أكاديمية طويق
Tuwaiq Academy

AI Diploma

أكاديمية طويق
Tuwaiq Academy

Machine Learning Algorithms and Applications

Semester One | Course Two

ROADMAP

أكاديمية طويق
Tuwaiq Academy

ML Road Map

Unit 1: Basic Data processing methods and regression

Data Preprocessing

- Missing Data
- Categorical Data
- Template For Preprocessing Data (General Steps)

Regression

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial regression

Unit 2: Advance Regression Algorithms

Regression Algorithms

- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression
- Evaluating Regression Model
- Regularisation Methods

Classification in Regression

- Logistic Regression

Unit 3: Classification

Classification

- K-Nearest Neighbors (K-NN)
- Support Vector Machine (SVM)
- Kernel SVM
- Naive Bayes
- Decision Tree Classification
- Random Forest Classification
- Evaluating Classification Model

Unit 4 : Clustering & Dimensionality Reduction

Clustering

- K-Means Clustering
- Hierarchical Clustering

Dimensionality Reduction

- Principal Component Analysis

Unit 5: Model Selection & Boosting

- Model Selection
- XGBoost

أكاديمية طويق
Tuwaiq Academy

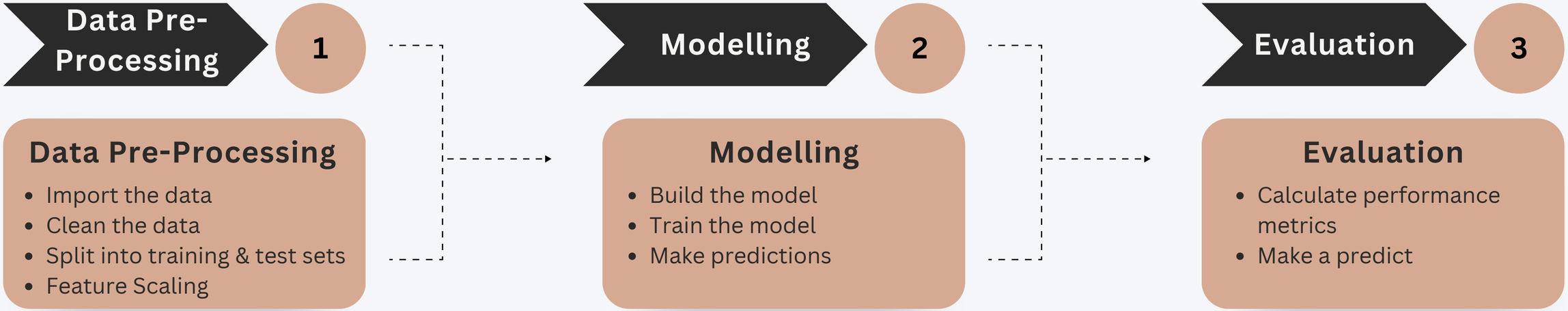
Machine Learning Algorithms and Applications

Semester One | Course Two

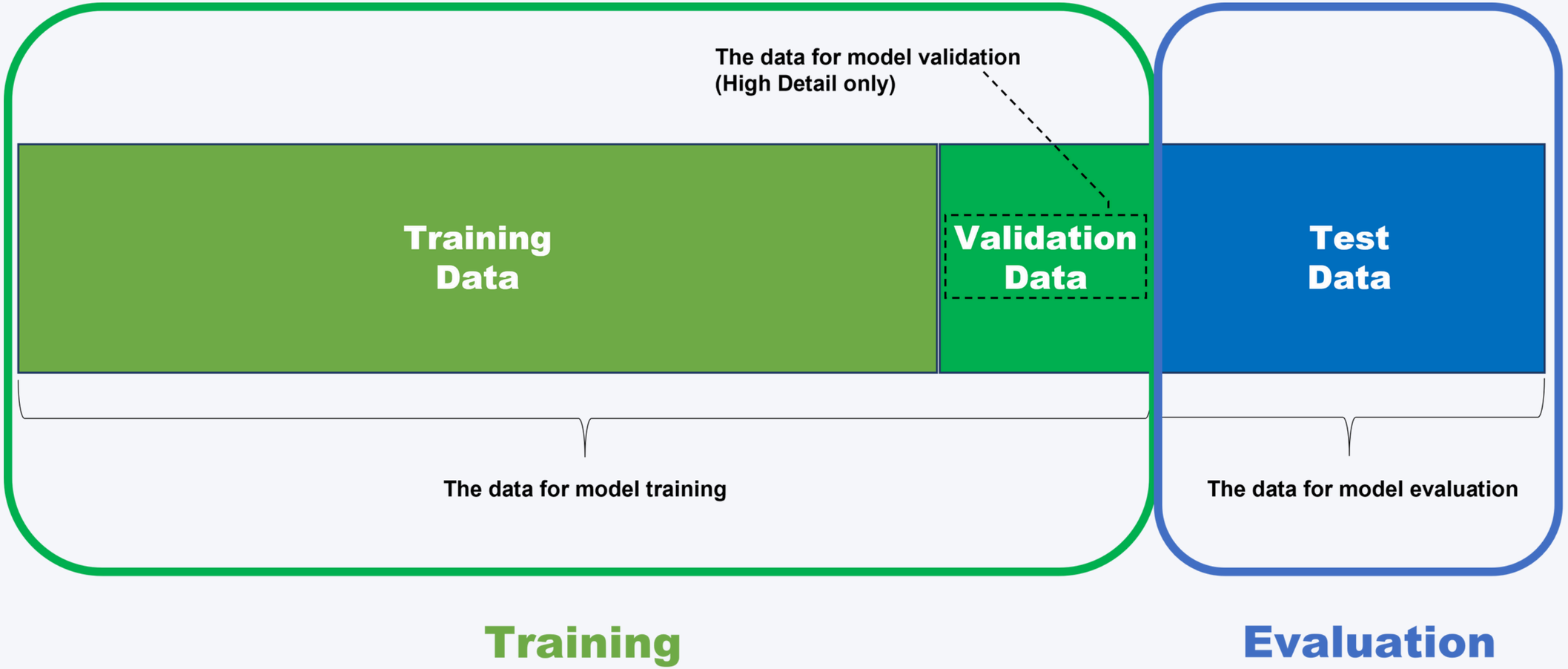
Unit 1 : Basic Data Processing Methods and Regression

Data Preprocessing

The Machine Learning Process



Training Set & Test Set



https://support.cognex.com/docs/deep-learning_311/web/EN/deep-learning/Content/deep-learning-Topics/training-blue-read/prepare-train-test.htm?TocPath=Training%20Blue%20Read%7C_____8

Feature Scaling

أكاديمية طويق
Tuwaiq Academy

- **Feature scaling** is a technique to standardize or normalize the range of independent variables (**features**).

