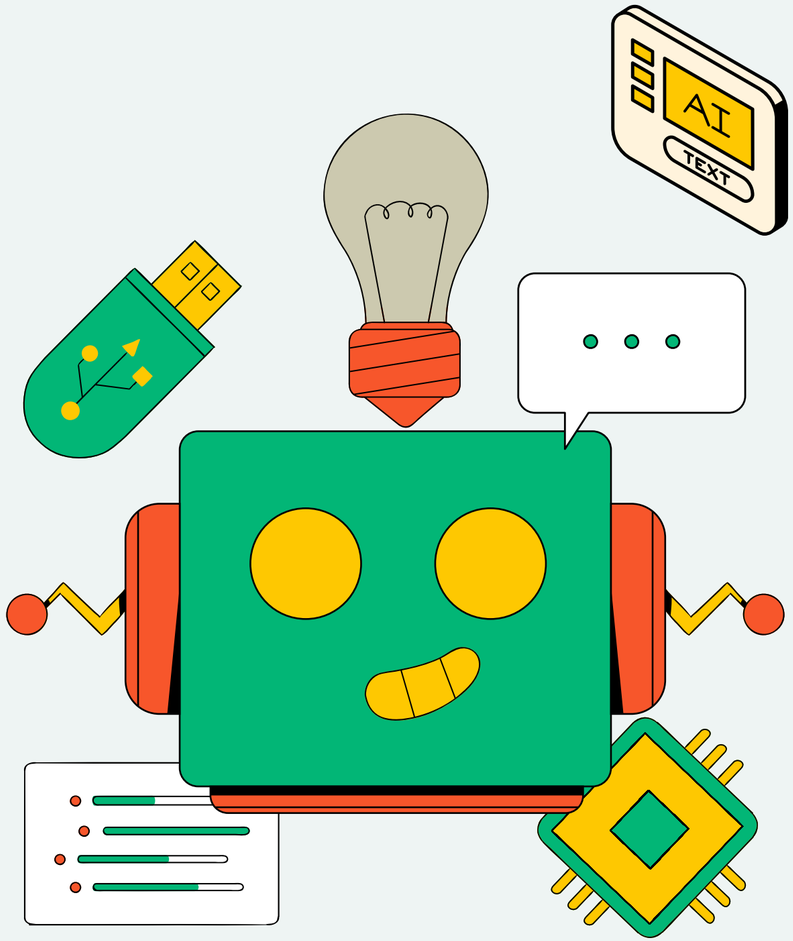
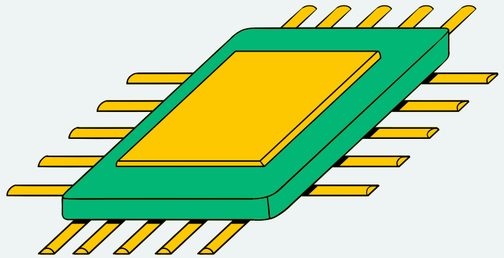


WHAT THE MACHINE LEARNING

CROSS VALIDATION

DONNÉES ENTRAÎNEMENT/TEST /VALIDATION

PRESENTED BY:



POURQUOI ÉVALUER CORRECTEMENT UN MODÈLE ?

Éviter les résultats trompeurs

Assurer une bonne généralisation sur de nouvelles données

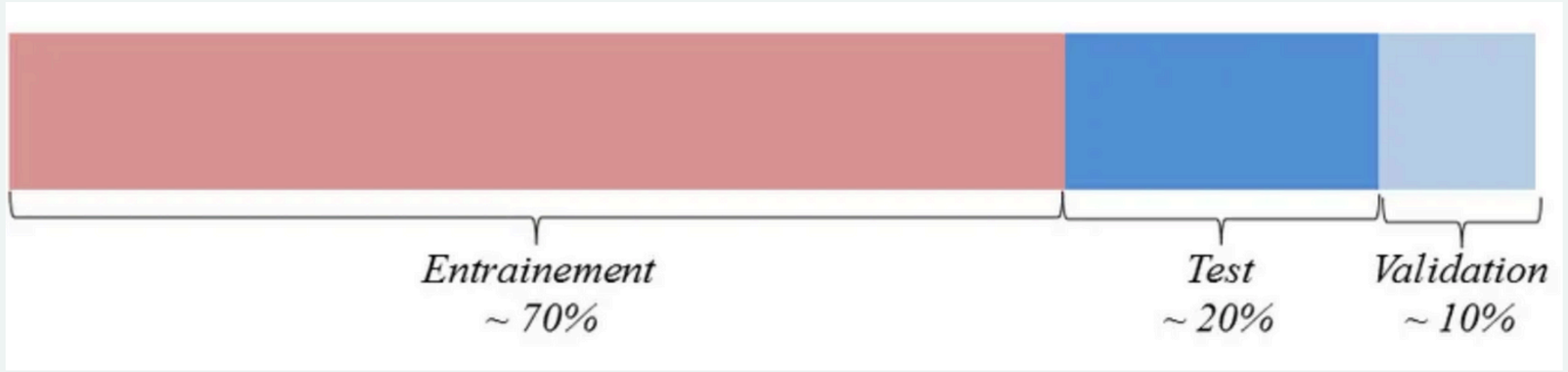
Choisir les bons paramètres et le bon modèle

