UNIVERSIDADE CATÓLICA PORTUGUESA

BRAGA

Machine Learning

Session 2 - T

Data Preprocessing

Degree in Applied Science

2024/2025

The importance of data

- Most Machine Learning courses primarily focus on algorithms, operating under the assumption of **high-quality** and **sufficient data** to create robust models;
- Real-world data often **contains errors**, highlighting the need for data curation. Even widely-used benchmark datasets frequently contain errors in assigned labels (labelerrors.com/);
- The quality of the output is determined by the quality of the input (**garbage-in, garbage-out**);
- The emerging trend of "**data-centric AI**" focuses on enhancing datasets to improve model outcomes.

Data Representation

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- We will assume datasets to be organized in the following way:
  - Data is organized in a **tabular format**;
  - Rows represent **samples** (aka records, entities, or examples) and columns **variables** (aka attributes or features);
  - Variables can be categorized as either **numerical** (or continuous) or **discrete** (or nominal).

Player	Minutes	Points	Rebounds	Assists
A	39	20	6	150
B	30	29	7	6
C	22	7	7	2
D	20	3	3	14
E	9	19	5	5
…	…	…	…	…

Raw vs Processed Data

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Raw data:**
  - The original source of the data;
  - Often challenging for direct analysis;
  - Preprocessing needs:
    - ❑ Handling missing values;
    - ❑ Rescaling variables;
    - ❑ Detecting and managing outliers;
    - ❑ Correcting errors;
    - ❑ …
  - May come from different sources:
    - ❑ Data integration from multiple sources;
    - ❑ Data cleaning;
    - ❑ Data transformations;
    - ❑ Data selection;
    - ❑ Data enrichment;
    - ❑ …

Player	Minutes	Points	Rebounds	Assists
A	39	20	6	150
B	30	29	7	6
C	22	7	7	2
D	20	three	3	14
E	-9	19	?	5
F	14	6	1	3
G	22		4	

Raw vs Processed Data

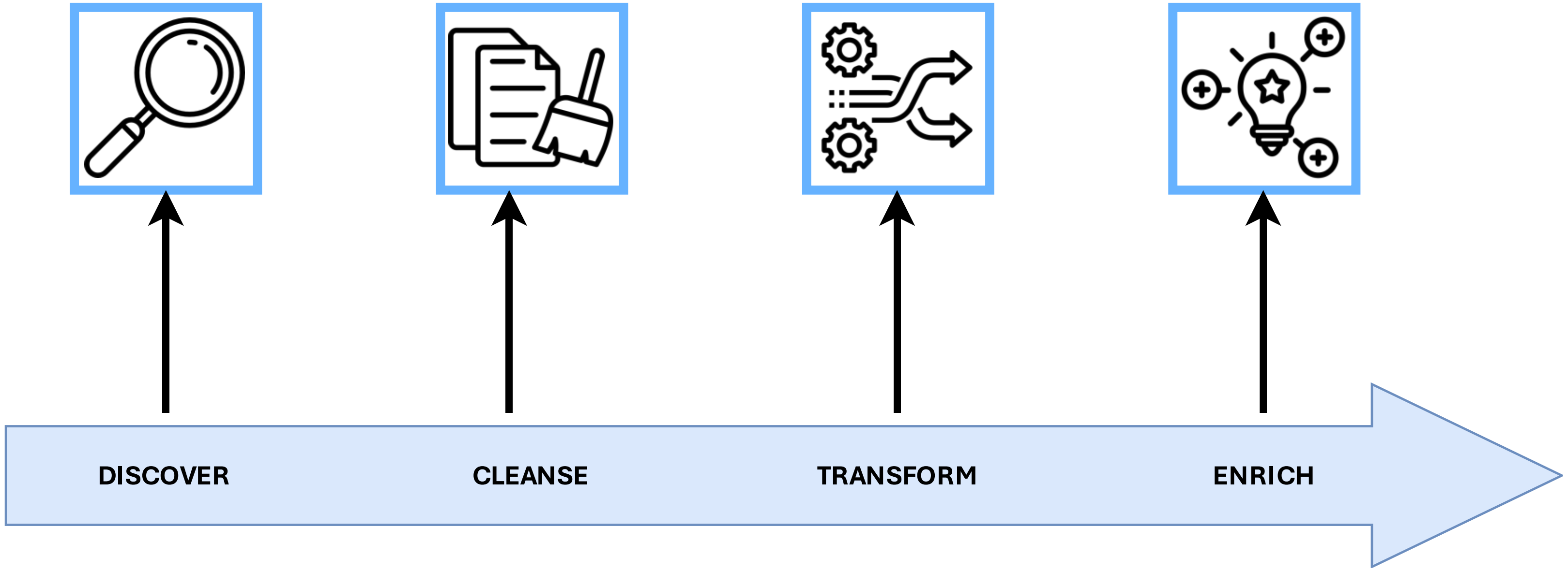
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Processed data:**
  - Data is in a structure ready for analysis;
  - Usually represented in a matrix or vector format.

