Machine learning

Mohammad-Reza A. Dehaqani

dehaqani@ut.ac.ir

Understanding Data Types

- Structured vs. Unstructured Data
- Numeric vs. Categorical Data
- Time-Series Data, Text Data, Image Data
- Labeled vs. Unlabeled Data

Structured Data

Structured data refers to information that is highly organized and is assigned to fixed fields in a database, such as rows and columns

Characteristics

- Organized in **rows and columns**.
- Follows a **schema** or data model.
- Easily accessible and queryable.

Examples:

- Customer information stored in a CRM system (name, email, purchase history)
- Financial data stored in accounting software (transaction amount, date, account number)

Unstructured Data

Data that does not have a predefined structure or organization. Typically stored in **raw form**.

Characteristics

- Lacks a structured format.
- Cannot be easily organized into rows/columns.
- Requires advanced processing techniques (e.g., Natural Language Processing, AI) to analyze.

Examples:

- Text files
- Emails
- Social media posts
- Audio and video files
- Customer reviews

Overview of Data Variable Types

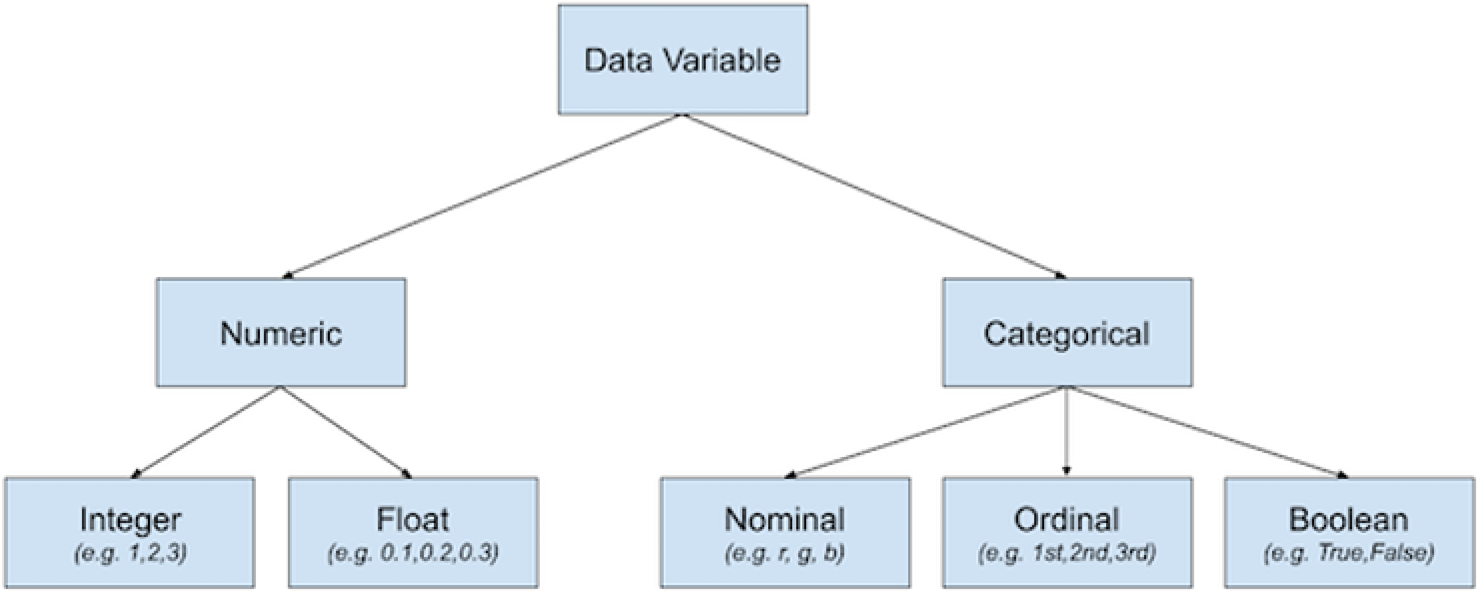


Figure 1: Overview of Data Variable Types.

