Machine Learning

Session 2 - Supervised classification

hadrien.salem@centralelille.fr

[introduction-to-data-science]

Introduction

What did we do last time?

Course outline

Intro to ML course

Session 1: Introduction to ML & Regression

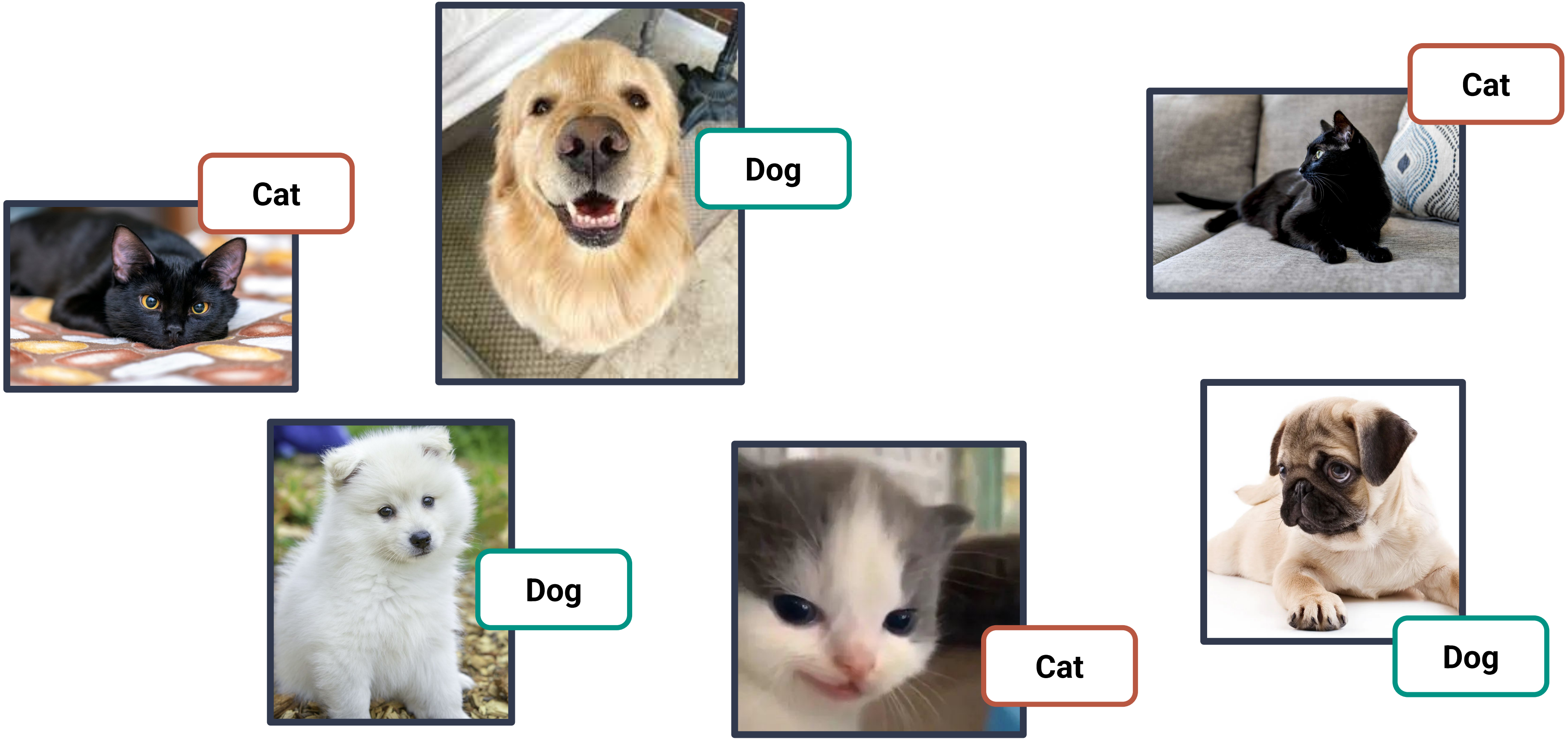
Session 2: Supervised classification

Session 3: Decision trees & Ensemble methods

↓

Deep Learning

What is classification?



