Introduction aux méthodes ensemblistes*

On va vous booster le gradient

Mélina Hillion
Unité SSP-Lab
Insee
melina.hillion@insee.fr

Olivier Meslin
Unité SSP-Lab
Insee
olivier.meslin@insee.fr

December 6, 2024

Abstract

A compléter

Keywords: machine learning • méthodes ensemblistes • formation

*Nous remercions Daffy Duck et Mickey Mouse pour leur contribution.

Sommaire

1. Introduction

Une bien belle introduction pour le site et le DT.

2. Aperçu des méthodes ensemblistes

Principe: Cette section propose une introduction intuitive aux méthodes ensemblistes. Elle s'adresse aux lecteurs qui souhaitent acquérir une compréhension générale du fonctionnement de ces techniques et identifier rapidement les situations concrètes dans lesquelles elles peuvent être utiles. L'objectif est d'en expliciter les principes-clés sans recourir au formalisme mathématique, afin de rendre le contenu accessible sans prérequis.

2.1. Que sont les méthodes ensemblistes?

Les méthodes ensemblistes sont des techniques d'apprentissage supervisé en *machine learning* développées depuis le début des années 1990. Leur objectif est de prédire une variable-cible y (appelée *target*) à partir d'un ensemble de variables prédictives $\mathbf{X}$ (appelées *features*), que ce soit pour des tâches de classification (prédire une catégorie) ou de régression (prédire une valeur numérique). Elles peuvent par exemple être utilisées pour prédire le salaire d'un salarié, la probabilité de réponse dans une enquête, le niveau de diplôme…

Plutôt que de s'appuyer sur un seul modèle complexe, les méthodes ensemblistes se caractérisent par la combinaison des prédictions de plusieurs modèles plus simples, appelés "apprenants faibles" (*weak learner* ou *base learner*), pour créer un modèle performant, dit "apprenant fort" (*strong learner*).

Le choix de ces modèles de base, ainsi que la manière dont leurs prédictions sont combinées, sont des facteurs déterminants de la performance finale. Le présent document se concentre sur les méthodes à base d'**arbres de décisions**, qui sont parmi les plus utilisées en pratique. Nous allons examiner les fondements de ces méthodes, leurs avantages et inconvénients, ainsi que les algorithmes les plus populaires.

2.2. Pourquoi utiliser des méthodes ensemblistes?

Les méthodes ensemblistes sont particulièrement bien adaptées à de nombreux cas d'usage de la statistique publique, pour deux raisons. D'une part, elles sont conçues pour s'appliquer à des *données tabulaires* (enregistrements en lignes, variables en colonnes), structure de données omniprésente dans la statistique publique. D'autre part, elles peuvent être mobilisées dans toutes les situations où le statisticien mobilise une régression linéaire ou une régression logistisque (imputation, repondération…).

Les méthodes ensemblistes présentent trois avantages par rapport aux méthodes économétriques traditionnelles (régression linéaire et régression logistique):

- Elles ont une **puissance prédictive supérieure**: alors que les méthodes traditionnelles supposent fréquemment l'existence d'une relation linéaire ou log-linéaire entre y

et $\mathbf{X}$, les méthodes ensemblistes ne font quasiment aucune hypothèse sur la relation entre y et $\mathbf{X}$, et se contentent d'approximer le mieux possible cette relation à partir des données disponibles. En particulier, les modèles ensemblistes peuvent facilement modéliser des **non-linéarités** de la relation entre y et $\mathbf{X}$ et des **interactions** entre variables explicatives *sans avoir à les spécifier explicitement* au préalable, alors que les méthodes traditionnelles supposent fréquemment l'existence d'une relation linéaire ou log-linéaire entre y et $\mathbf{X}$.

- Elles nécessitent **moins de préparation des données**: elles ne requièrent pas de normalisation des variables explicatives et peuvent s'accommoder des valeurs manquantes (selon des techniques variables selon les algorithmes).
- Elles sont généralement **moins sensibles aux valeurs extrêmes et à l'hétéroscédasticité** des variables explicatives que les approches traditionnelles.

Elles présentent par ailleurs deux inconvénients rapport aux méthodes économétriques traditionnelles. Premièrement, bien qu'il existe désormais de multiples approches permettent d'interpétrer partiellement les modèles ensemblistes, leur interprétabilité reste moindre que celle d'une régression linéaire ou logistique. Deuxièmement, les modèles ensemblistes sont plus complexes que les approches traditionnelles, et leurs hyperparamètres doivent faire l'objet d'une optimisation, par exemple au travers d'une validation croisée. Ce processus d'optimisation est généralement plus complexe et plus long que l'estimation d'une régression linéaire ou logistique. En revanche, les méthodes ensemblistes sont relativement simples à prendre en main, et ne requièrent pas nécessairement une puissance de calcul importante.

i Et par rapport au *deep learning*?

Si les approches de *deep learning* sont sans conteste très performantes pour le traitement du langage naturel, des images et du son, leur supériorité n'est pas établie pour les applications reposant sur des données tabulaires. Les comparaisons disponibles dans la littérature concluent en effet que les méthodes ensemblistes à base d'arbres sont soit plus performantes que les approches de *deep learning* (L. Grinsztajn, E. Oyallon, and G. Varoquaux [1], R. Shwartz-Ziv and A. Armon [2]), soit font jeu égal avec elles (D. McElfresh *et al.* [3]). Ces études ont identifié trois avantages des méthodes ensemblistes: elles sont peu sensibles aux variables explicatives non pertinentes, robustes aux valeurs extrêmes des variables explicatives, et capables d'approximer des fonctions très irrégulières. De plus, dans la pratique les méthodes ensemblistes sont souvent plus rapides à entraîner et moins gourmandes en ressources informatiques, et l'optimisation des hyperparamètres s'avère souvent moins complexe (R. Shwartz-Ziv and A. Armon [2]).

2.3. Comment fonctionnent les méthodes ensemblistes?

Cette section présente d'abord le modèle de base sur lesquelles sont construites les méthodes ensemblistes à base d'arbres: l'arbre de classification et de régression (CART) (Section 2.3.1). Bien que simples et intuitifs, les arbres CART sont souvent insuffisants en termes de performance lorsqu'ils sont utilisés isolément.

Elle introduit ensuite les **deux grandes familles de méthodes ensemblistes** décrites dans ce document: le *bagging* et les forêts aléatoires (Section 2.3.2), et le *gradient boosting* (Section 2.3.3).

2.3.1. Le modèle de base: l'arbre de classification et de régression

2.3.1.1. Qu'est-ce qu'un arbre CART?

Le modèle de base des méthodes ensemblistes est souvent un arbre de classification et de régression (CART, L. Breiman, J. Friedman, R. Olshen, and C. Stone [4]). Un arbre CART est un algorithme prédictif qui traite un problème de prédiction complexe en le décomposant en une série de décisions simples, organisées de manière hiérarchique. Ces décisions permettent de segmenter progressivement les données en régions homogènes au sein desquelles il est plus simple de faire des prédictions. Il s'agit d'un outil puissant pour explorer les relations entre les variables explicatives et la variable cible, sans recourir à des hypothèses *a priori* sur la forme de cette relation.

Trois caractéristiques essentielles définissent un arbre CART :

- L'arbre partitionne l'espace des variables explicatives X en régions (appelées feuilles ou *leaves*) les plus homogènes possible, au sens d'une mesure de l'hétérogénéité (par exemple, l'entropie ou l'erreur quadratique moyenne). Ces divisions vont permettre de regrouper des observations similaires pour faciliter la prédiction;
- Chaque région est définie par un ensemble de conditions, appelées régles de décision (*splitting rules*), appliquées successivement sur les variables explicatives. Par exemple, une première règle pourrait poser la question : "L'individu est-il en emploi ?", et subdiviser les données en deux groupes (oui/non). Une deuxième règle pourrait alors affiner la segmentation en posant la question : "L'individu est-il diplômé du supérieur ?". Une région spécifique serait ainsi définie par la condition combinée : "l'individu est en emploi et est diplômé du supérieur".
- Une fois l'arbre construit, chaque feuille produit une prédiction en se basant sur les données de la région correspondante. En classification, la prédiction est généralement la classe la plus fréquente parmi les observations de la région. En régression, la prédiction est souvent la moyenne des valeurs observées dans la région.

Deux conséquences importantes découlent de cette construction : - L'algorithme CART ne fait **aucune hypothèse *a priori*** sur la relation entre les variables explicatives $\mathbf{X}$ et la variable cible y. C'est une différence majeure avec les modèles économétriques standard, tels que la régression linéaire qui suppose une relation linéaire de la forme $E(y) = \mathbf{X}\beta$. - **L'arbre final est une fonction constante par morceaux**: la prédiction est **la même** pour toutes les observations situées dans la même région ; elle ne peut varier qu'entre régions.

Illustration, et représentation graphique (sous forme d'arbre et de graphique).

2.3.1.2. Avantages et limites des arbres CART

Les arbres CART présentent plusieurs avantages: leur principe est simple, ils sont aisément interprétables et peuvent faire l'objet de représentations graphiques intuitives. Par ailleurs, la flexibilité offerte par le partitionnement récursif assure que les arbres obtenus reflètent les corrélations observées dans les données d'entraînement.

Ils souffrent néanmoins de deux limites. D'une part, les arbres CART ont souvent un **pouvoir prédictif faible** qui en limite l'usage. D'autre part, ils sont **peu robustes et instables**: on dit qu'ils présentent une **variance élevée**. Ainsi, un léger changement dans les données (par exemple l'ajout ou la suppression de quelques observations) peut entraîner des modifications significatives dans la structure de l'arbre et dans la définition des régions utilisées pour la prédiction (feuilles). Les arbres CART sont notamment sensibles aux valeurs extrêmes, aux points aberrants et au bruit statistique. De plus, les prédictions des arbres CART sont sensibles

à de petites fluctuations des données d'échantillonnage: celles-ci peuvent aboutir à ce qu'une partie des observations change brutalement de feuille et donc de valeur prédite.

Ces limites motivent l'utilisation des deux familles de méthodes ensemblistes présentées dans la suite (le *bagging*, dont la *random forests*, et le *gradient boosting*), qui s'appuient sur un grand nombre d'arbres pour accroître à la fois la précision et la stabilité des prédictions. La différence essentielle entre ces deux familles portent sur la façon dont les arbres sont entraînés.

> **i** Les familles de méthodes ensemblistes
>
> Les méthodes ensemblistes basées sur des arbres de décision se répartissent en **deux grandes familles**, qui se distinguent selon la manière dont les modèles de base sont construits. Lorsque les modèles de base sont entraînés en parallèle et indépendamment les uns des autres, on parle de *bagging* (*Bootstrap Aggregating*). La *forêt aléatoire* (*random forest*) est une variante particulièrement performante du *bagging*. Lorsque les modèles de base sont *entraînés de manière séquentielle*, chaque modèle visant à corriger les erreurs des modèles précédents, on parle de *boosting*. Ce document aborde essentiellement le *gradient boosting*, qui est l'approche de *boosting* la plus utilisée actuellement.

2.3.2. Le *bagging* (Bootstrap Aggregating) et les forêts aléatoires

2.3.2.1. Le *bagging*

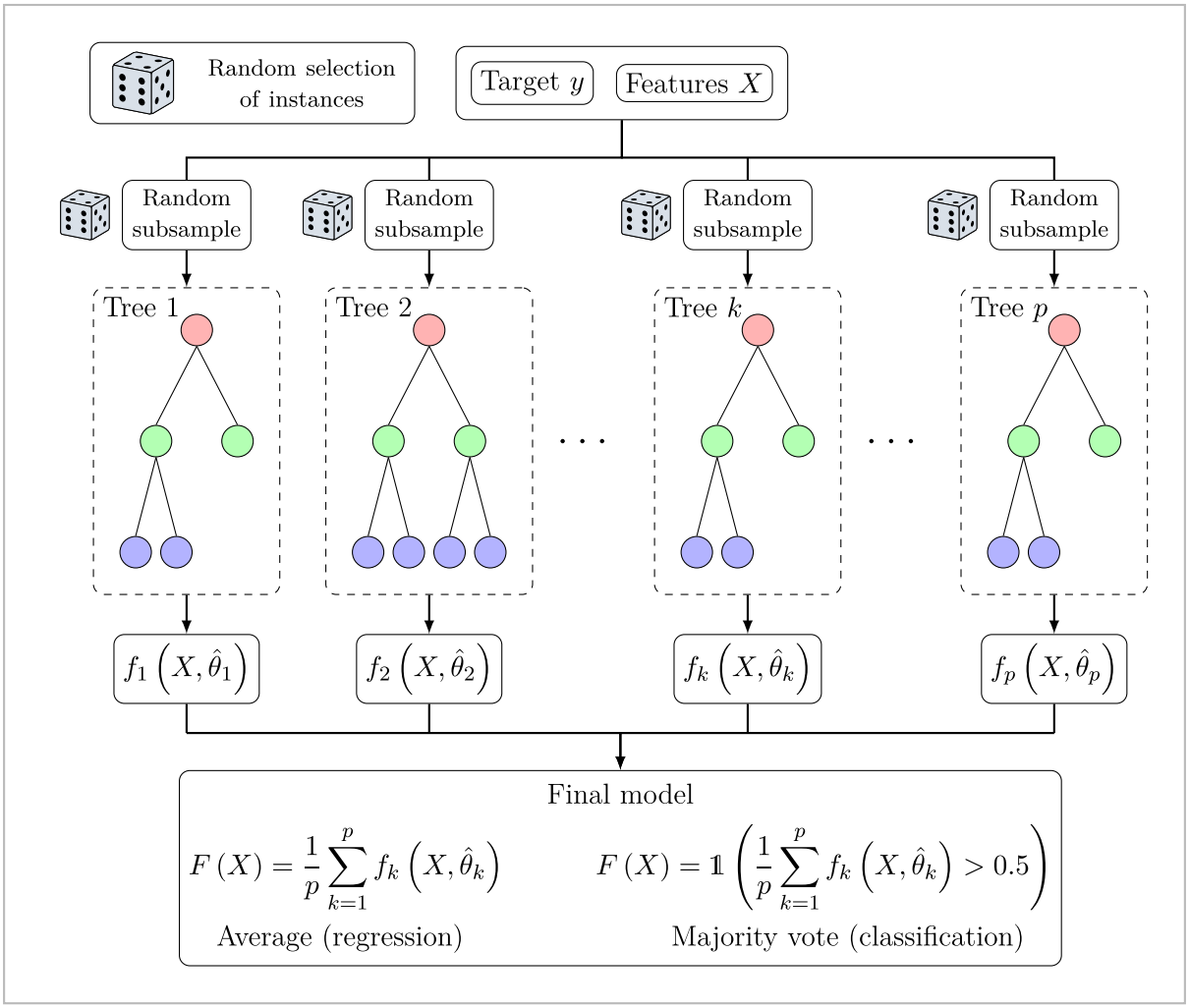
Le *bagging* (Bootstrap Aggregating) est une méthode ensembliste qui repose sur l'agrégation des prédictions de plusieurs modèles individuels, entraînés indépendamment les uns des autres, pour construire un modèle global plus performant (L. Breiman [5]). Cette approche constitue également le socle des forêts aléatoires, qui en sont une version améliorée.