- It ensures that no feature dominates the learning process due to larger numerical values.

Examples of why it's needed:

- Height (cm) and weight (kg) might have very different ranges.

X1	X2	X3	X4
$ 179.43	56.784	34.6181	3.55
$ 641.87	62.054	47.7306	1.692
$ 556.30	64.13	55.596	1.559
$ 578.47	63.377	52.7121	1.679
$ 591.16	61.553	46.1315	1.984
$ 242.03	58.29	39.2952	2.942
$ 364.66	59.93	42.4628	2.494
$ 190.68	57.271	36.2725	3.419
$ 547.23	63.763	54.1971	1.634
$ 359.69	59.375	41.5105	2.128
$ 438.08	60.484	43.493	2.47
$ 637.17	62.525	49.428	1.725

Feature Scaling

أكاديمية طويق
Tuwaiq Academy

Why Do We Need It?

- **Improves Model Performance:** Many ML algorithms rely on distance-based calculations (e.g., KNN, SVM, Gradient Descent).
- **Speeds Up Convergence:** For optimization algorithms, scaling helps them reach a solution faster.
- **Avoids Bias:** Prevents large values from dominating smaller ones.

Feature Scaling

أكاديمية طويق
Tuwaiq Academy

Normalization

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

$[0 , 1]$

Normalization

Rescales data between 0 and 1.

Suitable for: Algorithms sensitive to absolute magnitudes (e.g., KNN).).

Standardization

$$X' = \frac{X - \mu}{\sigma}$$

$[-3 , +3]$

Standardization

Rescales data to have a mean of 0 and a standard deviation of 1.

Suitable for: Gaussian-distributed features (e.g., Logistic Regression)

Feature Scaling

أكاديمية طويق
Tuwaiq Academy

Normalization

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

$[0 , 1]$

Normalization

Rescales data between 0 and 1.

Suitable for: Algorithms sensitive to absolute magnitudes (e.g., KNN).).

Standardization

$$X' = \frac{X - \mu}{\sigma}$$

$[-3 , +3]$

Standardization

Rescales data to have a mean of 0 and a standard deviation of 1.

Suitable for: Gaussian-distributed features (e.g., Logistic Regression)

Feature Scaling

Choosing the Right Scaling Method

- **Normalization**
  - Use when features have different ranges but need to be scaled between **0 and 1**.
  - **Example**: Min-Max Scaling.

- **Standardization**
  - Use when data needs a Gaussian distribution or involves **outliers**.

Feature Scaling

Example

70,000 $	45 yrs	**1**	**1**
10,000	1		
60,000 $	44 yrs	**0.444**	**0.8**
8,000	4		
52,000 $	40 yrs	**0**	**0**

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Normalization

Categorical Data

What is Categorical Data?

- Data that represents categories or groups.

Types:

- **Ordinal**: Has an order (e.g., small, medium, large).
- **Nominal**: No order (e.g., colors: red, green, blue).

Why Handle Categorical Data?

- Machine learning models require numerical data.

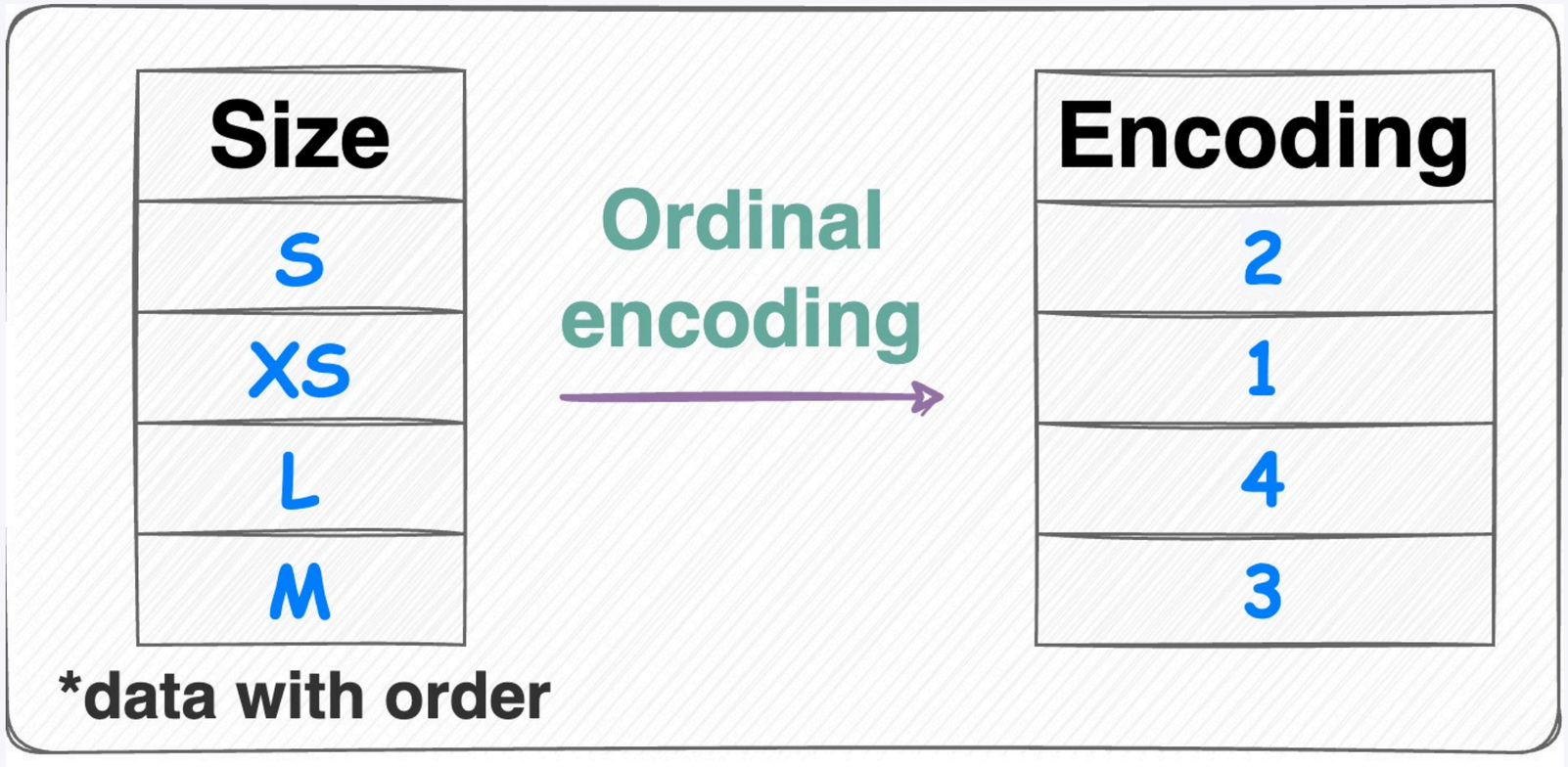
Example:

- Raw: ["Red", "Green", "Blue"]
- Encoded: [1, 2, 3] or One-Hot Encoding.

Categorical Data

How to deal with Categorical Data?

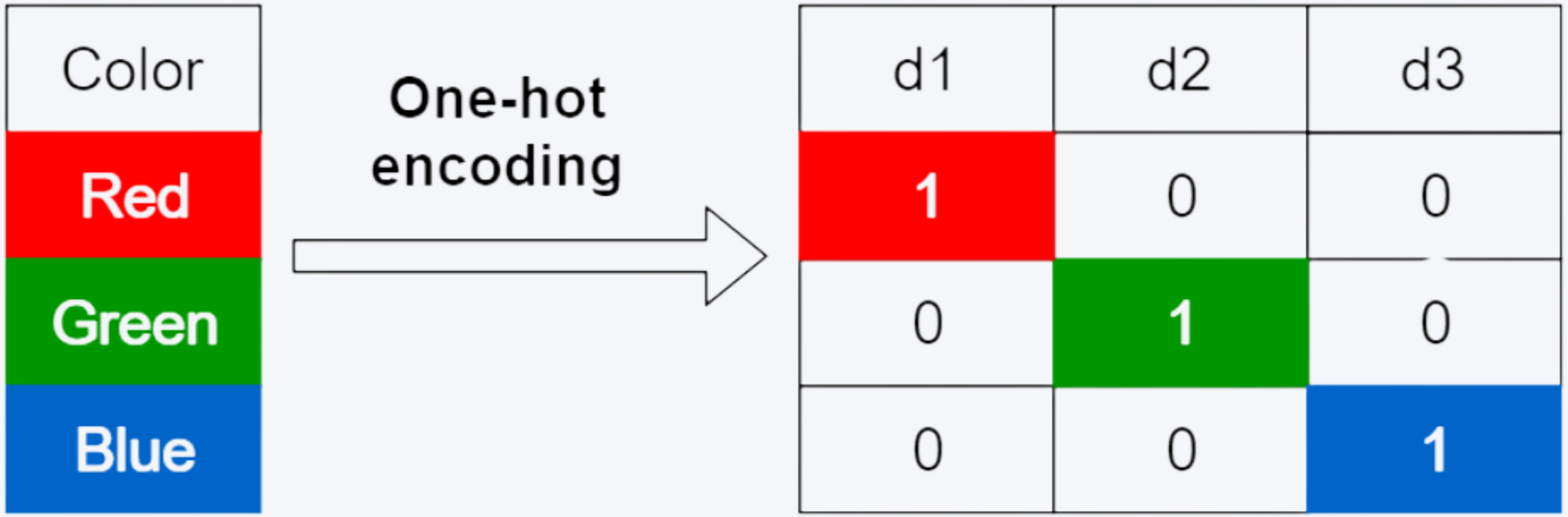
Ordinal Encoding



Categorical Data

How to deal with Categorical Data?