Numeric vs. Categorical Data

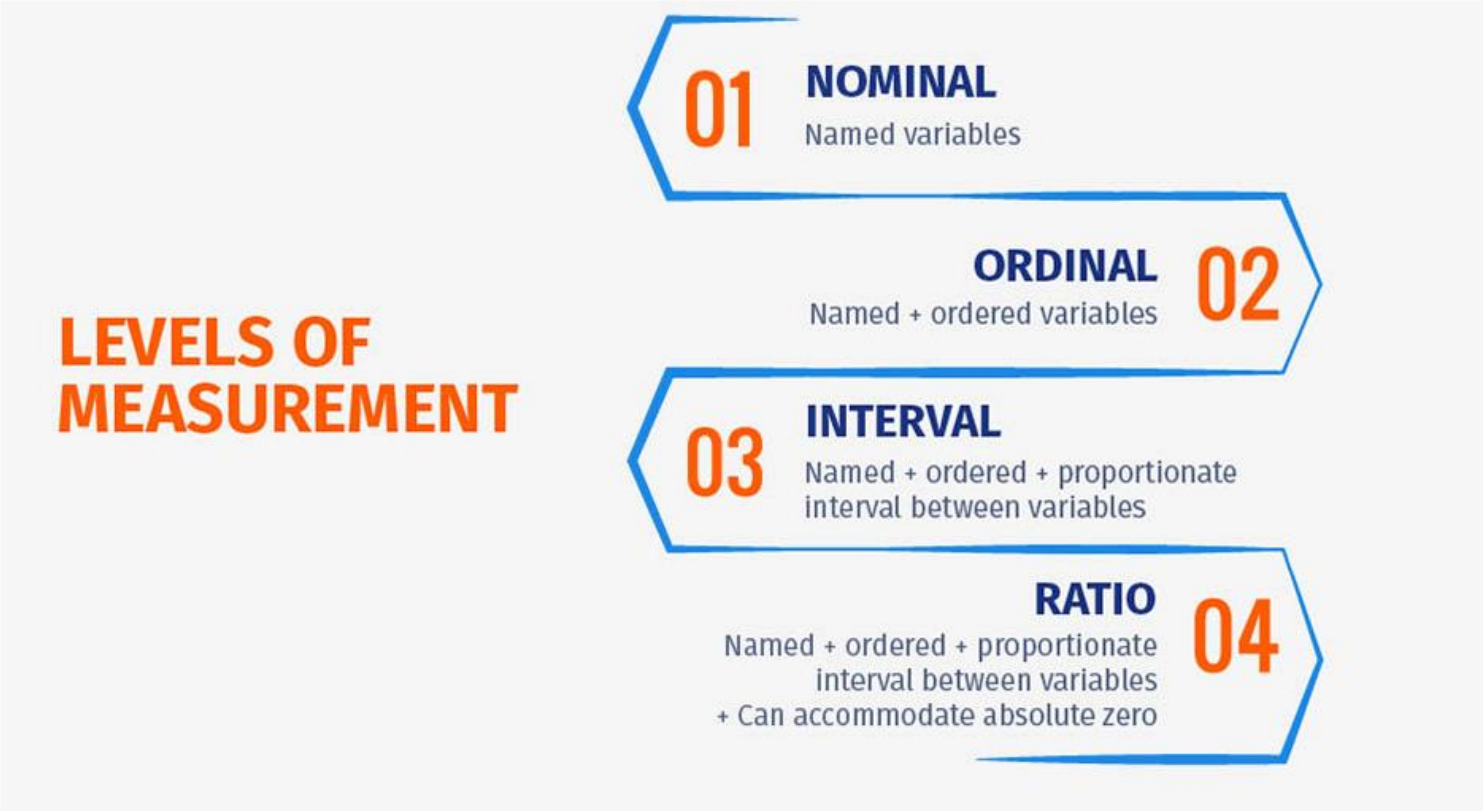
Numeric Data: Data that represents quantities and is often involved in mathematical operations.

