



Machine learning

Stanislas FRANCFORT

Panorama, données, première évaluation, validation croisée et régularisation.

3 séquences pour choisir le Machine Learning

- Apprentissage et data, modèles et algorithmes
- Protocoles indispensables pour évaluer des modèles
- complexité et régularisation.

Apprentissage

- **Supervisé** : prédire un prix ou reconnaître un chiffre avec exemples annotés.
- **Non supervisé** : regrouper des musiques sans étiquette fournie.
- **Auto-supervisé** : construire un signal d'apprentissage à partir des données.
- **Renforcement** : améliorer des décisions par interaction et récompenses.

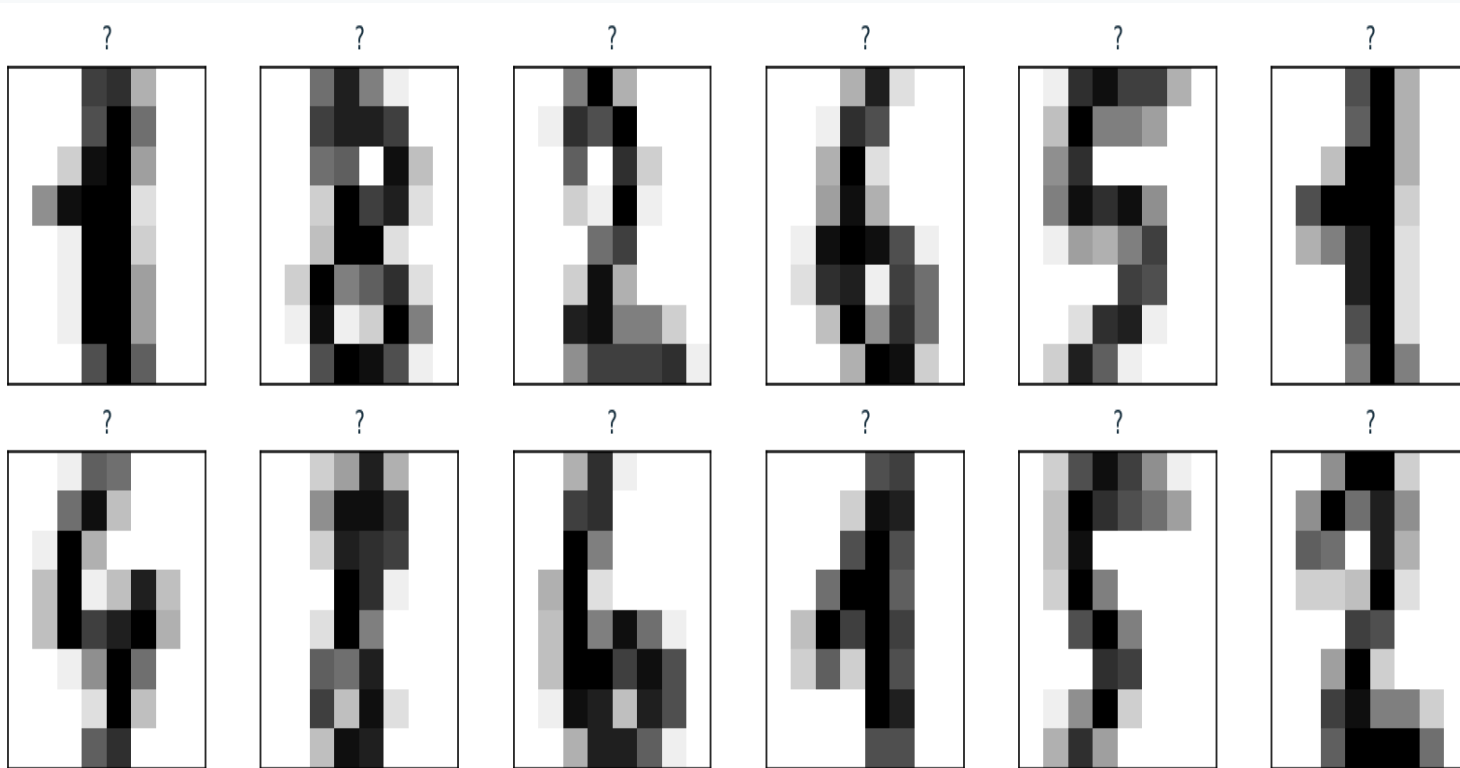
Tâche, modèle, apprentissage

- Tâche : reconnaître le chiffre représenté par une image.
- Modèle : associer 64 intensités de pixels à une des dix classes.
- Apprentissage : ajuster ses paramètres ou construire des règles sur des exemples.
- Évaluation : comparer aux étiquettes d'images non vues.

Familles de modèles : les idées clés

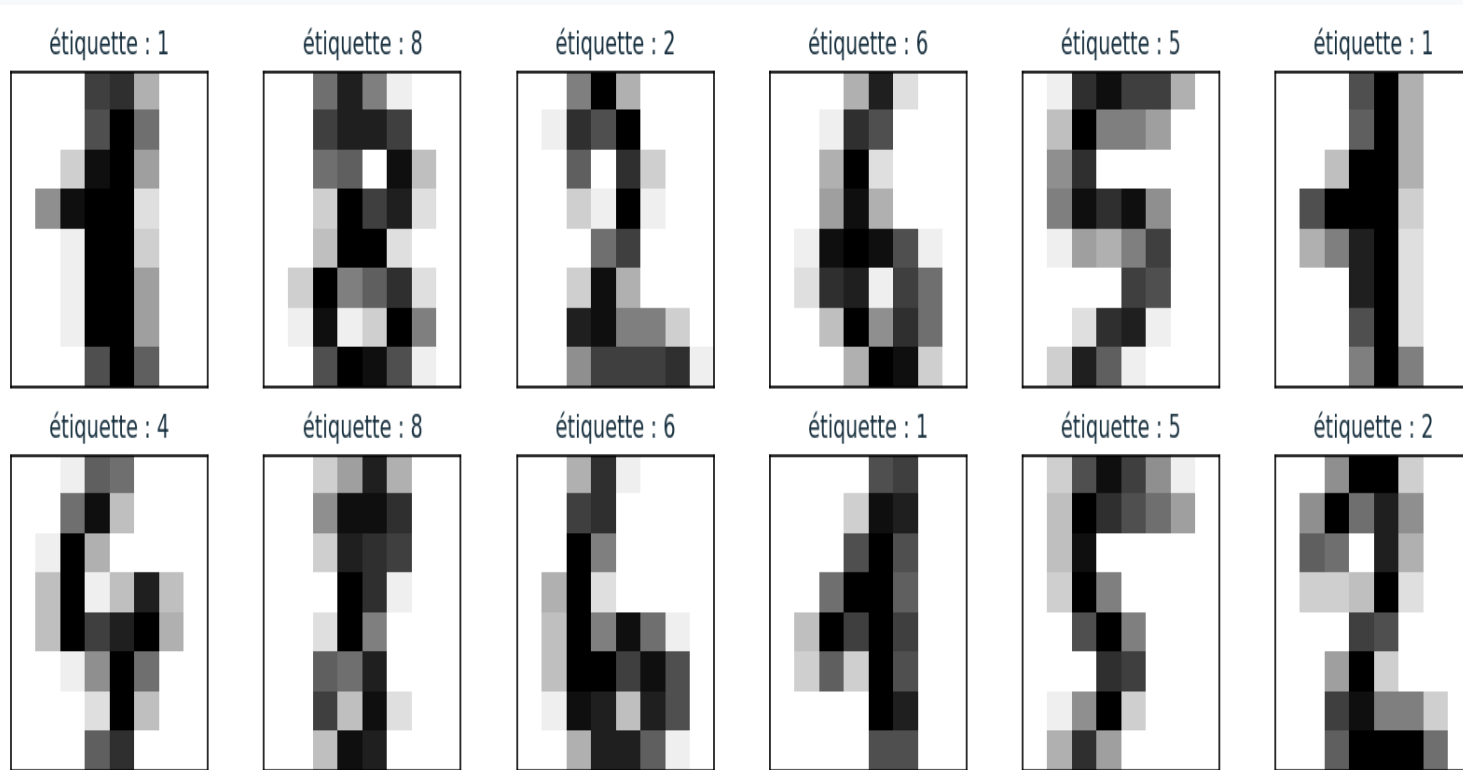
- Modèle linéaire : score pondéré ; k-NN : vote des voisins ; arbre : découpages.
- Forêt et boosting : ensembles d'arbres ; SVM : marge de séparation.
- Réseaux de neurones : couches de représentations apprises.

