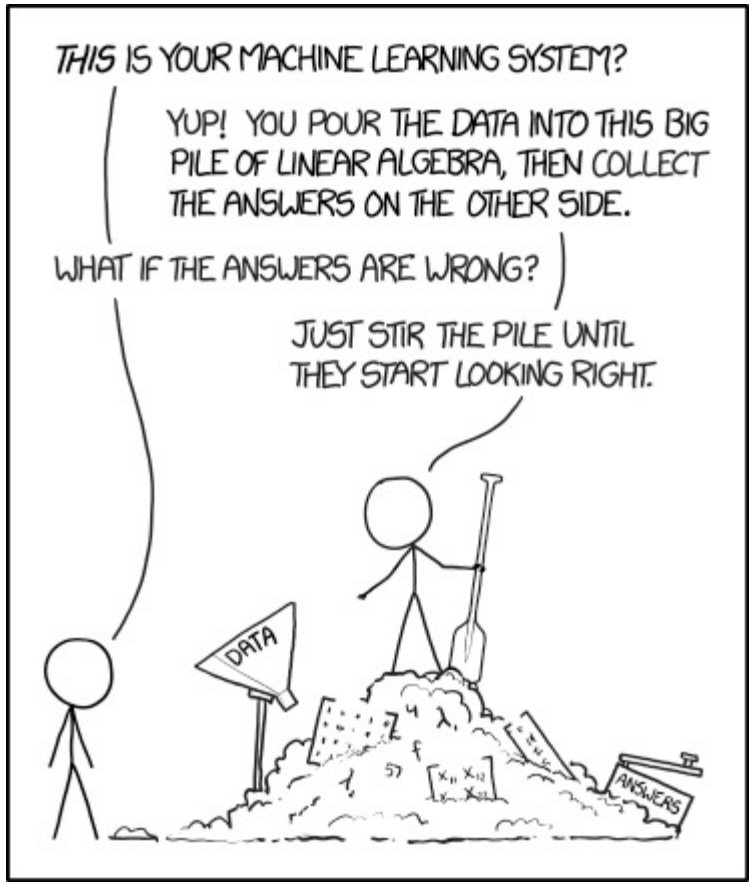
Apprentissage machine

Concepts intermédiaires



CONSORTIUM SANTÉ NUMÉRIQUE

IVADO

Yves Terrat, PhD
Consortium Santé Numérique, UDEM

But de l’apprentissage machine

Choisir et entrainer un algorithme prédictif qui va avoir une bonne capacité de

<u>généralisation</u>

Apprentissage machine (versus) statistiques

POINTS OF SIGNIFICANCE

Statistics versus machine learning

Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns.

Two major goals in the study of biological systems are inference and prediction. Inference creates a mathematical model of the data-generation process to formalize understanding or test a hypothesis about how the system behaves. Prediction aims at forecasting unobserved outcomes or future behavior, such as whether a mouse with a given gene expression pattern has a disease. Prediction makes it possible to identify best courses of action (e.g., treatment choice)

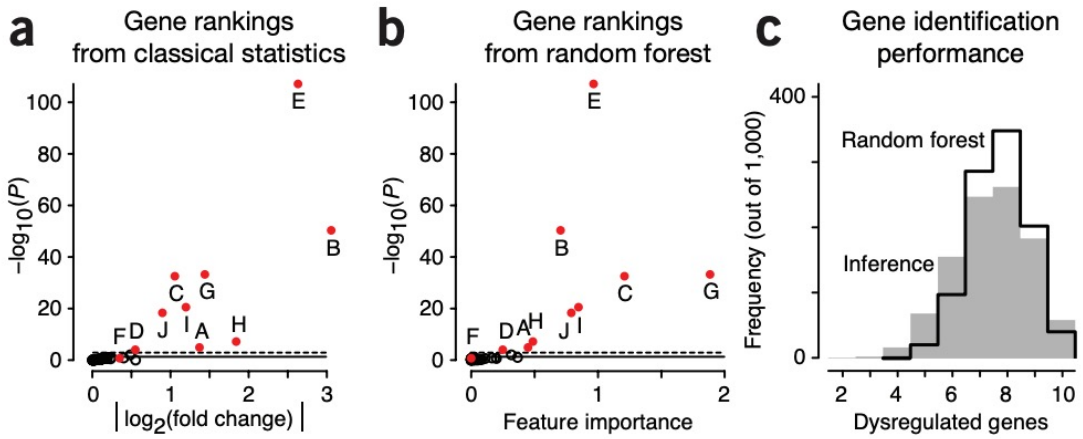


Figure 2 | Analysis of gene ranking by classical inference and ML. (**a**) Unadjusted log-scaled *P* values from statistical differential expression analysis as a function of effect size, measured by fold change in expression. (**b**) Log-scaled *P* values from **a** as a function of gene importance from random forest classification. In **a** and **b**, red circles identify the ten differentially expressed genes from **Figure 1**; the remaining genes are indicated by open circles. (**c**) Distribution of the number of dysregulated genes correctly identified in 1,000 simulations by inference (gray fill) and random forest (black line).

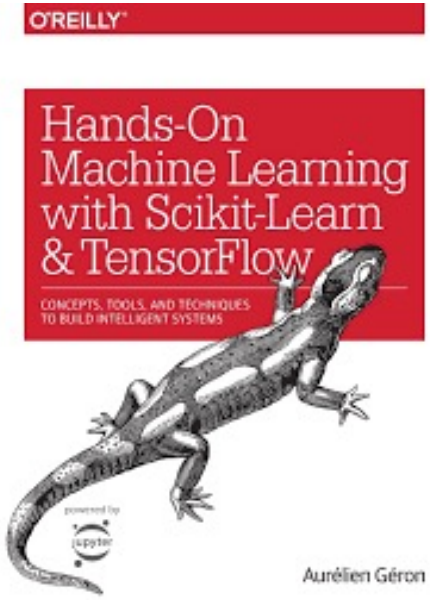
https://www.nature.com/articles/nmeth.4642

L’apprentissage machine n’est PAS magique

Sources d'information

stackoverflow

PubMed



EpiMed

DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

iid Institut intelligence et données

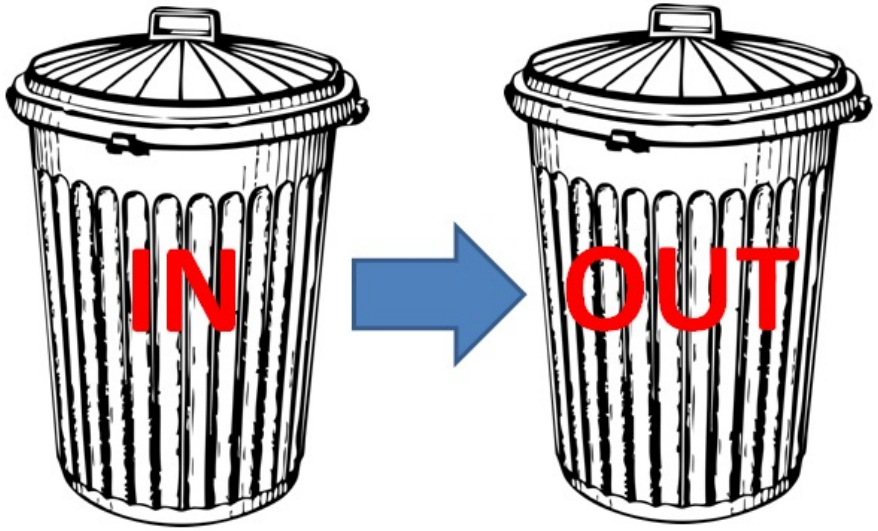
kaggle

tds

Les données

Collecte des données

« garbage in, garbage out »



Collecte des données

- *Où sont les données et sont-elles accessibles?*
- *Quelle « masse » de données ai-je besoin ?*
- *Quels types de données dois-je collecter ?*

« Nettoyage » et préparation des données

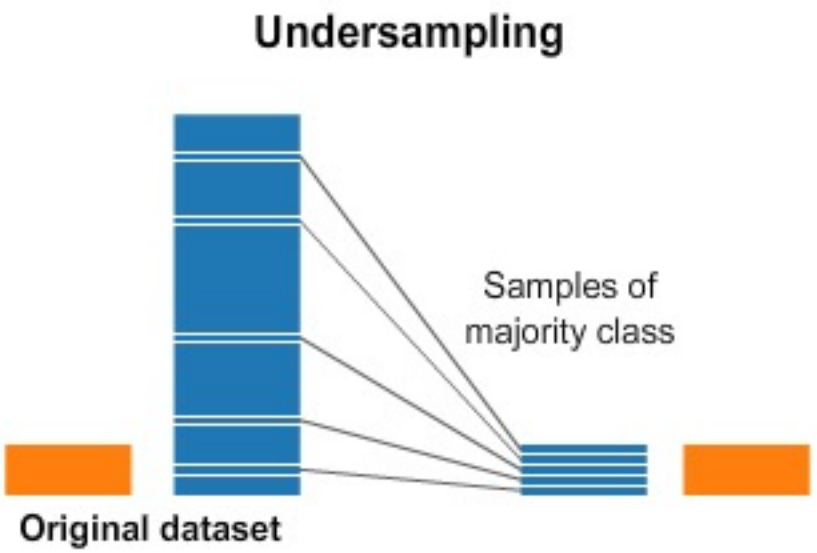
- *Encodage des données (LabelEncoder, OneHotencoder,…)*
- *Imputation des données manquantes (Simple, Iterative, KNN, …)*

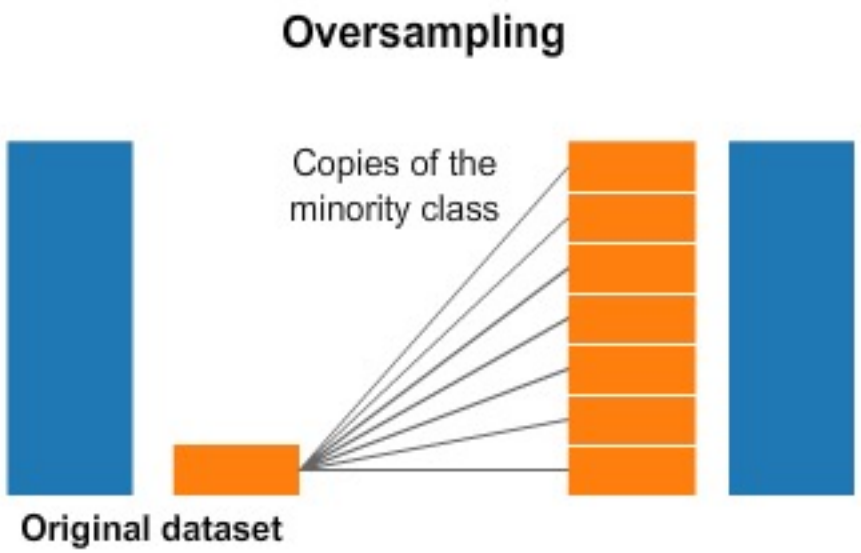
	V1	V2
S1	1	10
S2	2	NaN
S3	NaN	5

	V1	V2
S1	1	10
S2	2	**7.5**
S3	**1.5**	5

- *Élimination des données aberrantes* 🚩
- *Standardisation / Normalisation (MinMax, StandardScaler, …)*
- *Réduction de la dimension (Human, PCA, SelectKBest, …)*

« Nettoyage » et préparation des données





https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

« Feature selection » : pourquoi c'est important ?

npj | Digital Medicine

www.nature.com/npjdigitalmed

PERSPECTIVE OPEN

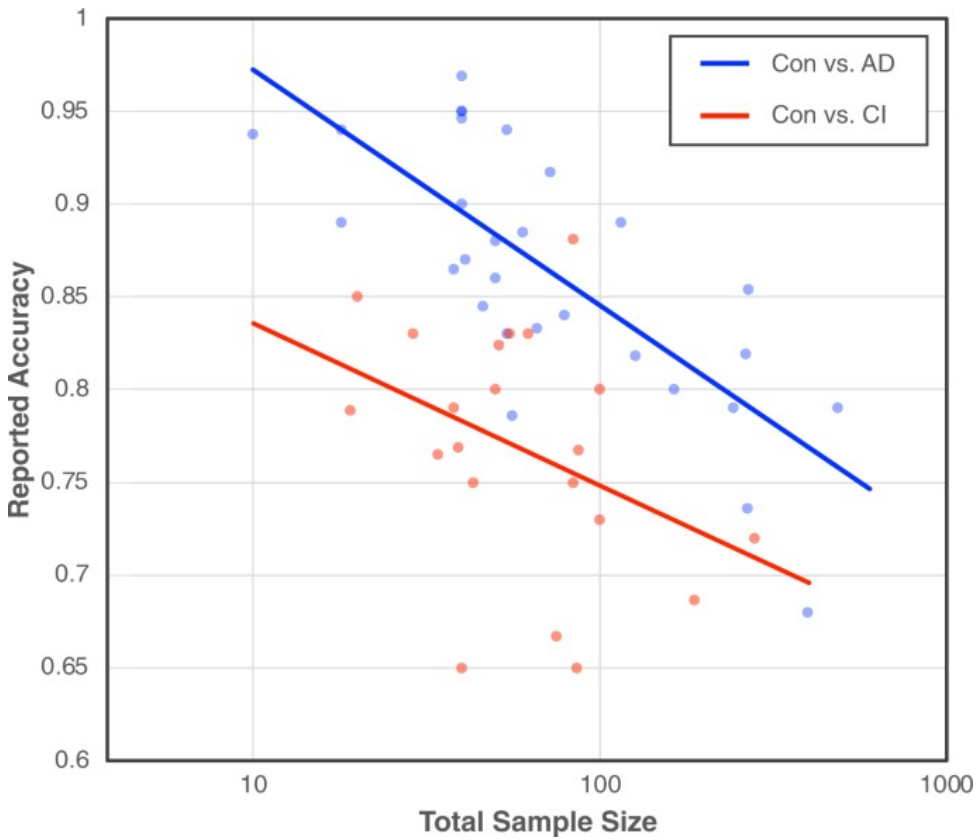
Check for updates

Digital medicine and the curse of dimensionality

Visar Berisha[1,2,3]✉, Chelsea Krantsevich[3,4], P. Richard Hahn[4], Shira Hahn[2,3], Gautam Dasarathy[1], Pavan Turaga[1,5] and Julie Liss[2,3]

Digital health data are multimodal and high-dimensional. A patient's health state can be characterized by a multitude of signals including medical imaging, clinical variables, genome sequencing, conversations between clinicians and patients, and continuous signals from wearables, among others. This high volume, personalized data stream aggregated over patients' lives has spurred interest in developing new artificial intelligence (AI) models for higher-precision diagnosis, prognosis, and tracking. While the promise of these algorithms is undeniable, their dissemination and adoption have been slow, owing partially to unpredictable AI model performance once deployed in the real world. We posit that one of the rate-limiting factors in developing algorithms that generalize to real-world scenarios is the very attribute that makes the data exciting—their high-dimensional nature. This paper considers how the large number of features in vast digital health data can challenge the development of robust AI models—a phenomenon known as "the curse of dimensionality" in statistical learning theory. We provide an overview of the curse of dimensionality in the context of digital health, demonstrate how it can negatively impact out-of-sample performance, and highlight important considerations for researchers and algorithm designers.