Intuitively, classification is giving objects the right label

$$\boxed{f^*(x) = \arg\max_k \mathbb{P}(C_k|x)}$$

Where f^* is a rule for classification, C_k are the **classes**, and x the **examples**

The goal of a classification algorithm is to find this rule

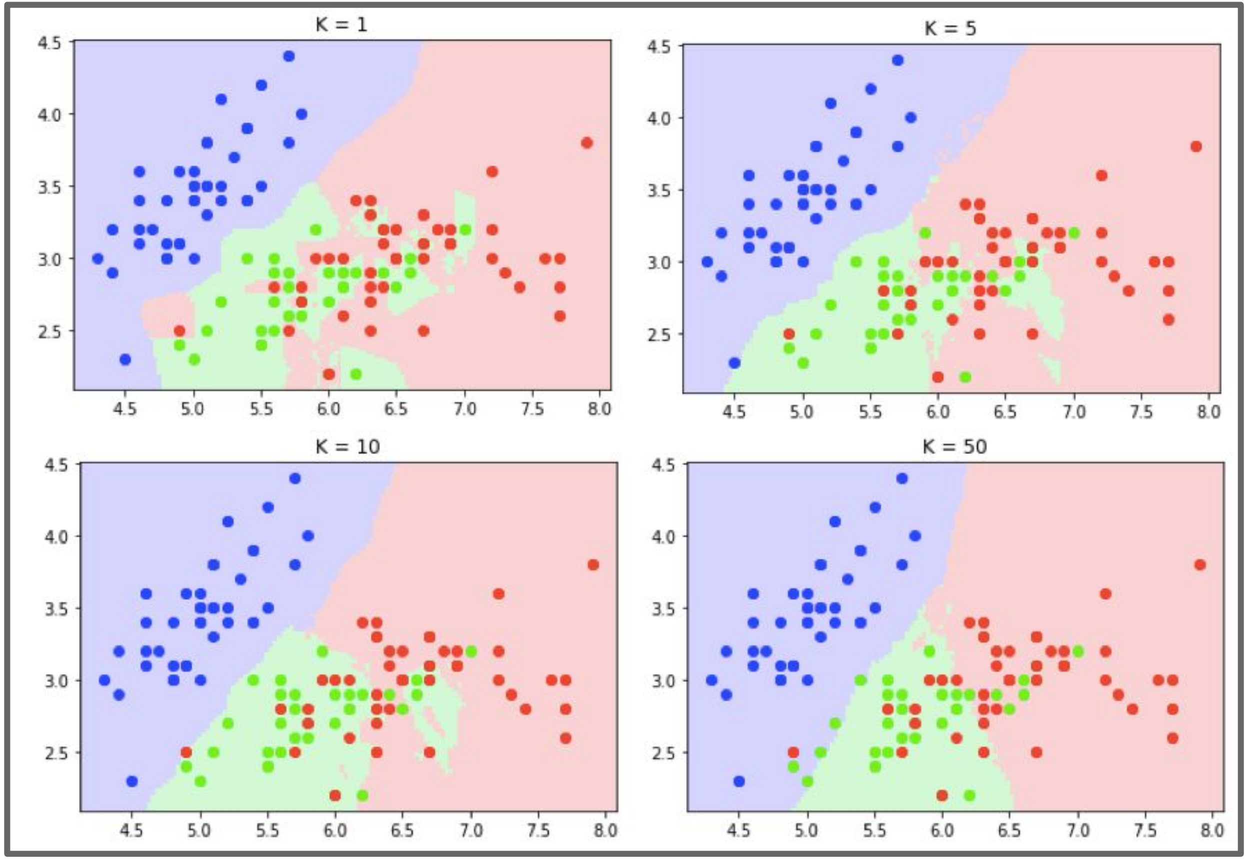
Formally, classification is finding the most probable class for an example