Le bagging offre deux avantages majeurs par rapport aux arbres de décision CART : une meilleure capacité prédictive et une plus grande stabilité des prédictions. Cette amélioration découle de la stratégie d'entraînement. Au lieu d'entraîner un seul modèle sur l'ensemble des données, le bagging procède en trois étapes principales:

- **Tirage de sous-échantillons aléatoires**: À partir du jeu de données initial, plusieurs sous-échantillons sont générés par échantillonnage aléatoire avec remise (*bootstrapping*). Chaque sous-échantillon a la même taille que le jeu de données original, mais peut contenir des observations répétées, tandis que d'autres peuvent être omises.
- **Entraînement parallèle**: Un arbre est entraîné sur chaque sous-échantillon de manière indépendante. Ces arbres sont habituellement assez complexes et profonds.
- **Agrégation des prédictions**: Les prédictions des modèles sont combinées pour produire le résultat final. En classification, la prédiction finale est souvent déterminée par

un vote majoritaire, tandis qu'en régression, elle correspond généralement à la moyenne des prédictions.

Figure 1 : Représentation schématique d'un algorithme de *bagging*



La Figure 1 propose une représentation schématique du *bagging*: d'abord, des sous-échantillons sont générés aléatoires avec remise à partir du jeu de données d'entraînement. Ensuite, des arbres de décision sont entraînés indépendamment sur ces sous-échantillons. Enfin, leurs prédictions sont agrégées pour obtenir les prédictions finales. On procède généralement au vote majoritaire (la classe prédite majoritairement par les arbres) dans un problème de classification, et à la moyenne dans un problème de régression.

L'efficacité du *bagging* provient de la réduction de la variance qui est permise par l'agrégation des prédictions. Chaque arbre est entraîné sur un sous-échantillon légèrement différent, sujet à des fluctuations aléatoires. L'agrégation des prédictions (par moyenne ou vote majoritaire) de tous les arbres réduit la sensibilité du modèle final aux fluctuations des données d'entraînement. Le modèle final est ainsi plus robuste et plus précis que chacun des arbres pris individuellement.

- Illustration avec un cas d'usage de classification en deux dimensions.

Malgré ses avantages, le *bagging* souffre d'une limite importante qui provient de la **corrélation entre les arbres**. En effet, malgré le tirage aléatoire des sous-échantillons, les arbres présentent souvent des structures similaires, car les règles de décision sous-jacentes restent généralement assez proches. Cette corrélation réduit l'efficacité de l'agrégation et limite les gains en performance.

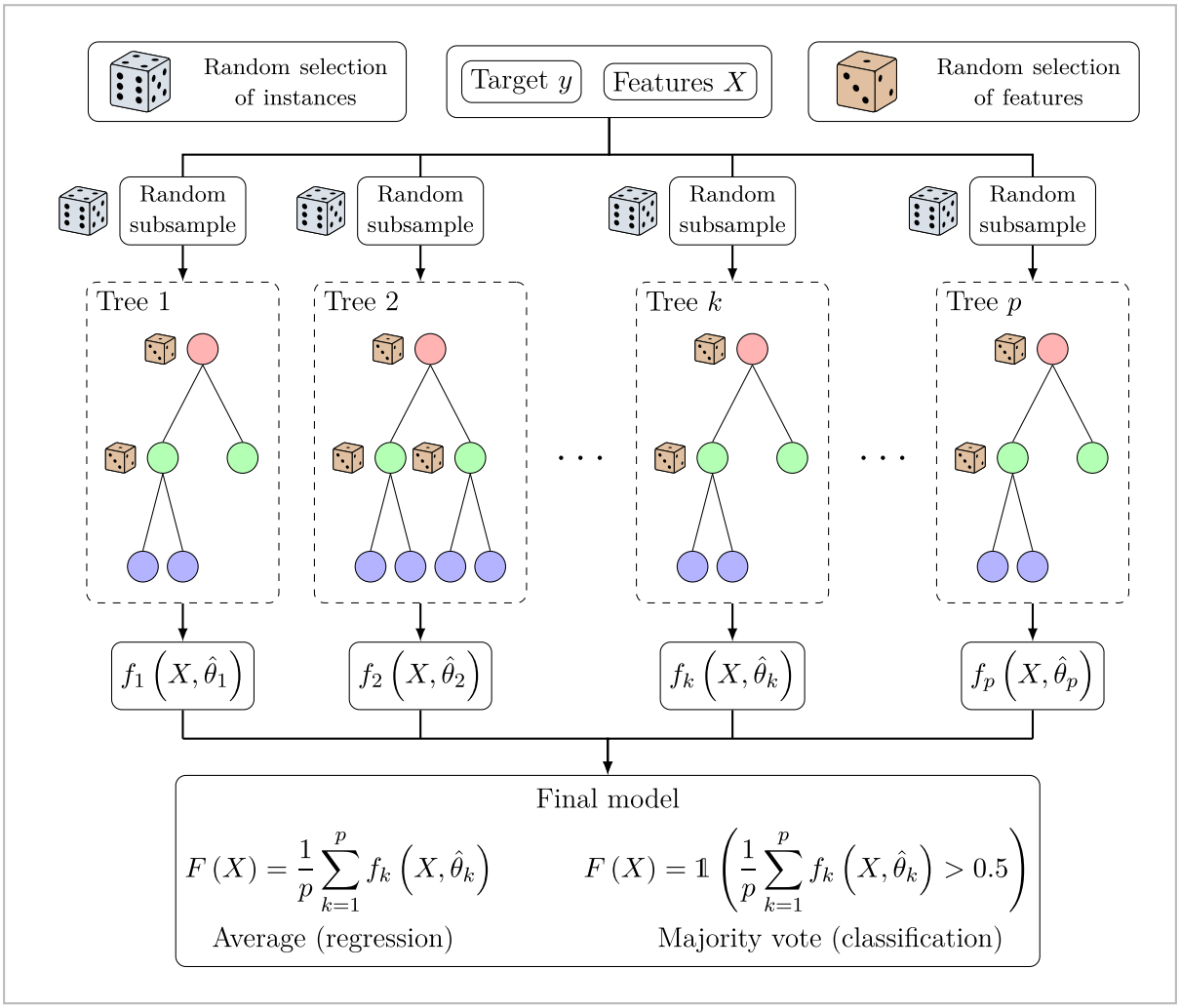
Pour réduire cette corrélation entre arbres, les forêts aléatoires introduisent une étape supplémentaire de randomisation. Leur supériorité prédictive explique pourquoi le *bagging* seul est rarement utilisé en pratique. Néanmoins, les forêts aléatoires tirent leur efficacité des principes fondamentaux du *bagging*.

2.3.2.2. Les forêts aléatoires (*random forests*)

Les forêts aléatoires (*random forests*, L. Breiman [6]) sont une variante du *bagging* qui vise à produire des modèles très performants en conciliant deux objectifs: maximiser le pouvoir prédictif des arbres pris isolément, et minimiser la corrélation entre ces arbres (le problème inhérent au *bagging*).

Pour atteindre ce second objectif, la forêt aléatoire introduit une nouvelle source de randomisation: la **sélection aléatoire de variables**. Lors de la construction de chaque arbre, au lieu d'utiliser toutes les variables disponibles pour déterminer la meilleure séparation à chaque nœud, un sous-ensemble aléatoire de variables est sélectionné. En limitant la quantité d'information à laquelle chaque arbre a accès au moment de chaque nouvelle division, cette étape supplémentaire contraint mécaniquement les arbres à être plus diversifiés (car deux arbres ne pourront plus nécessairement choisir les mêmes variables pour les mêmes séparations). Cela réduit significativement la corrélation entre les arbres, améliorant ainsi l'efficacité de l'agrégation. L'ensemble des prédictions devient ainsi plus précis et moins sujet aux fluctuations aléatoires.

Figure 2 : Représentation schématique d'un algorithme de forêt aléatoire



La Figure 2 propose une représentation schématique d'une forêt aléatoire. La logique d'ensemble reste la même que celle du *bagging*. L'échantillonnage *bootstrap* est inchangé, mais l'étape de construction de chaque arbre est modifiée pour n'utiliser, à chaque nouvelle division, qu'un sous-ensemble aléatoire de variables. L'agrégation des prédictions se fait ensuite de la même manière que pour le *bagging*.

Le principal enjeu de l'entraînement d'une forêt aléatoire est de trouver le bon arbitrage entre puissance prédictive des arbres individuels (que l'on souhaite maximiser) et corrélation entre les arbres (que l'on souhaite minimiser). L'optimisation des hyper-paramètres des forêts aléatoires (dont le plus important est le nombre de variables sélectionnées à chaque noeud) vise précisément à choisir le meilleur compromis possible entre pouvoir prédictif invividuel et diversité des arbres.

Les forêts aléatoires sont très populaires car elles sont faciles à implémenter, peu sensibles aux hyperparamètres (elles fonctionnent bien avec les valeurs par défaut de la plupart des

implémentations proposées en R ou en Python), et offrent de très bonnes performances dans de nombreux cas. Cependant, comme toute méthode d'apprentissage automatique, elles restent sujettes au surapprentissage (voir encadré), bien que dans une moindre mesure par rapport à d'autres techniques comme le *gradient boosting*.

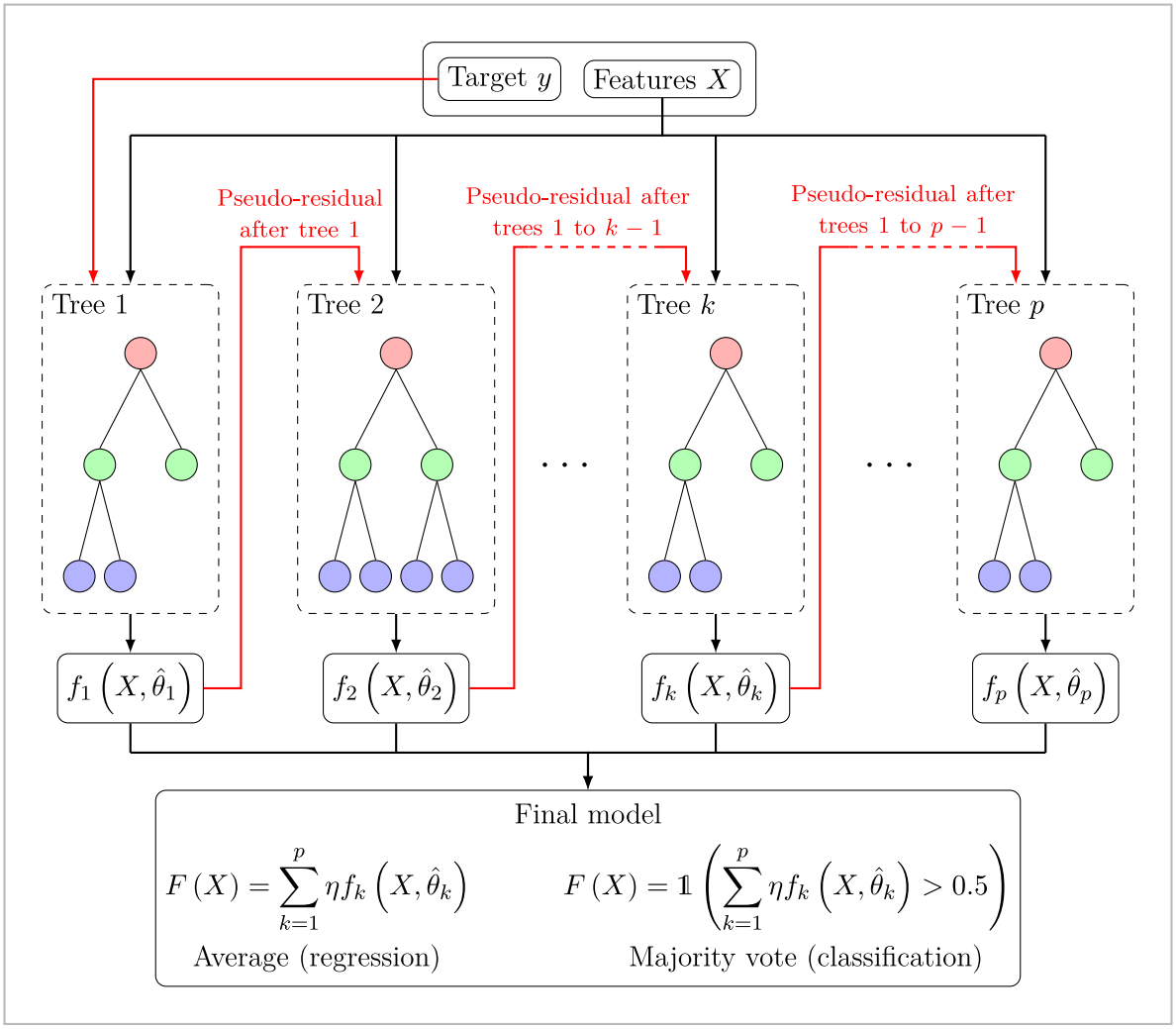
> **i** Qu'est-ce que le surapprentissage?
>
> Le surapprentissage (*overfitting*) est un phénomène fréquent en *machine learning* où un modèle apprend non seulement les relations sous-jacentes entre la variable cible et les variables explicatives, mais également le bruit présent dans les données d'entraînement. En capturant ces fluctuations aléatoires plutôt que les tendances générales, le modèle affiche une performance excellente mais trompeuse sur les données d'entraînement, et s'avère médiocre sur des données nouvelles ou de test, car il ne parvient pas à généraliser efficacement.