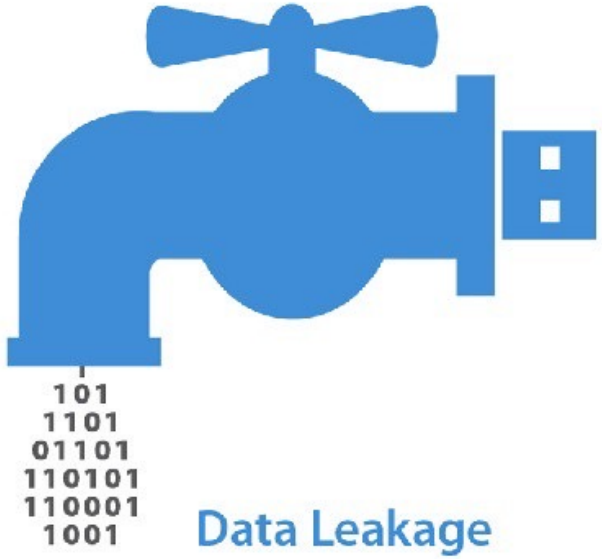
npj Digital Medicine (2021)4:153 ; https://doi.org/10.1038/s41746-021-00521-5



https://www.nature.com/articles/s41746-021-00521-5

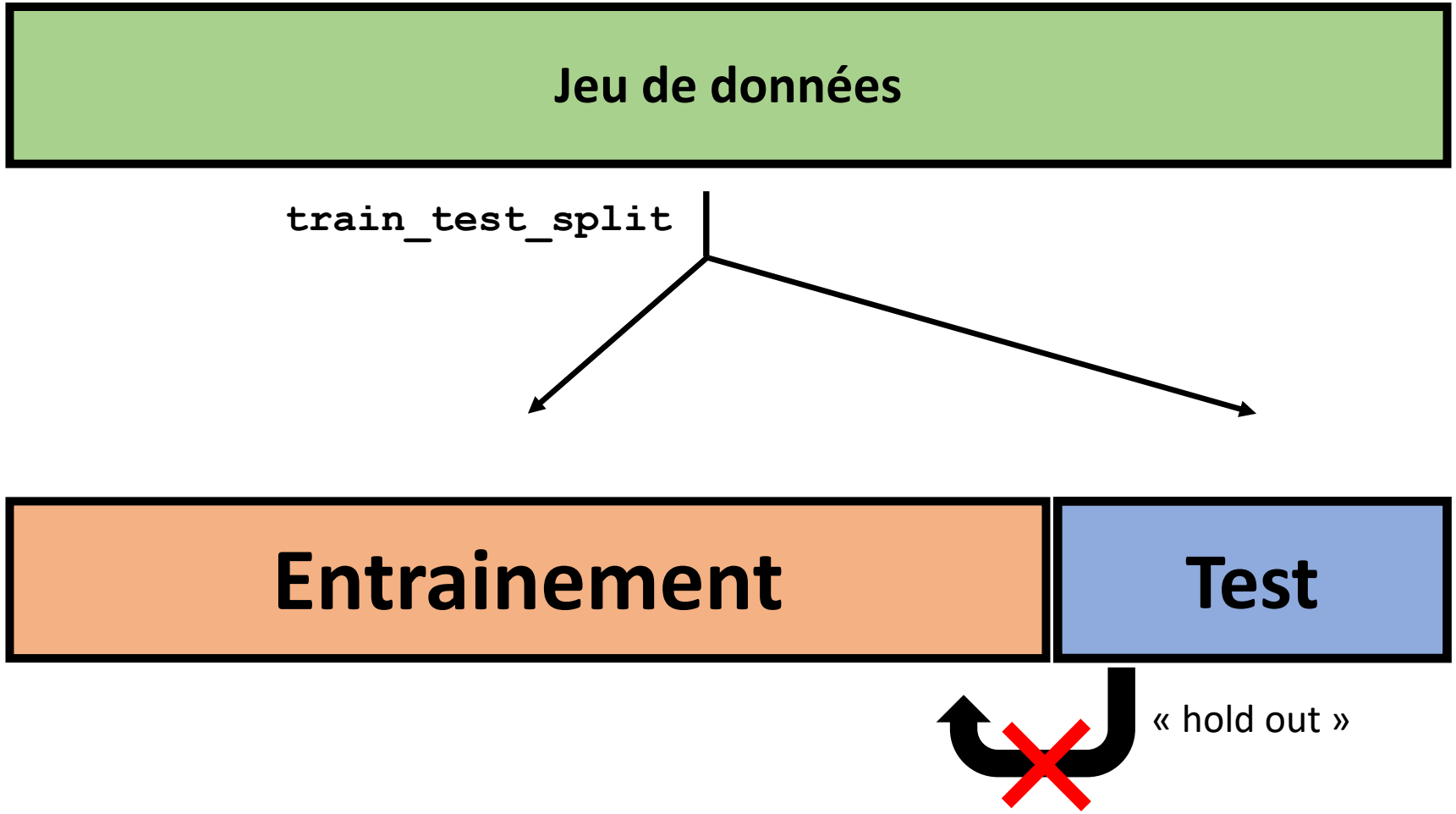
« Nettoyage » et préparation des données

Règle d’or : le futur ne doit pas informer le passé !



https://machinelearningmastery.com/data-leakage-machine-learning/

Identifier les fuites d'information (« Data leakage »)



Identifier les fuites d'information (« Data leakake »)

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

Variables

Échantillons

	V1	V2	V3	V4	V5
S1	1	10	20	4	9
S2	2	7	3	2	8
S3	12	2	9	3	7
S4	2	0	8	6	6
S5	7	3	7	9	5
S6	0	2	4	7	4
S7	0	1	5	8	2
S8	0	2	1	1	3
S9	0	3	4	2	0
S10	1.5	5	4	1	0
	μ1				
	σ1				

Les algorithmes

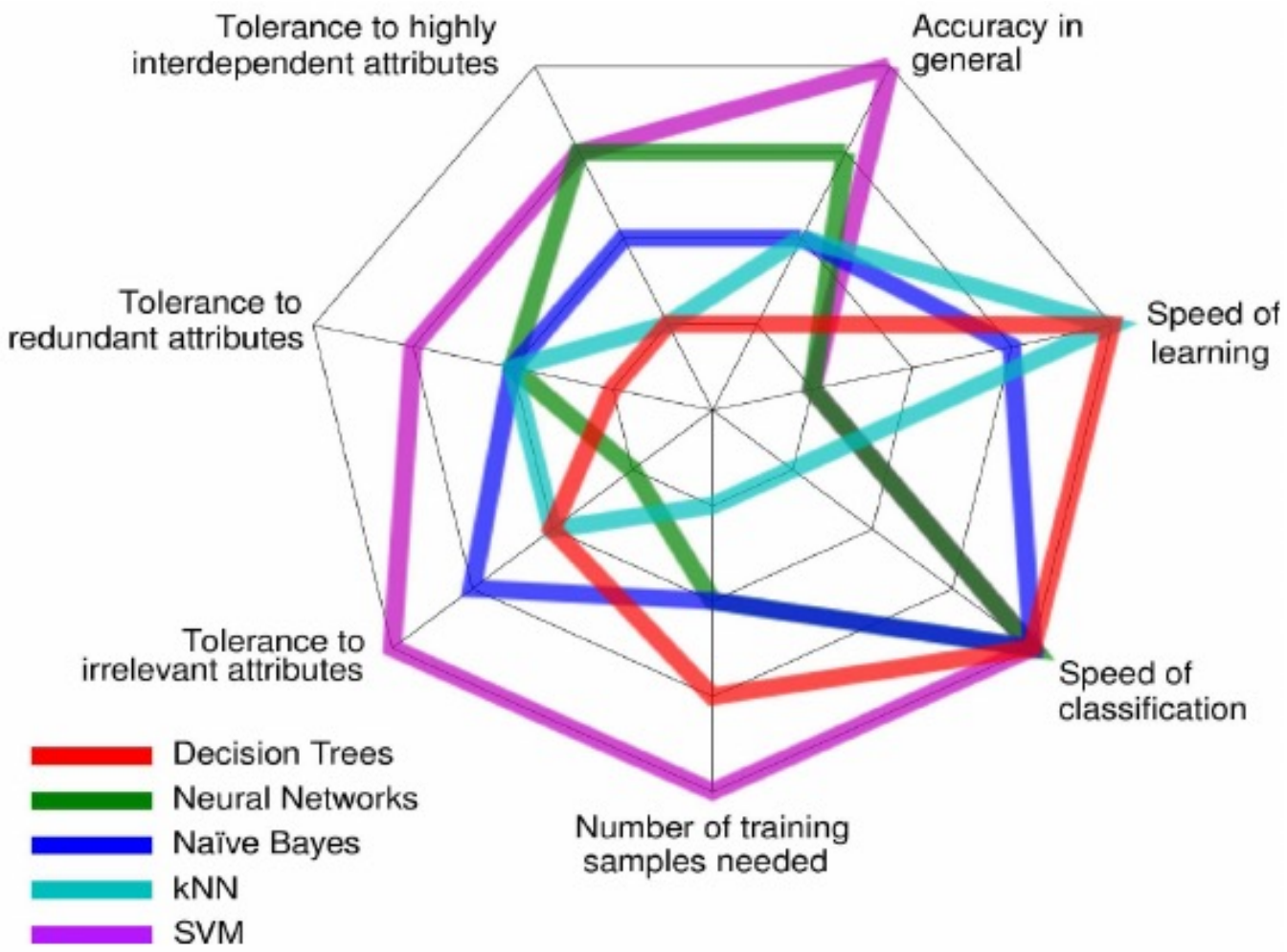
Choisir son algorithme

https://nbisweden.github.io/RaukR-2019/MachineLearning/Presentation_MachineLearning/Presentation_MachineLearning.html#1

Choisir son algorithme

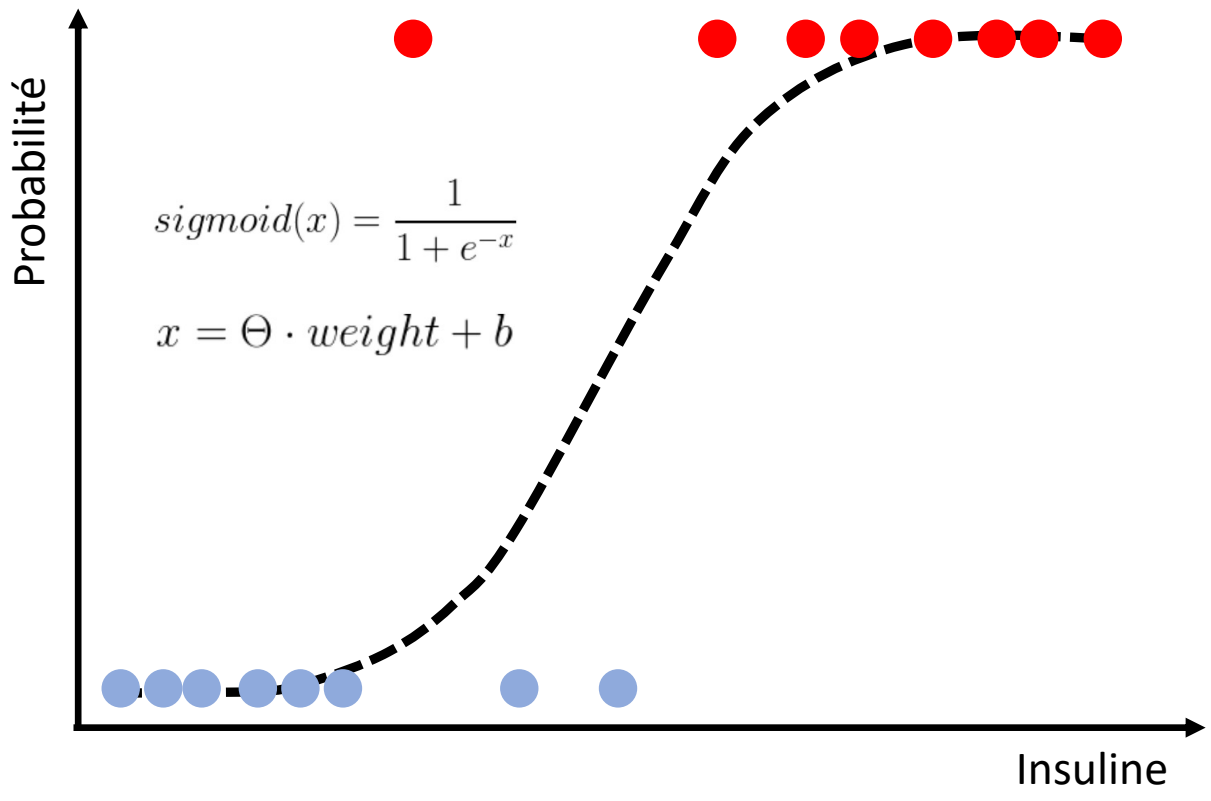
fonction(*question,*
domaine,
taille des données,
qualité des données,
accès à une infrastructure de calcul,
temps de calcul,
interprétabilité, …)

Choisir son algorithme : « performances »



https://www.researchgate.net/publication/229090754_Automated_Inference_System_for_End-To-End_Diagnosis_of_Network_Performance_Issues_in_Client-Terminal_Devices

Régression(!) logistique



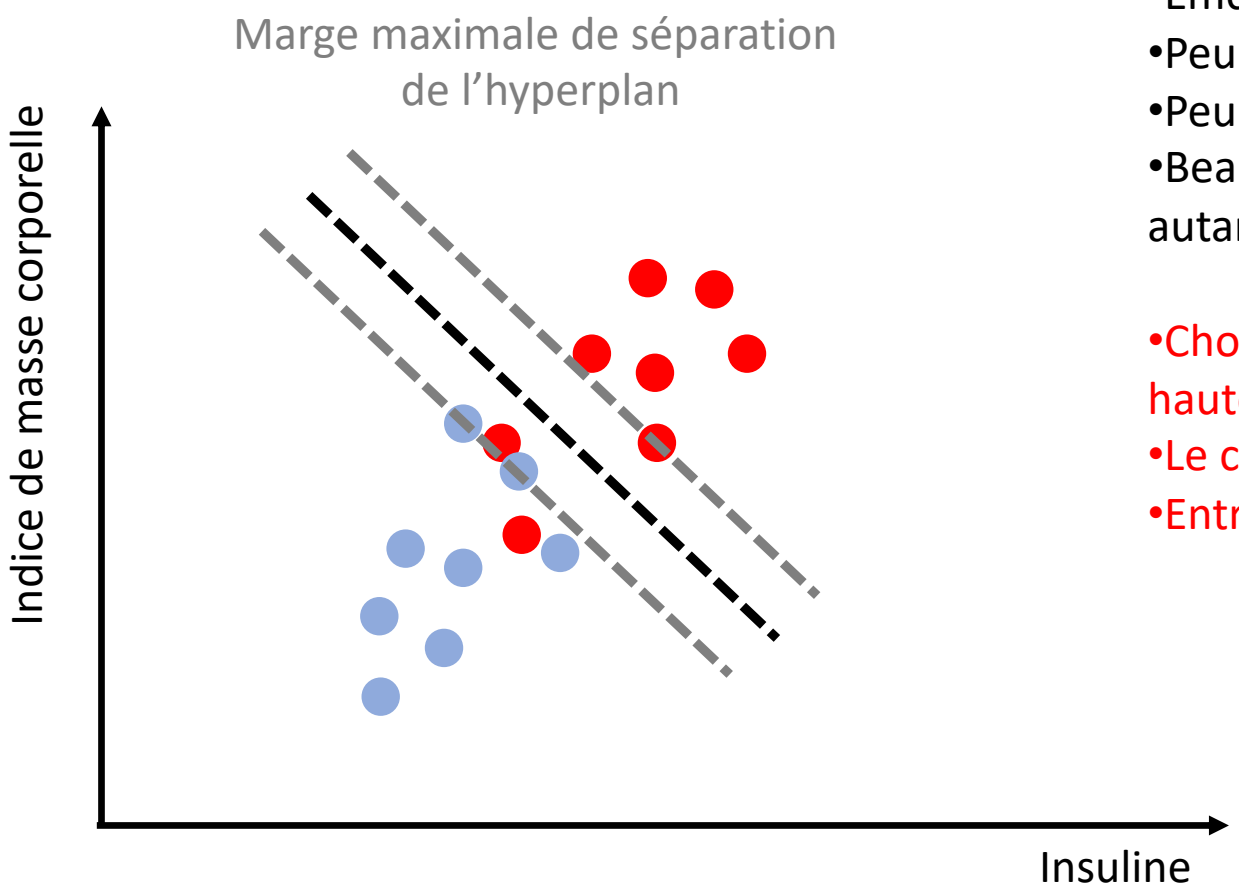
- Simple à implémenter
- Efficace
- Pas besoin de Normalisation/Standardisation
- Fonctionne très bien sans optimisation des hyperparamètres