- **Types**:
  - **Continuous Data(Float)**: Can take any value within a range (e.g., temperature).
  - **Discrete Data(Integer)**: Takes specific, distinct values (e.g., number of children).

Categorical Data: Data representing categories or labels.

- **Types**:
  - **Nominal**: Categories with no inherent order (e.g., colors).
  - **Ordinal**: Categories with an order (e.g., ratings: poor, fair, good).
  - **Boolean**: Values True and False.

Type of data	Nominal	Ordinal	Interval	Ratio
The sequence of variables is established	–	Yes	Yes	Yes
Mode	Yes	Yes	Yes	Yes
Median	–	Yes	Yes	Yes
Mean	–	–	Yes	Yes
Difference between variables can be evaluated	–	–	Yes	Yes
Addition and Subtraction of variables	–	–	Yes	Yes
Multiplication and Division of variables	–	–	–	Yes
Absolute zero	–	–	–	Yes



Interval-valued variables

- Standardize data
  - Calculate the **mean absolute deviation**:

$$s_f = \frac{1}{n}\left(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|\right)$$

where $m_f = \frac{1}{n}\left(x_{1f} + x_{2f} + \ldots + x_{nf}\right)$.

  - Calculate the standardized measurement (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation **is more robust** than using **standard deviation**

Similarity and Dissimilarity Between Objects

- **Distances** are normally used to measure the **similarity** or **dissimilarity** between two data objects
- Some popular ones include: ***Minkowski distance***

$$d(i,j) = \sqrt[q]{\left(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_d} - x_{j_d}|^q\right)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{id})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jd})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is **Manhattan distance**

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_d} - x_{j_d}|$$

Similarity and Dissimilarity Between Objects

- *If q = 2, d* is Euclidean distance

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_d}-x_{j_d}|^2)}$$

- Properties
  - $d(i,j) \geq 0$
  - $d(i,i) = 0$
  - $d(i,j) = d(j,i)$
  - $d(i,j) \leq d(i,k) + d(k,j)$
- Also, one can use **weighted distance**, **Pearson correlation coefficient**, or other dissimilarity measures

Binary Variables

- A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Distance measure for symmetric binary variables:

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- Distance measure for asymmetric binary variables:

$$d(i,j) = \frac{b+c}{a+b+c}$$

- Jaccard coefficient (***similarity*** measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i,j) = \frac{a}{a+b+c}$$

Nominal Variables

- A **generalization of the binary** variable in that it can take more than 2 states, e.g., red, yellow, blue, green
- Method 1: **Simple matching**
  - *m*: # of matches, *p*: total # of variables

$$d(i,j) = \frac{p-m}{p}$$

- Method 2: use a **large number** of binary variables
  - creating a **new binary variable** for each of the *M* nominal states

Ordinal Variables

- An ordinal variable can be **discrete** or **continuous**
  - **Order** is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by **their rank** $r_{if} \in \{1, \dots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

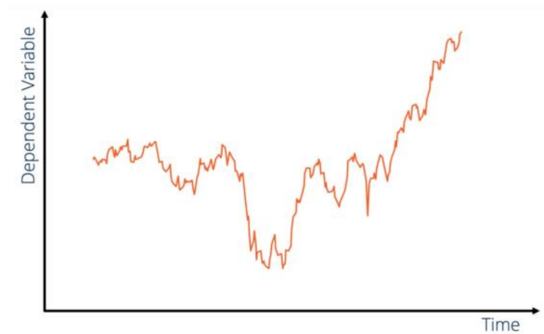
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the **dissimilarity** using methods for interval-scaled variables