Dataset digits : quel chiffre voyez-vous ?



- Les données et les étiquettes ne sont pas toujours évidentes.

Révéler les étiquettes



Les étiquettes du dataset sont affichées.

Le dataset “Digits”, les chiffres manuscrits

- 1 797 images manuscrites et 10 classes.
- Chaque image contient 8×8 pixels, soit 64 variables de 0 à 16.
- Tâche : classification multiclasse
- La représentation 8×8 est très réduite par rapport à une photo ordinaire.

Prépare les données de test avant de commencer

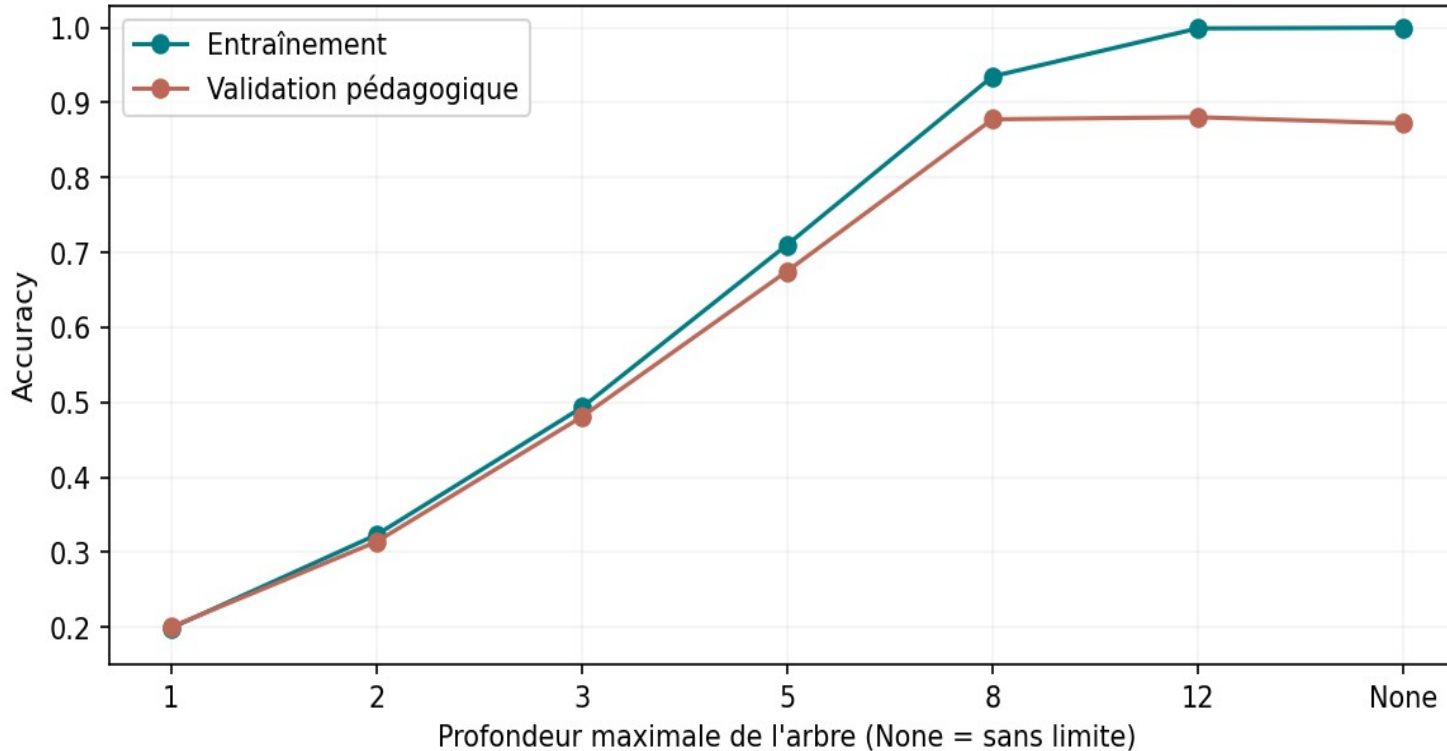
- Créer un ensemble de test final avant tout prétraitement apprenant.
- Utiliser uniquement les données prévues pour développer, comparer et régler.
- Vérifier si plusieurs lignes viennent d'un même objet ou d'une même personne.
- Ne pas utiliser le test comme une validation.