- Faible performance sur les données non linéaires
- Sensible à la multicolinéarité

Hyperparamètres :
- **penalty**
- **solver**
- **C**

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
https://towardsdatascience.com/quick-and-easy-explanation-of-logistics-regression-709df5cc3f1e

Séparateur à vastes marges (SVM)



- Efficace dans un espace de haute dimension
- Efficace pour les « fat data »
- Peu sensible aux outliners
- Peu gourmand en mémoire vive
- Beaucoup de noyaux différents adaptables à autant de situations

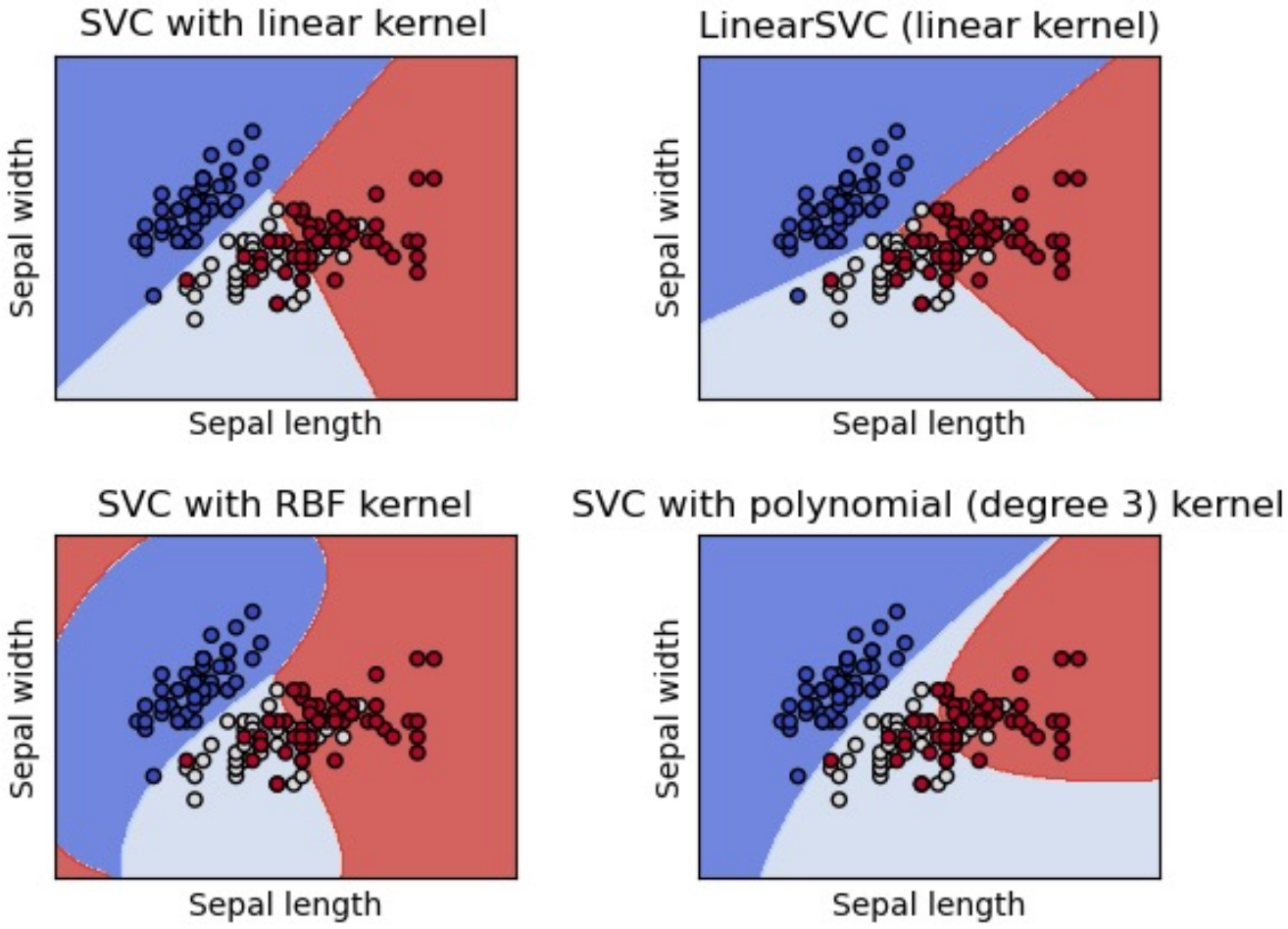
- Choix sensible de noyaux et régularisation en haute dimension
- Le calcul des probabilités est couteux
- Entrainement long

Hyperparamètres :

- **Kernel**
- **C (pénalité)**
- **gamma**

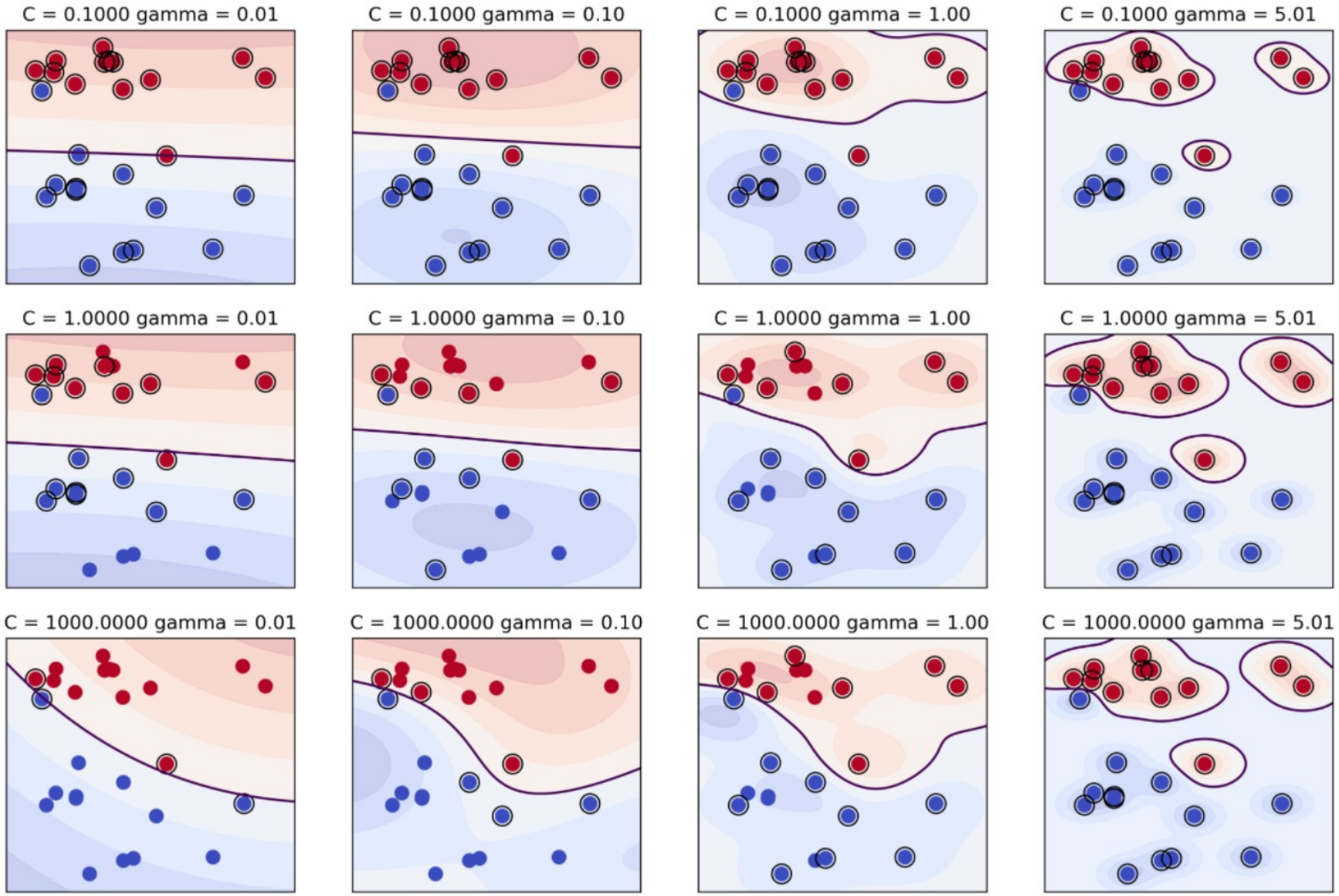
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Séparateur à vastes marges (SVM) : les noyaux



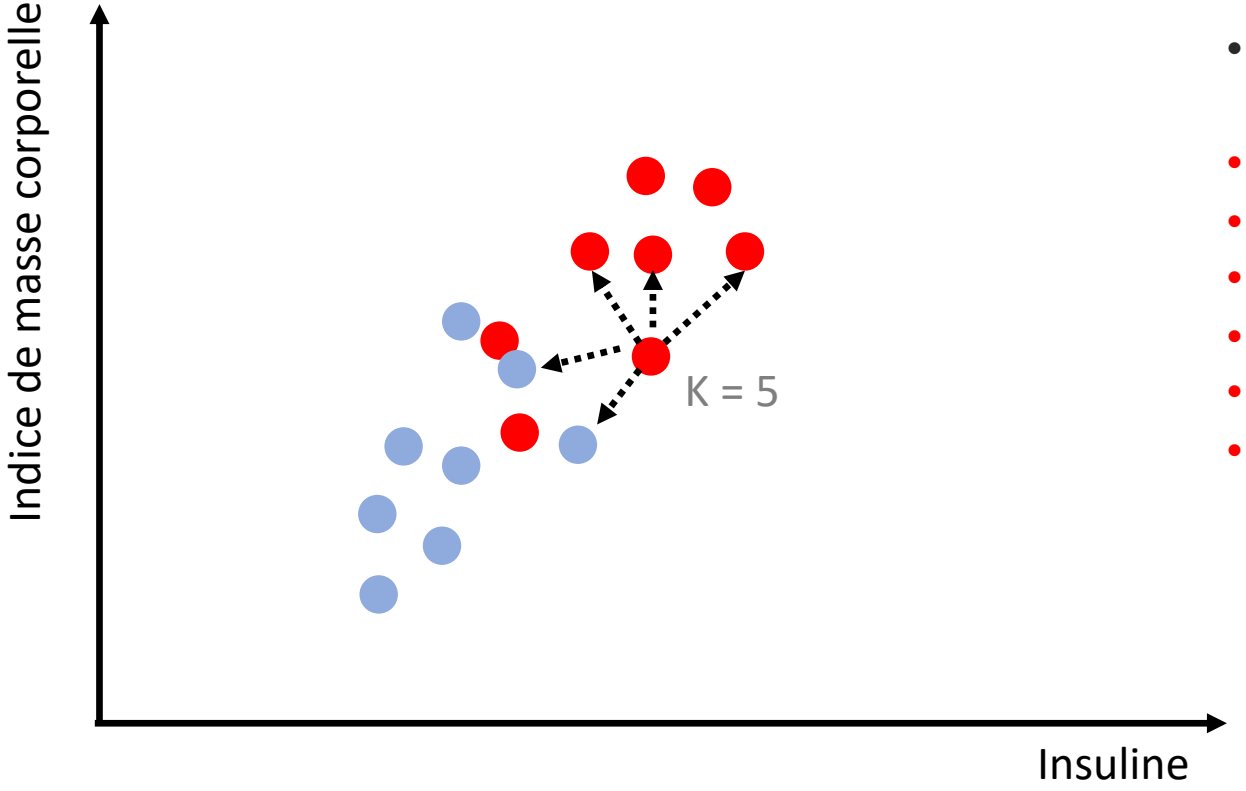
https://scikit-learn.org/stable/modules/svm.html

Séparateur à vastes marges (SVM) : C & gamma



https://amueller.github.io/aml/02-supervised-learning/07-support-vector-machines.html

K-nearest neighbors (KNN)



- Simple
- Non paramétrique
- Modèle adaptatif
- Efficace pour le multi-classe
- Un seul hyperparamètre à optimiser

- Lent pour les gros jeux de données
- Peu adapté à la haute dimension
- Normalisation requise
- Peu efficace sur les données débalancées
- Sensible aux outliners
- Sensible aux données manquantes

Hyperparamètre : k

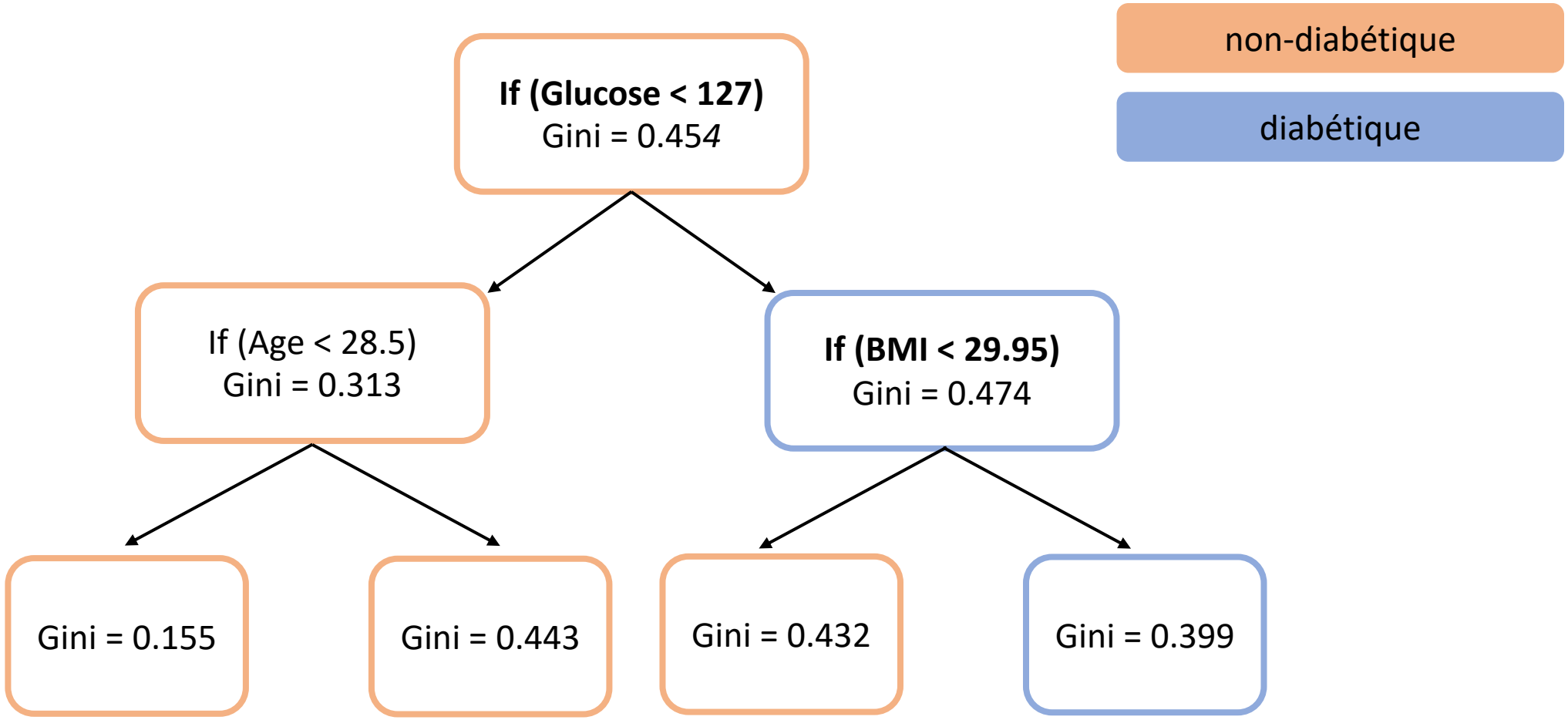
Decision Trees

$$\sum if\ else$$

- Pas besoin de normaliser ou standardiser
- Faible sensibilité aux données manquantes.
- Excellente explicabilité
- Visualisation aisée
- Sélection des variables « automatique »