2.3.3. Le *gradient boosting*

Contrairement aux forêts aléatoires qui combinent des arbres de décision complexes et indépendants, le *gradient boosting* construit un ensemble d'arbres plus simples et entraînés de manière séquentielle. Chaque arbre vise à corriger les erreurs commises par les arbres précédents, améliorant progressivement la précision du modèle global. Cette approche repose sur des fondements théoriques très différents de ceux du *bagging*.

La logique du *gradient boosting* est illustrée par La Figure 3:

Figure 3 : Représentation schématique d'un algorithme de *gradient boosting*



- Un premier modèle simple et peu performant est entraîné sur les données.
- Un deuxième modèle est entraîné de façon à corriger les erreurs du premier modèle (par exemple en pondérant davantage les observations mal prédites);
- Ce processus est répété en ajoutant des modèles simples, chaque modèle corrigeant les erreurs commises par l'ensemble des modèles précédents;
- Tous ces modèles sont finalement combinés (souvent par une somme pondérée) pour obtenir un modèle complexe et performant.

Le *gradient boosting* offre des performances élevées mais exige une attention particulière portée sur la configuration des hyperparamètres et sur la prévention du surapprentissage. En particulier, les hyperparamètres sont nombreux et, contrairement aux forêts aléatoires, nécessitent un ajustement minutieux pour obtenir des résultats optimaux. Une mauvaise configuration peut conduire à des performances médiocres ou à un surapprentissage. L'utilisation du *gradient*

boosting nécessite donc une bonne connaissance du fonctionnement des algorithmes. En outre, les algorithmes de *gradient boosting* peuvent être sensibles au bruit dans les données et aux erreurs dans la variable cible. Un prétraitement rigoureux des données est donc essentiel. Enfin, une validation rigoureuse sur un jeu de données de test indépendant (non utilisé pendant l'entraînement) est indispensable pour évaluer la qualité du modèle obtenu par *gradient boosting*.

2.4. Comparaison entre forêts aléatoires et *gradient boosting*

Les forêts aléatoires et le *gradient boosting* paraissent très similaires au premier abord: il s'agit de deux approches ensemblistes, qui construisent des modèles très prédictifs performants en combinant un grand nombre d'arbres de décision. Mais en réalité, ces deux approches présentent plusieurs différences fondamentales:

- Les deux approches reposent sur des **fondements théoriques différents**: la loi des grands nombres pour les forêts aléatoires, la théorie de l'apprentissage statistique pour le *boosting*.
- **Les arbres n'ont pas le même statut dans les deux approches**. Dans une forêt aléatoire, les arbres sont entraînés indépendamment les uns des autres et constituent chacun un modèle à part entière, qui peut être utilisé, représenté et interprété isolément. Dans un modèle de *boosting*, les arbres sont entraînés séquentiellement, ce qui implique que chaque arbre n'a pas de sens indépendamment de l'ensemble des arbres qui l'ont précédé dans l'entraînement. Par ailleurs, les arbres d'une forêt aléatoire sont relativement complexes et profonds (car ce sont des modèles à part entière), alors que dans le *boosting* les arbres sont plus souvent simples et peu profonds.
- Les **points d'attention lors de l'entraînement** des algorithmes sont différents: l'enjeu principal de l'entraînement d'une forêt aléatoire est trouver le bon arbitrage entre puissance prédictive des arbres et corrélation entre arbres, tandis que l'entraînement d'un algorithme de *gradient boosting* porte davantage sur la lutte contre le surapprentissage.
- **Complexité d'usage**: les forêts aléatoires s'avèrent plus faciles à prendre en main que le *gradient boosting*, car elles comprennent moins d'hyperparamètres dont l'optimisation est moins complexe.
- **Conditions d'utilisation**: il est possible d'évaluer la qualité d'une forêt aléatoire en utilisant les données sur lesquelles elle a été entraînée grâce à l'approche *out-of-bag*, alors que c'est impossible avec le *gradient boosting*, pour lequel il faut impérativement conserver un ensemble de test. Cette différence peut sembler purement technique en

apparence, mais elle s'avère importante en pratique dans de nombreuses situations, par exemple lorsque les données disponibles sont de taille restreinte ou lorsque les ressources informatiques disponibles ne sont pas suffisantes pour mener un exercice de validation croisée.

2.4.1. Quelle approche choisir?

Le point de départ recommandé est de commencer par entraîner une forêt aléatoire avec les hyperparamètres par défaut.

Principe: cette partie propose une présentation formalisée des méthodes ensemblistes, à destination des personnes souhaitant comprendre en détail le fonctionnement des algorithmes.

3. La brique élémentaire: l'arbre de décision

Les arbres de décision sont des outils puissants en apprentissage automatique, utilisés pour des tâches de classification et de régression. Ces algorithmes non paramétriques consistent à diviser l'espace des caractéristiques en sous-ensembles homogènes à l'aide de règles simples, afin de faire des prédictions. Malgré leur simplicité apparente, les arbres de décision sont capable de saisir des relations complexes et non linéaires entre les variables (ou *caractéristiques*) d'un jeu de données.

3.1. Le principe fondamental : partitionner pour prédire

Imaginez que vous souhaitiez prédire le prix d'une maison en fonction de sa superficie et de son nombre de pièces. L'espace des caractéristiques (superficie et nombre de pièces) est vaste, et les prix des maisons (la *réponse* à prédire) sont très variables. Pour prédire le prix des maisons, l'idée est de diviser cet espace en zones plus petites, où les maisons ont des prix similaires, et d'attribuer une prédiction identique à toutes les maisons situées dans la même zone.

3.1.1. Les défis du partitionnement optimal

L'objectif principal est de trouver la partition de l'espace des caractéristiques qui offre les meilleures prédictions possibles. Cependant, cet objectif se heurte à plusieurs difficultés, et la complexité du problème augmente rapidement avec le nombre de caractéristiques et la taille de l'échantillon:

- **Infinité des découpages possibles** : Il existe une infinité de façons de diviser l'espace des caractéristiques.
- **Complexité de la paramétrisation** : Il est difficile de représenter tous ces découpages avec un nombre limité de paramètres.
- **Optimisation complexe** : Même avec une paramétrisation, trouver le meilleur découpage nécessite une optimisation complexe, souvent irréaliste en pratique.

3.1.2. Les solutions apportées par les arbres de décision

Pour surmonter ces difficultés, les méthodes d'arbres de décision, et notamment la plus célèbre, l'algorithme CART (Classication And Regression Tree, L. Breiman, J. Friedman, R. Olshen, and C. Stone [4]), adoptent deux approches clés :

1. **Simplification du partitionnement de l'espace**

Au lieu d'explorer tous les découpages possibles, les arbres de décision partitionnent l'espace des caractéristiques en plusieurs régions distinctes (non chevauchantes) en appliquant des règles de décision simples. Les règles suivantes sont communément adoptées:

- **Découpages binaires simples** : À chaque étape, l'algorithme divise une région de l'espace en deux sous-régions en se basant sur une seule caractéristique (ou *variable*) et en définissant un seul seuil (ou *critère*) pour cette segmentation. Concrètement, cela revient à poser une question du type : "La valeur de la caractéristique X dépasse-t-elle un certain seuil ?" Par exemple : "La superficie de la maison est-elle supérieure à 100 m² ?". Les deux réponses possibles ("Oui" ou "Non") génèrent deux nouvelles sous-régions distinctes de l'espace, chacune correspondant à un sous-ensemble de données plus homogène.
- **Prédictions locales** : Lorsque l'algorithme s'arrête, une prédiction simple est faite dans chaque région. Il s'agit souvent de la moyenne des valeurs cibles dans cette région (régression) ou de la classe majoritaire (classification).

Ces règles de découpage rendent le problème d'optimisation plus simple mais également plus interprétable.

2. **Optimisation gloutonne (greedy)**

Plutôt que d'optimiser toutes les divisions simultanément, les arbres de décision utilisent une approche simplifiée, récursive et séquentielle :

- **Division étape par étape** : À chaque étape, l'arbre choisit la meilleure division possible sur la base d'un critère de réduction de l'hétérogénéité intra-région. En revanche, il ne prend pas en compte les étapes d'optimisation futures.
- **Critère local** : La décision est basée sur la réduction immédiate de l'impureté ou de l'erreur de prédiction (par exemple, la réduction de la variance pour la régression). Ce processus est répété pour chaque sous-région, ce qui permet d'affiner progressivement la partition de l'espace en fonction des caractéristiques les plus discriminantes.

Cette méthode dite "gloutonne" (*greedy*) s'avère efficace pour construire un partitionnement de l'espace des caractéristiques, car elle décompose un problème d'optimisation complexe en une succession de problèmes plus simples et plus rapides à résoudre. Le résultat obtenu n'est pas nécessairement un optimum global, mais il s'en approche raisonnablement et surtout rapidement.

Le terme "arbre de décision" provient de la structure descendante en forme d'arbre inversé qui émerge lorsqu'on utilise un algorithme glouton pour découper l'espace des caractéristiques en sous-ensemble de réponses homogènes de manière récursive. A chaque étape, deux nouvelles branches sont créées et forment une nouvelle partition de l'espace des caractéristiques.

Une fois entraîné, un arbre de décision est une fonction **constante par morceaux** défini sur l'espace des caractéristiques. En raison de leur nature **non-continue** et **non-différentiable**,

il est impossible d'utiliser des méthodes d'optimisation classiques reposant sur le calcul de gradients.

3.1.3. Terminologie et structure d'un arbre de décision

Nous présentons la structure d'un arbre de décision et les principaux éléments qui le composent.

- **Nœud Racine (Root Node)** : Le nœud racine est le point de départ de l'arbre de décision, il est situé au sommet de l'arbre. Il contient l'ensemble des données d'entraînement avant toute division. À ce niveau, l'algorithme cherche la caractéristique la plus discriminante, c'est-à-dire celle qui permet de diviser les données de manière à optimiser une fonction de perte (comme l'indice de Gini pour la classification ou la variance pour la régression).
- **Nœuds Internes (Internal Nodes)** : Les nœuds internes sont les points intermédiaires où l'algorithme CART applique des règles de décision pour diviser les données en sous-ensembles plus petits. Chaque nœud interne représente une question ou condition basée sur une caractéristique particulière (par exemple, "La superficie de la maison est-elle supérieure à 100 m² ?"). À chaque étape, une seule caractéristique (la superficie) et un seul seuil (supérieur à 100) sont utilisés pour faire la division.
- **Branches**: Les branches sont les connexions entre les nœuds, elles illustrent le chemin que les données suivent en fonction des réponses aux questions posées dans les nœuds internes. Chaque branche correspond à une décision binaire, "Oui" ou "Non", qui oriente les observations vers une nouvelle subdivision de l'espace des caractéristiques.
- **Nœuds Terminaux ou Feuilles (Leaf Nodes ou Terminal Nodes)** : Les nœuds terminaux, situés à l'extrémité des branches, sont les points où le processus de division s'arrête. Ils fournissent la prédiction finale.
    - En **classification**, chaque feuille correspond à une classe prédite (par exemple, "Oui" ou "Non").
    - En **régression**, chaque feuille fournit une valeur numérique prédite (comme le prix estimé d'une maison).

Figure illustrative : Une représentation visuelle de la structure de l'arbre peut être utile ici pour illustrer les concepts de nœuds, branches et feuilles.

3.1.4. Illustration

Supposons que nous souhaitions prédire le prix d'une maison en fonction de sa superficie et de son nombre de pièces. Un arbre de décision pourrait procéder ainsi :

1. **Première division** : "La superficie de la maison est-elle supérieure à 100 m² ?"

- Oui : Aller à la branche de gauche.
- Non : Aller à la branche de droite.

