acta
LES INSTITUTS TECHNIQUES AGRICOLES#

machine learning : principe et intérêt pour les méthodes d'appariement

Atelier Estimer l'effet d'une pratique agricole à partir d'un réseau d'observation : intérêt des scores de propension

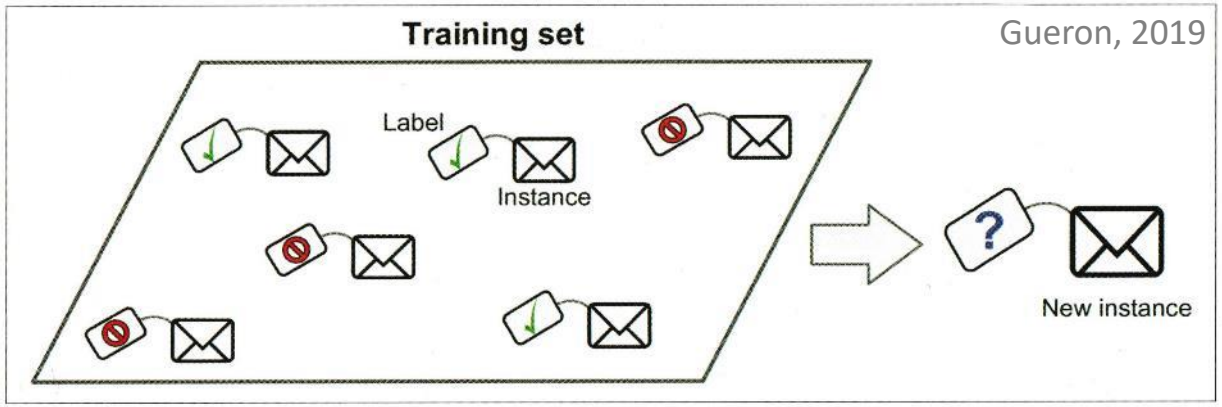
François Brun

Paris, 30 novembre 2023

Le principe du machine learning

Objectif du **machine learning :** « Construire de manière **automatique** une **fonction** qui fait correspondre une entrée à une sortie **de manière performante** en se **basant sur des exemples de paires** {entrée, sortie} »

- ***sortie*** (output) : une variable bien définie que l'on cherche à prédire => Y
- ***entrée*** (input) : un ensemble des variables explicatives (nombre de colonne) => X
- ***exemples de paires*** : le jeu de données pour apprendre. Le nombre d'exemple = le nombre d'individus = le nombre de ligne
- ***performante*** => évaluation



Gueron, 2019

Figure 1-5. A labeled training set for spam classification (an example of supervised learning)

Phase d’apprentissage des modèles

- **Diversité de modèles existants**

Une grande diversité d’algorithmes, depuis le modèle linaire au réseau de neurone

- **Notion d’ajustement**

les méthodes de machine learning consiste à ajuster les paramètres, la forme, d’un modèle à un jeu de données d’observations (aussi appelé jeu de données d’apprentissage) pour prédire de nouveaux individus.

Régression classique, limite et extension

- **Rappel : modèle de régression classique**

- Une régression linéaire classique
  - Y= a * X + b
  - a : la pente et b : la valeur à l'intercept

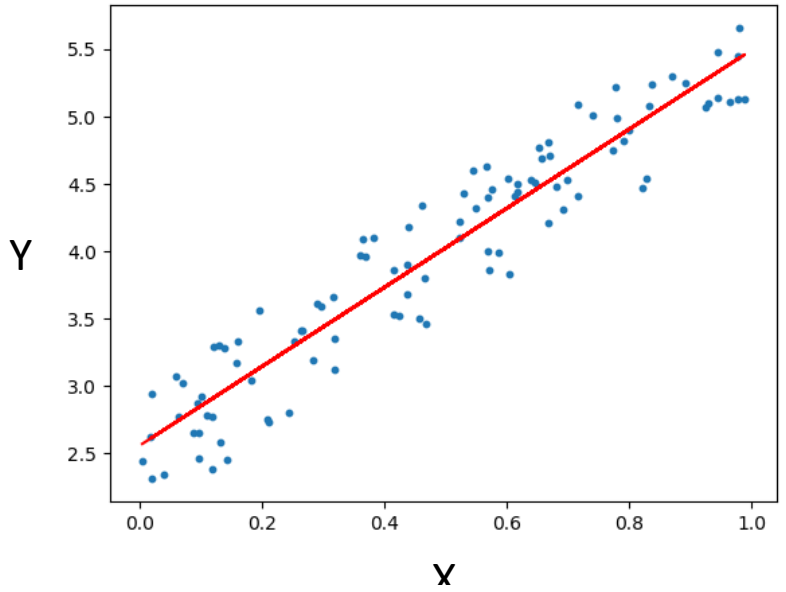
- Mais d'autres régressions :
  - Y= a1*X1 + a2*X2 + a12*X1*X2 + b
  - Etc…

- Généralisation : régression linéaire multiple

$$Y = X\theta + \varepsilon$$

$$\begin{matrix} & X_1 & X_2 & \dots & X_P \end{matrix}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ x_{21} & x_{22} & \dots & x_{2P} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_P \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}$$



P : nombre de variables explicatives

N : nombre d'individus / lignes

Mais

- Problème de la dimension, si trop de variables explicatives, notamment avec des corrélations
- pas évident de traiter conjointement variables continues et qualitatives
- Hypothèse forte sur la linéarité

la régression pénalisée, efficace pour sélectionner les variables

- Cas : n < p ou n ≈ p et certaines variables très corrélées
- **méthode à rétrécisseur** ou **de pénalisation** ou **de régularisation**