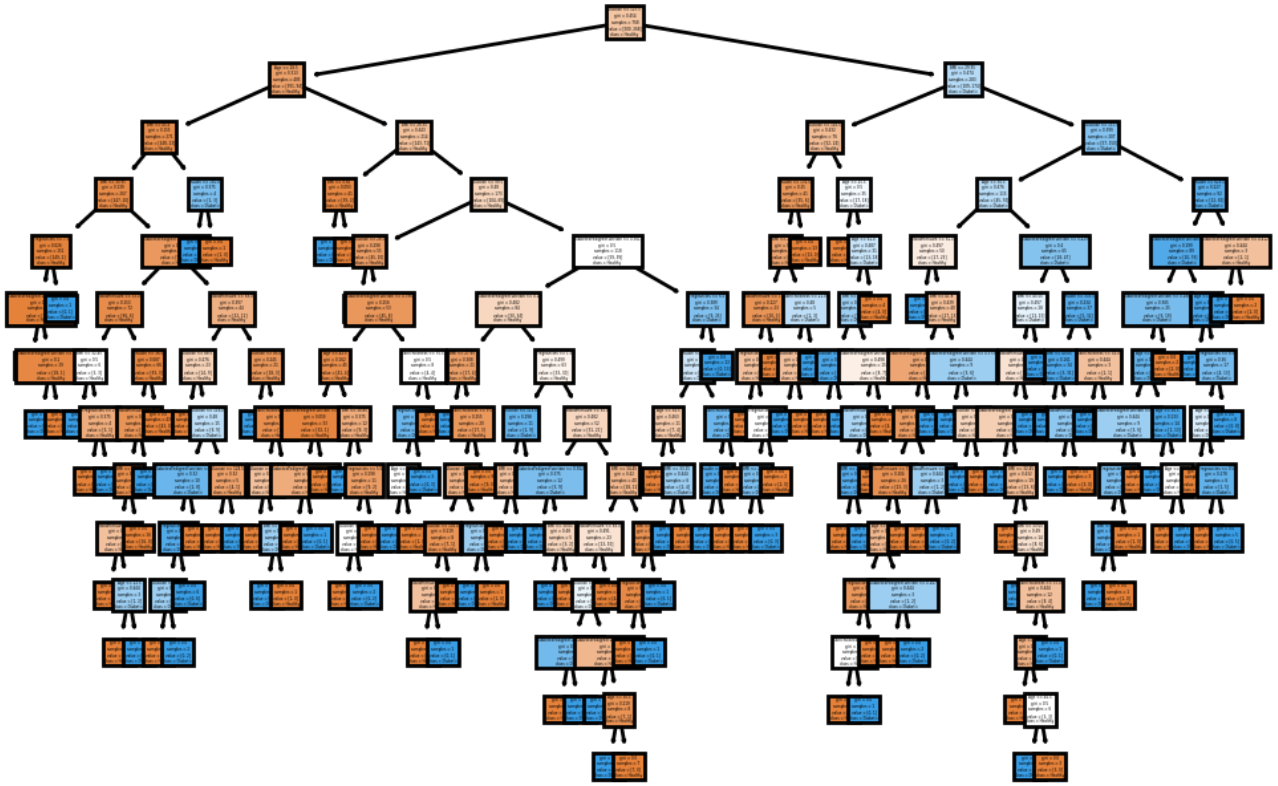
- Risque élevé de surentrainement
- Très sensible aux données
- Faible performance hors des méthodes d'ensemble
- Résultat non optimal

Decision Trees



https://stackoverflow.com/questions/50077562/why-is-my-decision-tree-creating-a-split-that-doesnt-actually-divide-the-sample

Decision Trees

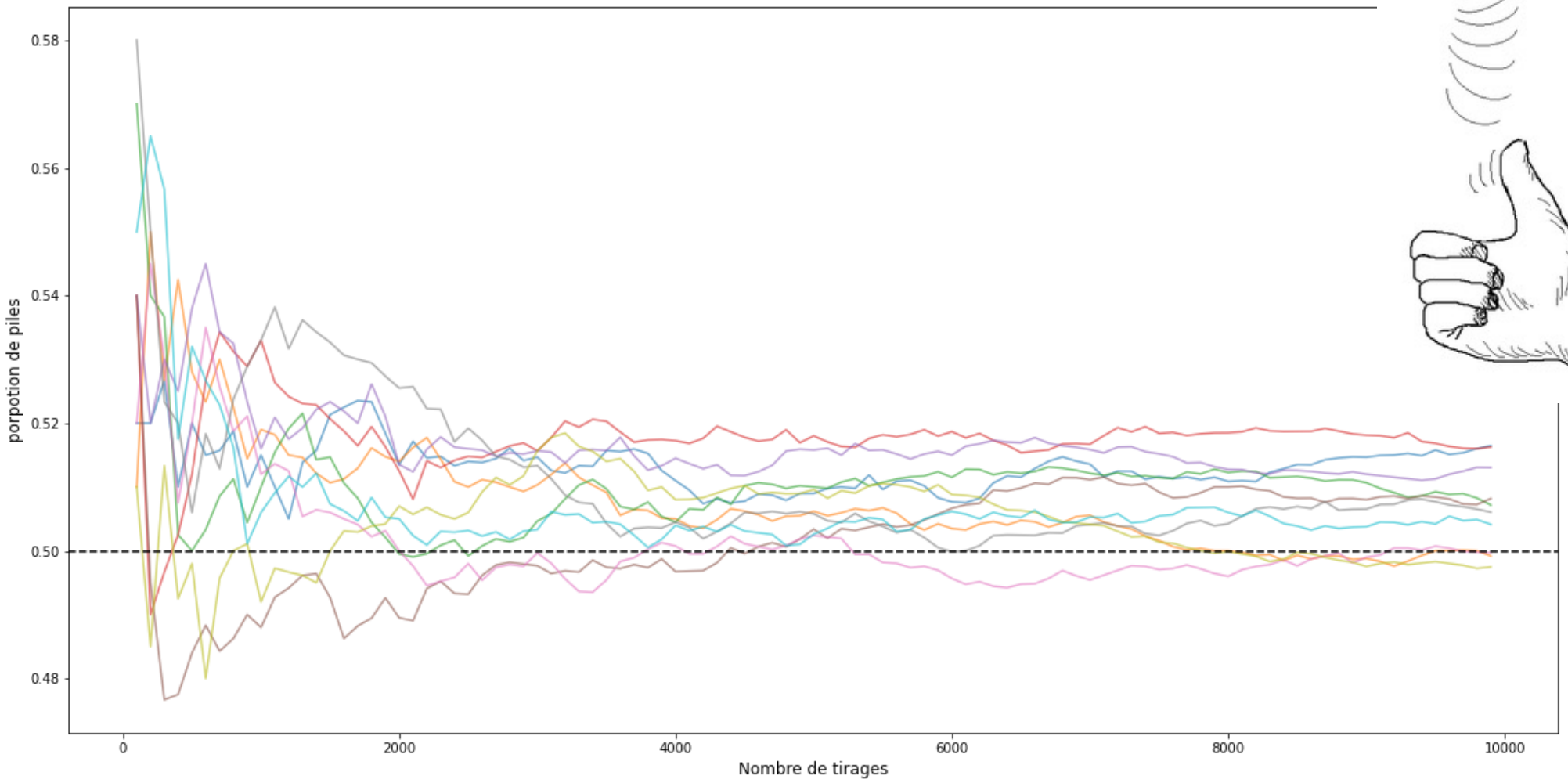


Hyperparamètres :

Criterion
max_depth
min_samples_split
min_samples_leaf
max_features
max_leaf_nodes
min_impurity_decrease

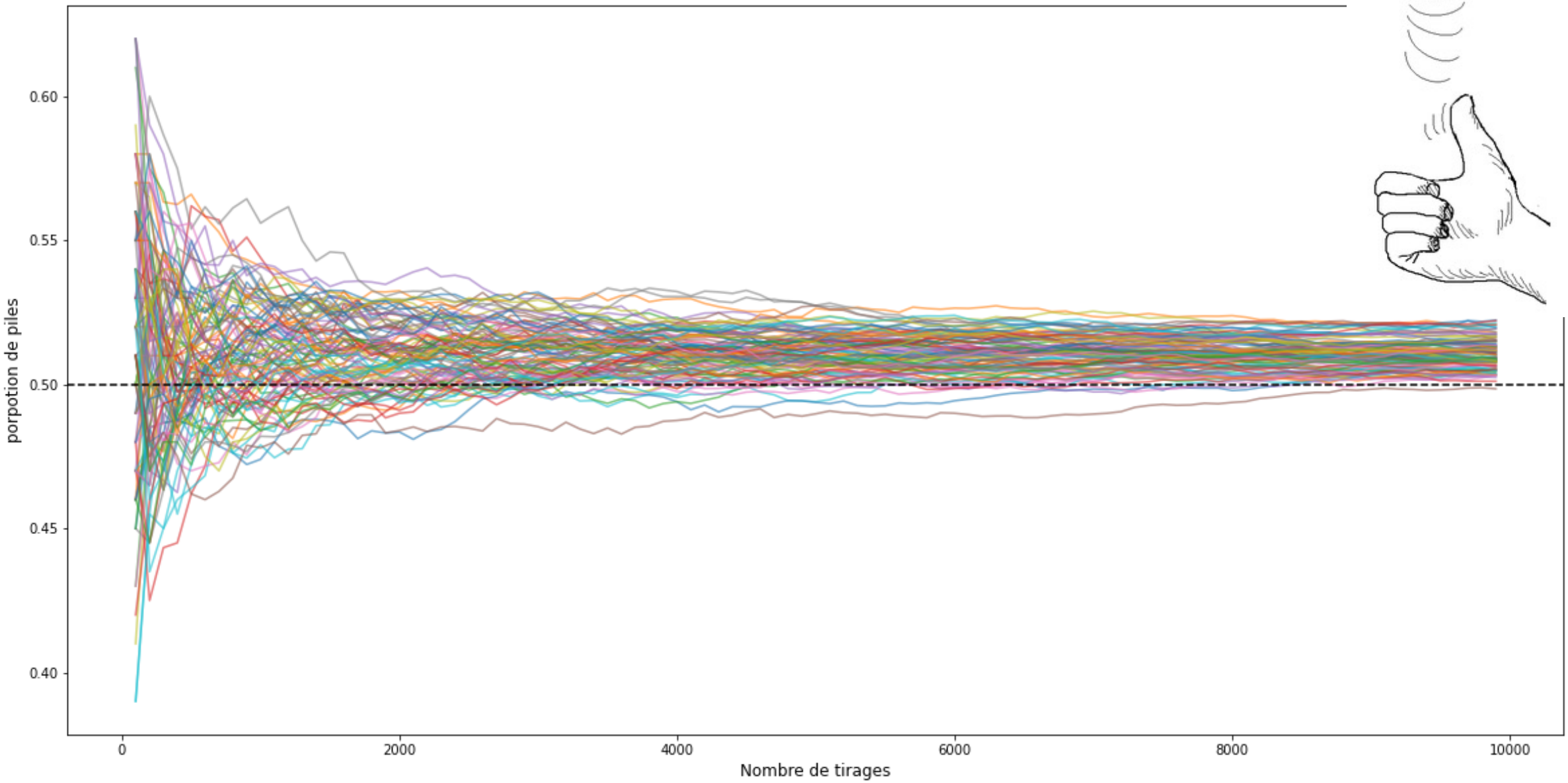
https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Les méthodes d’ensemble



Inspiré de « Hands on Machine Learning with Scikit-Learn & TensorFlow)

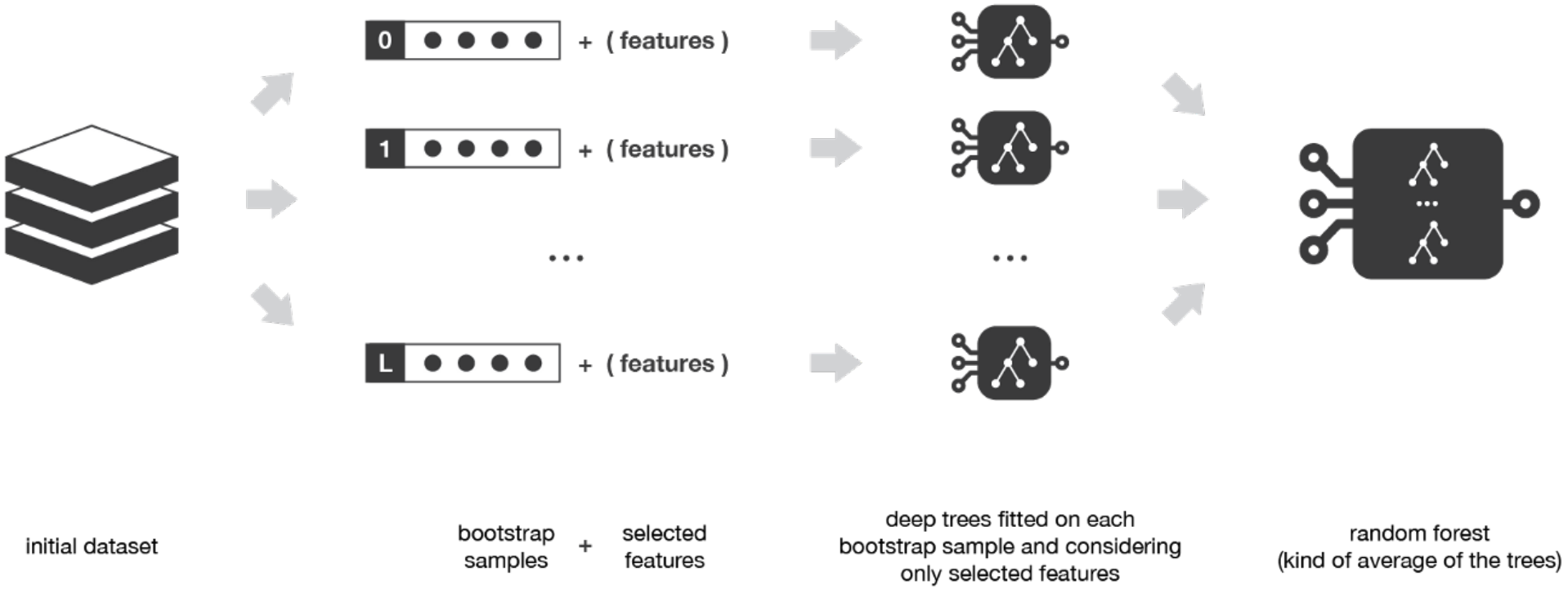
Les méthodes d’ensemble



Inspiré de « Hands on Machine Learning with Scikit-Learn & TensorFlow)