Families of classification models

While they are all classification models, they have different purposes

There are three main families of classification models

- **Discriminant functions**
  - The algorithm learns a function that finds the class directly
  - *Example*: K-nearest-neighbours
- **Discriminant models**
  - The algorithm models the decision boundary
  - *Example*: Support Vector Machines
- **Generative models**
  - The algorithm models the data distribution (meaning you can generate your own data)
  - *Example*: Gaussian Mixture Model

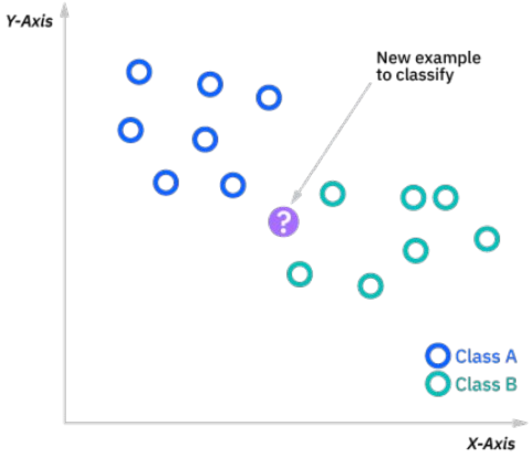


<u>When K is small</u>, the algorithm is very sensitive to **local variations** (risk of overfitting)

<u>When K is large</u>, the algorithm is **more stable**, but **does not take small variations into account** (risk of underfitting)