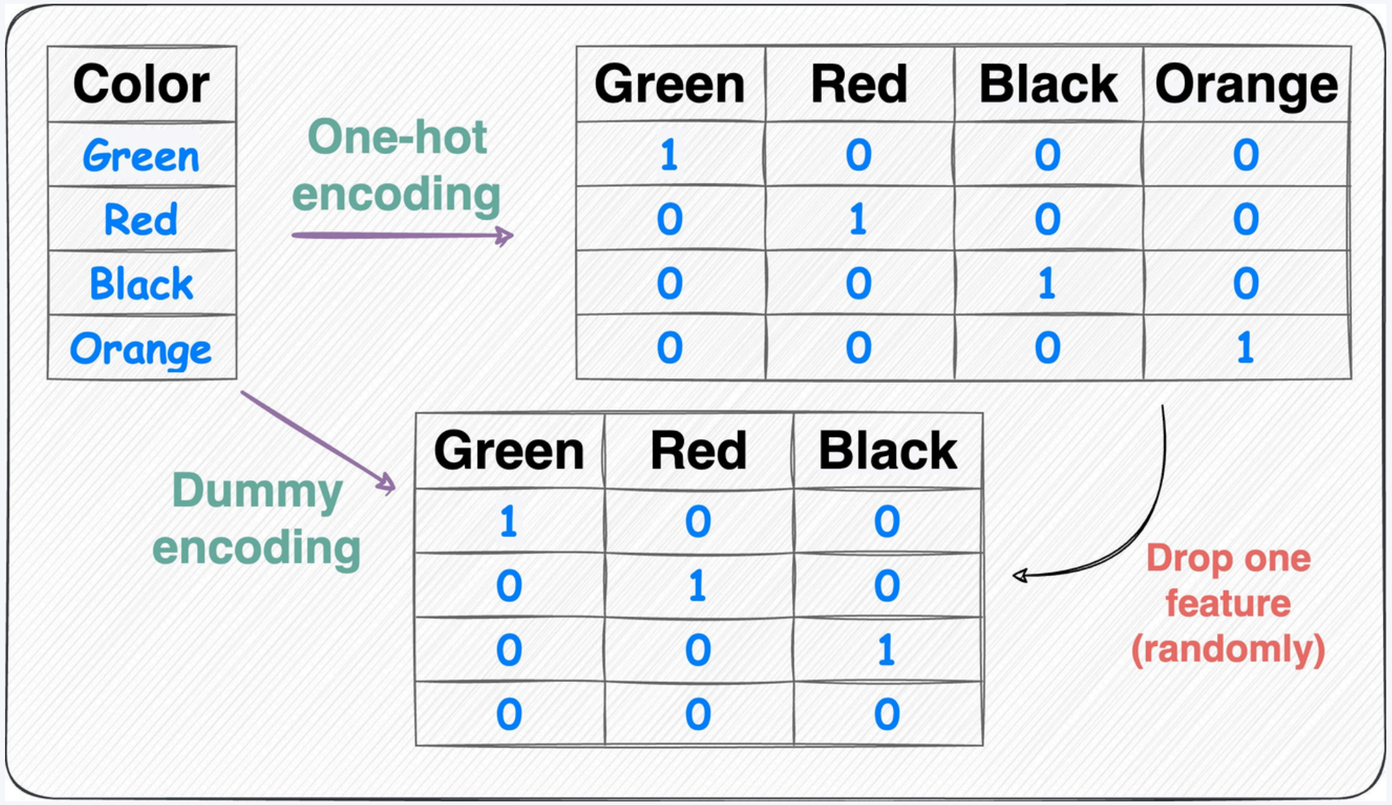
One Hot Encoding



Categorical Data

How to deal with Categorical Data?

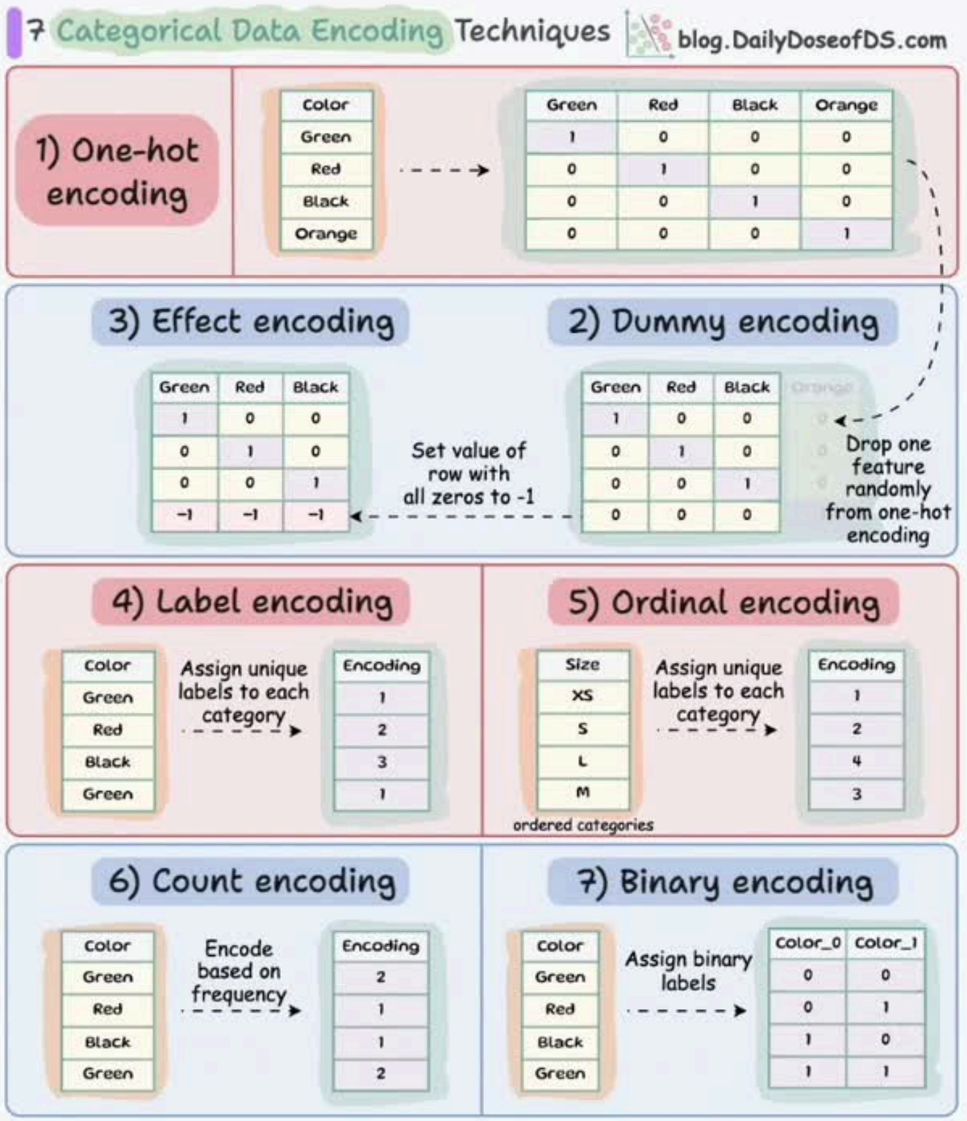
Dummy Encoding



Categorical Data

How to deal with Categorical Data?