2. **Deuxième division (branche de gauche)** : "Le nombre de pièces est-il supérieur à 4 ?"
    - Oui : Prix élevé (par exemple, plus de 300 000 €).
    - Non : Prix moyen (par exemple, entre 200 000 € et 300 000 €).
3. **Deuxième division (branche de droite)** : "Le nombre de pièces est-il supérieur à 2 ?"
    - Oui : Prix moyen (par exemple, entre 150 000 € et 200 000 €).
    - Non : Prix bas (par exemple, moins de 150 000 €).

Cet arbre utilise des règles simples pour diviser l'espace des caractéristiques (superficie et nombre de pièces) en sous-groupes homogènes et fournir une prédiction (estimer le prix d'une maison).

Figure illustrative

3.2. L'algorithme CART, un partitionnement binaire récursif

L'algorithme CART (Classification and Regression Trees) proposé par L. Breiman, J. Friedman, R. Olshen, and C. Stone [4] est une méthode utilisée pour construire des arbres de décision, que ce soit pour des tâches de classification ou de régression. L'algorithme CART fonctionne en partitionnant l'espace des caractéristiques en sous-ensembles de manière récursive, en suivant une logique de décisions binaires à chaque étape. Ce processus est itératif et suit plusieurs étapes clés.

3.2.1. Définir une fonction d'impureté adaptée au problème

La **fonction d'impureté** est une mesure locale utilisée dans la construction des arbres de décision pour évaluer la qualité des divisions à chaque nœud. Elle quantifie le degré d'hétérogénéité des observations dans un nœud par rapport à la variable cible (classe pour la classification, ou valeur continue pour la régression). Plus précisément, une mesure d'impureté est conçue pour croître avec la dispersion dans un nœud. Un nœud est dit **pur** lorsque toutes les observations qu'il contient appartiennent à la même classe (classification) ou présentent des valeurs similaires/identiques (régression).

L'algorithme CART utilise ce type de mesure pour choisir les divisions qui créent des sous-ensembles plus homogènes que le nœud parent. À chaque étape de construction, l'algorithme sélectionne la division qui réduit le plus l'impureté, afin de garantir des nœuds de plus en plus homogènes au fur et à mesure que l'arbre se développe.

Le choix de la fonction d'impureté dépend du type de problème :

- **Classification** : L'**indice de Gini** ou l'**entropie** sont très souvent utilisées pour évaluer la dispersion des classes dans chaque nœud.
- **Régression** : La **somme des erreurs quadratiques** (SSE) est souvent utilisée pour mesurer la variance des valeurs cibles dans chaque nœud.

3.2.1.1. Mesures d'impureté classiques pour les problèmes de classification

Dans le cadre de la classification, l'objectif est de partitionner les données de manière à ce que chaque sous-ensemble (ou région) soit le plus homogène possible en termes de classe prédite. Plusieurs mesures d'impureté sont couramment utilisées pour évaluer la qualité des divisions.

Propriété-définition d'une mesure d'impureté

Pour un nœud t contenant K classes, une **mesure d'impureté** $I(t)$ est une fonction qui quantifie l'hétérogénéité des classes dans ce nœud. Elle doit satisfaire les propriétés suivantes :

- **Pureté maximale** : Lorsque toutes les observations du nœud appartiennent à une seule classe, c'est-à-dire que la proportion $p_k = 1$ pour une classe k et $p_j = 0$ pour toutes les autres classes $j \neq k$, l'impureté est minimale et $I(t) = 0$. Cela indique que le nœud est **entièrement pur**, ou homogène.
- **Impureté maximale** : Lorsque les observations sont réparties de manière uniforme entre toutes les classes, c'est-à-dire que $p_k = \frac{1}{K}$ pour chaque classe k, l'impureté atteint son maximum. Cette situation reflète une **impureté élevée**, car le nœud est très hétérogène et contient une forte incertitude sur la classe des observations.

1. L'indice de Gini

L'**indice de Gini** est l'une des fonctions de perte les plus couramment utilisées pour la classification. Il mesure la probabilité qu'un individu sélectionné au hasard dans un nœud soit mal classé si on lui attribue une classe au hasard, en fonction de la distribution des classes dans ce nœud.

Pour un nœud t contenant K classes, l'indice de Gini $G(t)$ est donné par :

$$G(t) = 1 - \sum_{k=1}^{K} p_k^2 \tag{1}$$

où p_k est la proportion d'observations appartenant à la classe k dans le nœud t.

Critère de choix : L'indice de Gini est souvent utilisé parce qu'il est simple à calculer et capture bien l'homogénéité des classes au sein d'un nœud. Il privilégie les partitions où une classe domine fortement dans chaque sous-ensemble.

2. L'entropie (ou entropie de Shannon)

L'**entropie** est une autre mesure de l'impureté utilisée dans les arbres de décision. Elle mesure la quantité d'incertitude ou de désordre dans un nœud, en s'appuyant sur la théorie de l'information.

Pour un nœud t contenant K classes, l'entropie $E(t)$ est définie par :

$$E(t) = -\sum_{k=1}^{K} p_k \log(p_k) \tag{2}$$

où p_k est la proportion d'observations de la classe k dans le nœud t.

Critère de choix : L'entropie a tendance à être plus sensible aux changements dans les distributions des classes que l'indice de Gini, car elle attribut un poids plus élevé aux événements rares (valeurs de p_k très faibles). Elle est souvent utilisée lorsque l'erreur de classification des classes minoritaires est particulièrement importante.

3. Taux d'erreur

Le **taux d'erreur** est une autre mesure de l'impureté parfois utilisée dans les arbres de décision. Il représente la proportion d'observations mal classées dans un nœud.

Pour un nœud t, le taux d'erreur TE (t) est donné par :

$$\text{TE } (t) = 1 - \max(p_k) \tag{3}$$

où $\max(p_k)$ est la proportion d'observations appartenant à la classe majoritaire dans le nœud.

Critère de choix : Bien que le taux d'erreur soit simple à comprendre, il est moins souvent utilisé dans la construction des arbres de décision parce qu'il est moins sensible que l'indice de Gini ou l'entropie aux petits changements dans la distribution des classes.

3.2.1.2. Mesures d'impureté classiques pour les problèmes de régression

Dans les problèmes de régression, l'objectif est de partitionner les données de manière à réduire au maximum la variabilité des valeurs au sein de chaque sous-ensemble. Pour mesurer cette variabilité, la somme des erreurs quadratiques (SSE) est la fonction d'impureté la plus couramment employée. Elle évalue l'impureté d'une région en quantifiant à quel point les valeurs de cette région s'écartent de la moyenne locale.

1.Somme des erreurs quadratiques (SSE) ou variance

La **somme des erreurs quadratiques** (ou **SSE**, pour *Sum of Squared Errors*) est une mesure qui quantifie la dispersion des valeurs dans un nœud par rapport à la moyenne des valeurs dans ce nœud.

Formule : Pour un nœud t, contenant N observations avec des valeurs y_i, la SSE est donnée par :

$$\text{SSE } (t) = \sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{4}$$

où $\hat{y}$ est la moyenne des valeurs y_i dans le nœud t.

Propriété :

- Si toutes les valeurs de y_i dans un nœud sont proches de la moyenne $\hat{y}$, la SSE sera faible, indiquant une homogénéité élevée dans le nœud.
- En revanche, une SSE élevée indique une grande variabilité dans les valeurs, donc un nœud impur.

Critère de choix : La somme des erreurs quadratiques (SSE) est particulièrement sensible aux écarts élevés entre les valeurs observées et la moyenne prédite. En cherchant à minimiser la SSE, les modèles visent à former des nœuds dans lesquels les valeurs des observations sont aussi proches que possible de la moyenne locale.

3.2.2. Identifier la partition binaire maximisant la réduction de l'impureté

Une fois la mesure d'impureté définie, l'algorithme CART examine toutes les divisions binaires possibles de l'espace des caractéristiques. À chaque nœud, et pour chaque caractéristique, il cherche à identifier le **seuil optimal**, c'est-à-dire le seuil qui minimise le plus efficacement l'impureté des deux sous-ensembles générés. L'algorithme compare ensuite toutes les divisions potentielles (caractéristiques et seuils optimaux associés à chaque nœud) et sélectionne celle qui entraîne la réduction maximale de l'impureté.

Prenons l'exemple d'une caractéristique continue, telle que la superficie d'une maison :

- Si l'algorithme teste la règle "Superficie > 100 m²", il calcule la fonction de perte pour les deux sous-ensembles générés par cette règle ("Oui" et "Non").
- Ce processus est répété pour différentes valeurs seuils afin de trouver la partition qui minimise le plus efficacement l'impureté au sein des sous-ensembles.

3.2.3. Réitérer le processus jusqu'à atteindre un critère d'arrêt

L'algorithme CART poursuit le partitionnement de l'espace des caractéristiques en appliquant de manière récursive les mêmes étapes : identification de la caractéristique et du seuil optimal pour chaque nœud, puis sélection du partitionnement binaire qui maximise la réduction de l'impureté. Ce processus est répété jusqu'à ce qu'un **critère d'arrêt** soit atteint, par exemple :

- **Profondeur maximale de l'arbre** : Limiter le nombre de divisions successives pour éviter un arbre trop complexe.
- **Nombre minimum d'observations par feuille** : Empêcher la création de feuilles contenant très peu d'observations, ce qui réduirait la capacité du modèle à généraliser.
- **Réduction minimale de l'impureté à chaque étape**

3.2.4. Elagage (*pruning*)

3.2.5. Prédire

Une fois l'arbre construit, la prédiction pour une nouvelle observation s'effectue en suivant les branches de l'arbre, en partant du nœud racine jusqu'à un nœud terminal (ou feuille). À chaque nœud interne, une décision est prise en fonction des valeurs des caractéristiques de l'observation, ce qui détermine la direction à suivre vers l'un des sous-ensembles. Ce cheminement se poursuit jusqu'à ce que l'observation atteigne une feuille, où la prédiction finale est effectuée.

- En **classification**, la classe attribuée est celle majoritaire dans la feuille atteinte.
- En **régression**, la valeur prédite est généralement la moyenne des valeurs cibles des observations dans la feuille.

3.2.6. Critères de qualité et ajustements

Pour améliorer la performance de l'arbre, on peut ajuster les hyperparamètres tels que la profondeur maximale ou le nombre minimum d'observations dans une feuille. De plus, des techniques comme la **prédiction avec arbres multiples** (bagging, forêts aléatoires) permettent de surmonter les limites des arbres individuels, souvent sujets au surapprentissage.

3.3. Avantages et limites de cette approche

3.3.1. Avantages

- **Interprétabilité** : Les arbres de décision sont faciles à comprendre et à visualiser.
- **Simplicité** : Pas besoin de transformations complexes des données.
- **Flexibilité** : Ils peuvent gérer des caractéristiques numériques et catégorielles, ainsi que les valeurs manquantes.
- **Gestion des interactions** : Modèles non paramétriques, pas d'hypothèses sur les lois par les variables. Ils capturent naturellement les interactions entre les caractéristiques.

3.3.2. Limites

- **Surapprentissage** : Les arbres trop profonds peuvent surapprendre les données d'entraînement.
- **Optimisation locale** : L'approche gloutonne peut conduire à des solutions sous-optimales globalement (optimum local).
- **Stabilité** : De petits changements dans les données peuvent entraîner des changements significatifs dans la structure de l'arbre (manque de robustesse).

4. Le *bagging*

Le *bagging*, ou "bootstrap aggregating", est une méthode ensembliste qui vise à améliorer la stabilité et la précision des algorithmes d'apprentissage automatique en agrégeant plusieurs modèles (L. Breiman [5]). Chaque modèle est entraîné sur un échantillon distinct généré par une technique de rééchantillonnage (*bootstrap*). Ces modèles sont ensuite combinés pour produire une prédiction agrégée, souvent plus robuste et généralisable que celle obtenue par un modèle unique.

4.1. Principe du *bagging*

Le *bagging* comporte trois étapes principales:

- **L'échantillonnage bootstrap** : L'échantillonnage bootstrap consiste à créer des échantillons distincts en tirant aléatoirement avec remise des observations du jeu de données initial. Chaque échantillon *bootstrap* contient le même nombre d'observations que le jeu de données initial, mais certaines observations sont répétées (car sélectionnées plusieurs fois), tandis que d'autres sont omises.
- **L'entraînement de plusieurs modèles** : Un modèle (aussi appelé *apprenant de base* ou *weak learner*) est entraîné sur chaque échantillon bootstrap. Les modèles peuvent être des arbres de décision, des régressions ou tout autre algorithme d'apprentissage. Le *bagging* est particulièrement efficace avec des modèles instables, tels que les arbres de décision non élagués.
- **L'agrégation des prédictions** : Les prédictions de tous les modèles sont ensuite agrégées, en procédant généralement à la moyenne (ou à la médiane) des prédictions dans le cas de la régression, et au vote majoritaire (ou à la moyenne des probabilités prédites pour chaque classe) dans le cas de la classification, afin d'obtenir des prédictions plus précises et généralisables.

4.2. Pourquoi (et dans quelles situations) le *bagging* fonctionne

Certains modèles sont très sensibles aux données d'entraînement, et leurs prédictions sont très instables d'un échantillon à l'autre. L'objectif du *bagging* est de construire un prédicteur plus précis en agrégeant les prédictions de plusieurs modèles entraînés sur des échantillons (légèrement) différents les uns des autres.

L. Breiman [5] montre que cette méthode est particulièrement efficace lorsqu'elle est appliquée à des modèles très instables, dont les performances sont particulièrement sensibles aux variations du jeu de données d'entraînement, et peu biaisés.