Bagging (Bootstrap Agregating)

RandomForest , ExtraTrees



« heterogeneous ensembles model » : moins de variance

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Bagging (Bootstrap Agregating)

RandomForest , ExtraTrees

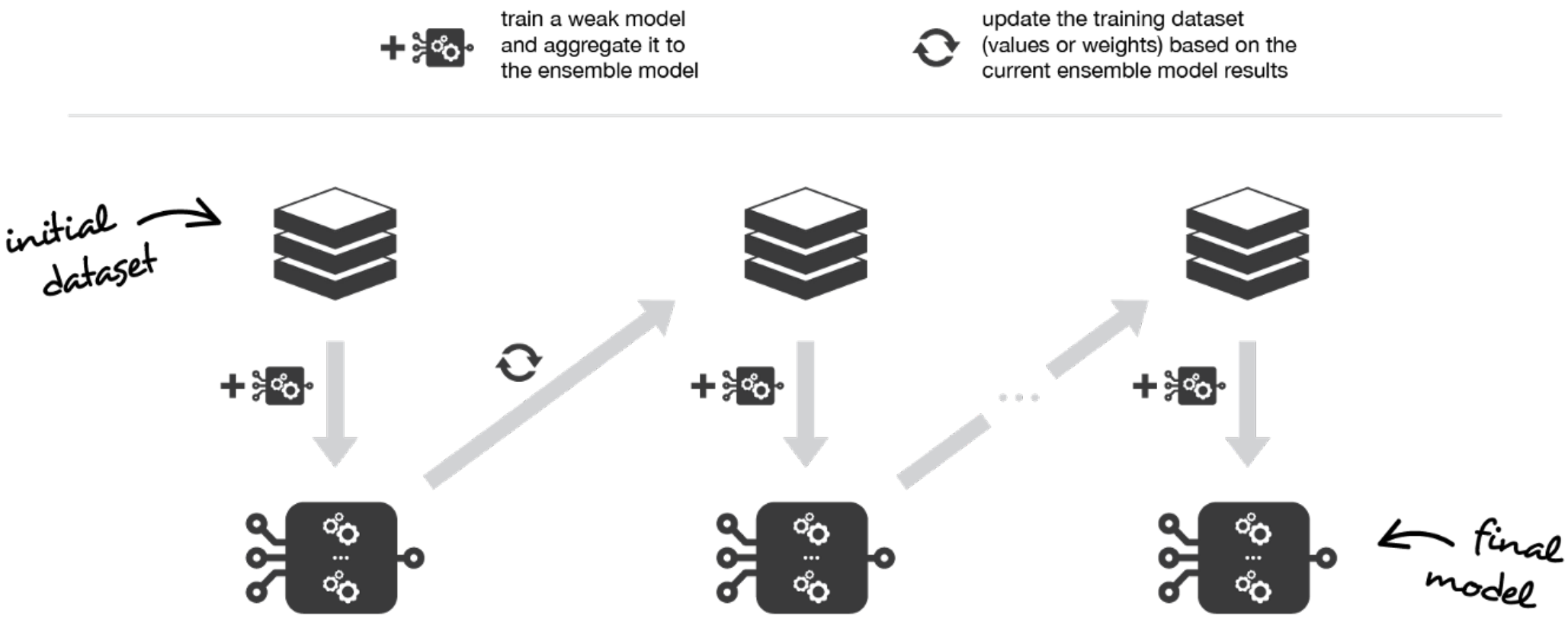
- Réduction de l'erreur et de la variance:
- Bonne performance sur les jeux de donnée débalancés:
- Peut gérer des données massives et de haute dimension
- Peu sensible aux données manquantes:
- Peu sensible aux outliners
- Peut être utilisé pour l'extraction de variables

- Entrainement lent
- Risque de biais avec les variables catégorielles

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Boosting

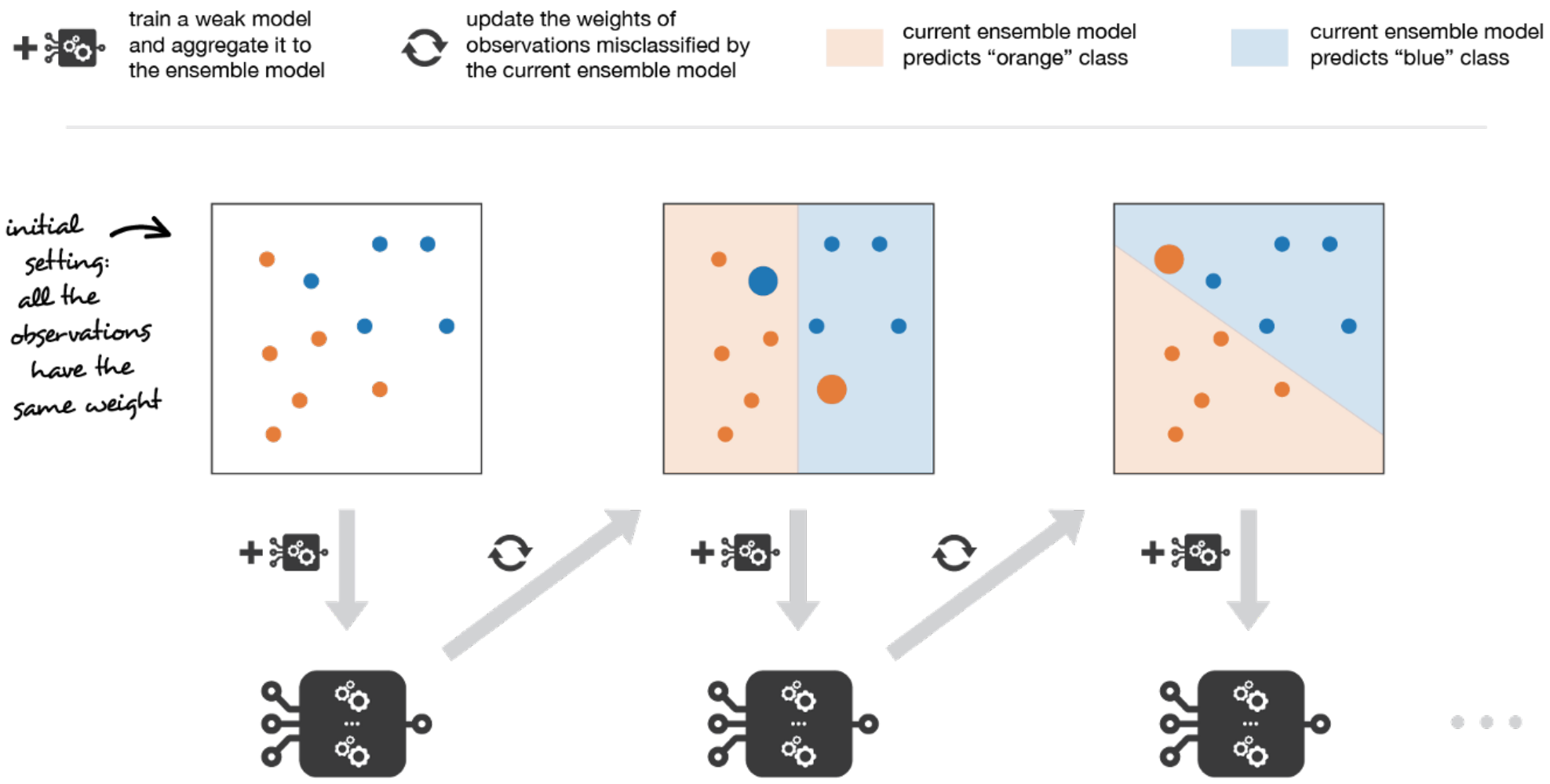
AdaBoost, GradientBoosting



« homogenous ensembles model » : moins de biais

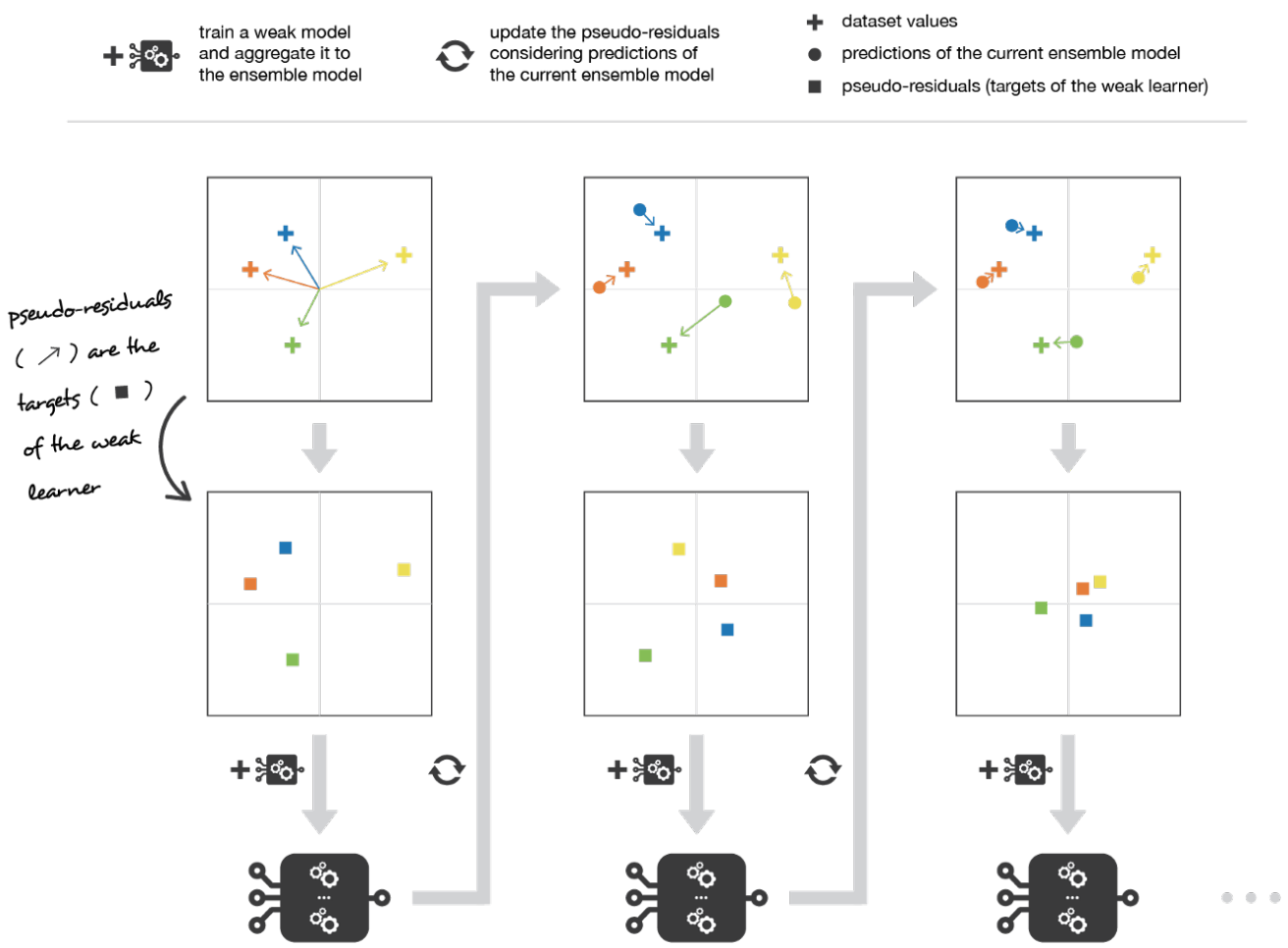
https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Adaptative Boosting



https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Gradient Boosting



https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

L’évaluation

Les métriques d'évaluation : matrice de confusion

Réalité

		Vrai	Faux
Prédictions du modèle	Vrai		
	Faux		

https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Les métriques d'évaluation : matrice de confusion

		Réalité	
		Vrai	Faux
Prédictions du modèle	Vrai	Vrais positifs	
	Faux		Vrais négatifs

https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Les métriques d'évaluation : matrice de confusion

		Réalité	
		Vrai	Faux
Prédictions du modèle	Vrai	Vrais positifs	Faux positifs
	Faux	Faux négatifs	Vrais négatifs

https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Les métriques d'évaluation

Exactitude = $\frac{Nb\ correct\ predictions}{Nb\ total\ predicions}$
(Accuracy)

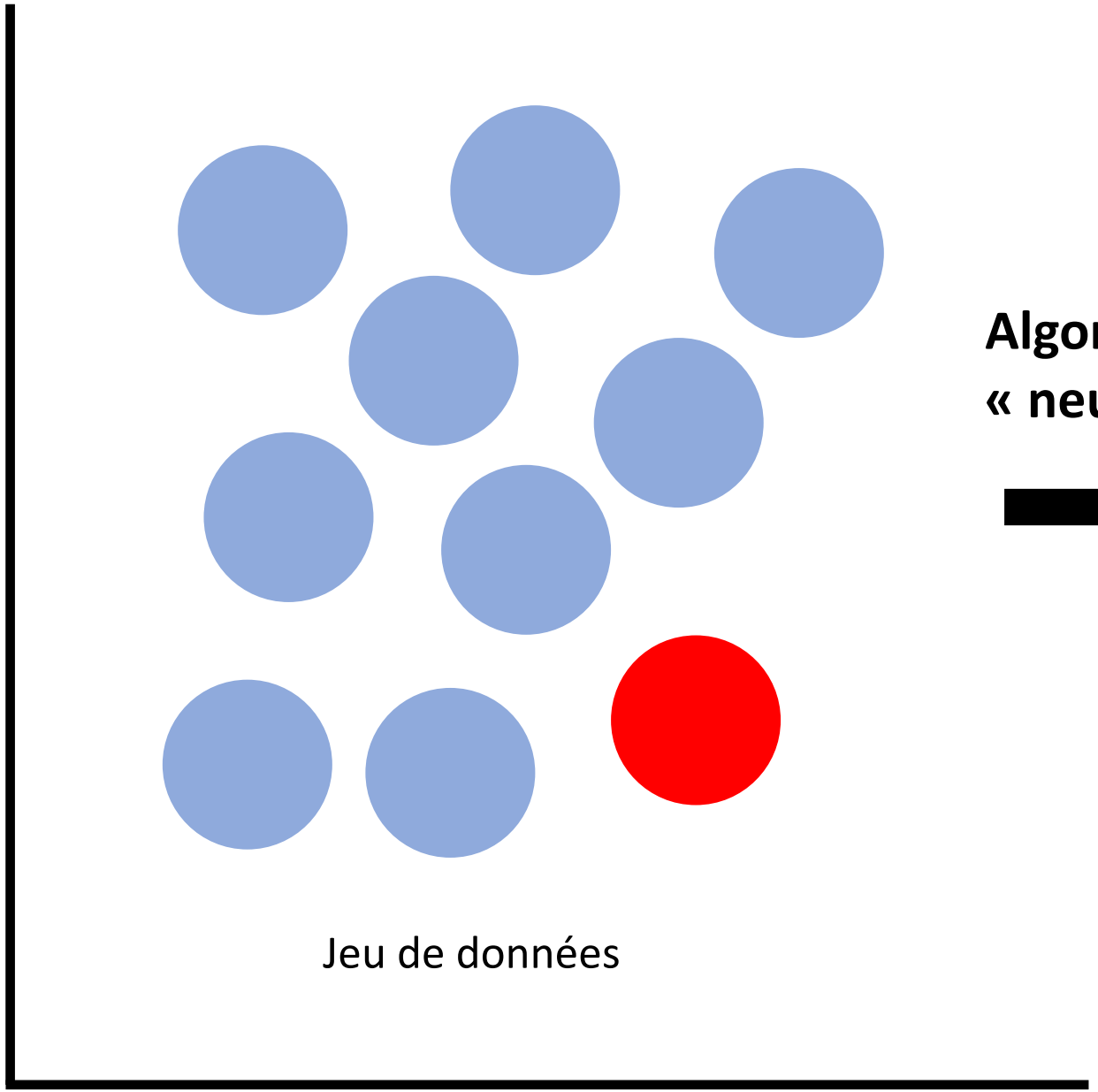
Spécificité = $\frac{True\ positives}{True\ positives + False\ positives}$
(Precision)