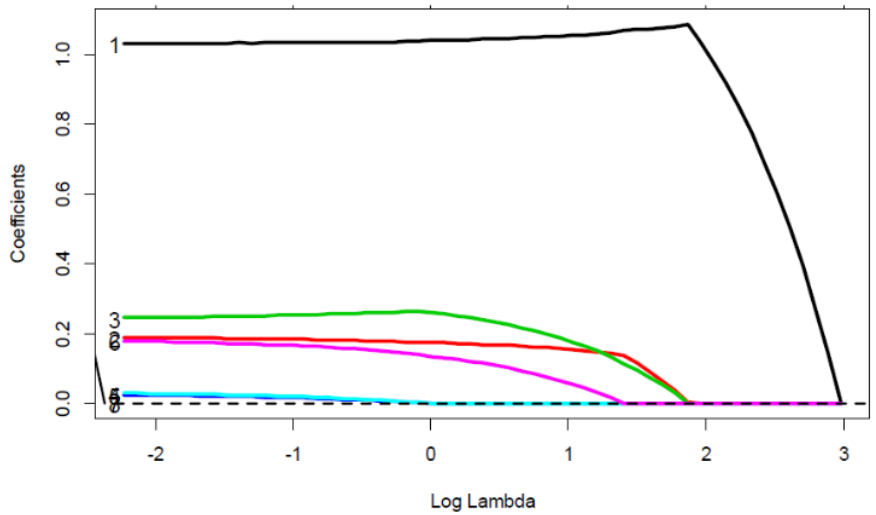
- Cas classique = estimation des paramètres sans contrainte en minimisant $(Yobs-Ypred)^2$

→

- Cas pénalisé
on va minimiser : $\boldsymbol{(Yobs-Ypred)^2 + \lambda G}$

⟹Conséquence : des valeurs de paramètres deviennent nulles…

- λ: mesure de la complexité du modèle
- *G*: dépend des paramètres β.



(exemple de LASSO)

Différentes formes de pénalisation

$$G = \sum_j \beta_j{}^2 \quad \Rightarrow \quad \text{Régression } \textbf{ridge}$$

$$G = \sum_j |\beta_j| \quad \Rightarrow \quad \text{Régression } \textbf{lasso}$$

$$G = \sum_j \alpha|\beta_j| + \sum_j (1-\alpha)\beta_j{}^2 \quad \Rightarrow \quad \text{Régression } \textbf{elastic net}$$

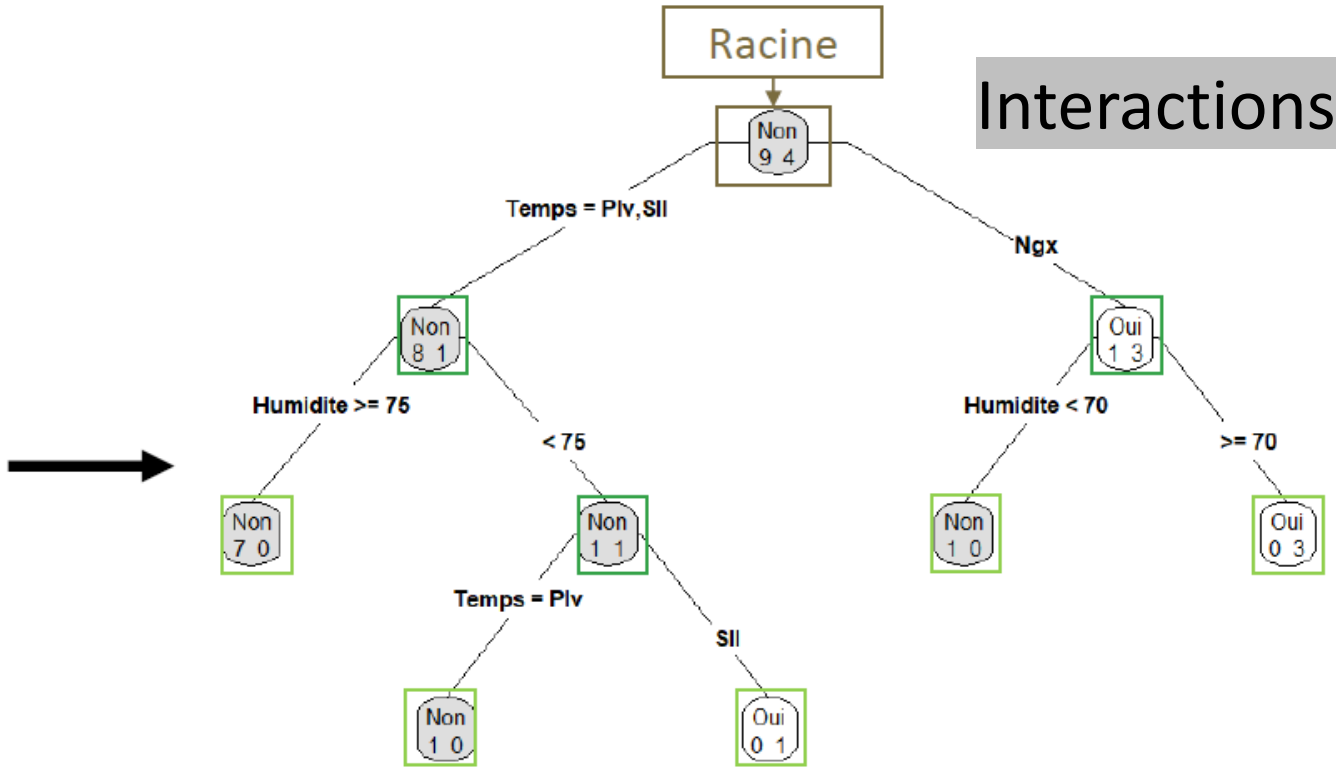
$$0 \le \alpha \le 1$$

Des méthodes d'ensemble.
Exemple des forêts aléatoires (random forest)

- **Le principe d'un arbre de décision**

- Facile à interpréter, mais qualité de prédiction…

Temps	Humidite	Joue
Nuageux	78	Oui
Nuageux	65	Non
Nuageux	90	Oui
Nuageux	75	Oui
Pluvieux	96	Non
Pluvieux	80	Non
Pluvieux	70	Non
Pluvieux	80	Non
Pluvieux	80	Non
Soleil	85	Non
Soleil	90	Non
Soleil	95	Non
Soleil	70	Oui