Specialized Data Types

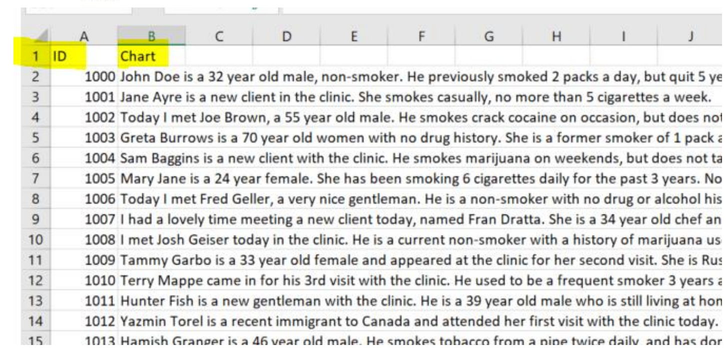
Time-Series Data

- stock prices
- sensor readings



Text Data

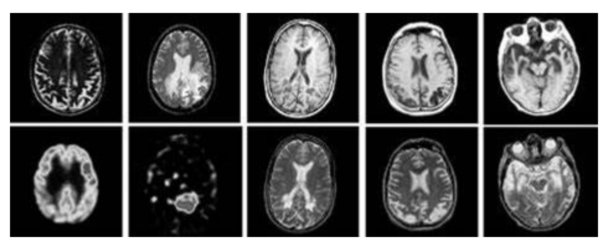
- Reviews
- social media posts



	A	B
1	ID	Chart
2	1000	John Doe is a 32 year old male, non-smoker. He previously smoked 2 packs a day, but quit 5 ye
3	1001	Jane Ayre is a new client in the clinic. She smokes casually, no more than 5 cigarettes a week.
4	1002	Today I met Joe Brown, a 55 year old male. He smokes crack cocaine on occasion, but does not
5	1003	Greta Burrows is a 70 year old women with no drug history. She is a former smoker of 1 pack a
6	1004	Sam Baggins is a new client with the clinic. He smokes marijuana on weekends, but does not ta
7	1005	Mary Jane is a 24 year female. She has been smoking 6 cigarettes daily for the past 3 years. No
8	1006	Today I met Fred Geller, a very nice gentleman. He is a non-smoker with no drug or alcohol his
9	1007	I had a lovely time meeting a new client today, named Fran Dratta. She is a 34 year old chef an
10	1008	I met Josh Geiser today in the clinic. He is a current non-smoker with a history of marijuana us
11	1009	Tammy Garbo is a 33 year old female and appeared at the clinic for her second visit. She is Rus
12	1010	Terry Mappe came in for his 3rd visit with the clinic. He used to be a frequent smoker 3 years a
13	1011	Hunter Fish is a new gentleman with the clinic. He is a 39 year old male who is still living at hon
14	1012	Yazmin Torel is a recent immigrant to Canada and attended her first visit with the clinic today.
15	1013	Hamish Granger is a 46 year old male. He smokes tobacco from a pipe twice daily, and has dor

Image Data

- Grayscale
- RGB



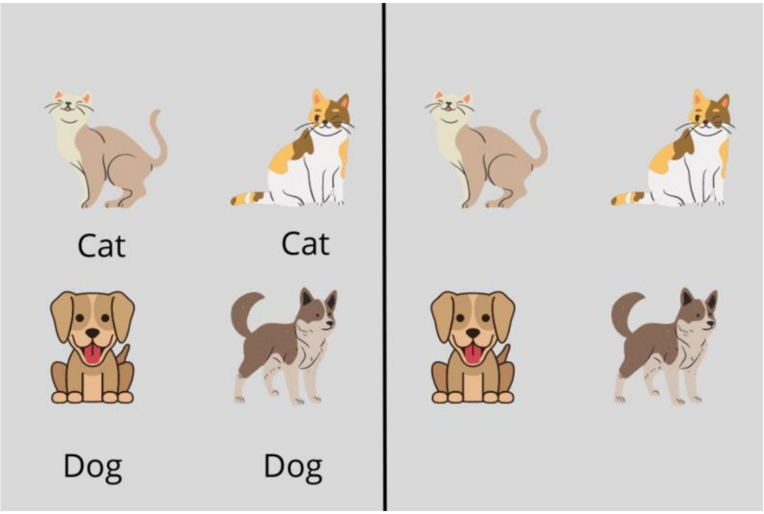
Labeled vs. Unlabeled Data

Labeled Data

- Contains both features and corresponding target labels (used for supervised learning).