⇒ When choosing the parameters, there is a compromise between the two

Common classification algorithms



Introduce a new example

Compute distances

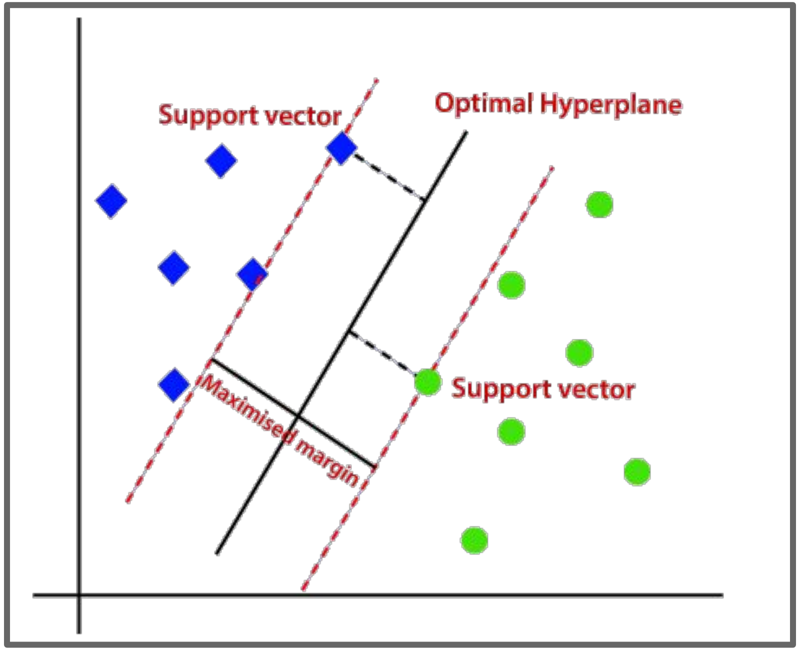
Majority vote

The K-nearest-neighbours (KNN) algorithm for data imputation

Source : IBM

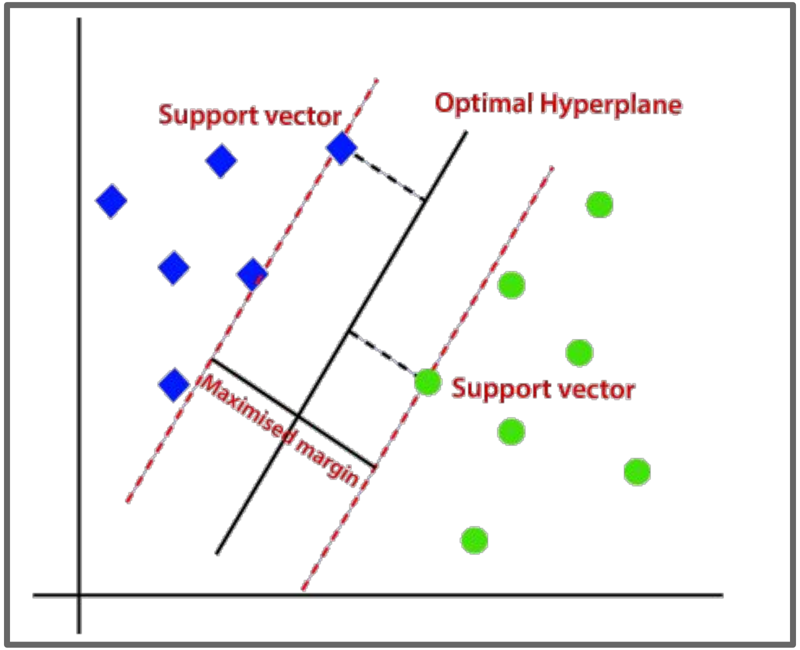
K-nearest neighbours

Decision boundary	Non-linear
Advantages	• Easy to use and understand • No assumptions
Disadvantages	• Slow for large datasets • Inefficient in high dimension



The objective is to **find a hyperplane** such that the **margin between the two classes is maximized**.

Data can be transformed into a higher-dimensional space if it is not linearly separable in the feature space. This is achieved with **kernels** (e.g. polynomial, sigmoid, etc.).



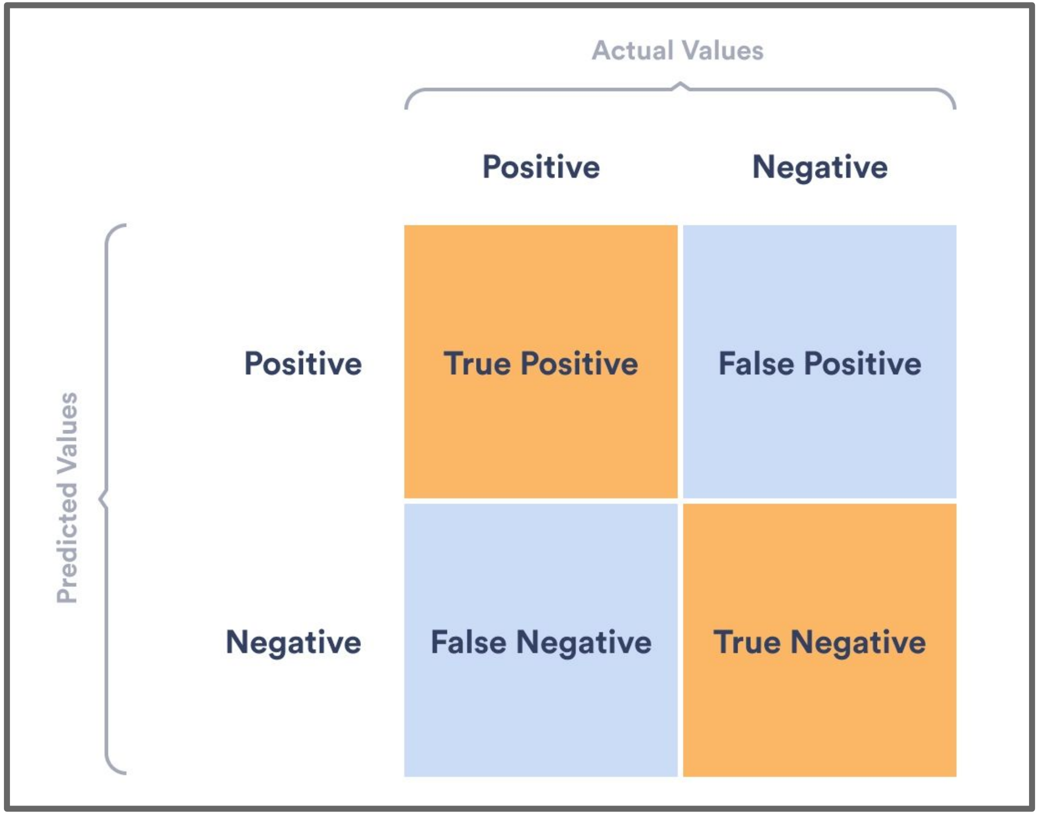
Decision boundary	Linear in the transformed space, can be non-linear in the feature space
Advantages	• Works well in high dimension • Robust to outliers • Low memory consumption
Disadvantages	• Slow for large datasets • Choosing a kernel can be difficult