Interactions

Nœuds

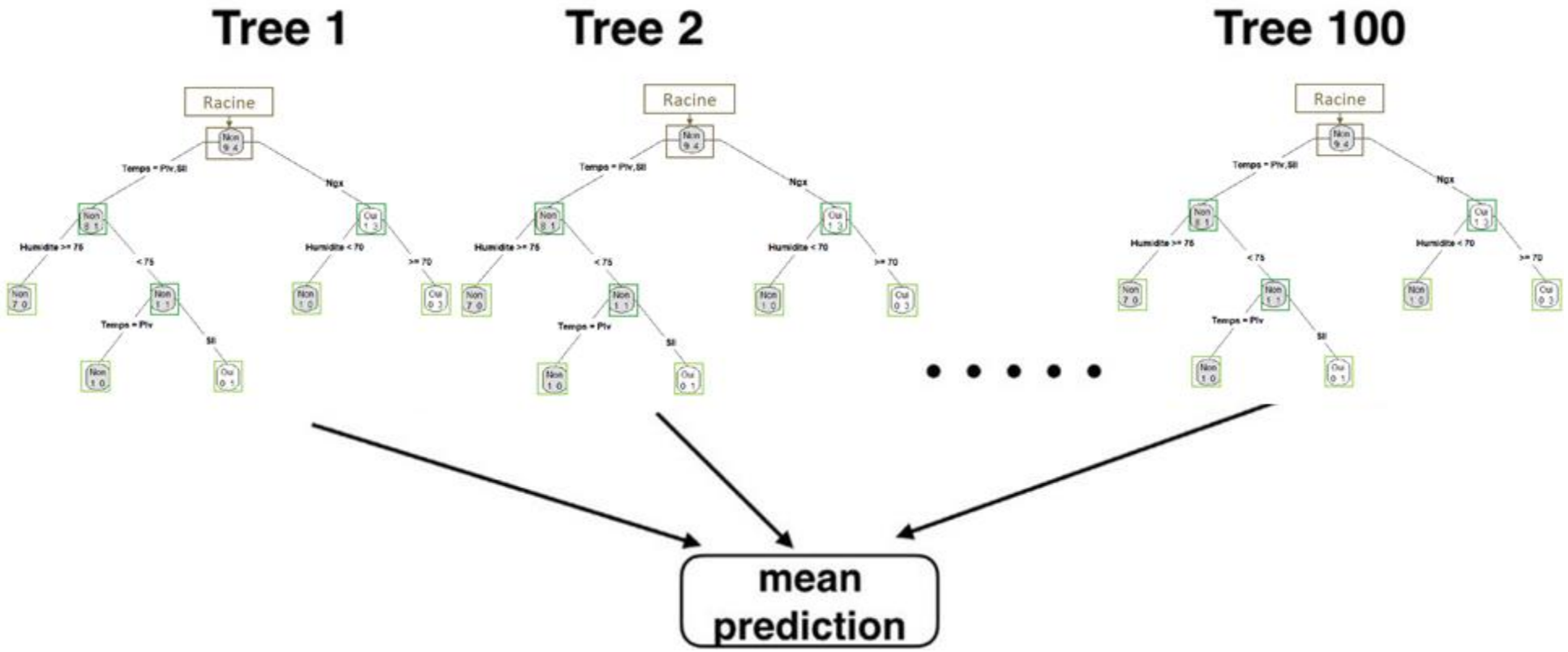
Nœuds terminaux / Feuilles

Méthodes d'ensemble (forêts aléatoires)

- **Le principe de forêt aléatoire (random forest)**

Idée : Un grand nombre d'arbres simples est plus performants qu'un arbre complexe.

➡ Principe : construction d'une multitude d'arbres indépendants (chacun ayant une vision partielle des données)



Méthodes d'ensemble (forêts aléatoires)

- **Le principe de forêt aléatoire (random forest)**

Principe : construction d'une multitude d'arbres indépendants (chacun ayant une vision partielle des données)

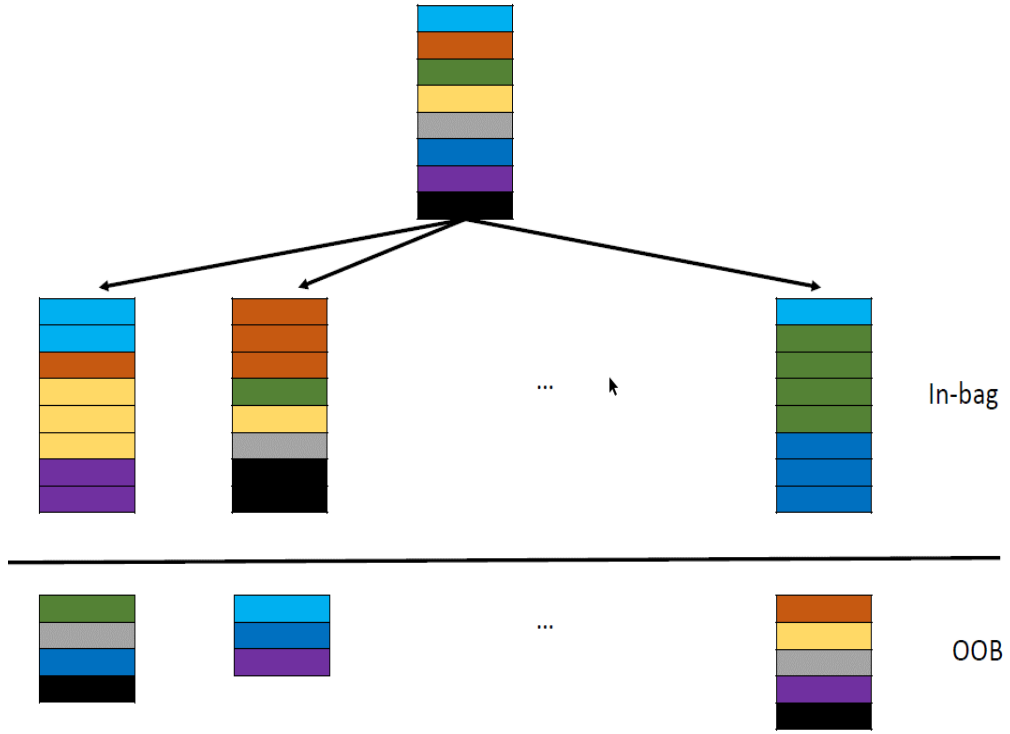
Random forest = bagging + feature sampling.

bagging = boostrap + agregation.

