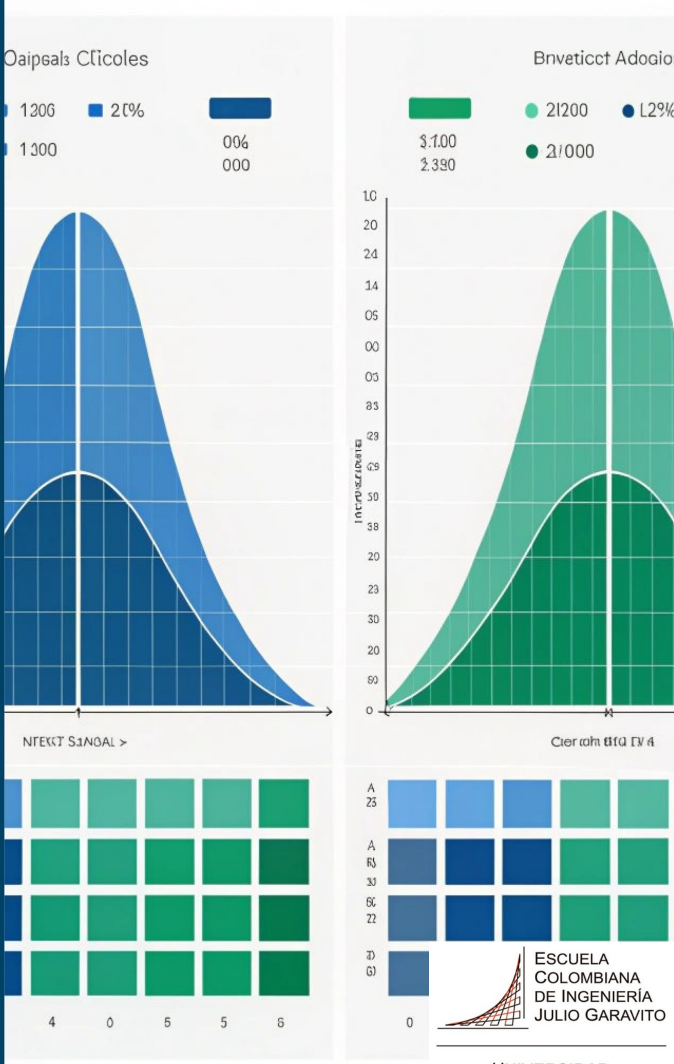
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

UNIVERSIDAD

Evaluation of Imputation Methods

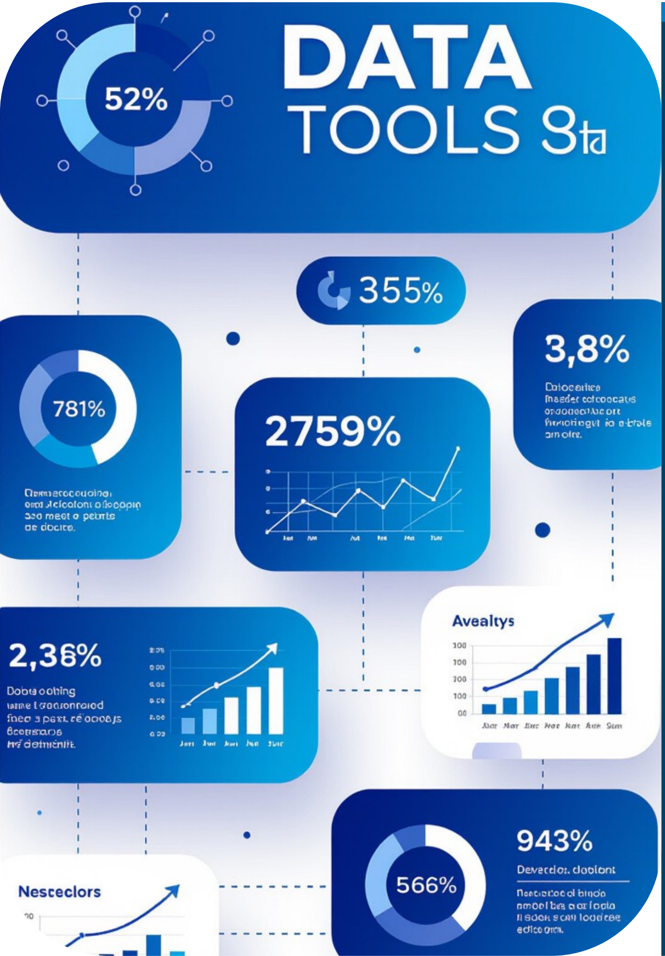
Evaluate imputation using:

- **RMSE/MAE** for numerical data
- **Classification accuracy** for categorical data
- **Visual inspection** (e.g., distribution plots) Use simulated missingness to benchmark methods.



ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

UNIVERSIDAD

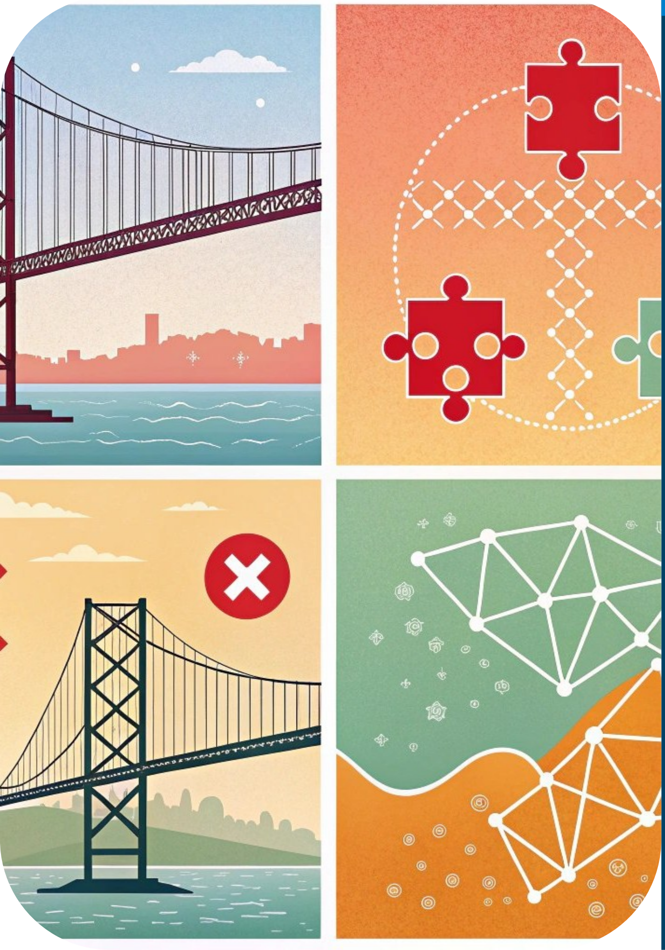


Tools and Libraries

Popular Python libraries for imputation:

- **pandas**: fillna()
- **scikit-learn**: SimpleImputer, KNNImputer
- **fancyimpute**: MICE, SoftImpute
- **PyTorch/TensorFlow**: For deep learning-based imputation
  Choose tools based on data size, type, and model requirements.

Escuela Colombiana de Ingeniería Julio Garavito
Universidad



Common Pitfalls

- Imputing test data with training statistics
- Ignoring the mechanism of missingness
- Overfitting with complex imputation models
- Not validating imputation impact on model performance

Always document and justify your imputation strategy.

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

UNIVERSIDAD

Summary

- Missing data is a critical issue in data mining and neural networks.
- Imputation methods range from simple to advanced.
- Choice depends on data type, missingness mechanism, and model goals.
- Evaluate and validate imputation strategies rigorously.

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
UNIVERSIDAD

INTER
NATIONAL
SUMMER
SCHOOL

UNIVERSIDADE DA CORUÑA

ESCUELA
COLOMBIANA
DE INGENIERÍA
JULIO GARAVITO

UNIVERSIDAD