Le test ne sert pas à choisir ou calibrer le modèle !

Le modèle « parfait » sur ses données, mais qui ne prédit rien

- Une “baseline” fixe un niveau de référence (ex : majoritaire).
- Un arbre non contraint peut mémoriser les exemples d'entraînement.
- Comparer la même métrique sur les données d'entraînement et de validation.
- Un excellent score d'entraînement ne certifie pas la généralisation.

La profondeur n'est pas toujours un progrès



■ Courbes calculées sur Digits avec un découpage interne fixe.

■ L'entraînement progresse ; la validation peut plafonner ou reculer.

Aucun résultat ici n'utilise le test final réservé.

Piège : confondre apprentissage et généralisation.

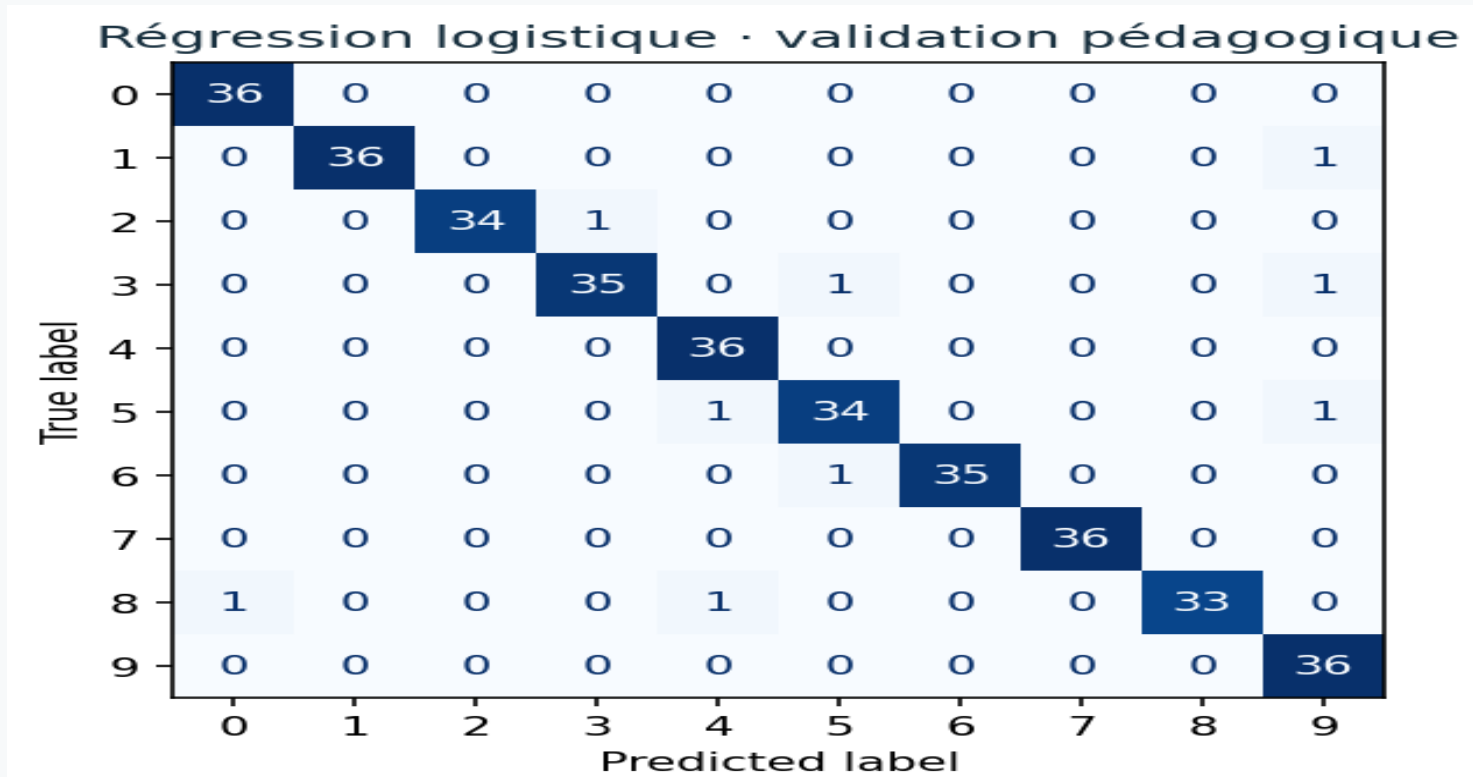
Trois façons de classer un chiffre

- **Régression logistique** : calculer un score pour chaque classe à partir des pixels.
- **k-NN** : regarder les images les plus proches selon une distance.
- **Arbre** : enchaîner des questions sur les intensités des pixels.
- Comparer aussi leurs erreurs et leurs besoins de prétraitement.

Représenter et prétraiter

- L'image 8×8 est aplatie en un vecteur de 64 nombres.
- **StandardScaler** apprend *moyenne* et *dispersion* sur le sous-ensemble d'entraînement.
- **Pipeline** refait cet apprentissage à l'intérieur de chaque pli.
- k-NN et modèles linéaires sont plus sensibles à l'échelle que les arbres.