Cette section vise à mieux comprendre comment (et sous quelles conditions) l'agrégation par *bagging* permet de construire un prédicteur plus performant.

Dans la suite, nous notons $\varphi(x, L)$ un prédicteur (d'une valeur numérique dans le cas de la *régression* ou d'un label dans le cas de la *classification*), entraîné sur un ensemble d'apprentissage L, et prenant en entrée un vecteur de caractéristiques x.

4.2.1. La régression: réduction de l'erreur quadratique moyenne par agrégation

Dans le contexte de la **régression**, l'objectif est de prédire une valeur numérique Y à partir d'un vecteur de caractéristiques x. Un modèle de régression $\phi(x, L)$ est construit à partir d'un ensemble d'apprentissage L, et produit une estimation de Y pour chaque observation x.

4.2.1.1. Définition du prédicteur agrégé

Dans le cas de la régression, le **prédicteur agrégé** est défini comme suit :

$ _A(x) = E_L[(x, L)] $

où $\phi_A(x)$ représente la prédiction agrégée, $E_L[.]$ correspond à l'espérance prise sur tous les échantillons d'apprentissage possibles L, chacun étant tiré selon la même distribution que le jeu de données initial, et $\phi(x, L)$ correspond à la prédiction du modèle construit sur l'échantillon d'apprentissage L.

4.2.1.2. La décomposition biais-variance

Pour mieux comprendre comment l'agrégation améliore la performance globale d'un modèle individuel $\phi(x, L)$, revenons à la **décomposition biais-variance** de l'erreur quadratique moyenne (il s'agit de la mesure de performance classiquement considérée dans un problème de régression):

$$E_L\left[(Y - \phi(x, L))^2\right] = \underbrace{(E_L[\phi(x, L) - Y])^2}_{\text{Biais}^2} + \underbrace{E_L\left[(\phi(x, L) - E_L[\phi(x, L)])^2\right]}_{\text{Variance}} \tag{5}$$

- Le **biais** est la différence entre la valeur observée Y que l'on souhaite prédire et la prédiction moyenne $E_L[\phi(x, L)]$. Si le modèle est sous-ajusté, le biais sera élevé.
- La **variance** est la variabilité des prédictions ($\phi(x, L)$) autour de leur moyenne ($E_L[\phi(x, L)]$). Un modèle avec une variance élevée est très sensible aux fluctuations au sein des données d'entraînement: ses prédictions varient beaucoup lorsque les données d'entraînement se modifient.

L'équation 5 illustre l'**arbitrage biais-variance** qui est omniprésent en *machine learning*: plus la complexité d'un modèle s'accroît (exemple: la profondeur d'un arbre), plus son biais sera plus faible (car ses prédictions seront de plus en plus proches des données d'entraînement), et

plus sa variance sera élevée (car ses prédictions, étant très proches des données d'entraînement, auront tendance à varier fortement d'un jeu d'entraînement à l'autre).

4.2.1.3. L'inégalité de Breiman (1996)

L. Breiman [5] compare l'erreur quadratique moyenne d'un modèle individuel avec celle du modèle agrégé et démontre l'inégalité suivante :

$ (Y - _A(x))^2 E_L[(Y - (x, L))^2] $ {#eq-inegalite-breiman1996}

- Le terme $(Y - \phi_A(x))^2$ représente l'erreur quadratique du **prédicteur agrégé** $\phi_A(x)$;
- Le terme $E_L\left[(Y - \phi(x, L))^2\right]$ est l'erreur quadratique moyenne d'un **prédicteur individuel** $\phi(x, L)$ entraîné sur un échantillon aléatoire L. Cette erreur varie en fonction des données d'entraînement.

Cette inégalité montre que **l'erreur quadratique moyenne du prédicteur agrégé est toujours inférieure ou égale à la moyenne des erreurs des prédicteurs individuels**. Puisque le biais du prédicteur agrégé est identique au biais du prédicteur individuel, alors l'inégalité précédente implique que la **variance du modèle agrégé** $\phi_A(x)$ est **toujours inférieure ou égale** à la variance moyenne d'un modèle individuel :

$ (_A(x)) = (E_L[(x, L)]) E_L[((x, L))] $

Autrement dit, le processus d'agrégation réduit l'erreur de prédiction globale en réduisant la **variance** des prédictions, tout en conservant un biais constant.

Ce résultat ouvre la voie à des considérations pratiques immédiates. Lorsque le modèle individuel est instable et présente une variance élevée, l'inégalité $Var(\phi_A(x)) \leq E_L[Var(\phi(x, L))]$ est forte, ce qui signifie que l'agrégation peut améliorer significativement la performance globale du modèle. En revanche, si $\phi(x, L)$ varie peu d'un ensemble d'entraînement à un autre (modèle stable avec variance faible), alors $Var(\phi_A(x))$ est proche de $E_L[Var(\phi(x, L))]$, et la réduction de variance apportée par l'agrégation est faible. Ainsi, **le *bagging* est particulièrement efficace pour les modèles instables**, tels que les arbres de décision, mais moins efficace pour les modèles stables tels que les méthodes des k plus proches voisins.

4.2.2. La classification: vers un classificateur presque optimal par agrégation

Dans le cas de la classification, le mécanisme de réduction de la variance par le *bagging* permet, sous une certaine condition, d'atteindre un **classificateur presque optimal** (*nearly optimal classifier*). Ce concept a été introduit par L. Breiman [5] pour décrire un modèle qui tend à classer une observation dans la classe la plus probable, avec une performance approchant celle du classificateur Bayésien optimal (la meilleure performance théorique qu'un modèle de classification puisse atteindre).

Pour comprendre ce résutlat, introduisons $Q(j \mid x) = E_L\left(1_{\varphi(x,L)=j}\right) = P(\varphi(x, L) = j)$, la probabilité qu'un modèle $\varphi(x, L)$ prédise la classe j pour l'observation x, et $P(j \mid x)$, la probabilité réelle (conditionnelle) que x appartienne à la classe j.

4.2.2.1. Définition : classificateur order-correct

Un classificateur $\varphi(x, L)$ est dit **order-correct** pour une observation x si, en espérance, il identifie **correctement la classe la plus probable**, même s'il ne prédit pas toujours avec exactitude les probabilités associées à chaque classe $Q(j \mid x)$.

Cela signifie que si l'on considérait tous les ensemble de données possibles, et que l'on évaluait les prédictions du modèle en x, la majorité des prédictions correspondraient à la classe à laquelle il a la plus grande probabilité vraie d'appartenir $P(j \mid x)$.

Formellement, un prédicteur est dit "order-correct" pour une entrée x si :

$ argmax_j Q(j|x) = argmax_j P(j|x) $

où $P(j \mid x)$ est la vraie probabilité que l'observation x appartienne à la classe j, et $Q(j \mid x)$ est la probabilité que x appartienne à la classe j prédite par le modèle $\varphi(x, L)$.

Un classificateur est **order-correct** si, pour **chaque** observation x, la classe qu'il prédit correspond à celle qui a la probabilité maximale $P(j \mid x)$ dans la distribution vraie.

4.2.2.2. Prédicteur agrégé en classification: le vote majoritaire

Dans le cas de la classification, le prédicteur agrégé est défini par le **vote majoritaire**. Cela signifie que si K classificateurs sont entraînés sur K échantillons distincts, la classe prédite pour x est celle qui reçoit **le plus de votes** de la part des modèles individuels.

Formellement, le classificateur agrégé $\varphi A(x)$ est défini par :

$\varphi A(x) = \text{argmax}_j \sum_L I(\phi(x, L) = j) = argmax_j Q(j \mid x)$

4.2.2.3. Performance globale: convergence vers un classificateur presque optimal

L. Breiman [5] montre que si chaque prédicteur individuel $\varphi(x, L)$ est order-correct pour une observation x, alors le prédicteur agrégé $\varphi A(x)$, obtenu par **vote majoritaire**, atteint la performance optimale pour cette observation, c'est-à-dire qu'il converge vers la classe ayant la probabilité maximale $P(j \mid x)$ pour l'observation x lorsque le nombre de prédicteurs individuels augmente. Le vote majoritaire permet ainsi de **réduire les erreurs aléatoires** des classificateurs individuels.

Le classificateur agrégé ϕA est optimal s'il prédit systématiquement la classe la plus probable pour l'observation x dans toutes les régions de l'espace.

Cependant, dans les régions de l'espace où les classificateurs individuels ne sont pas order-corrects (c'est-à-dire qu'ils se trompent majoritairement sur la classe d'appartenance), l'agrégation par vote majoritaire n'améliore pas les performances. Elles peuvent même se détériorer

par rapport aux modèles individuels si l'agrégation conduit à amplifier des erreurs systématiques (biais).

4.3. L'échantillage par bootstrap peut détériorer les performances théoriques du modèle agrégé

En pratique, au lieu d'utiliser tous les ensembles d'entraînement possibles L, le *bagging* repose sur un nombre limité d'échantillons bootstrap tirés avec remise à partir d'un même jeu de données initial, ce qui peut introduire des biais par rapport au prédicteur agrégé théorique.

Les échantillons bootstrap présentent les limites suivantes :

- Une **taille effective réduite par rapport au jeu de données initial**: Bien que chaque échantillon bootstrap présente le même nombre d'observations que le jeu de données initial, environ 1/3 des observations (uniques) du jeu initial sont absentes de chaque échantillon bootstrap (du fait du tirage avec remise). Cela peut limiter la capacité des modèles à capturer des relations complexes au sein des données (et aboutir à des modèles individuels sous-ajustés par rapport à ce qui serait attendu théoriquement), en particulier lorsque l'échantillon initial est de taille modeste.
- Une **dépendance entre échantillons** : Les échantillons bootstrap sont tirés dans le même jeu de données, ce qui génère une dépendance entre eux, qui réduit la diversité des modèles. Cela peut limiter l'efficacité de la réduction de variance dans le cas de la régression, voire acroître le biais dans le cas de la classification.
- Une **couverture incomplète de l'ensemble des échantillons possibles**: Les échantillons bootstrap ne couvrent pas l'ensemble des échantillons d'entraînement possibles, ce qui peut introduire un biais supplémentaire par rapport au prédicteur agrégé théorique.

4.4. Le *bagging* en pratique

4.4.1. Quand utiliser le *bagging* en pratique

Le *bagging* est particulièrement utile lorsque les modèles individuels présentent une variance élevée et sont instables. Dans de tels cas, l'agrégation des prédictions peut réduire significativement la variance globale, améliorant ainsi la performance du modèle agrégé. Les situations où le *bagging* est recommandé incluent typiquement:

- Les modèles instables : Les modèles tels que les arbres de décision non élagués, qui sont sensibles aux variations des données d'entraînement, bénéficient grandement du

bagging. L'agrégation atténue les fluctuations des prédictions dues aux différents échantillons.

- Les modèles avec biais faibles: En classification, si les modèles individuels sont order-corrects pour la majorité des observations, le *bagging* peut améliorer la précision en renforçant les prédictions correctes et en réduisant les erreurs aléatoires.

Inversement, le *bagging* peut être moins efficace ou même néfaste dans certaines situations :

- Les modèles stables avec variance faible : Si les modèles individuels sont déjà stables et présentent une faible variance (par exemple, la régression linéaire), le *bagging* n'apporte que peu d'amélioration, car la réduction de variance supplémentaire est minimale.
- La présence de biais élevée : Si les modèles individuels sont biaisés, entraînant des erreurs systématiques, le *bagging* peut amplifier ces erreurs plutôt que de les corriger. Dans de tels cas, il est préférable de s'attaquer d'abord au biais des modèles avant de considérer l'agrégation.
- Les échantillons de petite taille : Avec des ensembles de données limités, les échantillons bootstrap peuvent ne pas être suffisamment diversifiés ou représentatifs, ce qui réduit l'efficacité du *bagging* et peut augmenter le biais des modèles.

Ce qui qu'il faut retenir: le *bagging* peut améliorer substantiellement la performance des modèles d'apprentissage automatique lorsqu'il est appliqué dans des conditions appropriées. Il est essentiel d'évaluer la variance et le biais des modèles individuels, ainsi que la taille et la représentativité du jeu de données, pour déterminer si le *bagging* est une stratégie adaptée. Lorsqu'il est utilisé judicieusement, le *bagging* peut conduire à des modèles plus robustes et précis, exploitant efficacement la puissance de l'agrégation pour améliorer la performance des modèles individuels.

4.4.2. Comment utiliser le *bagging* en pratique

4.4.2.1. Combien de modèles agréger?

"Optimal performance is often found by *bagging* 50–500 trees. Data sets that have a few strong predictors typically require less trees; whereas data sets with lots of noise or multiple strong predictors may need more. Using too many trees will not lead to overfitting. However, it's important to realize that since multiple models are being run, the more iterations you perform the more computational and time requirements you will have. As these demands increase, performing k-fold CV can become computationally burdensome."

4.4.2.2. Evaluation du modèle: cross validation et échantillon Out-of-bag (OOB)

"A benefit to creating ensembles via *bagging*, which is based on resampling with replacement, is that it can provide its own internal estimate of predictive performance with the out-of-bag (OOB) sample (see Section 2.4.2). The OOB sample can be used to test predictive performance and the results usually compare well compared to k-fold CV assuming your data set is sufficiently large (say n≥1,000). Consequently, as your data sets become larger and your *bagging* iterations increase, it is common to use the OOB error estimate as a proxy for predictive performance."

4.5. Mise en pratique (exemple avec code)

Ou bien ne commencer les mises en pratique qu'avec les random forest ?

