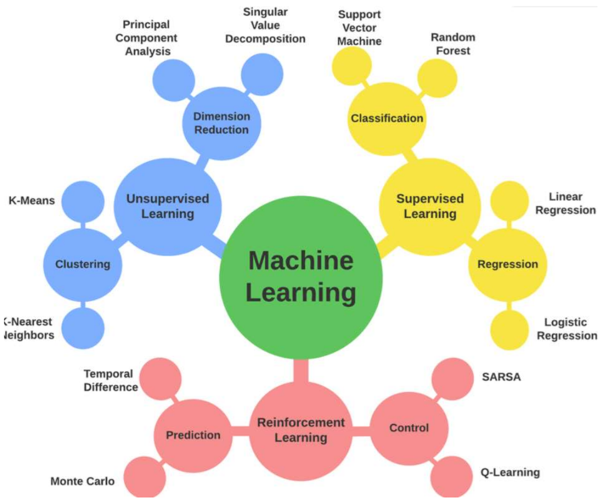
Các phương pháp học máy
Machine learning methods

4 TC: 2 LT – 2 TH

Giảng viên: **Tạ Hoàng Thắng**

tahoangthang@gmail.com

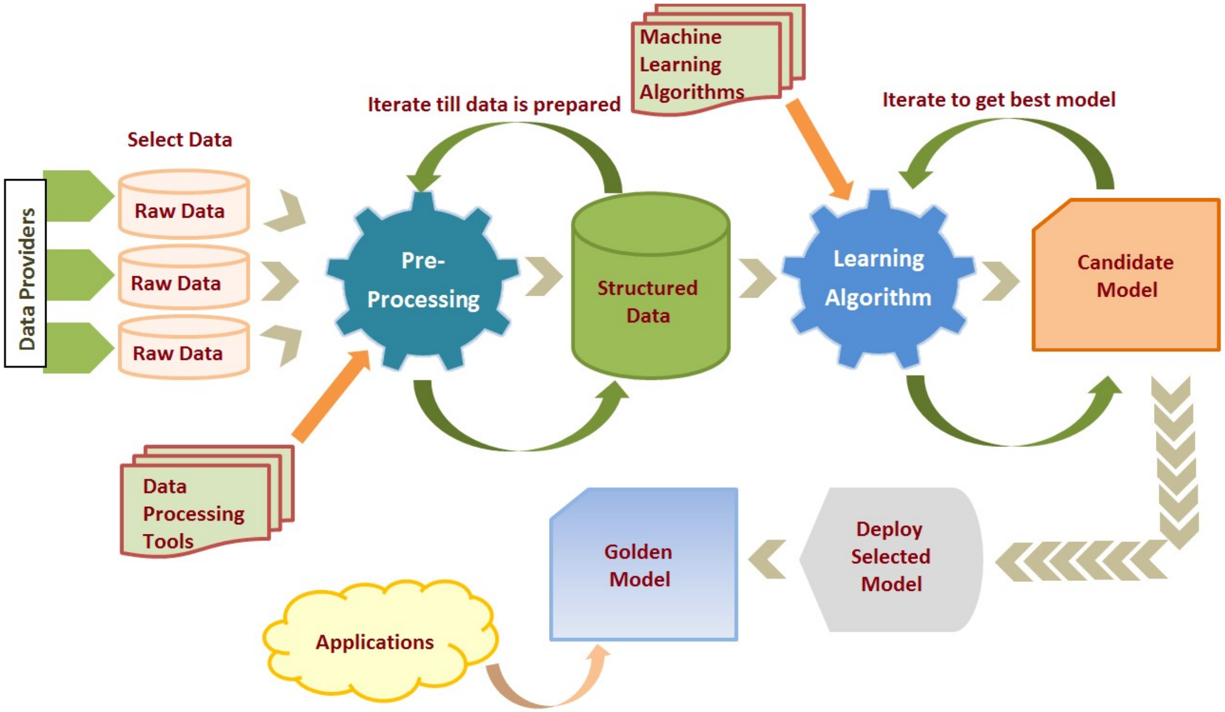
0975399307



Data Preprocessing

Data preprocessing

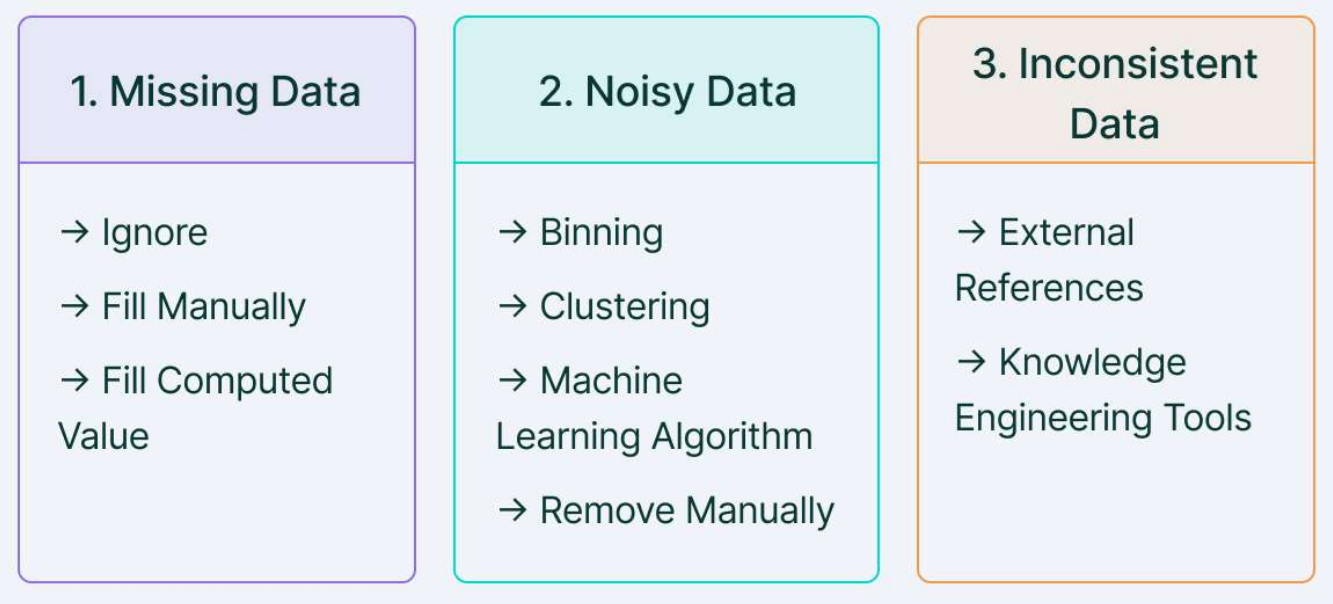
- is **a crucial step** in the data analysis pipeline, aimed at preparing raw data for analysis and modeling.



Data Preprocessing

Data preprocessing

- What we will do?



1. Missing Data

→ Ignore

→ Fill Manually

→ Fill Computed Value

2. Noisy Data

→ Binning

→ Clustering

→ Machine Learning Algorithm

→ Remove Manually

3. Inconsistent Data

→ External References

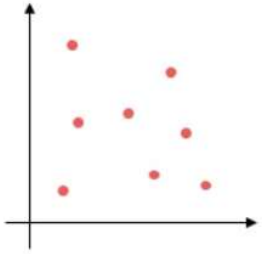
→ Knowledge Engineering Tools

Data Preprocessing

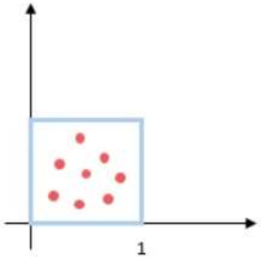
Data preprocessing steps:

- **Cleaning**: **Removing or correcting inaccuracies, missing values, or inconsistencies** in the data.
    - Use custom functions, pandas, and other packages.
- **Normalization/Standardization**: Adjusting the scale of features so they have similar ranges or distributions.
    - Help **algorithms perform better** (those sensitive to feature scaling)
        - Use MaxMinScaler, L1 Norm and L2 Norm (https://scikit-learn.org/stable/modules/preprocessing.html)