- **Bootstrap** :
  - On tire avec remise

Individus «In-Bag»: individus présents dans l'échantillon bootstrapé(~63%) = échantillon d'apprentissage

Individus «Out-Of-Bag» (OOB): individus non présents dans l'échantillon bootstrapé(~37%) = échantillon test



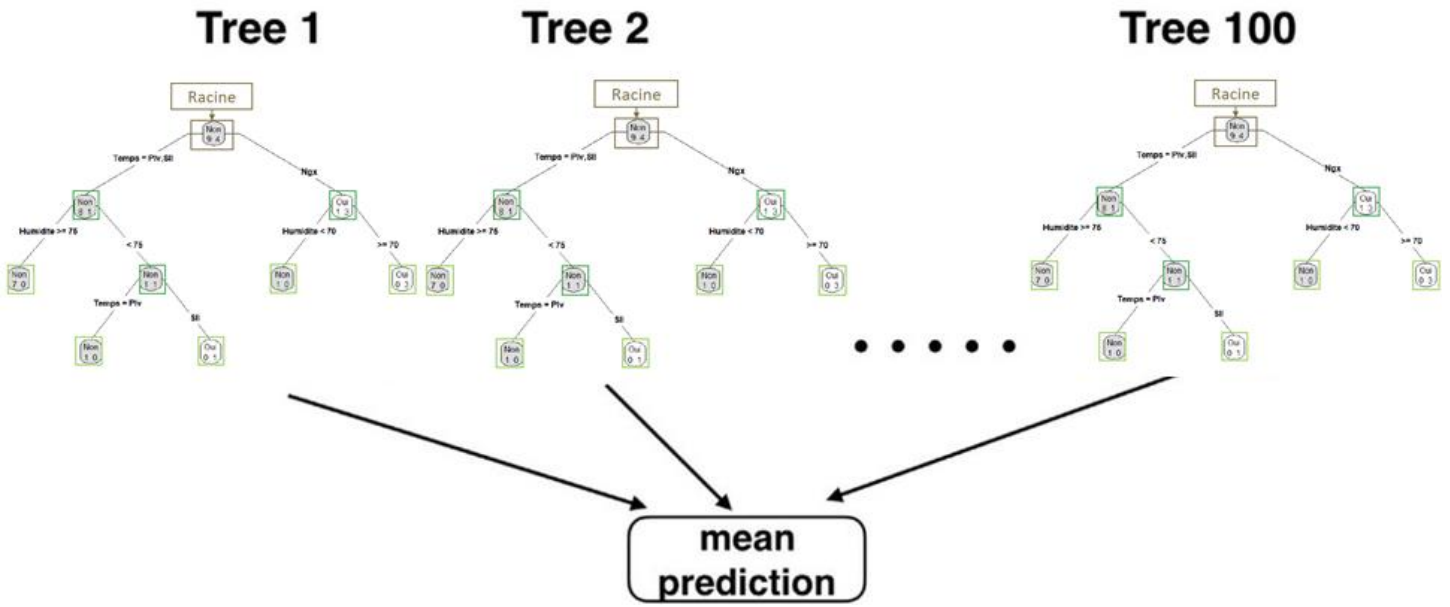
Méthodes d'ensemble (forêts aléatoires)

- **Le principe de forêt aléatoire (random forest)**

Principe : construction d'une multitude d'arbres indépendants (chacun ayant une vision partielle des données)

Random forest = bagging + feature sampling.

- **bagging :**
  - on tire n individus avec remise parmi les N du jeu de données (boostrap).
  - On entraîne un arbre à partir de cet échantillon.
  - Pour faire une prévision sur de nouvelles données, agréger les prédictions obtenues avec chacun des arbres.



Méthodes d'ensemble (forêts aléatoires)

- **Le principe de forêt aléatoire (random forest)**

Principe : construction d'une multitude d'arbres indépendants (chacun ayant une vision partielle des données)

Random forest = bagging + feature sampling.

- **feature sampling :** on choisit au hasard m variables parmi les p variables explicatives du jeu de données pour chaque noeud.

Nombre m fixé pendant toute la construction de la forêt

Classification : $m = \sqrt{p}$

Régression : $m = p/3$

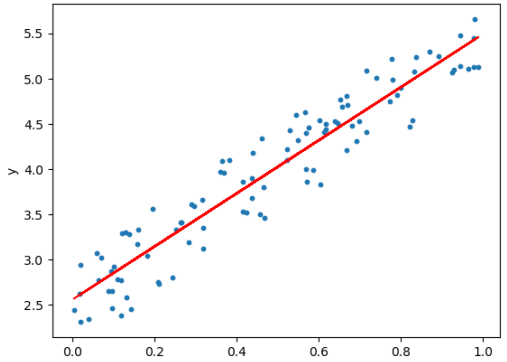
Phase d'apprentissage des modèles

- **Notion d'ajustement**

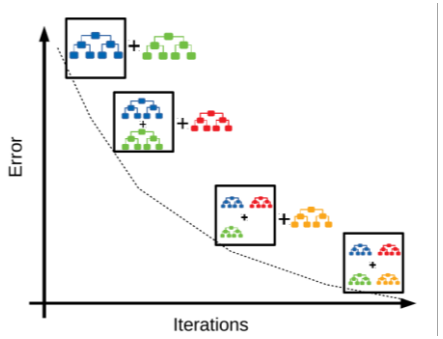
les méthodes de machine learning consiste à ajuster les paramètres, la forme, d'un modèle à un jeu de données d'observations (aussi appelé jeu de données d'apprentissage) pour prédire de nouveaux individus.

- **Diversité de modèles existants**

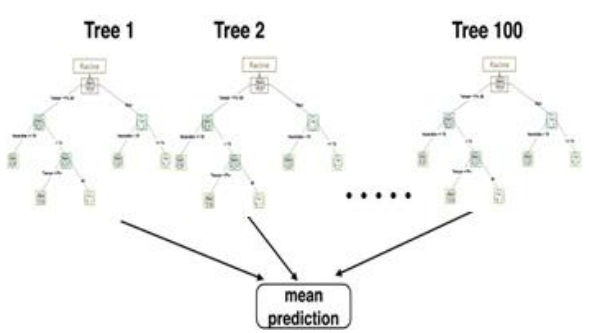
- **Méthodes de regression (avec ou sans pénalisation)**