Unlabeled Data

- Contains only features and no target labels (used for unsupervised learning).



Importance of Data Preprocessing

- Common Issues: Missing Data, Noisy Data, Duplicates
- Scaling and Normalizing Data
- Encoding Categorical Features
- Normalization vs. Standardization
- Data Splitting

Common Issues in Data

- **Missing Data**: Data that is not recorded or is missing from the dataset.
- **Noisy Data**: Data with errors or irrelevant features that may distort the model.
- **Outlier Data:** samples that are exceptionally far from the mainstream of the data.
- **Duplicate Data**: Repeated entries that can bias model training.

Cleaning Data

Handling Missing Data

- **Remove rows or columns** with too many missing values.
- **Impute missing values** using techniques like mean, median, or mode replacement.
- **Advanced Imputation**: Using machine learning models to predict missing values.

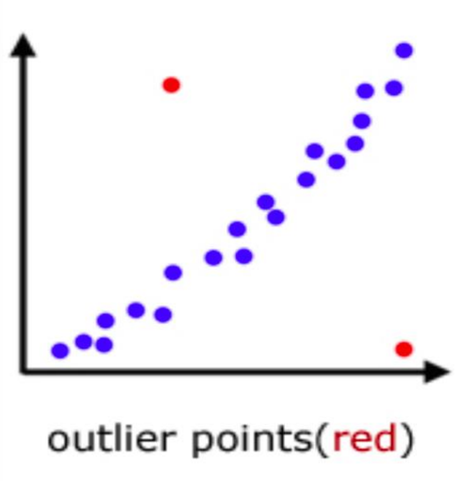
Handling Noisy Data

- **Smoothing techniques** (e.g., moving averages).

Cleaning Data

Handling Outlier Data

- Standard Deviation Method (Z-Score)
- Interquartile Range Method
- …



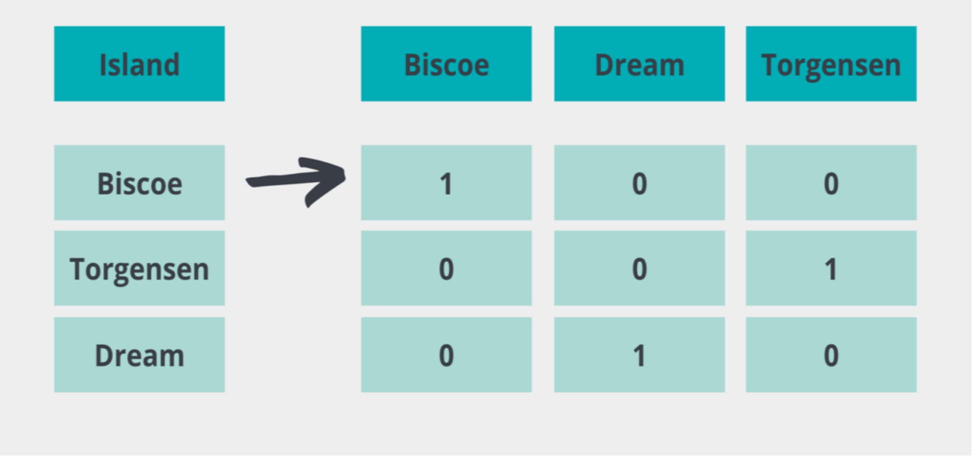
Handling Duplicates Data

- **Detect and remove** duplicate records.

Encoding Categorical Features

Categorical data must be encoded into a numerical format before use in machine learning models.