Support Vector Machines (SVMs)

Other methods for supervised classification

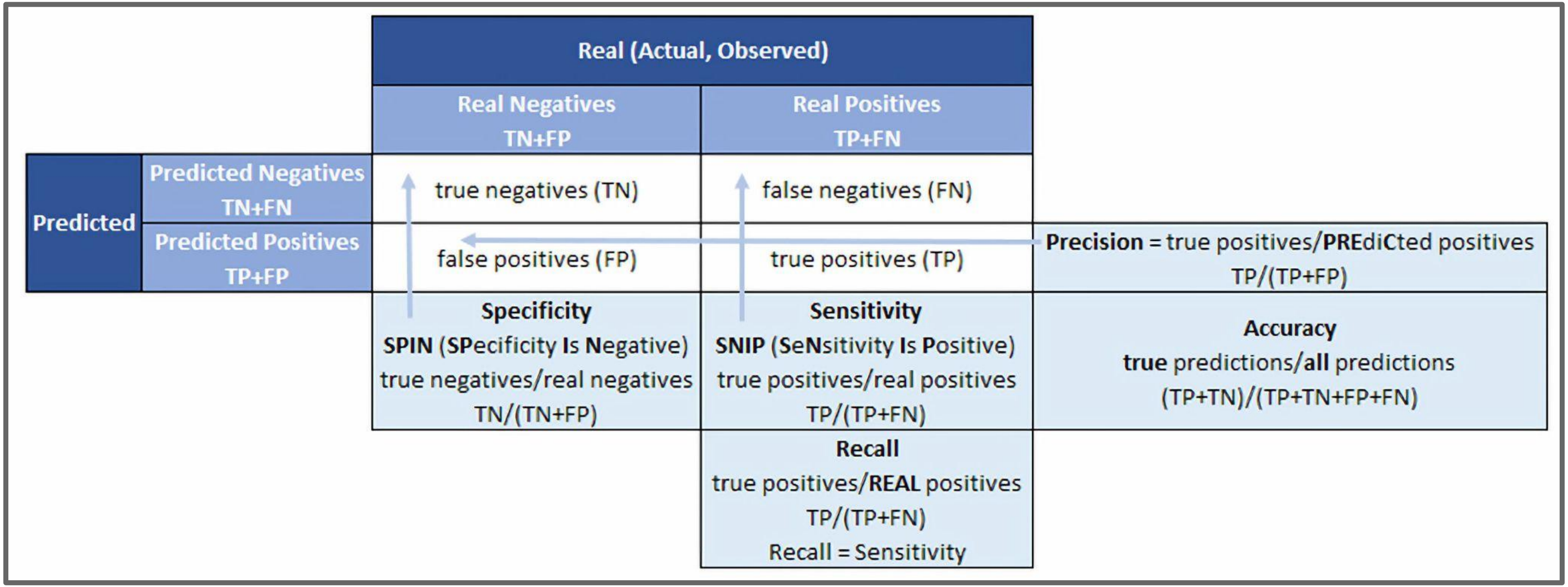
- **Logistic Regression**
    - THIS IS A CLASSIFICATION METHOD
    - Only works when data is linearly separable
    - Easy to use, good baseline method
- **Naive Bayes**
    - Assumes that features are independent
    - Non-linear decision boundary
      (computes class membership probabilities)
    - Low-cost, also good baseline
- **Linear / Quadratic discriminant analysis**
    - Assumes that data follows a normal distribution
    - Limited to linear / quadratic decision boundaries
    - Good baseline
- **And other algorithms we will study later**
    - Decision trees / Random Forests
    - Ensemble methods
    - Neural networks