4.6. Interprétation

5. La forêt aléatoire

La forêt aléatoire (*random forests*) est une méthode ensembliste puissante, largement utilisée pour les tâches de classification et de régression. Elle combine la simplicité des arbres de décision et l'échantillonnage des observations et des variables avec la puissance de l'agrégation pour améliorer les performances prédictives et réduire le risque de surapprentissage (*overfitting*).

5.1. Principe de la forêt aléatoire

La forêt aléatoire est une extension du *bagging*, présenté dans la section Section 4. Elle introduit un niveau supplémentaire de randomisation dans la construction des arbres, puisqu'à chaque nouvelle division (*noeud*), le critère de séparation est choisi en considérant uniquement un sous-ensemble de variables **sélectionné aléatoirement**. Cette randomisation supplémentaire **réduit la corrélation** entre les arbres, ce qui permet de diminuer la variance des prédiction du modèle agrégé.

Les forêts aléatoires reposent sur quatre éléments essentiels:

- **Les arbres CART**: Les modèles élémentaires sont des arbres CART non élagués, c'est-à-dire autorisés à pousser jusqu'à l'atteinte d'un critère d'arrêt défini en amont.
- **L'échantillonnage *bootstrap***: Chaque arbre est construit à partir d'un échantillon aléatoire du jeu de données d'entraînement tiré avec remise (ou parfois sans remise).
- **La sélection aléatoire de variables** : Lors de la construction d'un arbre, à chaque nœud de celui-ci, un sous-ensemble aléatoire de variables est sélectionné. La meilleure division est ensuite choisie parmi ces caractéristiques aléatoires.
- **L'agrégation des prédictions** : Comme pour le *bagging*, les prédictions de tous les arbres sont combinées. On procède généralement à la moyenne (ou à la médiane) des prédictions dans le cas de la régression, et au vote majoritaire (ou à la moyenne des probabilités prédites pour chaque classe) dans le cas de la classification.

5.2. Comment construit-on une forêt aléatoire?

L'entraînement d'une forêt aléatoire est très similaire à celui du *bagging* et se résume comme suit:

- Le nombre d'arbres à construire est défini *a priori*.
- Pour chaque arbre, on effectue les étapes suivantes:
  - Générer un échantillon *bootstrap* de taille fixe à partir des données d'entraînement.
  - Construire récursivement un arbre de décision à partir de cet échantillon:

  - À chaque nœud de l'arbre, un sous-ensemble de *features* est sélectionné aléatoirement.
  - Déterminer quel couple (variable, valeur) définit la règle de décision qui divise la population du nœud en deux sous-groupes les plus homogènes possibles.
  - Créer les deux nœuds-enfants à partir de cette règle de décision.
  - Arrêter la croissance de l'arbre selon des critères d'arrêt fixés *a priori*.

Pour construire la prédiction de la forêt aléatoire une fois celle-ci entraînée, on agrège les arbres selon une méthode qui dépend du problème modélisé:

- Régression: la prédiction finale est la moyenne des prédictions de tous les arbres.
- Classification: chaque arbre vote pour une classe, et la classe majoritaire est retenue.

Les principaux hyper-paramètres des forêts aléatoires (détaillés dans la section Section 7) sont les suivants: le nombre d'arbres, la méthode et le taux d'échantillonnage, le nombre (ou la proportion) de variables considérées à chaque nœud, le critère de division des nœuds (ou mesure d'hétérogénéité), et les critères d'arrêt (notamment la profondeur de l'arbre, le nombre minimal d'observations dans une feuille terminale, et le nombre minimal d'observations qu'un nœud doit comprendre pour être divisé en deux).

5.3. Pourquoi les forêts aléatoires sont-elles performantes?

Les propriétés théoriques des forêts aléatoires permettent de comprendre pourquoi (et dans quelles situations) elles sont particulièrement robustes et performantes.

5.3.1. Réduction de la variance par agrégation

L'agrégation de plusieurs arbres permet de réduire la variance globale du modèle, ce qui améliore la stabilité des prédictions. Lorsque les estimateurs sont (faiblement) biaisés mais caractérisés par une variance élevée, l'agrégation permet d'obtenir un estimateur avec un biais similaire mais une variance réduite. La démonstration est identique à celle présentée dans la section Section 4.

5.3.2. Convergence et limite théorique au surapprentissage

Bien qu'elle s'avèrent très performantes en pratique, **il n'est pas prouvé à ce stade que les forêts aléatoires convergent vers une solution optimale** lorsque la taille de l'échantillon tend vers l'infini (G. Louppe [7]). Plusieurs travaux théoriques ont toutefois fourni des preuves de convergence pour des versions simplifiées de l'algorithme (par exemple, G. Biau [8]).

Par ailleurs, une propriété importante des forêts aléatoires démontrée par L. Breiman [6] est que leur erreur de généralisation, c'est-à-dire l'écart entre les prédictions du modèle et les

résultats attendus sur des données jamais vues (donc hors de l'échantillon d'entraînement), diminue à mesure que le nombre d'arbres augmente et converge vers une valeur constante. Autrement dit, **la forêt aléatoire ne souffre pas d'un surapprentissage croissant avec le nombre d'arbres**. La conséquence pratique de ce résultat est qu'inclure un (trop) grand nombre d'arbres dans le modèle n'en dégrade pas la qualité, ce qui contribue à la rendre particulièrement robuste. En revanche, une forêt aléatoire peut souffrir de surapprentissage si ses autres hyperparamètres sont mal choisis (des arbres trop profonds par exemple).

5.3.3. Facteurs influençant l'erreur de généralisation

L'erreur de généralisation des forêts aléatoires est influencée par deux facteurs principaux :

- **La puissance prédictrice des arbres individuels** : Les arbres doivent être suffisamment prédictifs pour contribuer positivement à l'ensemble, et idéalement sans biais.
- **La corrélation entre les arbres** : Moins les arbres sont corrélés, plus la variance de l'ensemble est réduite, car leurs erreurs tendront à se compenser. Inversement, des arbres fortement corrélés auront tendance à faire des erreurs similaires, donc agréger un grand nombre d'arbres n'apportera pas grand chose.

On peut mettre en évidence ces deux facteurs dans le cas d'une forêt aléatoire utilisée pour une tâche de régression (où l'objectif est de minimiser l'erreur quadratique moyenne). Dans ce cas, la variance de la prédiction du modèle peut être décomposée de la façon suivante:

$$\text{Var}\left(\hat{f}(x)\right) = \rho(x)\sigma(x)^2 + \frac{1-\rho(x)}{M}\sigma(x)^2 \tag{6}$$

où $\rho(x)$ est le coefficient de corrélation moyen entre les arbres individuels, $\sigma(x)^2$ est la variance d'un arbre individuel, M est le nombre d'arbres dans la forêt. Cette décomposition fait apparaître l'influence de la corrélation entre les arbres sur les performance de la forêt aléatoire:

- **Si $\rho(x)$ est proche de 1** (forte corrélation entre les arbres) : la première composante $\rho\sigma^2$ domine et la réduction de variance est moindre lorsque le nombre d'arbres augmente.
- **Si $\rho(x)$ est proche de 0** (faible corrélation entre les arbres) : la seconde composante $\frac{1-\rho}{M}\sigma^2$ et la variance est davantage réduite avec l'augmentation du nombre d'arbres M.

L'objectif de l'entraînement des forêts aléatoires est donc de minimiser la corrélation entre les arbres tout en maximisant leur capacité à prédire correctement, ce qui permet de réduire la variance globale sans augmenter excessivement le biais. La sélection aléatoires des caractéristiques (*features*) à chaque nœud joue un rôle majeur dans cet arbitrage entre puissance prédictive des arbres et corrélation entre arbres.

5.4. Evaluation des performances par l'erreur *Out-of-Bag* (OOB)

La forêt aléatoire présente une particularité intéressante et très utile en pratique: **il est possible d'évaluer les performances d'une forêt aléatoire directement à partir des données d'entraînement**, grâce à l'estimation de l'erreur *Out-of-Bag* (OOB). Cette technique repose sur le fait que chaque arbre est construit à partir d'un échantillon *bootstrap*, c'est-à-dire un échantillon tiré avec remise. Cela implique qu'une part conséquente des observations ne sont pas utilisées pour entraîner un arbre donné. Ces observations laissées de côté forment un **échantillon dit *out-of-bag***, que l'on peut utiliser pour évaluer la performance de chaque arbre. On peut donc construire pour chaque observation du jeu d'entraînement une prédiction qui agrège uniquement les prédictions des arbres pour lesquels cette observation est *out-of-bag*; cette prédiction n'est pas affectée par le surapprentissage (puisque cette observation n'a jamais été utilisée pour entraîner ces arbres). De cette façon, il est possible d'évaluer correctement la performance de la forêt aléatoire en comparant ces prédictions avec la variable-cible à l'aide d'une métrique bien choisie.

La procédure d'estimation de l'erreur OOB se déroule comme ceci:

1. **Entraînement de la forêt aléatoire**: la forêt aléatoire est entraînée sur les données d'entraînement selon la procédure détaillée ci-dessus.
2. **Prédiction *out-of-bag*** : Pour chaque observation (x_i, y_i) des données d'entraînement, on calcule la prédiction de tous les arbres pour lesquels elle fait partie de l'échantillon *out-of-bag*.
3. **Agrégation des prédictions** : La prédiction finale est obtenue en agrégeant les prédictions selon la procédure standard détaillée ci-dessus (moyenne pour la régression, vote majoritaire pour la classification).
4. **Calcul de l'erreur OOB** : L'erreur OOB est ensuite calculée en comparant les prédictions avec la variable-cible y sur toutes les observations, à l'aide d'une métrique (précision, rappel, AUC, erreur quadratique moyenne, score de Brier…).

L'utilisation de l'erreur OOB présente de multiples avantages:

- **Approximation de l'erreur de généralisation**: L'erreur OOB est en général considérée comme une bonne approximation de l'erreur de généralisation, comparable à celle obtenue par une validation croisée.
- **Pas besoin de jeu de validation séparé** : L'un des principaux avantages de l'erreur OOB est qu'elle ne nécessite pas de réserver une partie des données pour la validation. Cela est particulièrement utile lorsque la taille du jeu de données est limitée, car toutes les données peuvent être utilisées pour l'entraînement tout en ayant une estimation

fiable de la performance. Ceci dit, il est malgré tout recommandé de conserver un ensemble de test si la taille des données le permet, car il arrive que l'erreur OOB sous

- **Gain de temps** : Contrairement à la validation croisée qui requiert de réentraîner plusieurs fois le modèle pour un jeu donné d'hyperparamètres, l'erreur OOB ne nécessite qu'un seul entraînement du modèle. Cela induit un gain de temps appréciable lors de l'optimisation des hyperparamètres.

5.5. Interprétation et importance des variables

Les forêts aléatoires sont des modèles d'apprentissage performants, mais leur complexité interne les rend difficiles à interpréter, ce qui leur vaut souvent le qualificatif de "boîtes noires". Comprendre l'influence des variables explicatives sur les prédictions est crucial pour interpréter les résutlats et être en mesure d'extraire des connaissances.

L'objectif des **méthodes d'interprétabilité** (ou d'importance des variables) est d'identifier les variables les plus influentes sur la variable cible, de comprendre les mécanismes prédictifs sous-jacents, et potentiellement d'extraire des règles de décision simples et transparentes. Plusieurs méthodes d'importance des variables existent, mais il est important de comprendre leurs forces et faiblesses.

5.5.1. Mesures d'importance classiques (et leurs biais)

- **Réduction moyenne de l'impureté** (*Mean Decrease in Impurity - MDI*) : Cette méthode quantifie l'importance d'une variable par la somme des réductions d'impureté qu'elle induit dans tous les arbres de la forêt. Plus spécifiquement, pour chaque variable, on s'intéresse à la moyenne des réductions d'impureté qu'elle a engendrées dans tous les nœuds de tous les arbres où elle est impliquée. Les variables présentant la réduction moyenne d'impureté la plus élevée sont considérées comme les prédicteurs les plus importants.

La MDI présente des biais importants. Elle est notamment sensible aux variables catégorielles avec de nombreuses modalités, qui peuvent apparaître artificiellement importantes (même si leur influence réelle est faible), ainsi qu'aux variables avec une échelle de valeurs plus étendues, qui obtiennent des scores plus élevés, indépendamment de leur importance réelle. Elle est également fortement biaisée en présence de variables explicatives corrélées, ce qui conduit à surestimer l'importance de variables redondantes. Les interactions entre variables ne sont pas non plus prises en compte de manière adéquate.

- **Importance par permutation** (*Mean Decrease Accuracy - MDA*) : Cette méthode évalue l'importance d'une variable en mesurant la diminution de précision du modèle après permutation aléatoire de ses valeurs. Plus spécifiquement, pour chaque variable,

les performances du modèle sont comparées avant et après la permutation de ses valeurs. La différence moyenne de performance correspond à la MDA. L'idée est que si l'on permute aléatoirement les valeurs d'une variable (cassant ainsi sa relation avec la cible), une variable importante entraînera une hausse significative de l'erreur de généralisation.

Comme la MDI, la MDA présente des biais lorsque les variables sont corrélées. En particulier, la MDA peut surévaluer l'importance de variables qui sont corrélées à d'autres variables importantes, même si elles n'ont pas d'influence directe sur la cible (C. Bénard, S. Da Veiga, and E. Scornet [9]).