Player	Minutes	Points	Rebounds	Assists
A	39	20	6	14
B	30	29	7	6
C	22	7	7	2
D	20	3	3	14
E	9	19	5	5
F	14	6	1	3
G	22	14	4	7

Data Preparation

- The differnet stages involved in data preparation, leading from raw to processed data, heavily rely on the **domain of the data** and the unique **characteristics of the datasets**.



Data Structure and Content Verification

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Exploring the structure and content of data involves various tasks, including:
  - Checking the **number of samples and variables**;
  - Verifying **data types** (numerical, discrete);
  - Checking **value ranges** for numerical variables or the set of **possible values** for discrete variables;
  - Identifying **missing or unknown values**;
  - Identifying **duplicates**;
  - ...

Data Sumarization

- Characterization of large datasets using **global metrics**;
- Aims to identify preprocessing needs like **handling outliers, missing values, and errors**;
- Tasks include checking **value distributions** and applying **summary statistics** to variables.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333	1.000000
std	0.828066	0.435866	1.765298	0.762238	0.819232
min	4.300000	2.000000	1.000000	0.100000	0.000000
25%	5.100000	2.800000	1.600000	0.300000	0.000000
50%	5.800000	3.000000	4.350000	1.300000	1.000000
75%	6.400000	3.300000	5.100000	1.800000	2.000000
max	7.900000	4.400000	6.900000	2.500000	2.000000

Data Transformation

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Data transformation can play a crucial role in preparing datasets for effective use in machine learning applications.
  - Typical operations:
    - **Feature engineering:** Creating new features from existing ones to enhance the predictive power of the model;
    - **Imputation:** Handling missing values by filling them in with estimated or imputed values;
    - **Feature Scaling/Normalization:** Rescaling features to a similar scale to ensure that no single feature dominates the learning process.
    - **Feature Encoding:** Converting categorical variables into numerical representations that machine learning algorithms can interpret;
    - **Dimensionality Reduction:** Reducing the number of features in the dataset while preserving important information.

Missing Values

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Missing values refers to values or information that are **not stored or absent** for certain variables within a given dataset.

- In Pandas, missing values are typically represented by **NaN**, which stands for "Not a Number."

- **Why is data missing from the dataset?**
  - Past data can become corrupted due to inadequate maintenance practices;
  - Recording failures from human error;
  - Intentional omission;
  - ...