Sensibilité = $\frac{True\ positives}{True\ positives + False\ negatives}$
(Recall)

F1 = $\frac{2\ * Precision\ * Recall}{Precision + Recall}$
(F1)

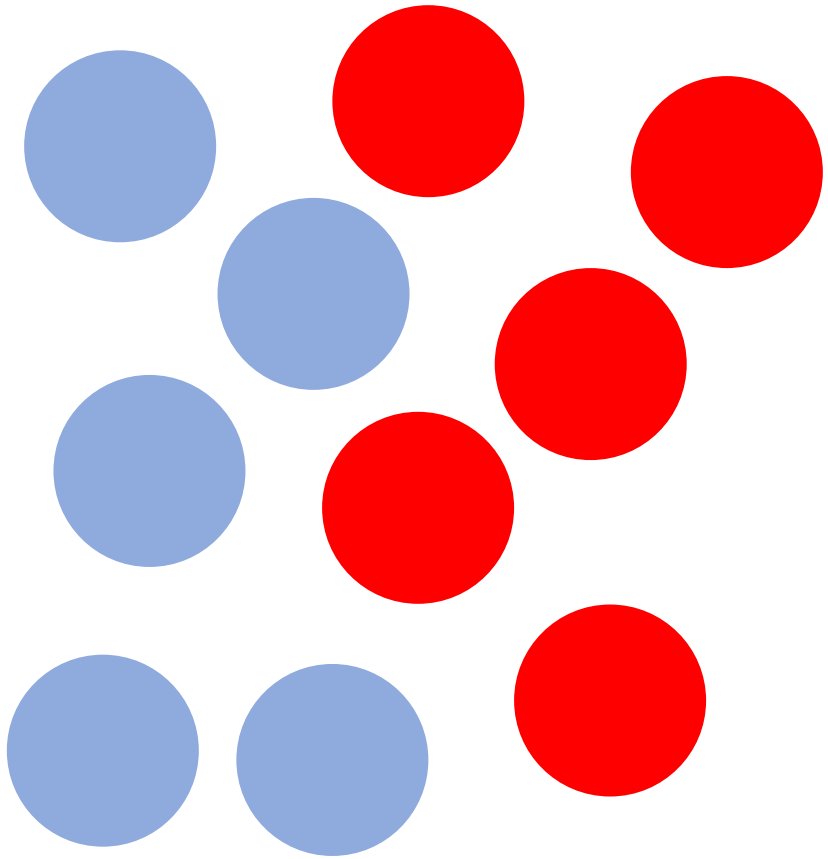
https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021

Les métriques d'évaluation



Les métriques d'évaluation

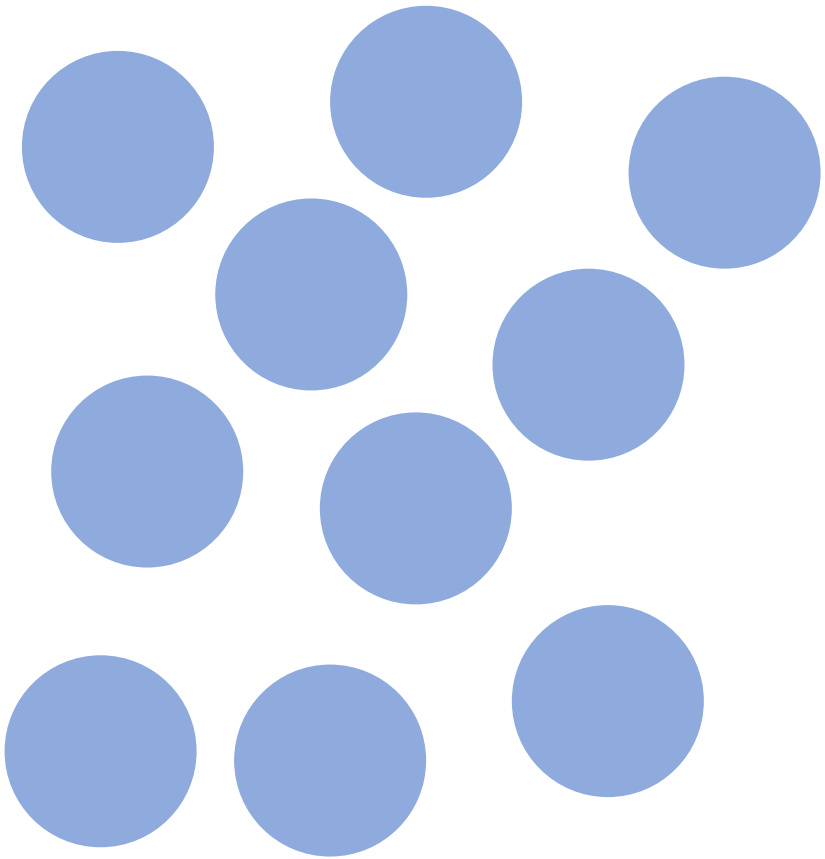
« Paradoxe de l'exactitude »



Jeu de données

Algorithme

« neuneu »