Plusieurs stratégies peuvent aider à réduire les biais d'interprétation :

- Prétraitement des variables: Standardisation des variables, regroupement des modalités rares des variables catégorielles, réduction de la cardinalité des variables catégorielles.
- Analyse des corrélations: Identification et gestion des variables fortement corrélées, qui peuvent fausser les mesures d'importance.
- Choix de méthodes robustes: Privilégier les méthodes moins sensibles aux biais, comme les CIF ou la Sobol-MDA, et, le cas échéant, SHAFF pour les valeurs de Shapley. Ces méthodes sont présentées dans la sectio suivante.

5.5.2. Méthodes d'importance avancées

Pour pallier les limites des méthodes traditionnelles, des approches plus sophistiquées ont été développées.

- **Valeurs de Shapley**: Les valeurs de Shapley permettent de quantifier la contribution de chaque variable explicative à la variance expliquée de la variable cible, en tenant compte des interactions entre les variables. Elles attribuent à chaque variable une contribution marginale moyenne à la performance du modèle, en considérant toutes les combinaisons possibles de sous-ensembles de variables. Cependant, l'estimation des valeurs de Shapley est computationnellement coûteuse (complexité exponentielle avec le nombre de variables). Des méthodes approximatives existent, mais peuvent introduire des biais. L'algorithme SHAFF (C. Bénard, G. Biau, S. Da Veiga, and E. Scornet [10]) propose une solution rapide et précise à ce problème, en tirant parti des propriétés des forêts aléatoires.
- **Conditional Inference Forests** (CIF): Les CIF (C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn [11]), implémentées dans le package party de R (cforest), corrigent certains biais de la MDI en utilisant des tests statistiques conditionnels pour sélectionner les variables et les seuils de coupure dans les arbres. Elles sont particulièrement robustes face aux variables hétérogènes et aux corrélations entre variables. Couplées

à un échantillonnage sans remise, les CIF fournissent des mesures d'importance plus fiables.

- **Sobol-MDA**: La Sobol-MDA combine l'idée de la MDA avec une approche basée sur les indices de Sobol, permettant de gérer efficacement les variables dépendantes. Au lieu de permuter les valeurs, elle projette la partition des arbres sur le sous-espace excluant la variable dont on souhaite mesurer l'importance, simulant ainsi son absence. Elle est plus efficace en calcul que les méthodes MDA classiques tout en fournissant une mesure d'importance cohérente, convergeant vers l'indice de Sobol total (la mesure appropriée pour identifier les covariables les plus influentes, même avec des dépendances) (C. Bénard, S. Da Veiga, and E. Scornet [9]).

5.6. Le *boosting*

5.6.1. Introduction

Le fondement théorique du *boosting* est un article de de 1990 (R. Shapire [12]) qui a démontré théoriquement que, sous certaines conditions, il est possible de transformer un modèle prédictif peu performant en un modèle prédictif très performant. Plus précisément, un modèle ayant un pouvoir prédictif arbitrairement élevé (appelé *strong learner*) peut être construit en combinant des modèles simples dont les prédictions ne sont que légèrement meilleures que le hasard (appelé *weak learners*). Le *boosting* est donc une méthode qui combine une approche ensembliste reposant sur un grand nombre de modèles simples avec un entraînement séquentiel: chaque modèle simple (souvent des arbres de décision peu profonds) tâche d'améliorer la prédiction globale en corrigeant les erreurs des prédictions précédentes à chaque étape. Bien qu'une approche de *boosting* puisse en théorie mobiliser différentes classes de *weak learners*, en pratique les *weak learners* utilisés par les algorithmes de *boosting* sont presque toujours des arbres de décision.

S'il existe plusieurs variantes, les algorithmes de *boosting* suivent la même logique :

- Un premier modèle simple et peu performant est entraîné sur les données.
- Un deuxième modèle est entraîné de façon à corriger les erreurs du premier modèle (par exemple en pondérant davantage les observations mal prédites);
- Ce processus est répété en ajoutant des modèles simples, chaque modèle corrigeant les erreurs commises par l'ensemble des modèles précédents;
- Tous ces modèles sont finalement combinés (souvent par une somme pondérée) pour obtenir un modèle complexe et performant.

En termes plus techniques, les algorithmes de *boosting* partagent trois caractéristiques communes:

- Ils visent à **trouver une approximation** $\hat{F}$“)`{=typst} d'une fonction inconnue F^* : $\mathbf{x} \mapsto y$ à partir d'un ensemble d'entraînement $(y_i, \mathbf{x_i})_{i=1,\dots,n}$;
- Ils supposent que la fonction F^* peut être approchée par une **somme pondérée de modèles simples** f de paramètres θ:

$$F(\mathbf{x}) = \sum_{m=1}^{M} \beta_m f(\mathbf{x}, \boldsymbol{\theta}_m) \tag{7}$$

- ils reposent sur une **modélisation additive par étapes**, qui décompose l'entraînement de ce modèle complexe en une **séquence d'entraînements de petits modèles**. Chaque étape de l'entraînement cherche le modèle simple f qui améliore la puissance prédictive du modèle complet, sans modifier les modèles précédents, puis l'ajoute de façon incrémentale à ces derniers:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \hat{\beta}_m f\left(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_\mathbf{m}\right) \tag{8}$$

METTRE ICI UNE FIGURE EN UNE DIMENSION, avec des points et des modèles en escalier qui s'affinent.

5.6.2. Les premières approches du *boosting*

5.6.2.1. Le *boosting* par repondération: Adaboost

Dans les années 1990, de nombreux travaux ont tâché de proposer des mise en application du *boosting* (L. Breiman [13], A. J. Grove and D. Schuurmans [14]) et ont comparé les mérites des différentes approches. Deux approches ressortent particulièrement de cette littérature: Adaboost (Adaptive Boosting, Y. Freund and R. E. Schapire [15]) et la *Gradient Boosting Machine* (J. H. Friedman [16]). Ces deux approches reposent sur des principes très différents.

Le principe d'Adaboost consiste à pondérer les erreurs commises à chaque itération en donnant plus d'importance aux observations mal prédites, de façon à obliger les modèles simples à se concentrer sur les observations les plus difficiles à prédire. Voici une esquisse du fonctionnement d'AdaBoost:

- Un premier modèle simple est entraîné sur un jeu d'entraînement dans lequel toutes les observations ont le même poids.
- A l'issue de cette première itération, les observations mal prédites reçoivent une pondération plus élevé que les observations bien prédites, et un deuxième modèle est entraîné sur ce jeu d'entraînement pondéré.

- Ce deuxième modèle est ajouté au premier, puis on repondère à nouveau les observations en fonction de la qualité de prédiction de ce nouveau modèle.
- Cette procédure est répétée en ajoutant de nouveaux modèles et en ajustant les pondérations.

L'algorithme Adaboost a été au coeur de la littérature sur le *boosting* à la fin des années 1990 et dans les années 2000, en raison de ses performances sur les problèmes de classification binaire. Il a toutefois été progressivement remplacé par les algorithmes de *gradient boosting* inventé quelques années plus tard.

5.6.2.2. L'invention du *boosting boosting* : la *Gradient Boosting Machine*

La *Gradient Boosting Machine* (GBM) propose une approche assez différente: elle introduit le *gradient boosting* en reformulant le *boosting* sous la forme d'un problème de descente de gradient. Voici une esquisse du fonctionnement de la *Gradient Boosting Machine*:

- Un premier modèle simple est entraîné sur un jeu d'entraînement, de façon à minimiser une fonction de perte qui mesure l'écart entre la variable à prédire et la prédiction du modèle.
- A l'issue de cette première itération, on calcule la dérivée partielle (*gradient*) de la fonction de perte par rapport à la prédiction en chaque point de l'ensemble d'entraînement. Ce gradient indique dans quelle direction et dans quelle ampleur la prédiction devrait être modifiée afin de réduire la perte.
- A la deuxième itération, on ajoute un deuxième modèle qui va tâcher d'améliorer le modèle complet en prédisant le mieux possible l'opposé de ce gradient.
- Ce deuxième modèle est ajouté au premier, puis on recalcule la dérivée partielle de la fonction de perte par rapport à la prédiction de ce nouveau modèle.
- Cette procédure est répétée en ajoutant de nouveaux modèles et en recalculant le gradient à chaque étape.

L'approche de *gradient boosting* proposée par J. H. Friedman [16] présente deux grands avantages. D'une part, elle peut être utilisée avec n'importe quelle fonction de perte différentiable, ce qui permet d'appliquer le gradient boosting à de multiples problèmes (régression, classification binaire ou multiclasse, *learning-to-rank*...). D'autre part, elle offre souvent des performances comparables ou supérieures aux autres approches de *boosting*. Le *gradient boosting* d'arbres de décision (*Gradient boosted Decision Trees* - GBDT) est donc devenue l'approche de référence en matière de *boosting*. En particulier, les implémentations modernes du *gradient boosting* comme XGBoost, LightGBM, et CatBoost sont des extensions et améliorations de la *Gradient Boosting Machine*.

5.6.3. La mécanique du *gradient boosting*

Depuis la publication de J. H. Friedman [16], la méthode de *gradient boosting* a connu de multiples développements et raffinements, parmi lesquels XGBoost (T. Chen and C. Guestrin [17]), LightGBM (G. Ke *et al.* [18]) et CatBoost (L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin [19]). S'il existe quelques différences entre ces implémentations, elles partagent néanmoins la même mécanique d'ensemble, que la section qui suit va présenter en détail en s'appuyant sur l'implémentation proposée par XBGoost.[17, .]

Choses importantes à mettre en avant:

- Le boosting est fondamentalement différent des forêts aléatoires. See ESL, chapitre 10.
- Toute la mécanique est indépendante de la fonction de perte choisie. En particulier, elle est applicable indifféremment à des problèmes de classification et de régression.
- Les poids sont calculés par une formule explicite, ce qui rend les calculs extrêmement rapides.
- Comment on interprète le gradient et la hessienne: cas avec une fonction de perte quadratique.
- Le boosting est fait pour overfitter; contrairement aux RF, il n'y a pas de limite à l'overfitting. Donc lutter contre le surapprentissage est un élément particulièrement important de l'usage des algorithmes de boosting.

5.6.3.1. Le modèle à entraîner

On veut entraîner un modèle comprenant K arbres de régression ou de classification:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i) \tag{9}$$

Chaque arbre f est défini par trois paramètres:

- sa structure qui est une fonction $q : \mathbb{R}^m \to \{1, ..., T\}$ qui à un vecteur d'inputs $\mathbf{x}$ de dimension m associe une feuille terminale de l'arbre);
- son nombre de feuilles terminales T;
- les valeurs figurant sur ses feuilles terminales $\mathbf{w} \in \mathbb{R}^T$ (appelées poids ou *weights*).

Le modèle est entraîné avec une **fonction-objectif** constituée d'une **fonction de perte** l et d'une **fonction de régularisation** Ω. La fonction de perte mesure la distance entre la prédiction $hat(y)$ et la vraie valeur y et présente généralement les propriétés suivantes: elle est convexe et dérivable deux fois, et atteint son minimum lorsque $hat(y) = y$. La fonction de régularisation pénalise la complexité du modèle. Dans le cas présent, elle pénalise les arbres avec un grand nombre de feuilles (T élevé) et les arbres avec des poids élevés (w_t élevés en valeur absolue).

$$\mathcal{L}(\phi) = \underbrace{\sum_i l(\hat{y}_i, y_i)}_{\substack{\text{Perte sur les} \\ \text{observations}}} + \underbrace{\sum_k \Omega(f_k)}_{\substack{\text{Fonction de} \\ \text{régularisation}}} \text{ avec } \Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{t=1}^{T} w_j^2 \tag{10}$$

5.6.3.2. Isoler le k-ième arbre

La fonction-objectif introduite précédemment est très complexe et ne peut être utilisée directement pour entraîner le modèle, car il faudrait entraîner tous les arbres en même temps. On va donc reformuler donc cette fonction objectif de façon à isoler le k-ième arbre, qui pourra ensuite être entraîné seul, une fois que les $k-1$ arbres précédents auront été entraînés. Pour cela, on note $\hat{y}_i^{(k)}$ la prédiction à l'issue de l'étape k: $\hat{y}_i^{(k)} = \sum_{j=1}^{t} f_j(\mathbf{x}_i)$, et on \$1 la fonction-objectif $\mathcal{L}^{(k)}$ au moment de l'entraînement du k-ième arbre:

$$\begin{aligned} \mathcal{L}^{(t)} &= \sum_{i=1}^{n} l\Big(y_i, \hat{y}_i^{(t)}\Big) + \sum_{k=1}^{t} \Omega(f_k) \\ &= \sum_{i=1}^{n} l\Big(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\Big) + \Omega(f_t) + constant \end{aligned} \tag{11}$$

5.6.3.3. Faire apparaître le gradient

Une fois isolé le k-ième arbre, on fait un développement limité d'ordre 2 de $l\Big(y_i, \hat{y}_i^{(k-1)} + f_k(\mathbf{x}_i)\Big)$ au voisinage de $\hat{y}_i^{(k-1)}$, en considérant que la prédiction du k-ième arbre $f_k(\mathbf{x}_i)$ est

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[l\Big(y_i, \hat{y}_i^{(t-1)}\Big) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \tag{12}$$

avec

$$g_i = \frac{\partial l\Big(y_i, \hat{y}_i^{(t-1)}\Big)}{\partial \hat{y}_i^{(t-1)}} \text{ et } h_i = \frac{\partial^2 l\Big(y_i, \hat{y}_i^{(t-1)}\Big)}{\partial \hat{y}_i^{((t-1))^2}} \tag{13}$$

Les termes g_i et h_i désignent respectivement la dérivée première (le gradient) et la dérivée seconde (la hessienne) de la fonction de perte par rapport à la variable prédite. Il est important de noter que les termes (A) et (B) sont constants car les $k-1$ arbres précédents ont déjà été entraînés et ne sont pas modifiés par l'entraînement du k-ième arbre. On peut donc retirer ces termes pour obtenir la fonction-objectif simplifiée qui sera utilisée pour l'entraînement du k-ième arbre.