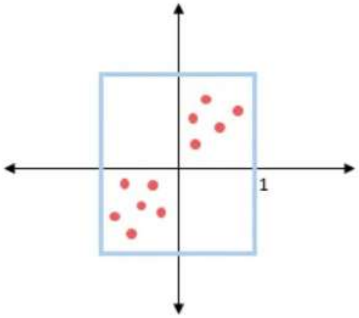
Data Cleaning



Actual Data



After normalizing

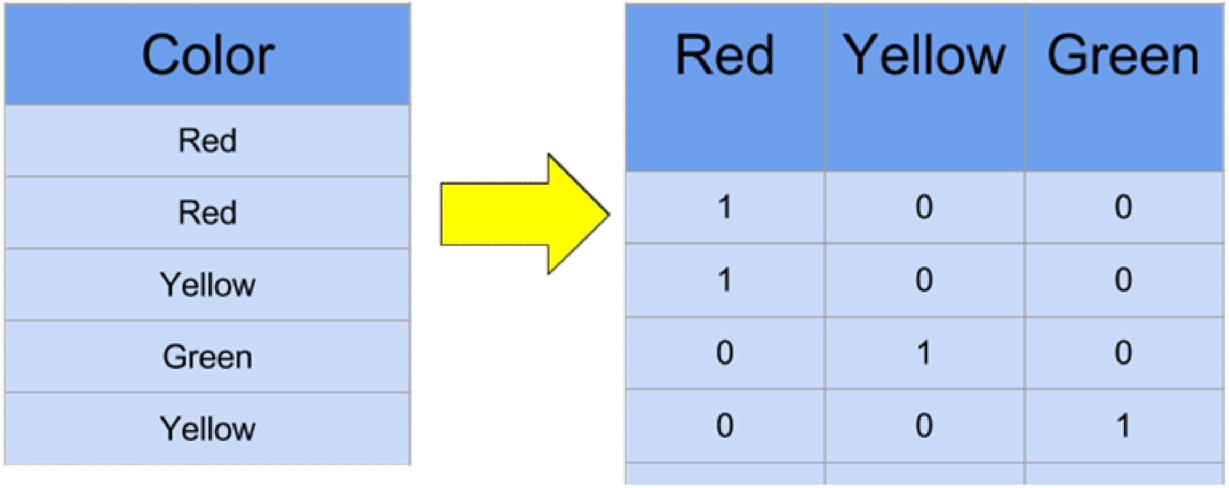


After standardization

Data Preprocessing

Data preprocessing steps:

- **Encoding**: Converting categorical variables into numerical formats.
  - This is necessary because many machine learning algorithms require numerical input.
    - One hot encoding



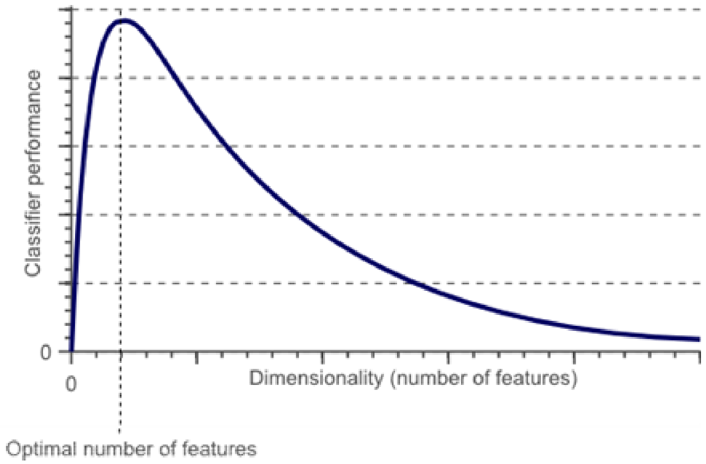
Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Data Preprocessing

Data preprocessing steps:

- **Feature Selection/Engineering**: Choosing relevant features or creating new ones from existing data to improve model performance and reduce complexity.
  - Curse of dimensionality: **challenges and issues that arise when working with high-dimensional data**.