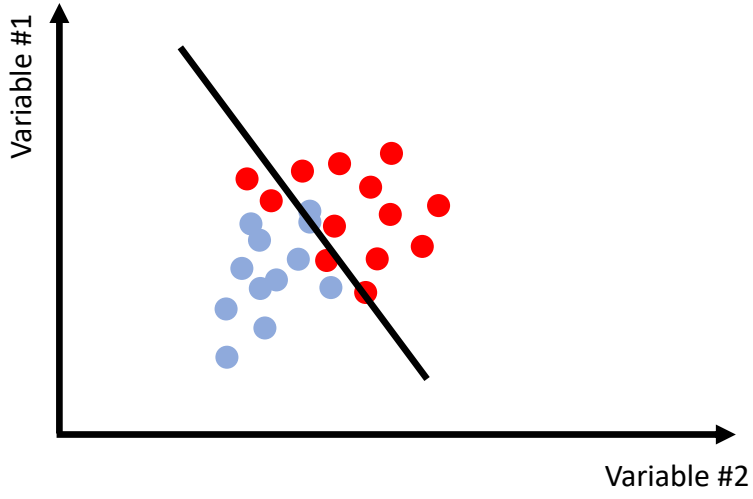


Prédiction

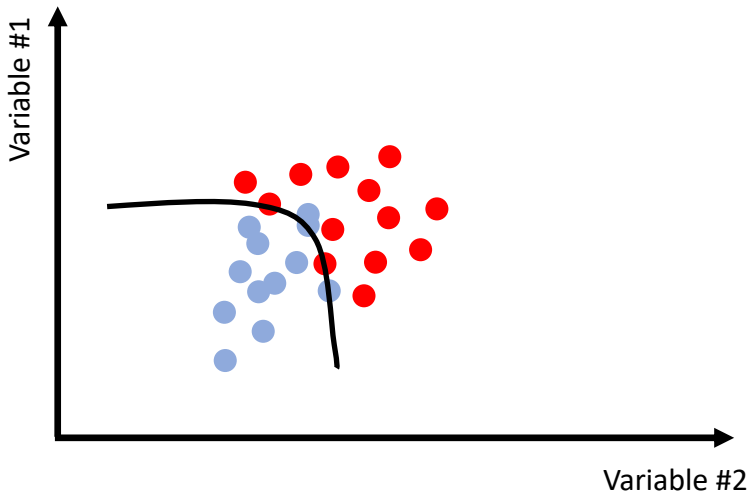
Accuracy = 50%

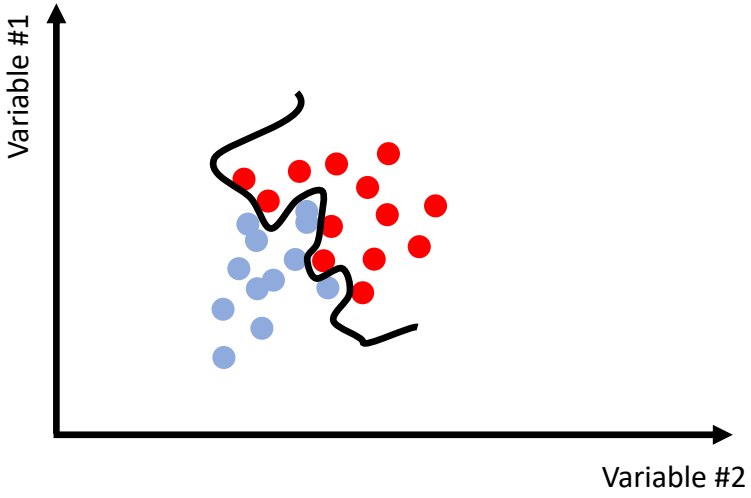
L’entrainement et la sélection

Sur et sous-apprentissage



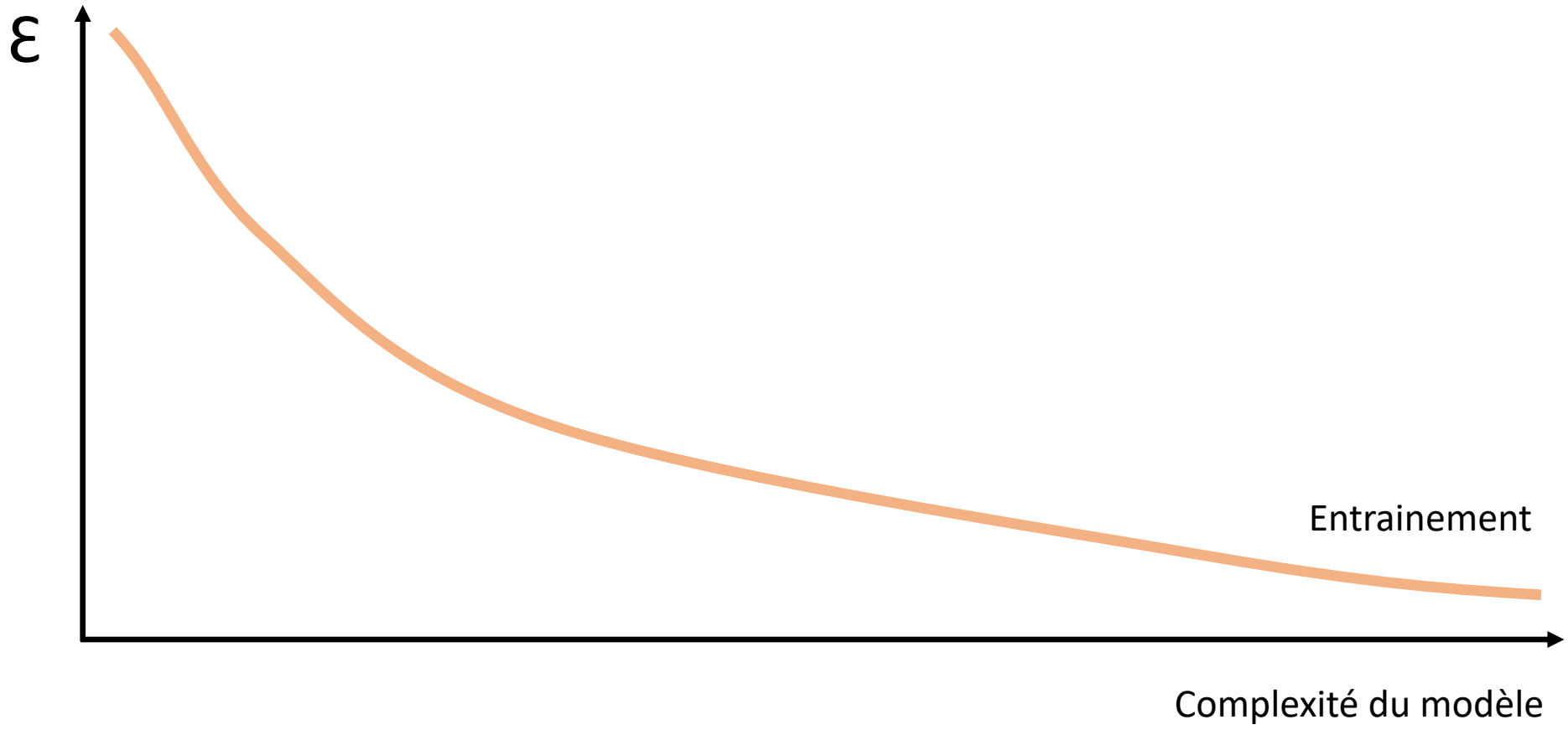
sous-apprentissage



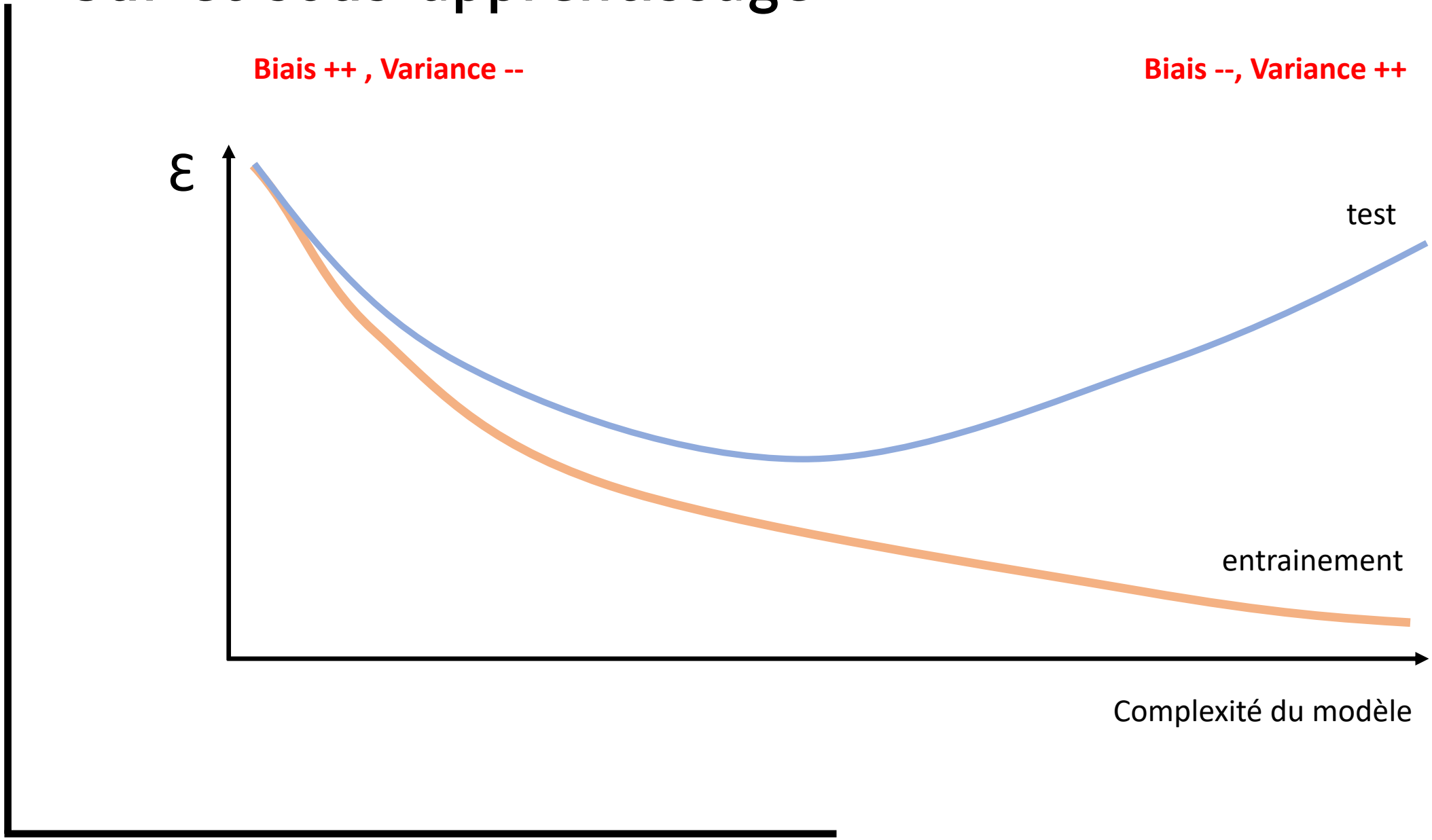


sur-apprentissage

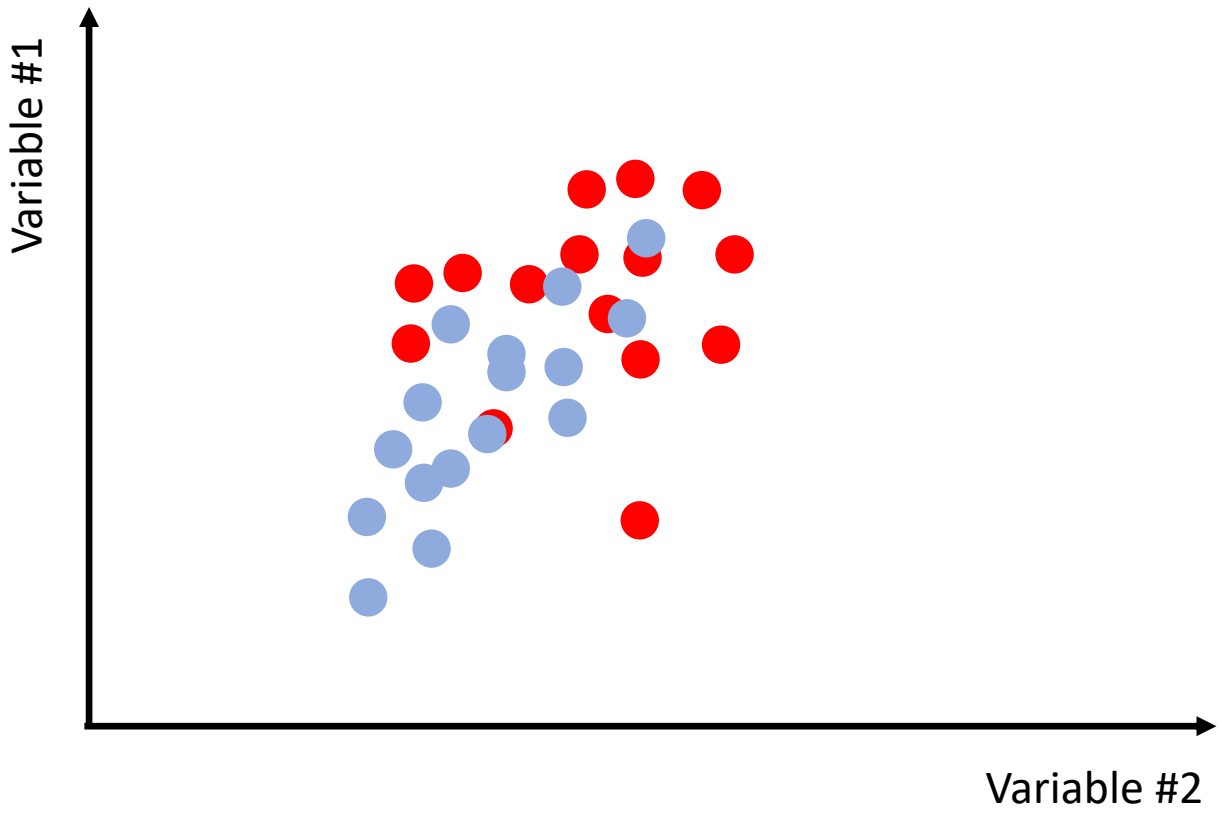
Sur et sous-apprentissage



Sur et sous-apprentissage

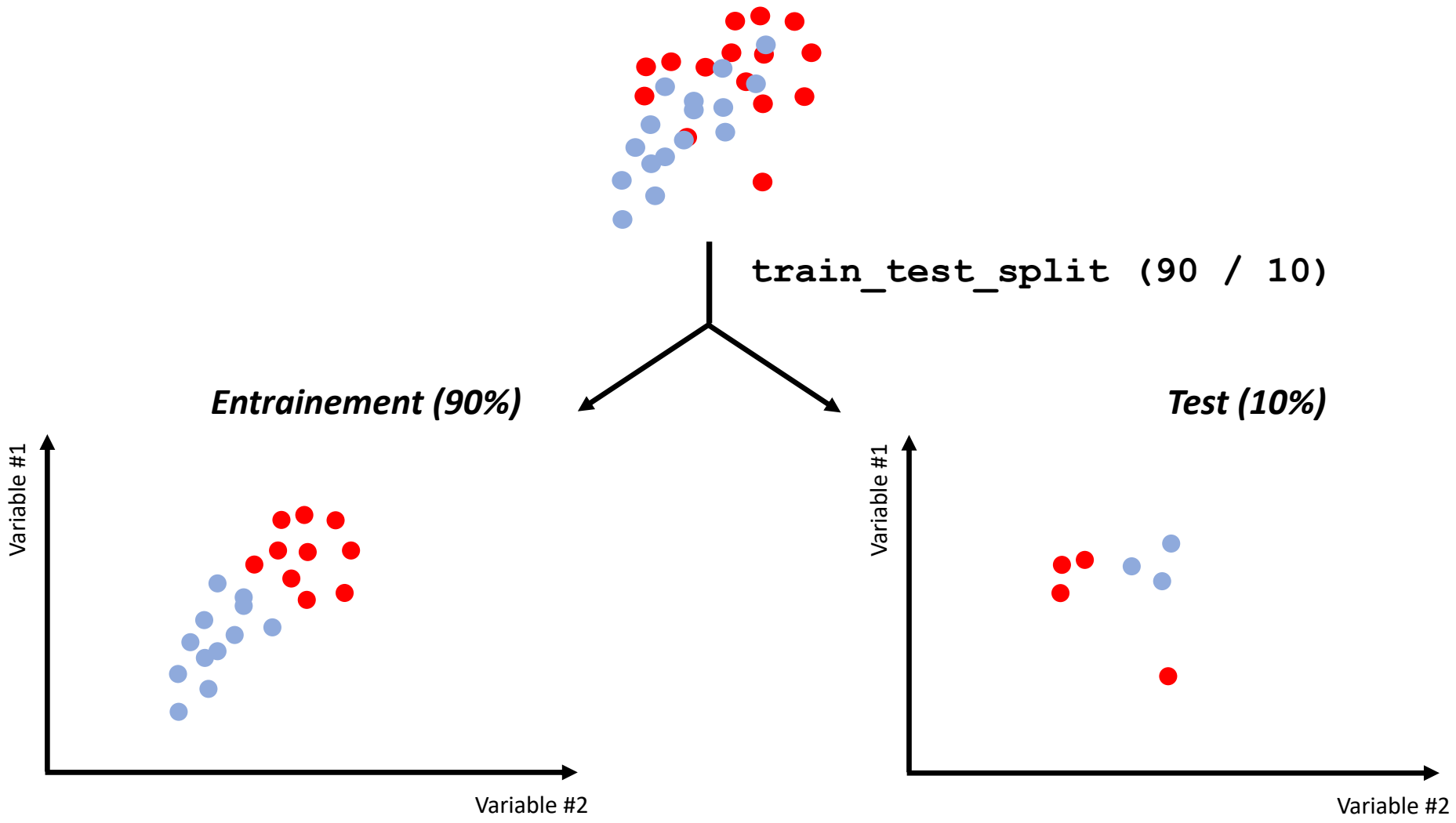


Pourquoi la validation croisée ?



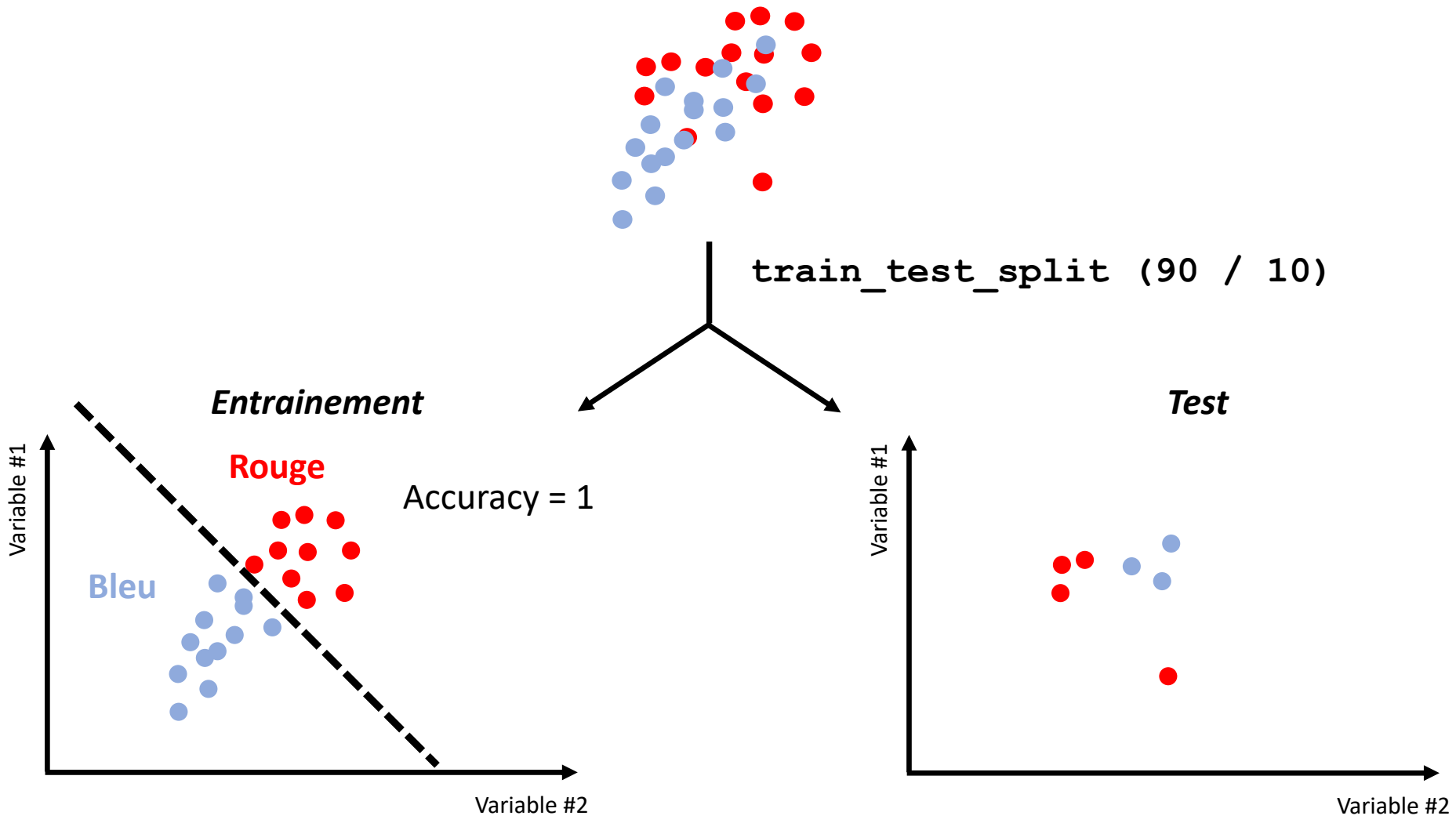
https://scikit-learn.org/stable/modules/cross_validation.html

Pourquoi la validation croisée ?