Evaluating a classification algorithm



Confusion matrices allow you to analyse how well each class is handled

[Image source]

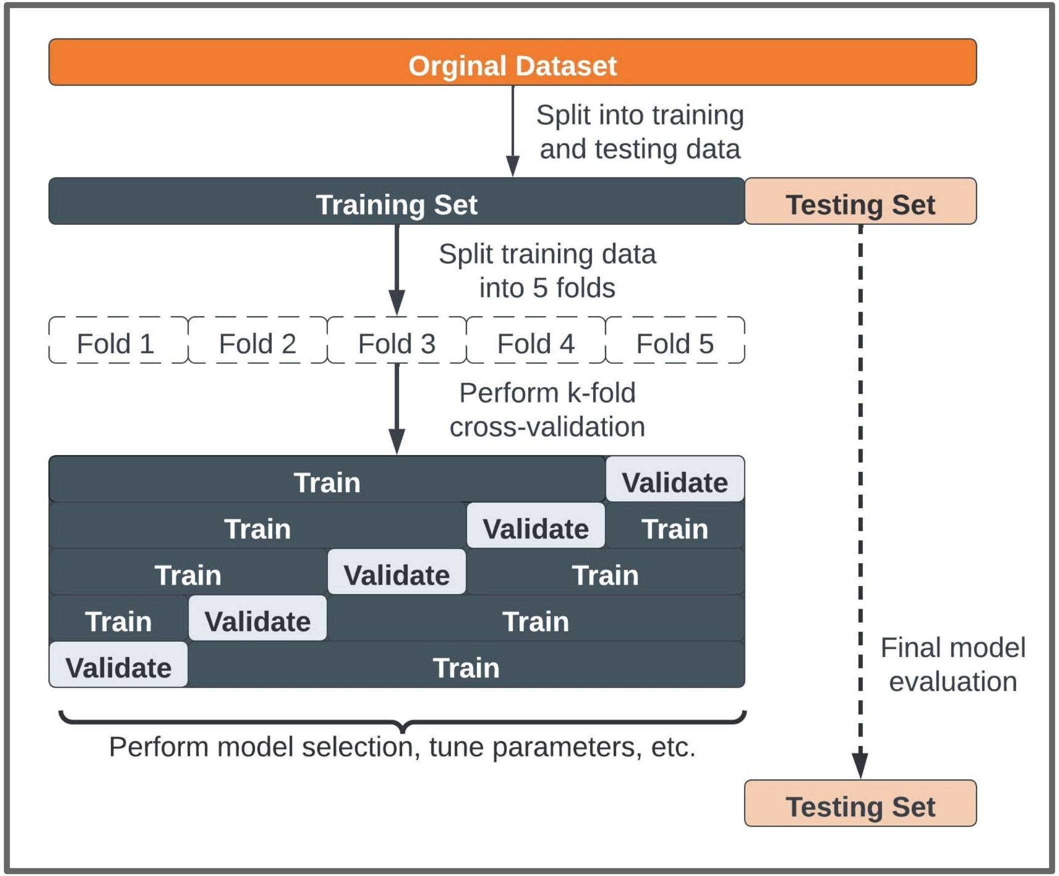


Several performance indicators can be computed from the confusion matrix

Image source

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-Score is a synthesis of the previous metrics and can be more meaningful than accuracy



Cross-validation can allows for a better estimation of the model's performance

Image source

Practical work

The notebook contains all the necessary instructions

Debrief

Debrief

What did we learn today?

What could we have done better?

What are we doing next time?

Machine Learning

Session 2 - Supervised classification

hadrien.salem@centralelille.fr

[introduction-to-data-science]