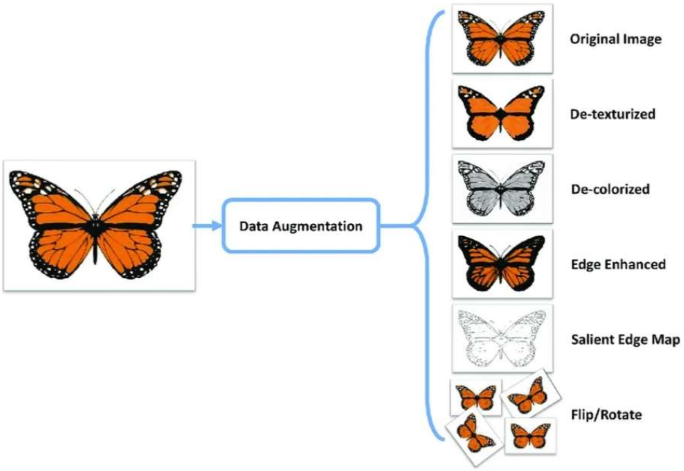


Data Preprocessing

Data preprocessing steps:

- **Splitting Data**: Dividing the dataset into training, validation, and test sets.
    - Help in assessing the model's performance on **unseen data and prevents overfitting**.

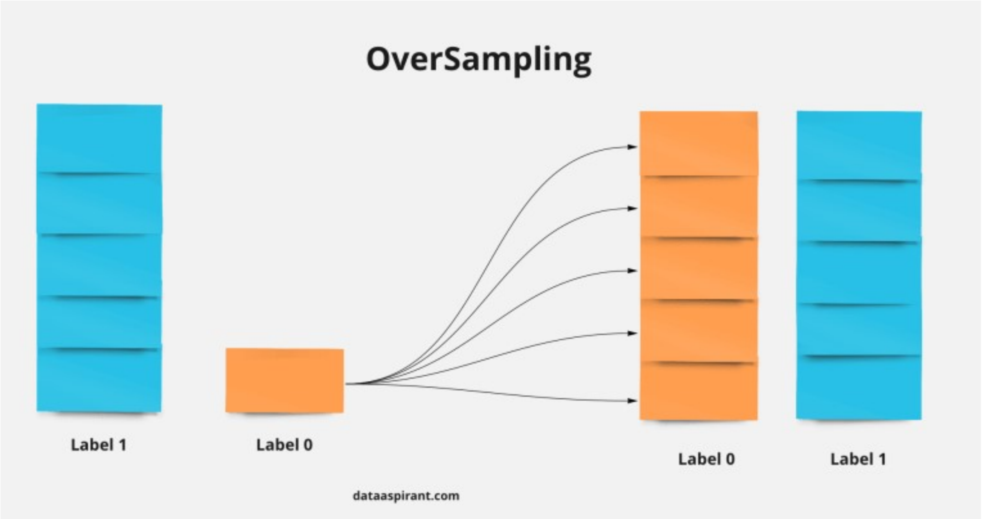
- **Handling Imbalanced Data**: Techniques like **resampling** or using different metrics to address imbalances between classes in classification problems.
    - Data augmentation
    - Oversampling
    - Undersampling



Data Preprocessing

Data preprocessing steps:

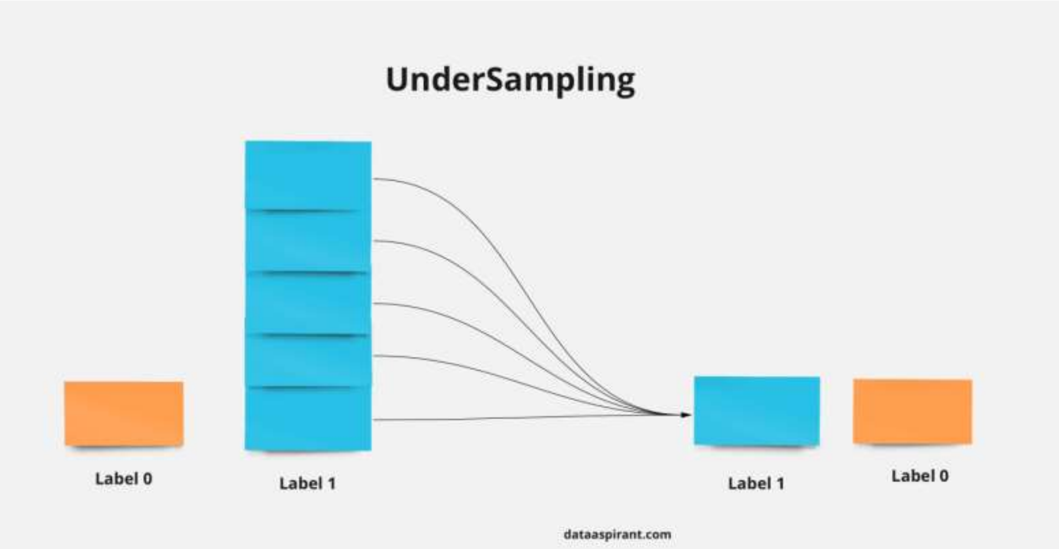
- **Handling Imbalanced Data**: Techniques like **resampling** or using different metrics to address imbalances between classes in classification problems.
  - Oversampling



Data Preprocessing

Data preprocessing steps:

- **Handling Imbalanced Data**: Techniques like **resampling** or using different metrics to address imbalances between classes in classification problems.
  - Undersampling



Data Preprocessing

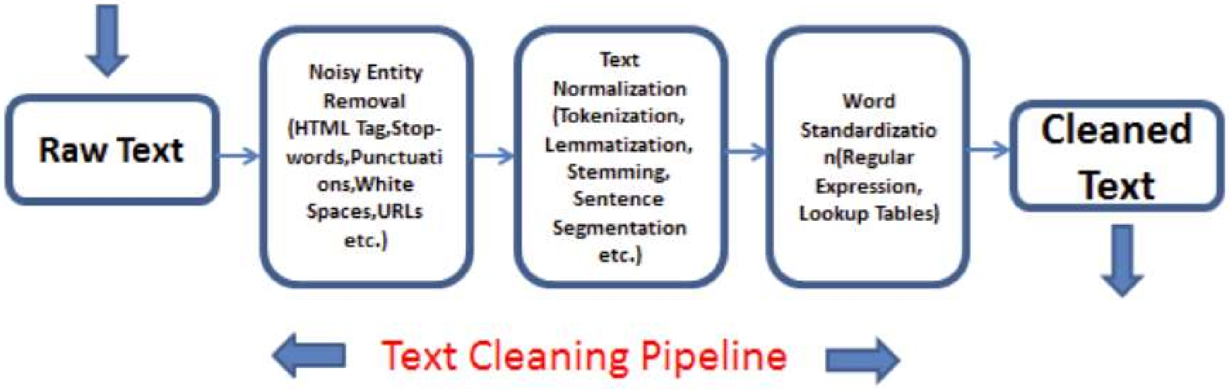
Why we do data preprocessing?

- Improves accuracy: **Clean and well-prepared data leads to better model performance** and more accurate predictions.
- Reduces noises: Removing **irrelevant or erroneous data helps in reducing noise** and enhancing signal quality.
- Ensures compatibility: Formatting data correctly ensures that it is compatible with different algorithms and tools.
- Saves time: Preprocessed **data speeds up the training process** and reduces the likelihood of errors.

Data Preprocessing

Packages for data preprocessing

- Numeric data: pandas, numpy, scikit-learn, torch
  - https://www.geeksforgeeks.org/data-processing-with-pandas/
- Text data: spaCy + NLTK
  - https://soshace.com/2023/04/05/nlp-preprocessing-using-spacy/
  - https://iq.opengenus.org/text-preprocessing-in-spacy



Raw Text

Noisy Entity Removal (HTML Tag,Stop-words,Punctuations,White Spaces,URLs etc.)

Text Normalization (Tokenization, Lemmatization, Stemming, Sentence Segmentation etc.)

Word Standardization(Regular Expression, Lookup Tables)

Cleaned Text

Text Cleaning Pipeline