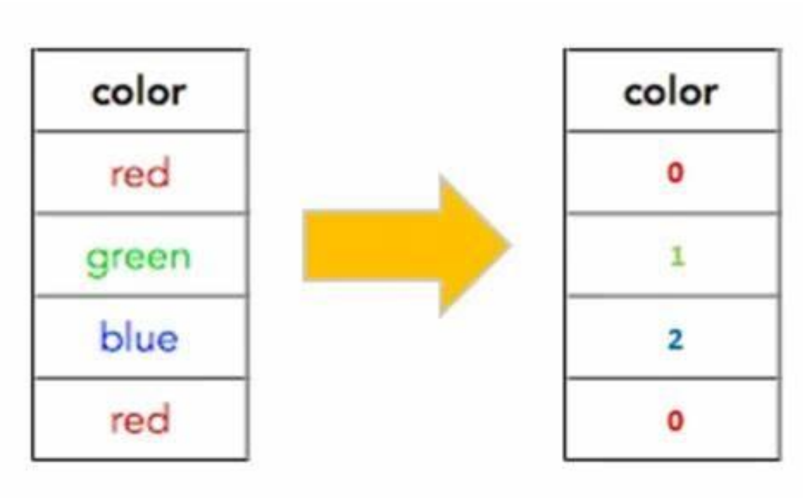
- **One-Hot Encoding**: Converts categorical features into binary vectors. Each category is represented as a separate feature with binary values (0 or 1).
    - Example: **Red**, **Green**, **Blue** becomes **Red: [1, 0, 0], Green: [0, 1, 0], Blue: [0, 0, 1].**



Encoding Categorical Features

Categorical data must be encoded into a numerical format before use in machine learning models.

- **Label Encoding**: Assigns a unique integer to each category. Example:
  - **Red** = 0,
  - **Green** = 1
  - **Blue** = 2



Normalization vs. Standardization

Normalization: Rescales data to a fixed range, typically [0, 1].

- **Min-Max Normalization**:

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

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

- **Z-Score Standardization**:

$$X_{std} = \frac{X - \mu}{\sigma}$$

Feature Normalization

- Linear scaling to **unit range**:
  - Given a lower bound l and an upper bound u for a feature **x**∈R,

$$\tilde{x} = \frac{x - l}{u - l}$$

- Linear scaling to **unit variance**:
  - A feature **x**∈R can be transformed to a random variable with zero mean and unit variance as

$$\tilde{x} = \frac{x - \mu}{\sigma}$$

where μ and s are the sample mean and the sample standard deviation of that feature, respectively

Feature Normalization

- Normalization using the **cumulative distribution function**:
  - Given a random variable $\mathbf{x} \in R$ with cumulative distribution function $F_x(x)$, the random variable $\tilde{x}$ resulting from the transformation $\tilde{x} = F_X(x)$ will be uniformly distributed in [0, 1].
- **Rank normalization**:
  - Given the sample for a feature as $x_1, \ldots, x_n \in R$, first we find the **order statistics** $x^{(1)}, \ldots, x^{(n)}$ and then replace each pattern's feature value by its corresponding norn

$$\tilde{x}_i = \frac{\underset{x_1,\ldots,x_n}{\text{rank}}(x_i) - 1}{n - 1}$$

- where x_i is the feature value for the i'th pattern. This procedure uniformly maps all feature values to the [0, 1] range

Training, Testing and Validation Data

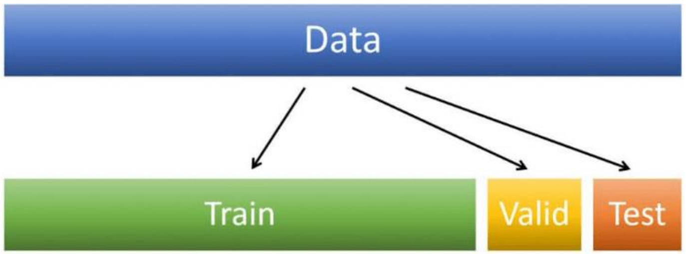
- **Training Data**: Used to train the model, allowing it to learn patterns.
- **Testing Data**: Used to evaluate model performance after training.
- **Validation Data**: Used to tune hyperparameters and validate the model during training to prevent overfitting.