- **Méthodes d'ensemble**

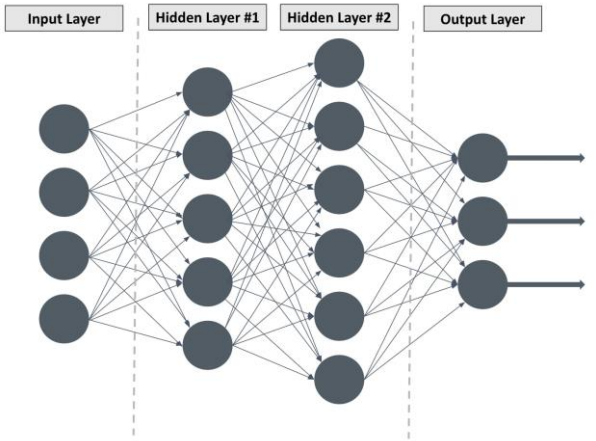


Gradient Boosting



Random Forest

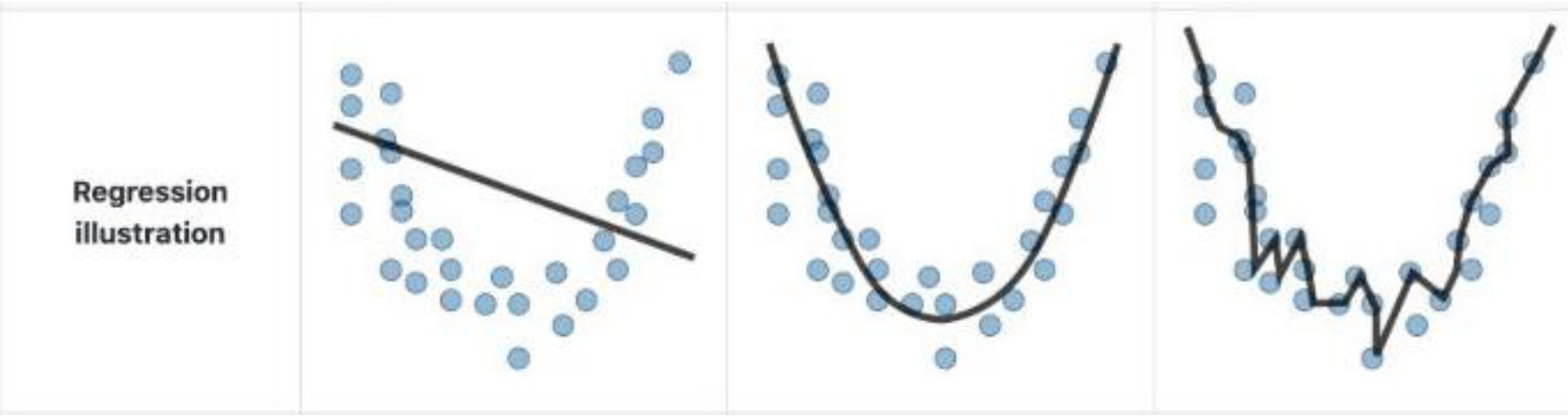
- **Deep learning**



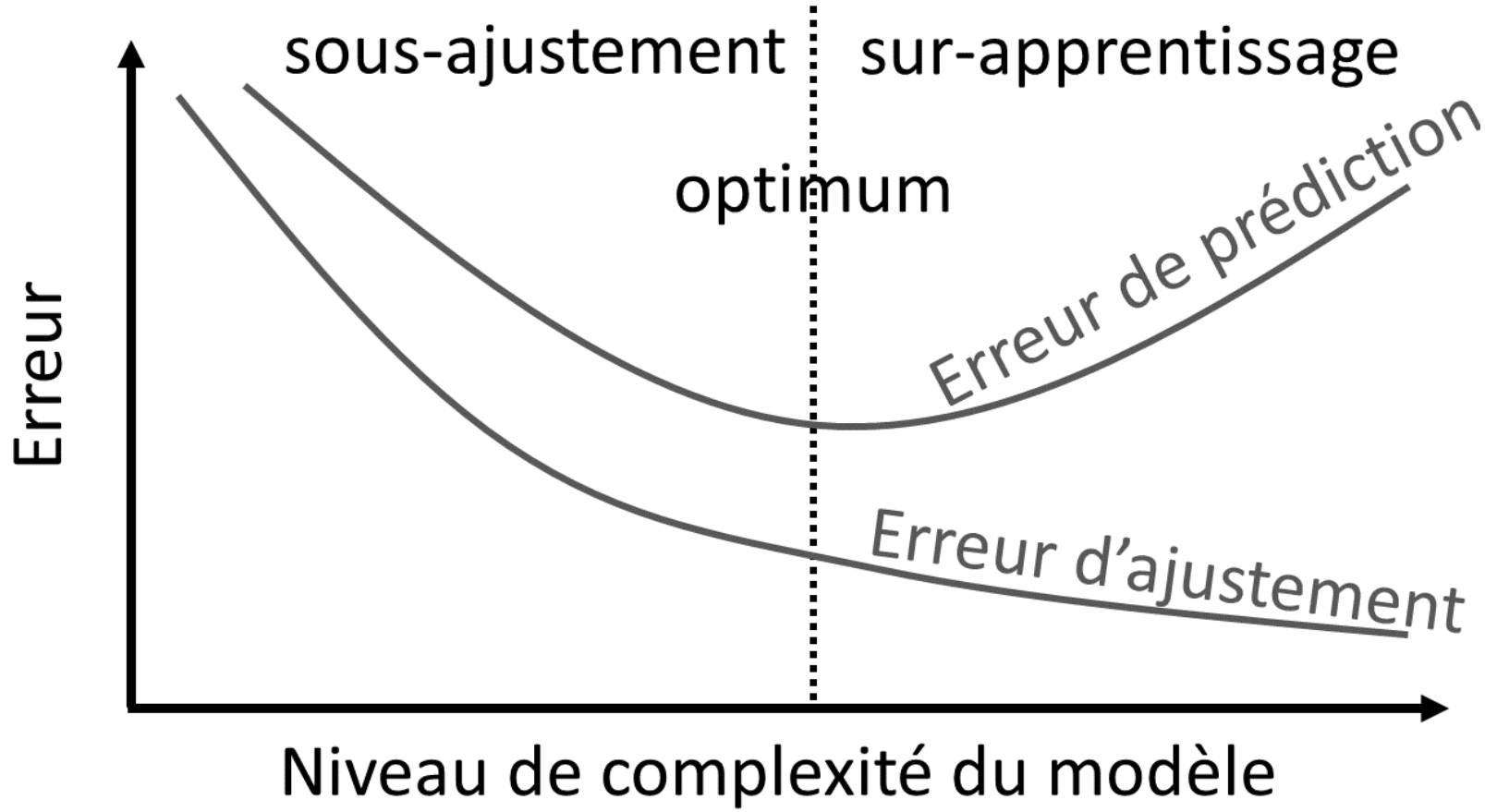
Ne pas se limiter à un modèle, tester plusieurs méthodes puis comparer pour sélectionner le/les meilleurs modèles

Un risque majeur: le sur-apprentissage ou sur-ajustement (overfitting)

Rappel : les méthodes d'apprentissage machine consiste à ajuster les paramètres, la forme, d'un modèle à un jeu de données d'observations pour prédire de nouveaux individus.