Missing Values

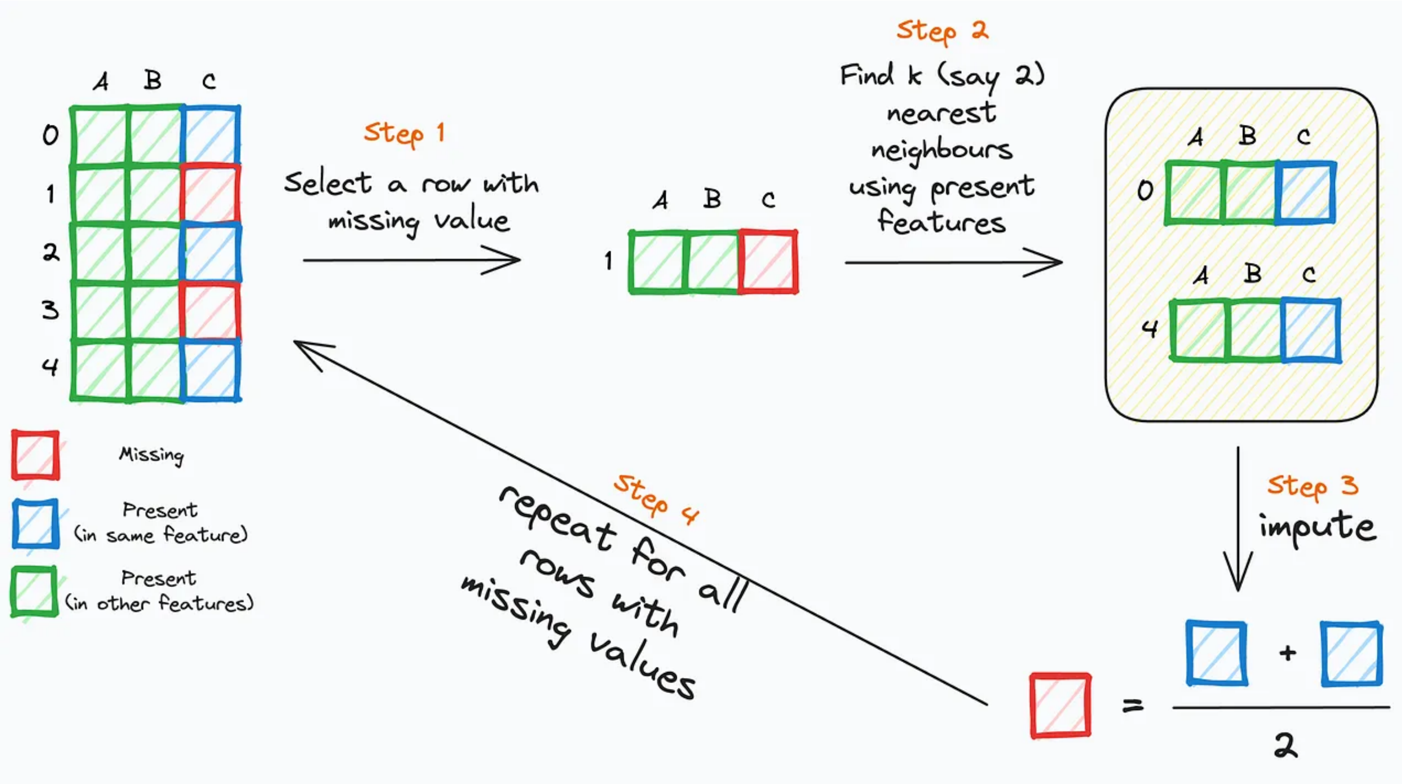
- **Why do I need to care about missing values?**
  - Some machine learning algorithms fail with missing values;
  - Failure to handle missing values can lead to biased models and inaccurate results;
  - Missing data can reduce precision in statistical analysis;
  - ...

Missing Values

- **Addressing missing values** involves various approaches that can be specific to different data types and analysis scenarios:
  - **Keep missing values** in the data if the analysis methods can handle them;
  - **Disregarding missing** values by **removing rows and/or columns** containing them.
  - **Substituting missing values** with other values, which can be achieved through methods such as:
    - ❑ **Imputation** with a constant value (specific value, mean, median, mode) per column (or row).
    - ❑ Utilizing more sophisticated techniques like **k-nearest neighbors**, leveraging information from neighboring data points.

Missing Values

- Handling missing values with the **k-nearest neighbor algorithm:**



https://www.blog.dailydoseofds.com/p/the-most-overlooked-problem-with-768

Resources

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- McCallum, Q. E. (2013). Bad Data Handbook Mapping the World of Data Problems. O'Reilly.

- Rattenbury, T., Hellerstein, J. M., Heer, J. M., Kandel, S., & Carreras, C. (2017). Principles of data wrangling: Practical Techniques for Data Preparation. O'Reilly.