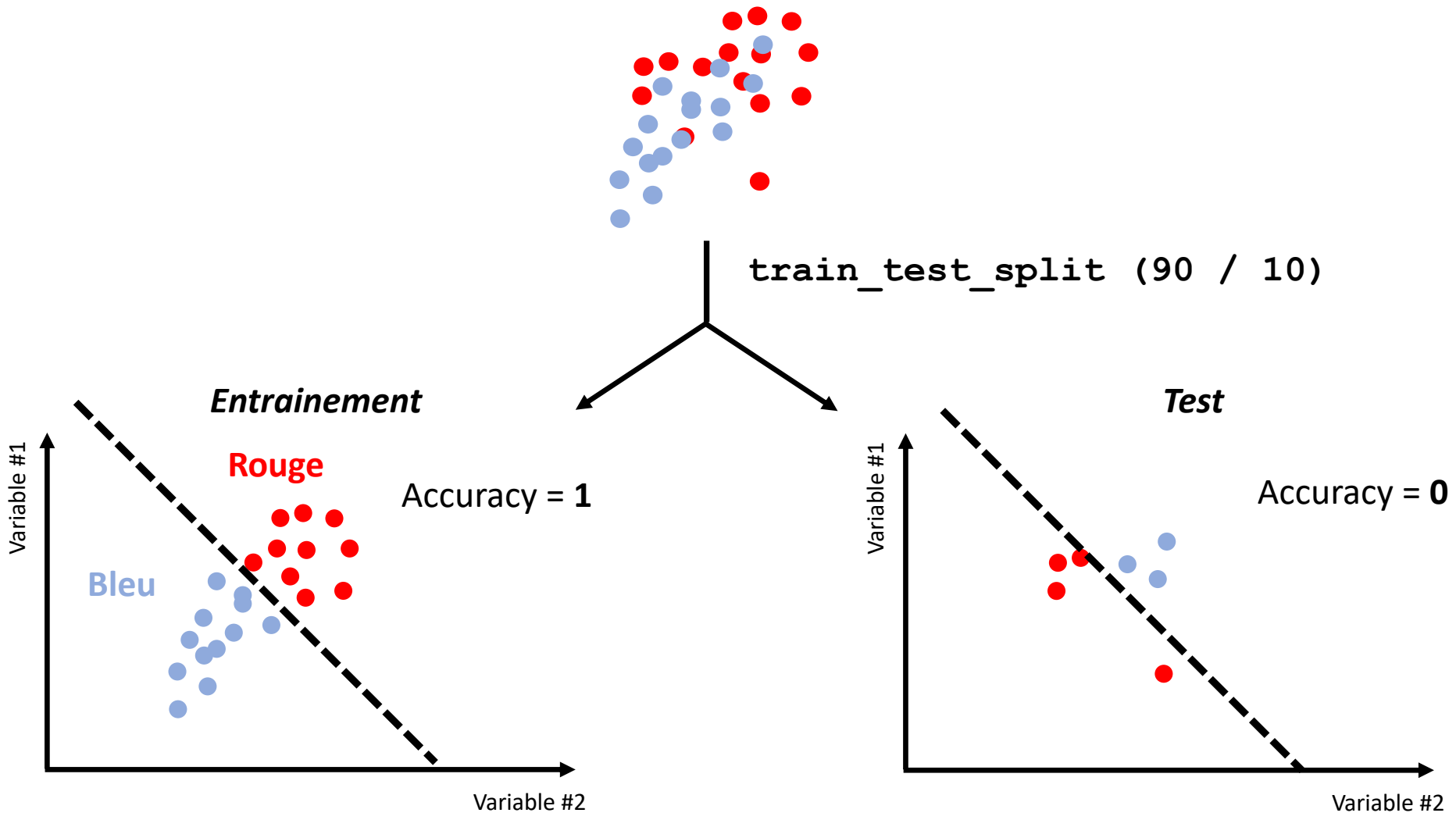
https://scikit-learn.org/stable/modules/cross_validation.html

Pourquoi la validation croisée ?



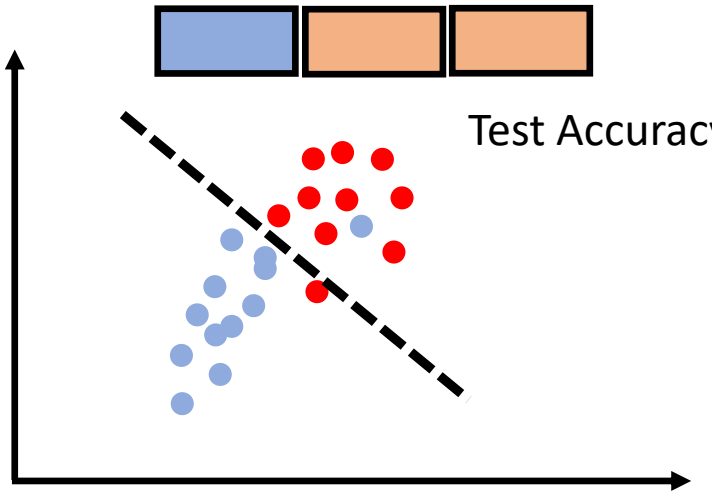
https://scikit-learn.org/stable/modules/cross_validation.html

Pourquoi la validation croisée ?

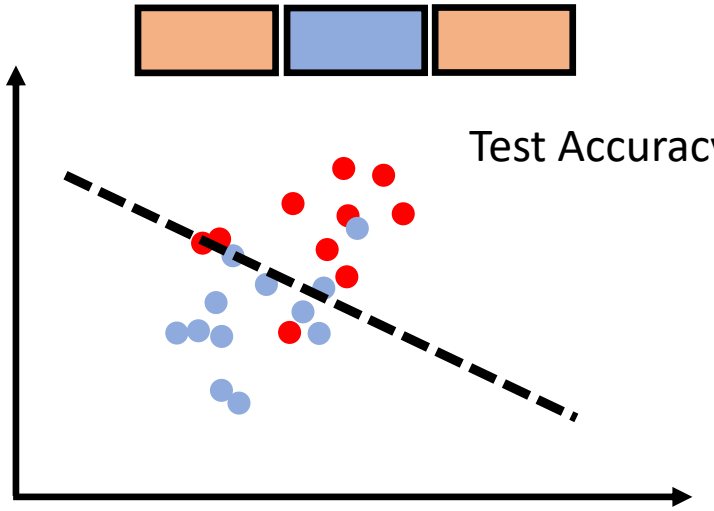


https://scikit-learn.org/stable/modules/cross_validation.html

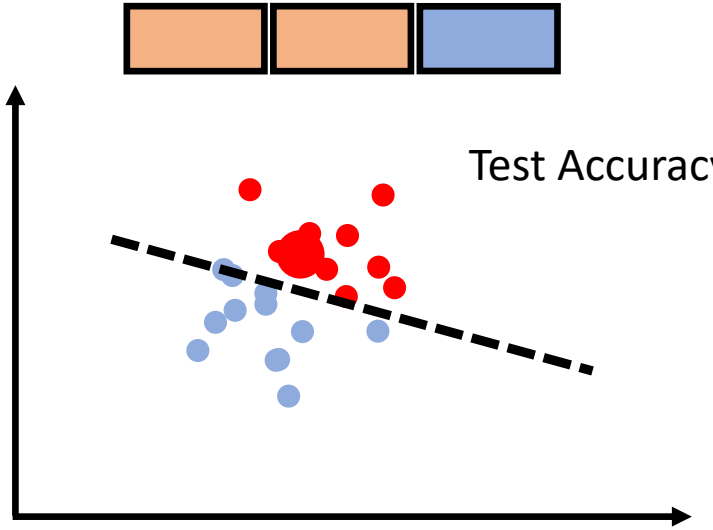
Validation croisée à k-plis (Kfolds cross-validation)



Test Accuracy = 0.95



Test Accuracy = 0.91



Test Accuracy = 1

$\mu = (0.95 + 0.91 + 1) / 3 = 0.9533$

$\sigma = 0.045$

https://scikit-learn.org/stable/modules/cross_validation.html

Les avantages (et désavantages) de la validation croisée

- Détection du sur-apprentissage
- Estimation de la performance moyenne du modèle et de sa robustesse
- Utilisation de la totalité de l'information pour la sélection du modèle
- Ajustement des hyperparamètres

- Tous ces avantages ont un coût : le calcul

https://scikit-learn.org/stable/modules/cross_validation.html

Pas une mais des validations croisées

Leave One Out (LOO)

Leave P Out (LPO)

Stratified k-fold

Leave P Groups Out

Repeated K-Fold

Stratified Shuffle Split

Leave One Group Out

Group Shuffle Split

Random permutations cross-validation a.k.a. Shuffle & Split

Time Series Split

https://scikit-learn.org/stable/modules/cross_validation.html

L’optimisation des hyperparamètres

GridSearchCV

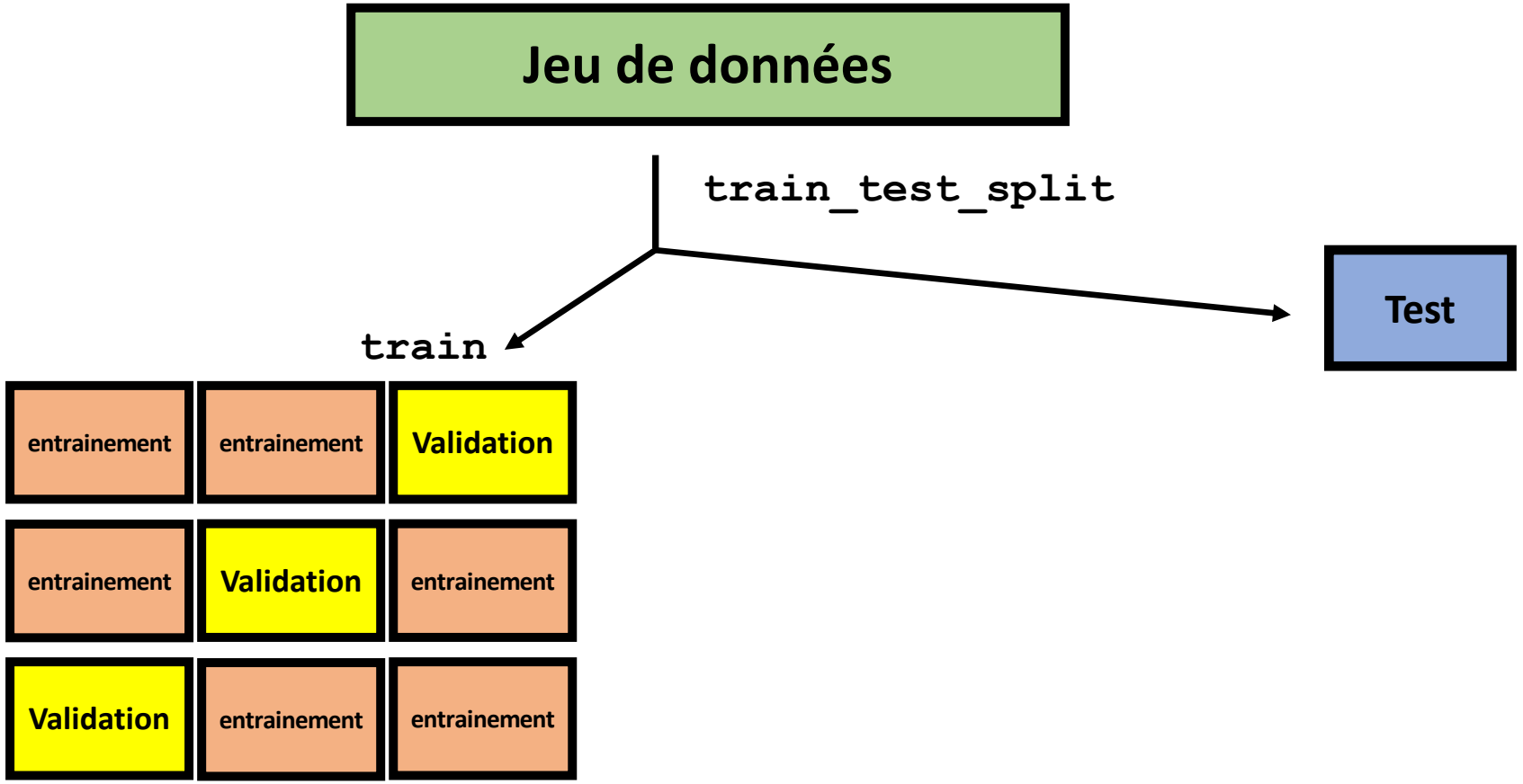
SVM { C : [0.1, 1] , gamma : [1,10] }

SVM { **C : 0.1** , **gamma : 1** } SVM { **C : 0.1** , **gamma : 10** } SVM { **C : 1** , **gamma : 1** } SVM { **C : 1** , **gamma : 10** }

GridSearchCV = Force brute.
Alternatives : Random Search, Bayesian optimisation

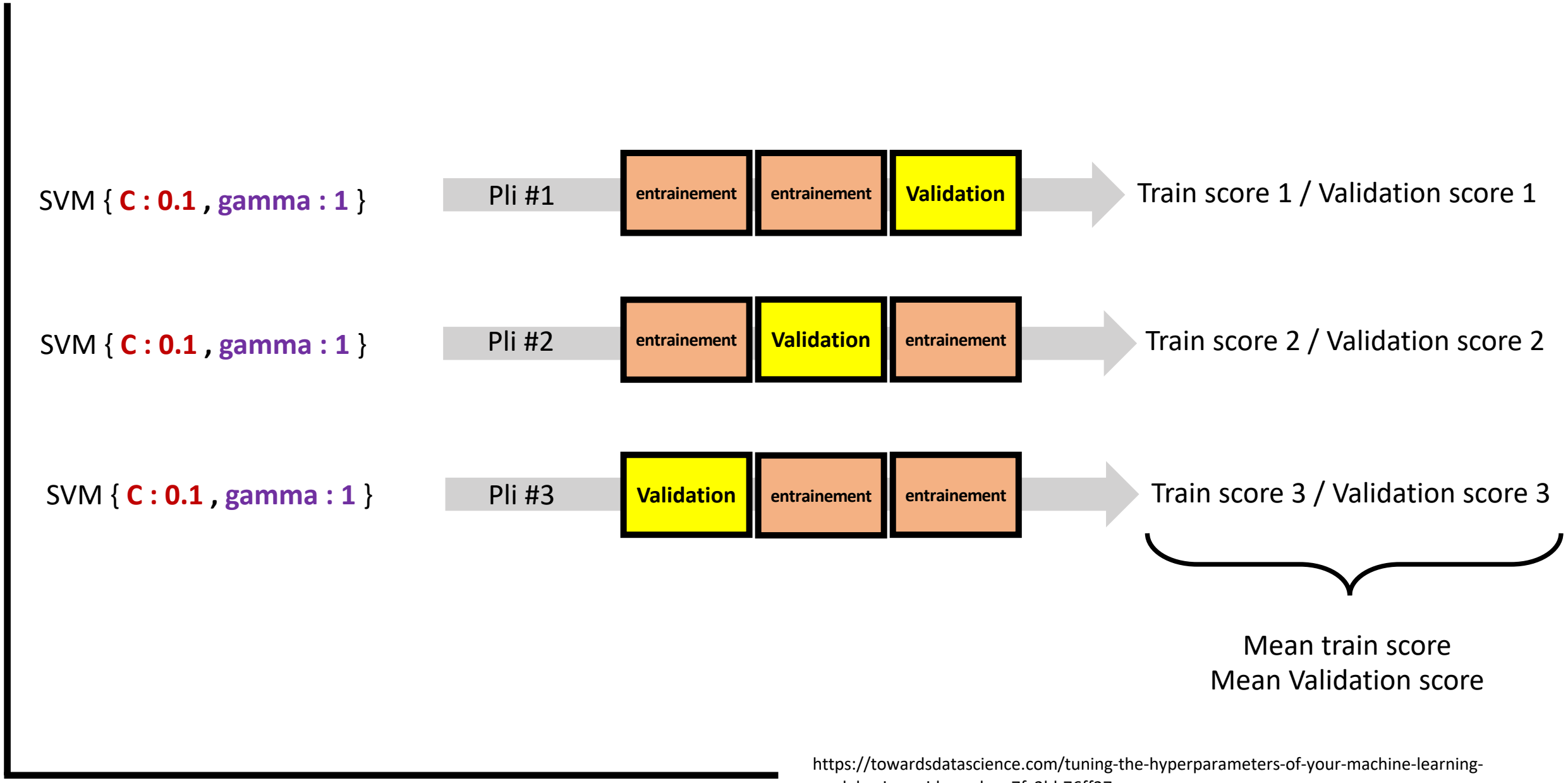
https://towardsdatascience.com/tuning-the-hyperparameters-of-your-machine-learning-model-using-gridsearchcv-7fc2bb76ff27

GridSearchCV



https://towardsdatascience.com/tuning-the-hyperparameters-of-your-machine-learning-model-using-gridsearchcv-7fc2bb76ff27

GridSearchCV



https://towardsdatascience.com/tuning-the-hyperparameters-of-your-machine-learning-model-using-gridsearchcv-7fc2bb76ff27

Passons à la pratique !