Other Type Encoding



https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

Handling Missing Data

What is Missing Data?

Missing or incomplete values in the dataset.

Types of Missing Data:

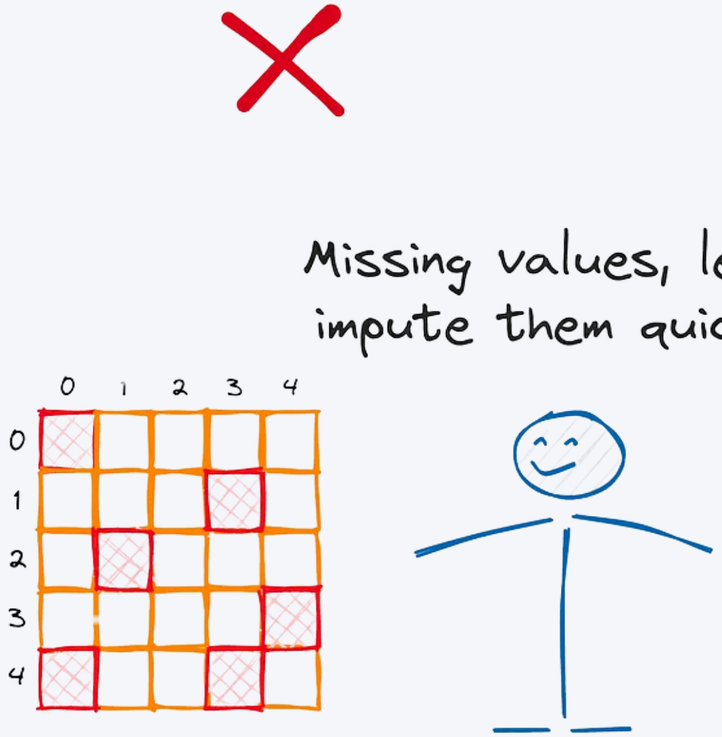
- Missing Completely at Random (MCAR).
- Missing at Random (MAR).
- Not Missing at Random (NMAR).

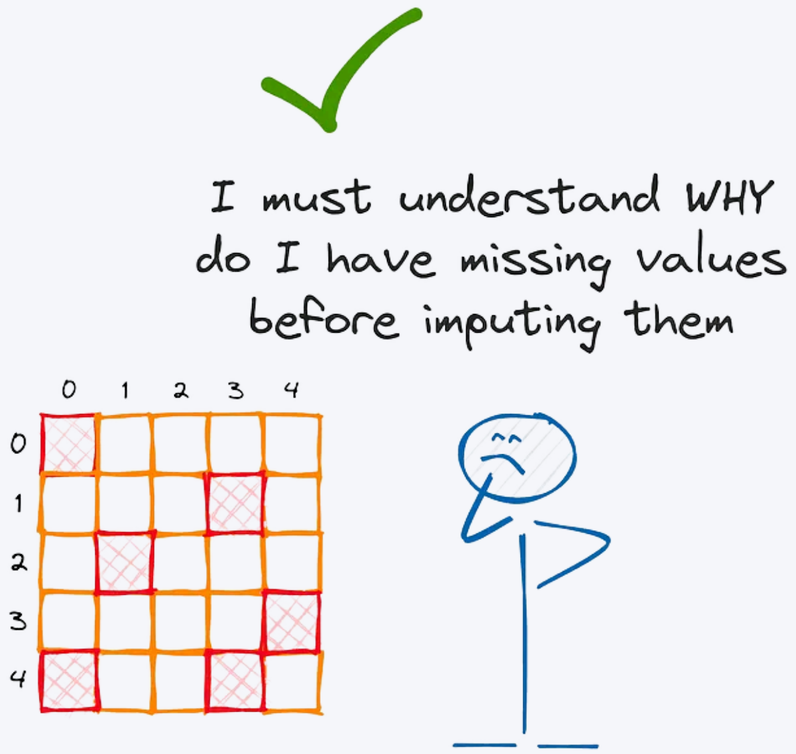
Challenges:

- Leads to biased models or errors in predictions.

Handling Missing Data

Why we have Missing Data?