Un risque majeur: le sur-apprentissage ou sur-ajustement (overfitting)



Comparaison des modèles et évaluation des performances

- **Indicateurs de performance :**

REGRESSION

- MSE : moyenne des carrés des erreurs / RMSE : racine carrée du MSE

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Yobs_i - Ypred_i)^2$$

Comparaison des modèles et évaluation des performances

- **Procédure d'évaluation des performances**

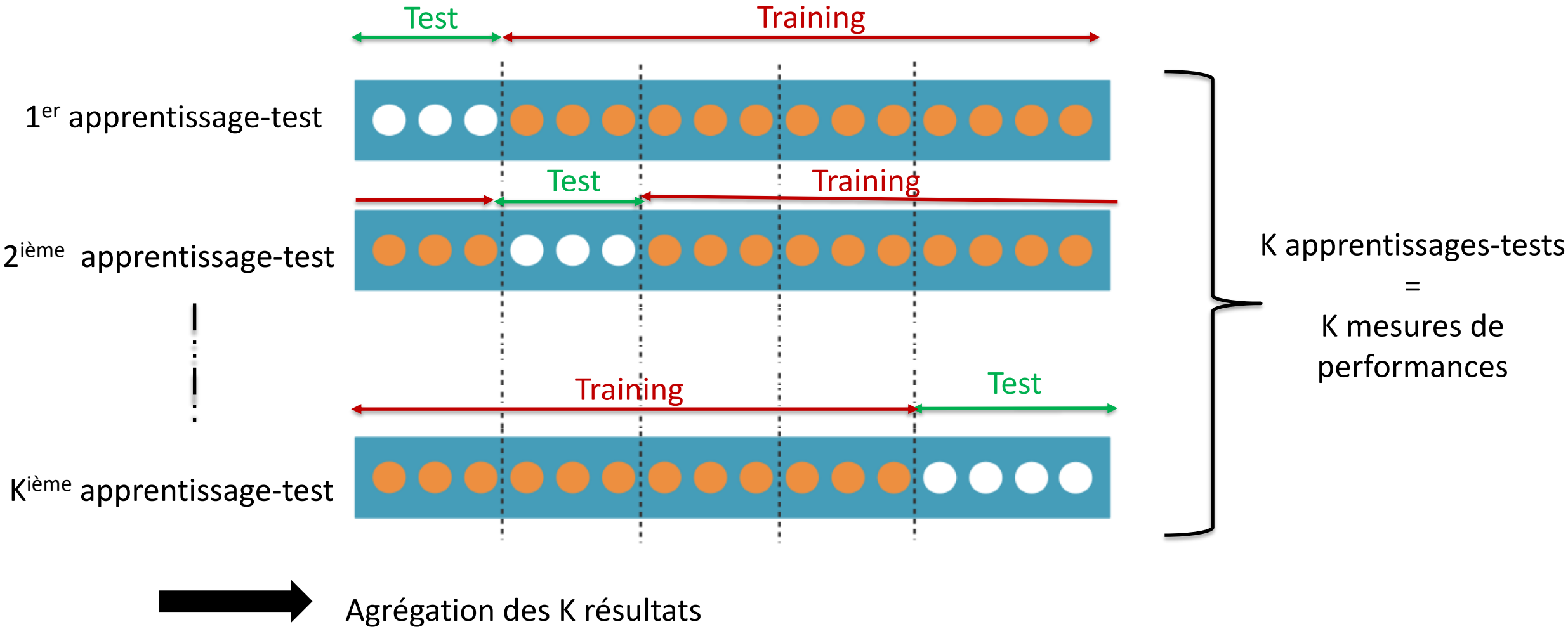
Pour **ne pas surestimer** les performances, il faut toujours évaluer les performances de prédiction sur un **jeu non utilisé pour l'apprentissage**

➡ Séparation aléatoire des données : proportion à choisir suivant la quantité de données (exemples : 70%-30%, 90%-10%)

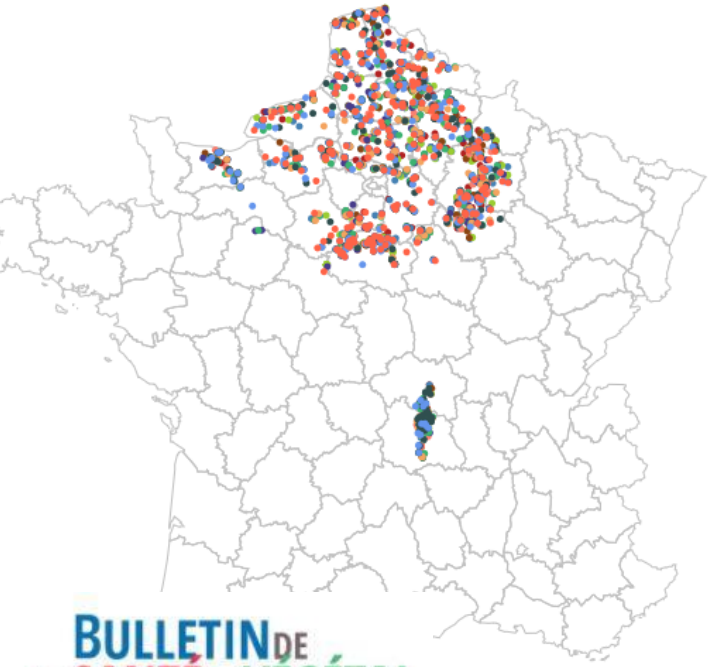
Jeux d'apprentissage (training set)	Jeux de test (test set)