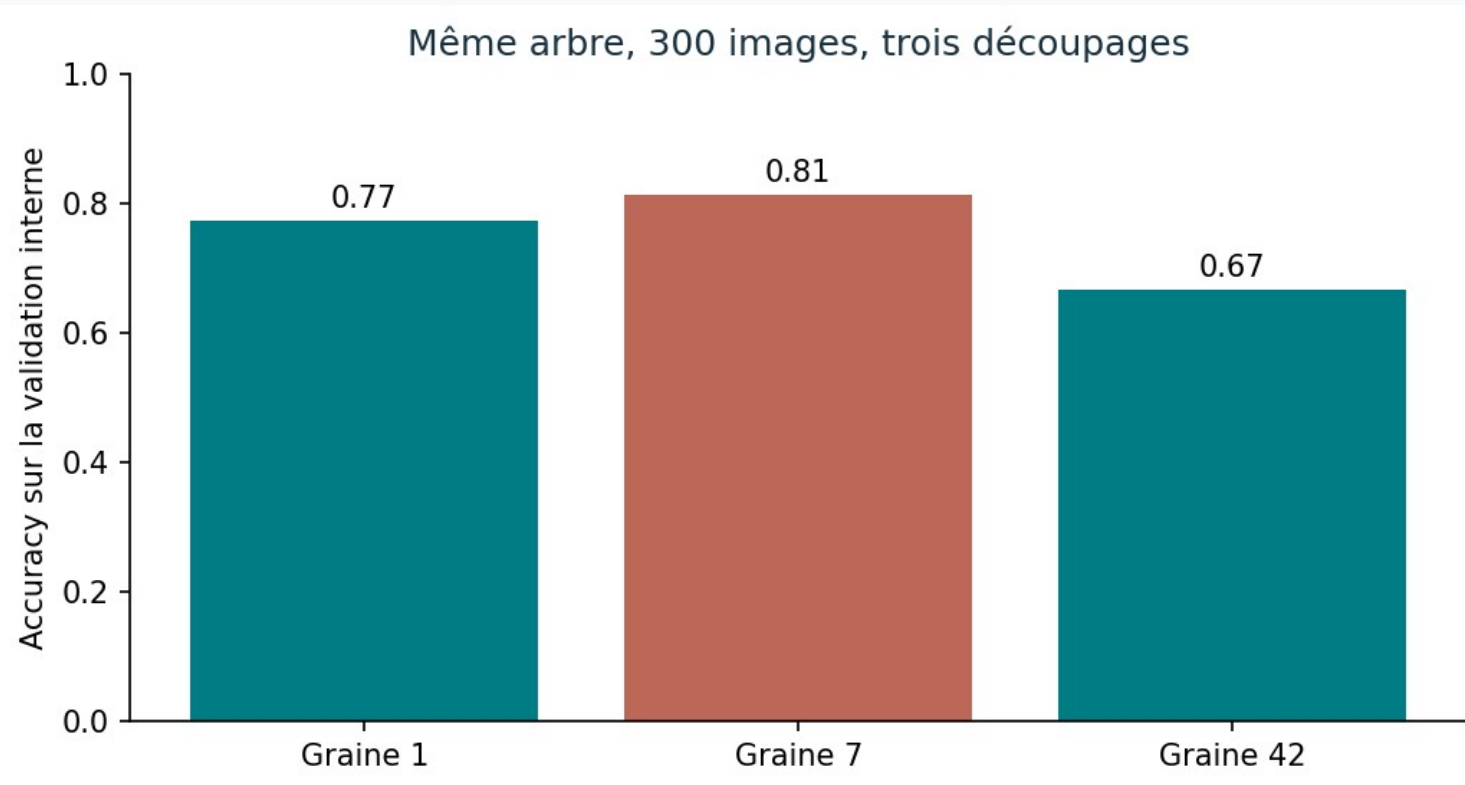
L'accuracy ne montre pas tout



- La diagonale représente les bonnes prédictions.
 - Les autres cases montrent quelles classes sont confondues.
- Revoir quelques images avant d'expliquer l'erreur.

Piège : un score global ne décrit pas toutes les erreurs.

Même modèle, trois découpages



- Arbre de profondeur 5 sur 300 images Digits.
- Seule la graine du découpage change.
- Un résultat isolé peut avantager ou pénaliser le modèle.

Comparer plusieurs plis avant d'annoncer un gagnant.

Un seul découpage peut tromper

- Le score dépend des observations tombées dans la validation.
- Une validation croisée à cinq plis répète la comparaison cinq fois.
- Utiliser les mêmes plis pour tous les modèles.
- Rapporter moyenne et dispersion plutôt qu'un seul nombre.

Choisir le bon type de plis

- **StratifiedKFold** conserve approximativement la proportion des classes.
- Des copies d'une data ne doivent pas apparaître dans plusieurs plis.
- Données temporelles ou groupes exigent parfois d'autres séparations.
- La méthode de validation doit simuler l'usage futur.

Comparer avec les mêmes plis

```
cv = StratifiedKFold(5, shuffle=True, random_state=42)
modele = make_pipeline(StandardScaler(),
                       LogisticRegression(max_iter=1000))
scores = cross_validate(modele, X_dev, y_dev,
                        cv=cv, scoring='accuracy')
```

- La normalisation reste dans le Pipeline.
- Comparer moyenne et écart-type de scores['test_score'].