Data Splitting

Splitting the data correctly ensures that models are trained, validated, and tested on appropriate subsets of data.

1- Best Practices for Splitting Data

- **Training Set**: Typically 60-70% of the dataset, used to train the model.
- **Validation Set**: Typically 10-20%, used to tune model hyperparameters and prevent overfitting.
- **Test Set**: Typically 10-20%, used for final model evaluation.



Data Splitting

2- Cross-Validation

Cross-validation is a robust technique to assess model performance by splitting the data into multiple training and validation subsets.

- **K-Fold Cross-Validation**
- **Leave-One-Out Cross-Validation (LOOCV)**

Data Splitting

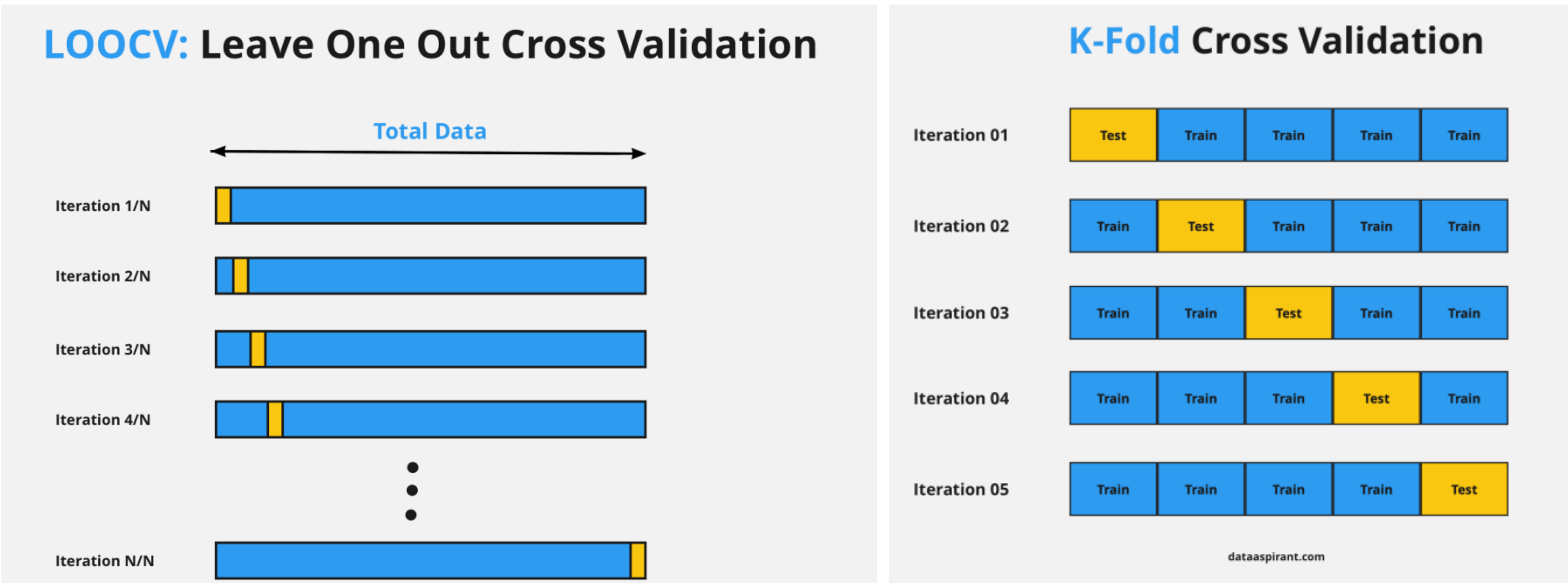


Figure 2: K-Fold Cross-Validation vs Leave-One-Out Cross-Validation