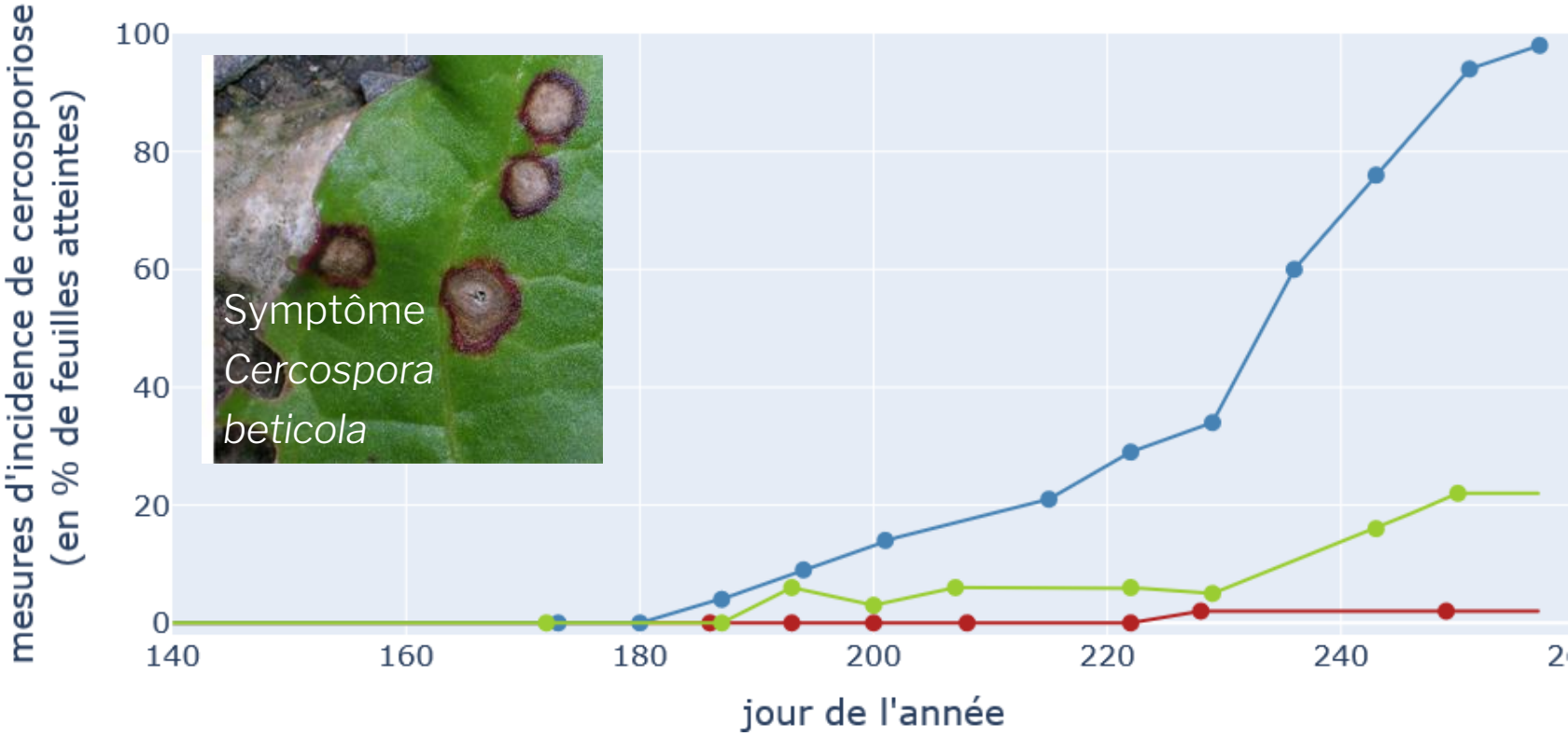
Comparaison des modèles et évaluation des performances

- **Principe de la validation croisée (cross-validation)**



Exemple d'application « classique » de prédiction avec le machine learning : cercosporiose de la betterave





Collectées sur 2009-2020 via Vigicultures pour le BSV

Exemples de 3 séries temporelles d'observation de l'incidence (% de feuilles atteintes)

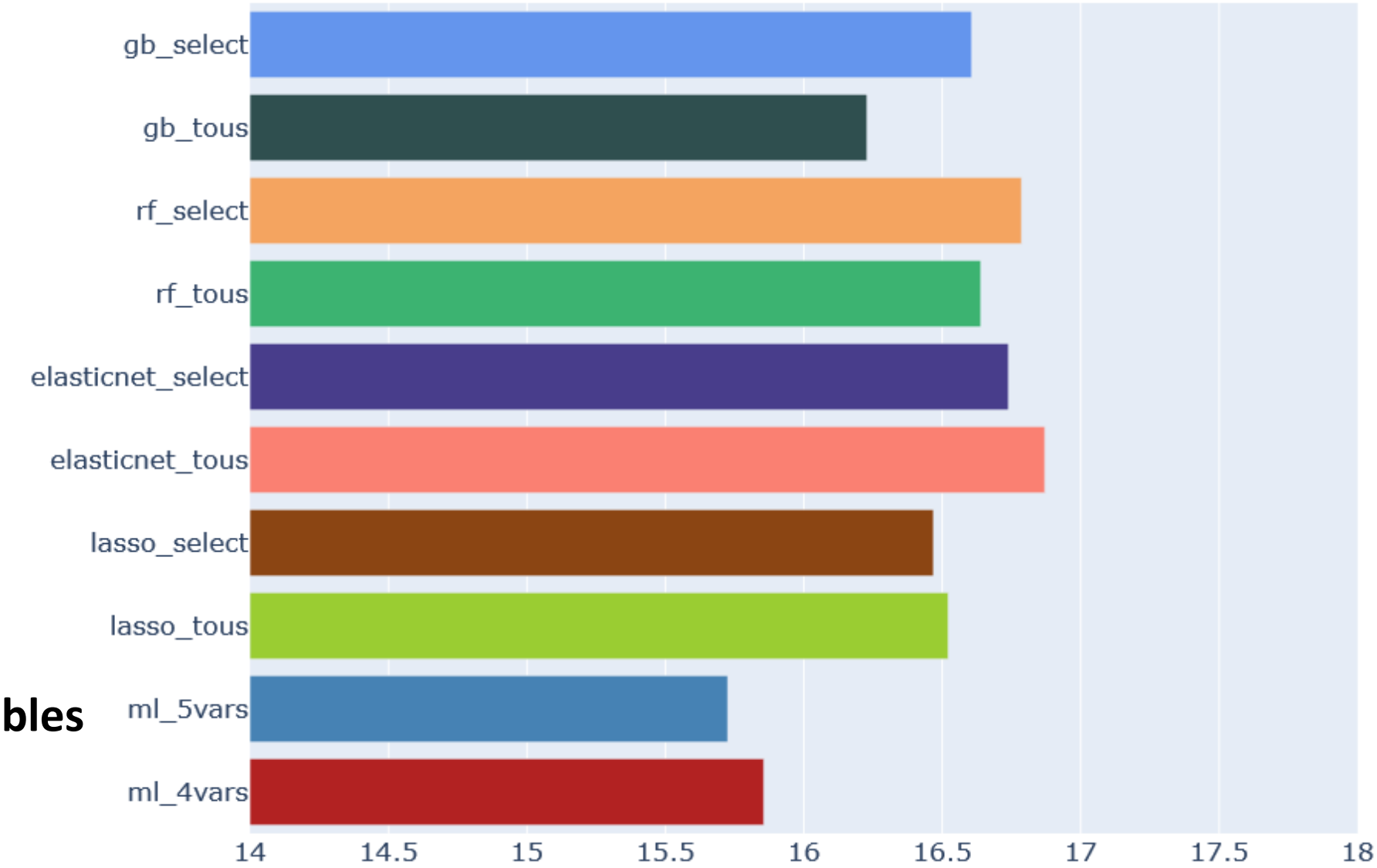
Comparaison des modèles et évaluation des performances