https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

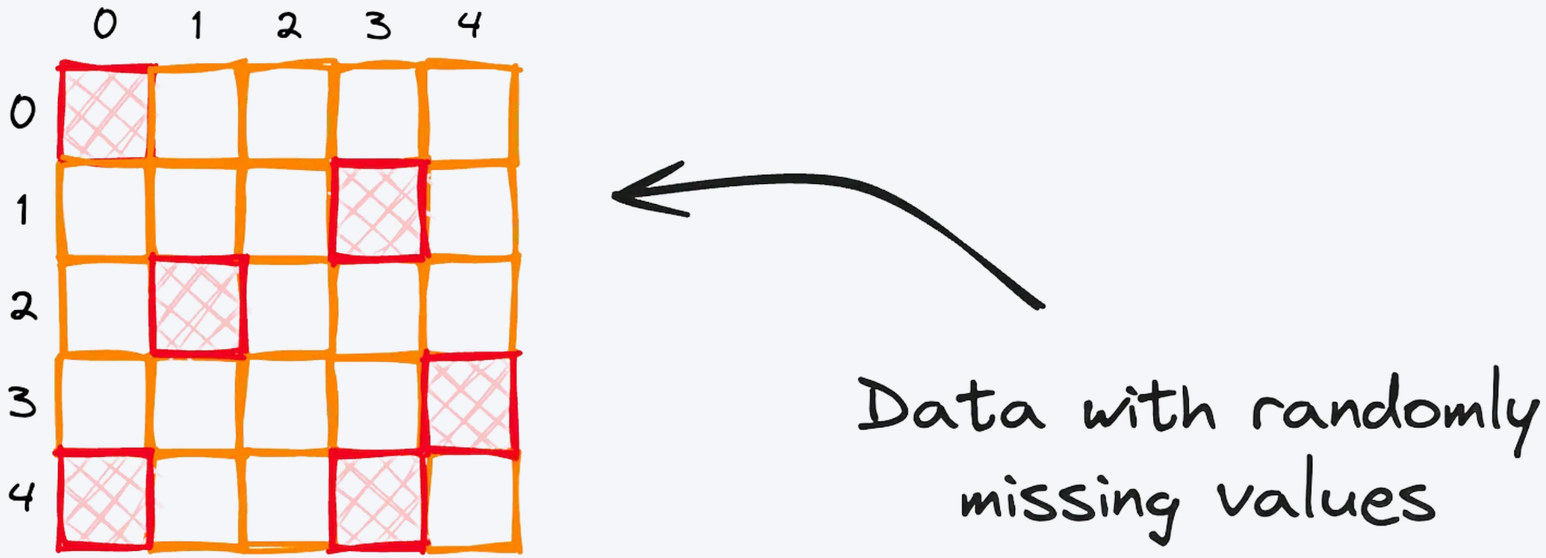
Handling Missing Data

أكاديمية طويق
Tuwaiq Academy

1) Missing completely at random (MCAR)

MCAR is a situation in which the **data is genuinely missing at random and has no relation to any observed or unobserved variables.**

In other words, the missing data points follow no recognized pattern.



https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

Handling Missing Data

أكاديمية طويق
Tuwaiq Academy

1) Missing completely at random (MCAR)

Unrealistic Assumption: MCAR is rarely realistic in real-world datasets because missing data is often influenced by observed or unobserved factors.

Influencing Factors: Missing data may arise due to:

- Human behavior (e.g., omitting sensitive information).
- Survey administration errors.
- External events influencing responses.

Selective Missingness: Certain groups or individuals may be more likely to leave responses blank, leading to patterns in missingness.

Need for Context: Assuming MCAR requires:

- Understanding the full data collection process.
- Consulting domain experts.
- Collaboration between data scientists and data engineers.

Practical Approach: If MCAR seems reasonable after analysis or expert input, simple imputation techniques can be used for handling missing values.

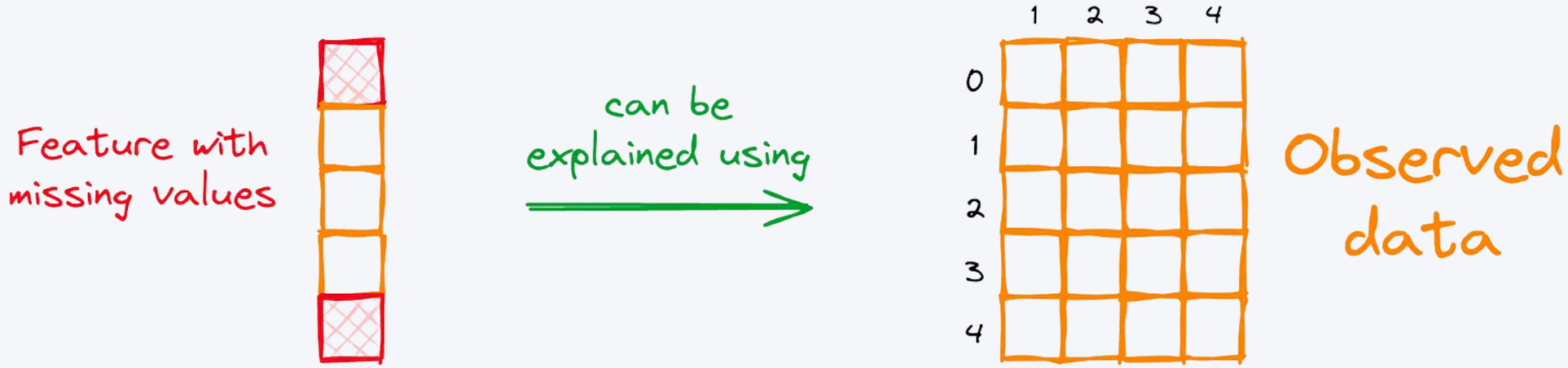
https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

Handling Missing Data

أكاديمية طويق
Tuwaiq Academy

2) Missing at random (MAR)

MAR is a situation in which the missingness of one feature can be explained by other observed features in the dataset.



https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

Handling Missing Data

أكاديمية طويق
Tuwaiq Academy

2) Missing at random (MAR)

- Missing at Random (MAR) assumes that the **probability of missing data is related to observed features but not to the missing feature itself.**

Practical Observation: MAR is more commonly seen in real-world datasets compared to MCAR.

- Missingness can be estimated using available data, allowing for accurate imputation through statistical methods.
- To identify MAR, examine if the probability of missingness changes based on other observed features.

- **Example**: In an academic survey, students with higher grades might avoid reporting study hours, showing a relationship between grades (observed feature) and missingness.

Imputation Techniques for MAR:

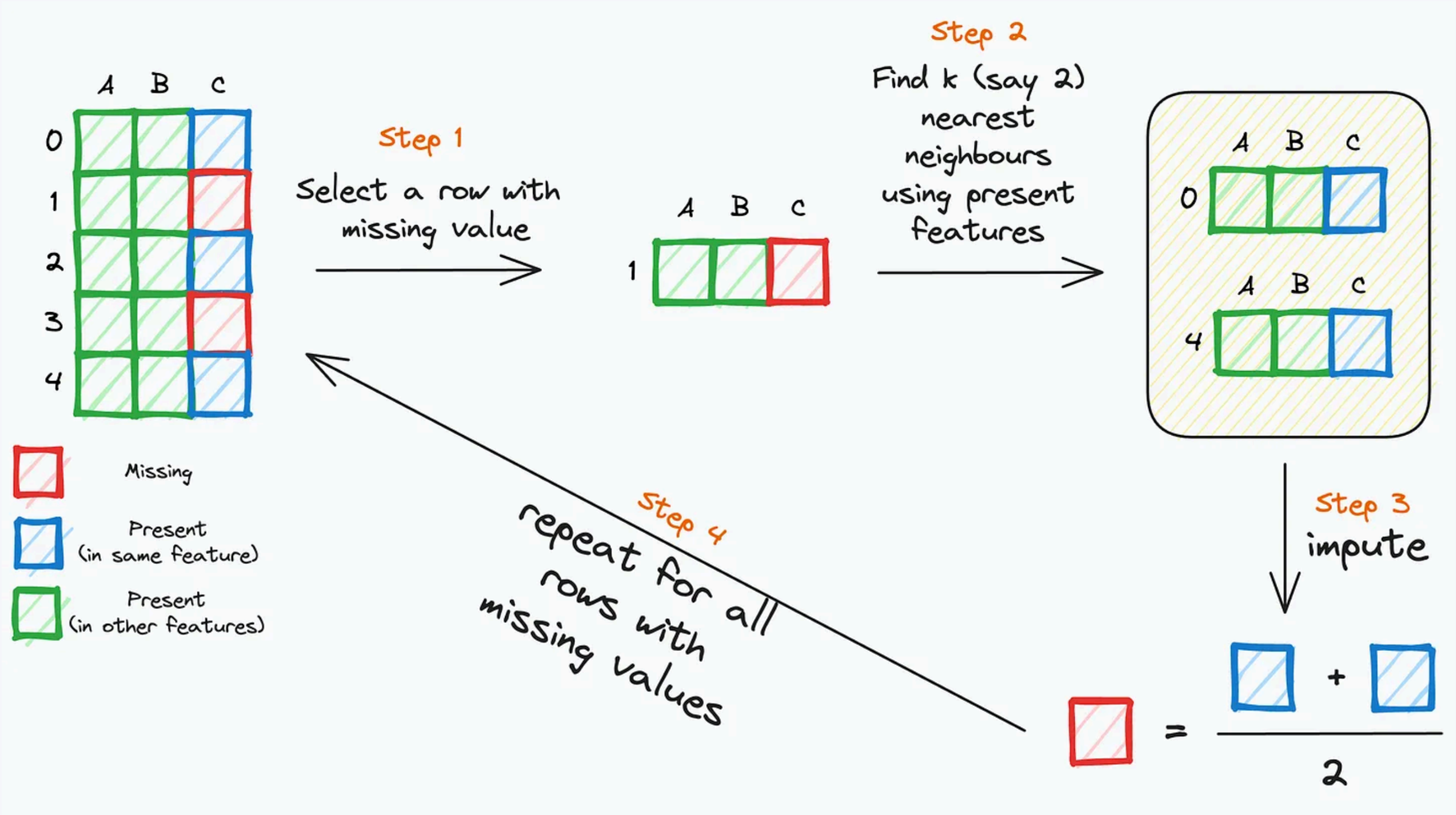
- kNN Imputation
- Miss Forest

These methods leverage observed features to fill in missing values effectively.