SÉPARATION CLASSIQUE DES DONNÉES

Entraînement

Validation

Test



HYPER PARAMETRE

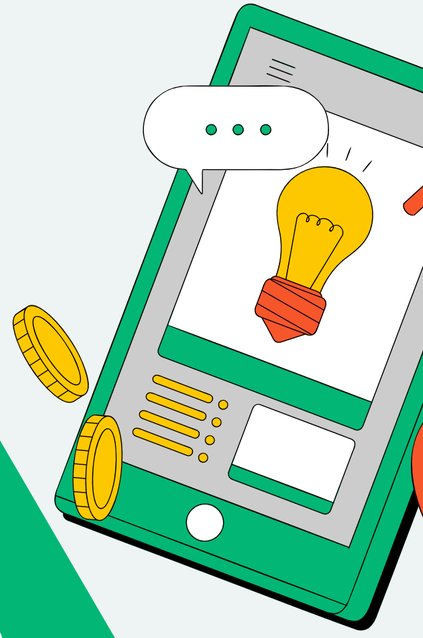
Taux d’apprentissage (learning rate)

Nombre d’arbres (Random Forest)

Nombre d’époques (epochs)

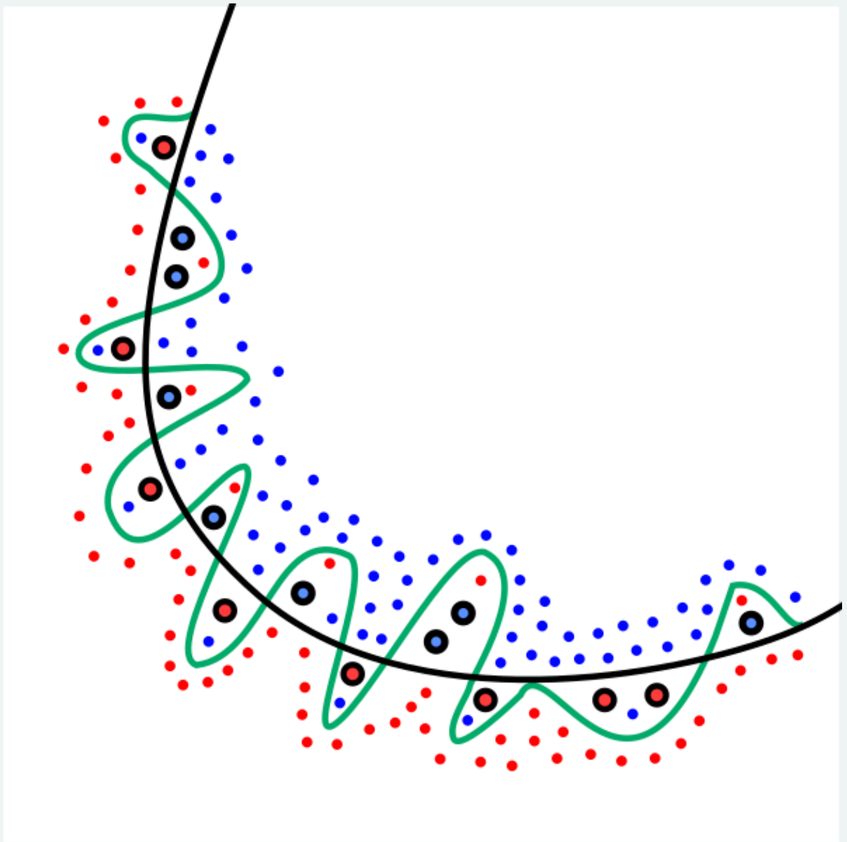
Profondeur maximale d’un arbre (Decision Tree)

Taille du batch (batch size)