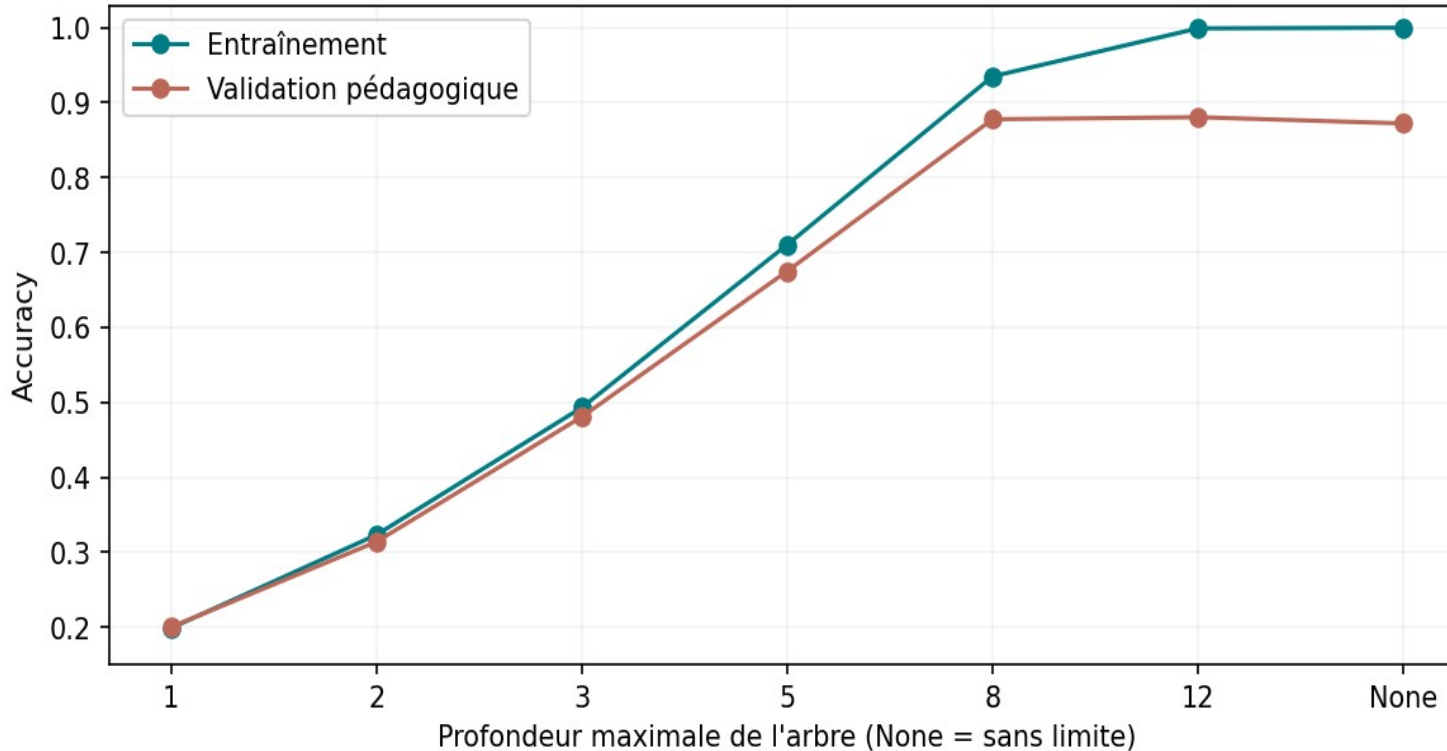
Sous-apprentissage ou surapprentissage ?

- Trop simple : le modèle manque des régularités utiles.
- Trop flexible : il apprend aussi le bruit des exemples.
- L'écart entraînement/validation aide au diagnostic.
- Le test final ne sert jamais à régler cette complexité.

Trois réglages de complexité

- Arbre : limiter `max_depth` réduit les découpages possibles.
- k-NN : augmenter `k` lisse les décisions locales.
- Logistique : diminuer `C` renforce la pénalisation des coefficients.
- Pénalisation explicite et limite de complexité sont deux stratégies différentes.

Relire les courbes avant de conclure



- La profondeur maximale n'est pas forcément la meilleure.
- Trop de contrainte peut aussi diminuer le score.
- Tester quelques valeurs sur les mêmes plis.

Quelques take away

- Baseline et protocole sont plus importants que la puissance des modèles.
- Le Pipeline protège les transformations apprenantes.
- On commence par la validation croisée ; le test vient ensuite.

Références

[S1] scikit-learn · Digits

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html

[S2] scikit-learn · Pièges

https://scikit-learn.org/stable/common_pitfalls.html

[S3] scikit-learn · Validation croisée

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