https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

Handling Missing Data

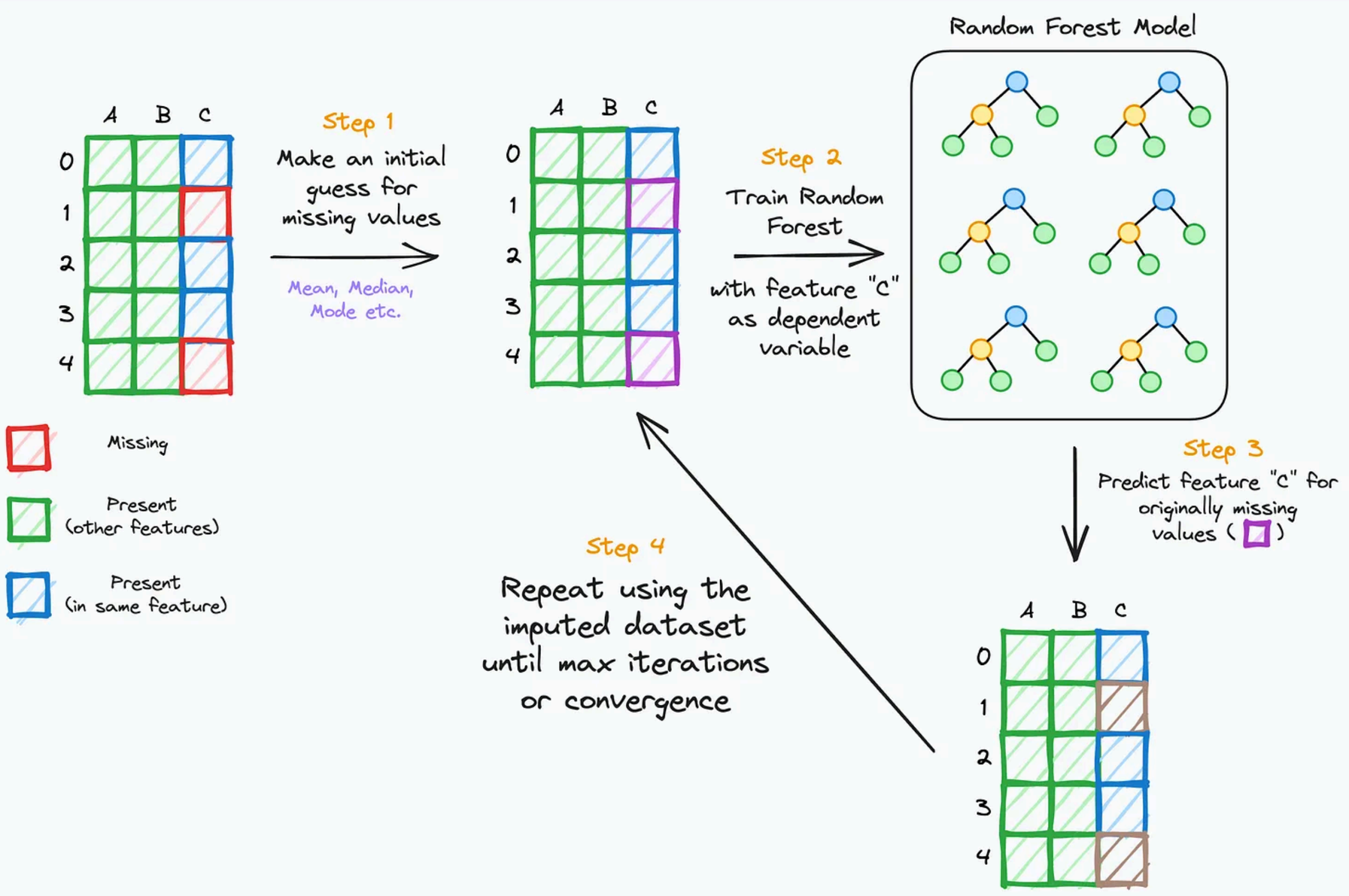
kNN Imputation



https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

Handling Missing Data

Miss Forest



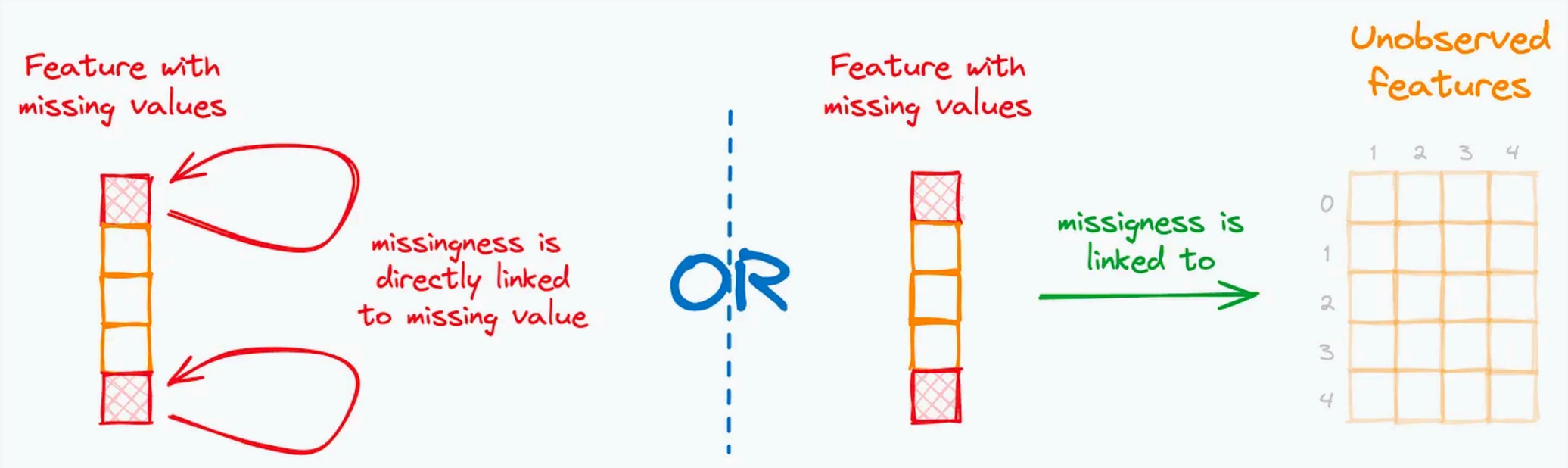
https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

Handling Missing Data

أكاديمية طويق
Tuwaiq Academy

3) Missing not at random (MNAR)

- **MNAR** is the most complicated situation of all three.
- In MNAR, missingness is either attributed to the missing value itself or the feature(s) that we didn't collect data for.



https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

Handling Missing Data

أكاديمية طويق
Tuwaiq Academy

3) Missing not at random (MNAR)

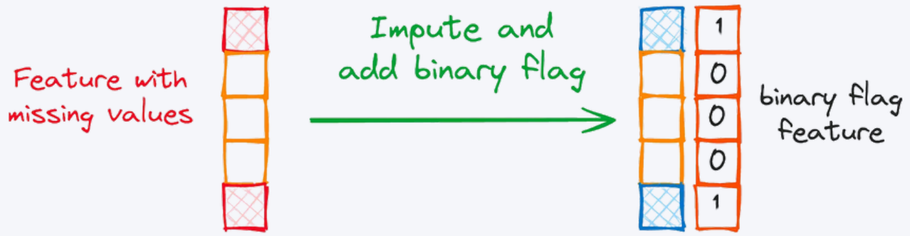
- **MNAR is the most complicated situation of all three.**
- **In MNAR, missingness is either attributed to the missing value itself or the feature(s) that we didn't collect data for.**

Unlike MCAR (no pattern) or MAR (related to observed features), MNAR has a clear missingness pattern tied to the missing variable.

Example: In a health survey, individuals with high stress levels might avoid disclosing their stress level due to stigma, creating a non-random missingness pattern.

Challenge: Since missingness is dependent on the missing variable, it's difficult to address without collecting additional data or having domain expertise.

Preserving Missingness Patterns: Add a binary indicator feature to flag whether a value was imputed. This allows machine learning algorithms to recognize and learn from the missingness pattern.



https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

Data Preprocessing Template

أكاديمية طويق
Tuwaiq Academy

Steps to Follow:

- Load the data.
- Check for missing data and handle it.
- Encode categorical data.
- Split dataset into training and test sets.
- Feature scale numerical data.

Input Raw Data → Handle Missing Data → Encode Categories → Train/Test Split → Scale Features → ML Algorithm

https://blog.dailydoseofds.com/p/7-categorical-data-encoding-techniques

أكاديمية طويق
Tuwaiq Academy

Hands-On Code

Data Preprocessing Template