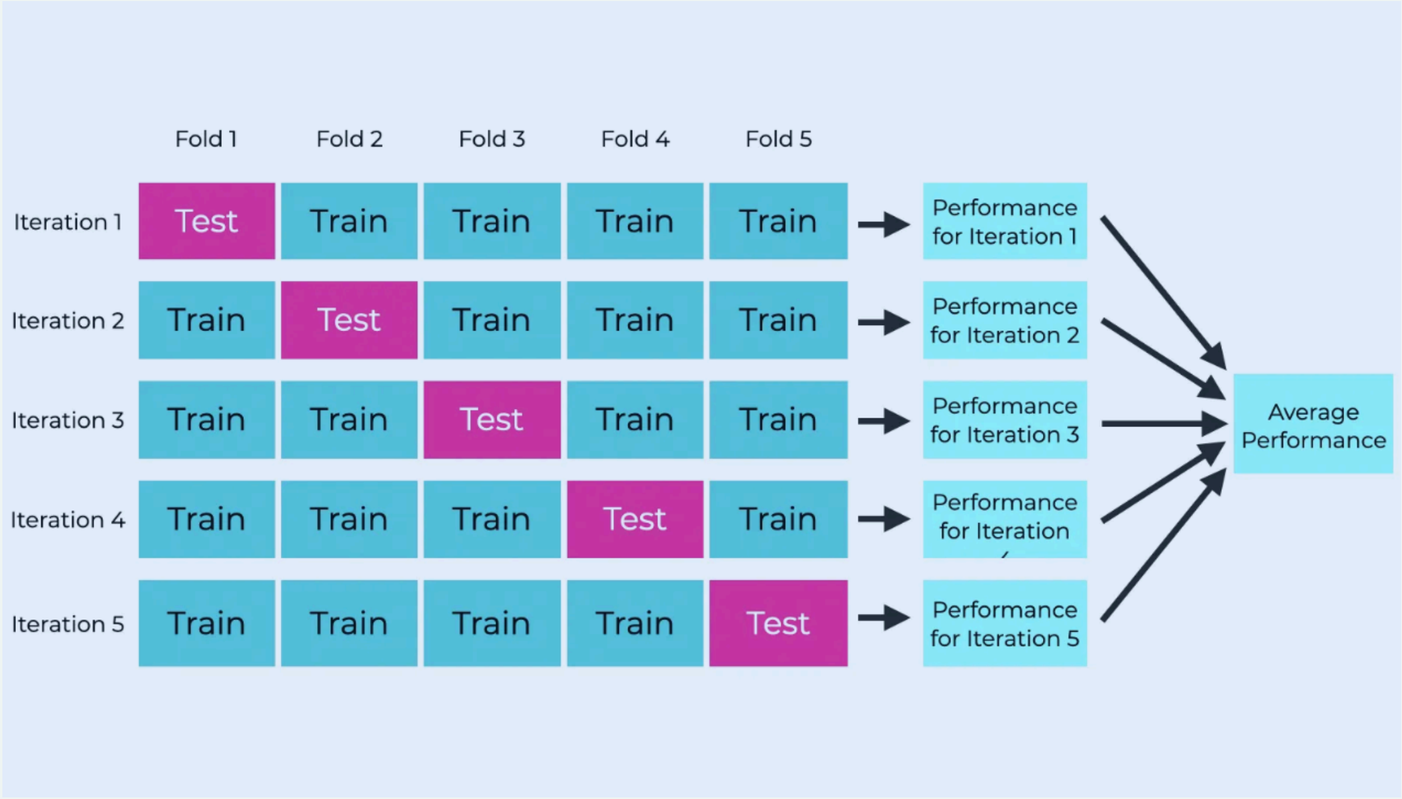
SURAPPRENTISSAGE

Le surapprentissage, est une analyse statistique qui correspond trop précisément à une collection particulière d'un ensemble de données. Ainsi, cette analyse peut ne pas correspondre à des données supplémentaires ou ne pas prévoir de manière fiable les observations futures.



LA VALIDATION CROISÉE

Pour détecter un surapprentissage, on utilise la validation croisée. On divise les données en k sous-ensembles : à chaque itération, on entraîne le modèle sur k-1 parts, et on le teste sur la dernière. Ce processus est répété k fois avec un fold de validation différent à chaque fois.



k = 5 ou 10 offre un bon compromis entre précision de l'évaluation et coût computationnel.

AVANTAGES ET LIMITES

DONNEES

pour de l'entraînement et pour le test

FIABILITE

Moyenne des scores

REDUCTION DES BIAIS

Moins de risque de sur- ou sous-estimer les performances

COÛT

Coût élevé du à l'entrainement k fois