ETUDE DE CAS

Meilleur RMSE = 15,73 jours

Meilleur R2 = 0,33

Modèle informatif, mais il reste pas mal d'erreur



2 modèles retenus

→ **Modèle linéaire à 5 variables**

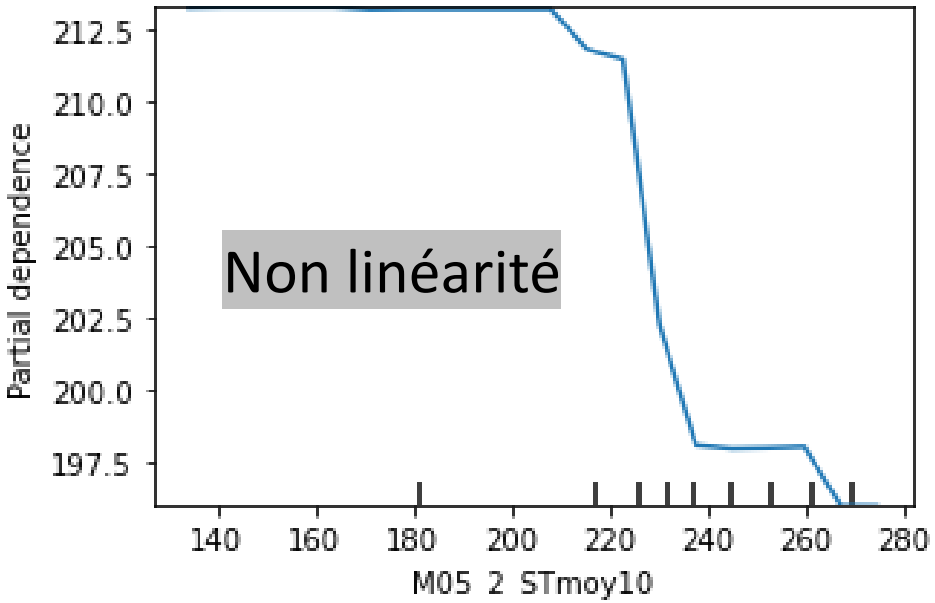
→ **Gradient boosting sans sélection de variables**

Interprétabilité des modèles

- **Le machine learning ne cherche pas à trouver les causes d'un phénomène mais à le prédire.**

- **l'interprétation n'est pas au cœur, mais c'est un aspect intéressant de l'interaction avec les experts**

- **Etude des effets des variables explicatives**
  - Régression classique: étude des signes des coefficients
  - Forêts aléatoires : graphe de dépendances partielles



- **Etude de l'importance des variables explicatives**
  - Forêts aléatoires : mesure basée sur le nombre de fois où la variable a été utilisée comme critère de séparation au niveau d'un nœud.

machine learning pour estimer les scores de propension

- Score de propension : prédire la probabilité d'appartenance à la classe traité / non traité en fonction des autres covariables
- Mais comment bien déterminer la forme de la relation entre les covariables et l'attribution du traitement ?

- Modéliser pour prédire pour des individus existants, pas de nouveaux individus
- Machine learning => différents algorithmes
  - Différentes formes, pour représenter les interactions, les non-linéarités. Pas mal de « souplesse » (bcp de paramètre)
  - mais si pas de garde-fou, risque de surajustement. Il faut donc garder une procédure de validation croisée pour trouver un modèle qui doit rester robuste.

machine learning pour estimer les scores de propension

- Intérêt de l'apprentissage automatique lors que les données sont complexes (bcp de X) et/ou de grande taille (bcp d'individu)

  => améliorer l'appariement des scores de propension

  - mieux identifier les facteurs de confusion pertinents (plus de forme de modélisation + interactions)
  - sélectionner l'algorithme le plus approprié (stratégie de comparaison)

  => réduire le risque de biais

  => améliorer la précision des estimations

Dans le cadre de la double robustesse

$$\hat{E}[Y^{a=1}] = \frac{1}{n}\sum_{i=1}^{n}\left[g(A=1, L_i) + \frac{A_i}{f(L_i)}\left(Y_i - g(A=1, L_i)\right)\right]$$

Predicted effect of *A*=1 on Y

Probability of *A*=1 estimated as a function of *L*

Error of prediction of *Y*

Le machine learning pour estimer g(A, L) est aussi possible

Mise en œuvre sous R

- Les principales fonctions d'algo sous R
  - **randomForest** : forêt aléatoire
  - **gbm** : Generalized Boosted Regression Models

```
library(randomForest)
fit <- randomForest(Y ~ ., data = df)
```

- Les packages facilitant la mise en œuvre
  - **caret** : Classification And REgression Training
  - **SuperLearner**

=> faciliter la mise en œuvre de manière homogène des différents algorithme (extraction des résultats, définition du plan de validation croisée,…)

Pour aller plus loin…

- Data science pour l'agriculture et l'environnement - Méthodes et applications avec R et Python.

https://www.editions-ellipses.fr/accueil/13446-data-science-pour-lagriculture-et-lenvironnement-methodes-et-applications-avec-r-et-python-9782340045774.html

7ième session de la formation « Data Science pour l'agriculture »
à distance, en 6 demi-journée, du 18/03/2024 au 22/03/2024
https://www.acta.asso.fr/formations/data-science-pour-lagriculture-2024