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_i^2 \tag{14}$$

Cette expression est importante car elle montre qu'on est passé d'un problème complexe où il fallait entraîner un grand nombre d'arbres simultanément (équation 10) à un problème beaucoup plus simple dans lequel il n'y a qu'un seul arbre à entraîner.

5.6.3.4. Calculer les poids optimaux

A partir de l'expression précédente, il est possible de faire apparaître les poids w_j du k-ième arbre. Pour une structure d'arbre donnée ($q : \mathbb{R}^m \to \{1, ..., T\}$), on définit $I_j = \{i \mid q(\mathbf{x}_i) = j\}$ l'ensemble des observations situées sur la feuille j puis on réorganise $\tilde{\mathcal{L}}^{(k)}$:

$$\begin{aligned}
\tilde{\mathcal{L}}^{(t)} &= \sum_{j=1}^{T} \sum_{i \in I_j} \Big[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \Big] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_i^2 \\
&= \sum_{j=1}^{T} \sum_{i \in I_j} \Big[g_i w_j + \frac{1}{2} h_i w_j^2 \Big] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_i^2 \qquad (15) \\
&= \sum_{j=1}^{T} \Big[w_j \sum_{i \in I_j} g_i + \frac{1}{2} w_j^2 \sum_{i \in I_j} h_i + \lambda \Big] + \gamma T
\end{aligned}$$

Dans la dernière expression, on voit que la fonction de perte simplifiée se reformule comme une combinaison quadratique des poids w_j, dans laquelle les dérivées première et seconde de la fonction de perte interviennent sous forme de pondérations. Tout l'enjeu de l'entraînement devient donc de trouver les poids optimaux w_j qui minimiseront cette fonction de perte, compte tenu de ces opérations.

Il se trouve que le calculs de ces poids optimaux est très simple: pour une structure d'arbre donnée ($q : \mathbb{R}^m \to \{1, ..., T\}$), le poids optimal w_j^{} de la feuille j est donné par l'équation:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \qquad (16)$$

Par conséquent, la valeur optimale de la fonction objectif pour l'arbre q est égale à

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \qquad (17)$$

Cette équation est utile car elle permet de comparer simplement la qualité de deux arbres, et de déterminer lequel est le meilleur.

5.6.3.5. Construire le k-ième arbre

Dans la mesure où elle permet de comparer des arbres, on pourrait penser que l'équation 17 est suffisante pour choisir directement le k-ième arbre: il suffirait d'énumérer les arbres possibles, de calculer la qualité de chacun d'entre eux, et de retenir le meilleur. Bien que cette approche soit possible théoriquement, elle est inemployable en pratique car le nombre d'arbres possibles est

extrêmement élevé. Par conséquent, le k-ième arbre n'est pas défini en une fois, mais construit de façon gloutonne:

REFERENCE A LA PARTIE CART/RF?

- on commence par le noeud racine et on cherche le *split* qui réduit au maximum la perte en séparant les données d'entraînement entre les deux noeuds-enfants.
- pour chaque noeud enfant, on cherche le *split* qui réduit au maximum la perte en séparant en deux la population de chacun de ces noeuds.
- Cette procédure recommence jusqu'à que l'arbre ait atteint sa taille maximale (définie par une combinaison d'hyperparamètres d$1s dans la partie **référence à ajouter**).

5.6.3.6. Choisir les *splits*

Traduire split par critère de partition?

Reste à comprendre comment le critère de partition optimal est choisi à chaque étape de la construction de l'arbre. Imaginons qu'on envisage de décomposer la feuille I en deux nouvelles feuilles I_L et I_R (avec $I = I_L \cup I_R$), selon une condition logique reposant sur une variable et une valeur de cette variable (exemple: $x_6 > 11$). Par application de l'équation 17, le gain potentiel induit par ce critère de partition est égal à:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{18}$$

Cette dernière équation est au coeur de la mécanique du *gradient boosting* car elle permet de comparer les critères de partition possibles. Plus précisément, l'algorithme de détermination des critère de partition (*split finding algorithm*) consiste en une double boucle sur les variables et les valeurs prises par ces variables, qui énumère un grand nombre de critères de partition et mesure le gain associé à chacun d'entre eux avec l'équation 18. Le critère de partition retenu est simplement celui dont le gain est le plus élevé.

L'algorithme qui détermine les critère de partition est un enjeu de performance essentiel dans le *gradient boosting*. En effet, utiliser l'algorithme le plus simple (énumérer tous les critères de partition possibles, en balayant toutes les valeurs de toutes les variables) s'avère très coûteux dès lors que les données contiennent soit un grand nombre de variables, soit des variables continues prenant un grand nombre de valeurs. C'est pourquoi les algorithmes de détermination des critère de partition ont fait l'objet de multiples améliorations et optimisations visant à réduire leur coût computationnel sans dégrader la qualité des critères de partition.

5.6.3.7. La suite

5.6.3.7.1. Les moyens de lutter contre l'*overfitting*:

- le *shrinkage*;

- le subsampling des lignes et des colonnes;
- les différentes pénalisations.

5.6.3.7.2. Les hyperparamètres

Hyperparamètre	Description	Valeur par défaut
`booster`	Le type de *weak learner* utilisé	`'gbtree'`
`learning_rate`	Le taux d'apprentissage	0.3
`max_depth`	La profondeur maximale des arbres	6
`max_leaves`	Le nombre maximal de feuilles des arbres	0
`min_child_weight`	Le poids minimal qu'une feuille doit contenir	1
`n_estimators`	Le nombre d'arbres	100
`lambda` ou `reg_lambda`	La pénalisation L2	1
`alpha` ou `reg_alpha`	La pénalisation L1	0
`gamma`	Le gain minimal nécessaire pour ajouter un noeud supplémentaire	0
`tree_method`	La méthode utilisée pour rechercher les splits	`'hist'`
`max_bin`	Le nombre utilisés pour discrétiser les variables continues	0
`subsample`	Le taux d'échantillonnage des données d'entraînement	1
`sampling_method`	La méthode utilisée pour échantillonner les données d'entraînement	`'uniform'`
`colsample_bytree` `colsample_bylevel` `colsample_bynode`	Taux d'échantillonnage des colonnes par arbre, par niveau et par noeud	1, 1 et 1
`scale_pos_weight`	Le poids des observations de la classe positive (classification uniquement)	1
`sample_weight`	La pondération des données d'entraînement	1
`enable_categorical`	Activer le support des variables catégorielles	`False`
`max_cat_to_onehot`	Nombre de modalités en-deça duquel XGBoost utilise le *one-hot-encoding*	A COMPLETER

Hyperparamètre	Description	Valeur par défaut
`max_cat_threshold`	Nombre maximal de catégories considérées dans le partitionnement optimal des variables catégorielles	A COMPLETER

5.6.3.8. La préparation des données

- les variables catégorielles:
  - ordonnées: passer en integer;
  - non-ordonnées: OHE ou approche de Fisher.
- les variables continues:
  - inutile de faire des transformations monotones.
  - Utile d'ajouter des transformations non monotones.

5.6.3.9. Les fonctions de perte

5.6.4. Liste des hyperparamètres d'une RF

Source: P. Probst, M. N. Wright, and A.-L. Boulesteix [20]

- structure of each individual tree:
  - dudu
  - dudu
  - dudu
- structure and size of the forest:
- The level of randomness (je dirais plutôt :)

6. Une bien belle section

Principe: Conseils + mise en oeuvre pratique.

6.1. Préparation des données

6.1.1. Préparation des données (Feature Engineering)

- **Valeurs manquantes** : Les forêts aléatoires peuvent gérer les données manquantes, mais une imputation préalable peut améliorer les performances.
- **Variables catégorielles** : Utiliser un encodage adapté (one-hot encoding, ordinal encoding) en fonction de la nature des données. Convertir en variables continues dès que c'est possible.
- **Échelle des Variables** : Pas nécessaire de normaliser, les arbres sont invariants aux transformations monotones.

6.1.2. Process: utiliser les pipelines scikit, pour expliciter la structure du modèle complet et réduire les risques d'erreur

6.1.3. Train-test

Pas indispensable pour RF, mais souhaitable. Indispensable pour GB.

6.2. Evaluation des performances du modèle et optimisation des hyper-paramètres

6.2.1. Estimation de l'erreur par validation croisée

La validation croisée est une méthode d'évaluation couramment utilisée en apprentissage automatique pour estimer la capacité d'un modèle à généraliser les prédictions à de nouvelles données. Bien que l'évaluation par l'erreur *Out-of-Bag* (OOB) soit généralement suffisante pour les forêts aléatoires, la validation croisée permet d'obtenir une évaluation plus robuste, car moins sensible à l'échantillon d'entraînement, notamment sur des jeux de données de petite taille.

Concrètement, le jeu de donné est divisé en k sous-ensembles, un modèle est entraîné sur $k-1$ sous-ensembles et testé sur le sous-ensemble restant. L'opération est répétée k fois de manière à ce que chaque observation apparaisse au moins une fois dans l'échantillon test. L'erreur est ensuite moyennée sur l'ensemble des échantillons test.

Procédure de validation croisée:

La validation croisée la plus courante est la validation croisée en k sous-échantillons (*k-fold cross-validation*):

- **Division des données** : Le jeu de données est divisé en k sous-échantillons égaux, appelés folds. Typiquement, k est choisi entre 5 et 10, mais il peut être ajusté en fonction de la taille des données.
- **Entraînement et test** : Le modèle est entraîné sur k - 1 sous-échantillons et testé sur le sous-échantillon restant. Cette opération est répétée k fois, chaque sous-échantillon jouant à tour de rôle le rôle de jeu de test.
- **Calcul de la performance** : Les k performances obtenues (par exemple, l'erreur quadratique moyenne pour une régression, ou l'accuracy (*exactitude*) pour une classification) sont moyennées pour obtenir une estimation finale de la performance du modèle.

Avantages de la validation croisée:

- **Utilisation optimale des données** : En particulier lorsque les données sont limitées, la validation croisée maximise l'utilisation de l'ensemble des données en permettant à chaque échantillon de contribuer à la fois à l'entraînement et au test.
- **Réduction de la variance** : En utilisant plusieurs divisions des données, on obtient une estimation de la performance moins sensible aux particularités d'une seule division.

Bien que plus coûteuse en termes de calcul, la validation croisée est souvent préférée lorsque les données sont limitées ou lorsque l'on souhaite évaluer différents modèles ou hyperparamètres avec précision.

Leave-One-Out Cross-Validation (LOOCV) : Il s'agit d'un cas particulier où le nombre de sous-échantillons est égal à la taille du jeu de données. En d'autres termes, chaque échantillon est utilisé une fois comme jeu de test, et tous les autres échantillons pour l'entraînement. LOOCV fournit une estimation très précise de la performance, mais est très coûteuse en temps de calcul, surtout pour de grands jeux de données.

6.2.2. Choix des hyper-paramètres du modèle

L'estimation Out-of-Bag (OOB) et la validation croisée sont deux méthodes clés pour optimiser les hyper-paramètres d'une forêt aléatoire. Les deux approches permettent de comparer les performances obtenues pour différentes combinaisons d'hyper-paramètres et de sélectionner celles qui maximisent les performances prédictives, l'OOB étant souvent plus rapide et moins coûteuse, tandis que la validation croisée est plus fiable dans des situations où le surapprentissage est un risque important (P. Probst, M. N. Wright, and A.-L. Boulesteix [20]).

Il convient de définir une stratégie d'optimisation des hyperparamètres pour ne pas perdre de temps à tester trop de jeux d'hyperparamètres. Plusieurs stratégies existent pour y parvenir, les principales sont exposées dans la section Section 7. Les implémentations des forêts

aléatoires disponibles en R et en Python permettent d'optimiser aisément les principaux hyper-paramètres des forêts aléatoires.

6.2.2.1. Méthodes de recherche exhaustives

- **Recherche sur grille** (Grid Search): Cette approche simple explore toutes les combinaisons possibles d'hyperparamètres définis sur une grille. Les paramètres continus doivent être discrétisés au préalable. La méthode est exhaustive mais coûteuse en calcul, surtout pour un grand nombre d'hyperparamètres.
- **Recherche aléatoire** (Random Search): Plus efficace que la recherche sur grille, cette méthode échantillonne aléatoirement les valeurs des hyperparamètres dans un espace défini. Bergstra et Bengio (2012) ont démontré sa supériorité pour les réseaux neuronaux, et elle est également pertinente pour les forêts aléatoires. La distribution d'échantillonnage est souvent uniforme.

6.2.2.2. Optimisation séquentielle/itérative basée sur un modèle (SMBO)

La méthode SMBO (Sequential model-based optimization) est une approche plus efficace que les précédentes car elle s'appuie sur les résultats des évaluations déjà effectuées pour guider la recherche des prochains hyper-paramètres à tester (P. Probst, M. N. Wright, and A.-L. Boulesteix [20]).

Voici les étapes clés de cette méthode:

- Définition du problème: On spécifie une mesure d'évaluation (ex: AUC pour la classification, MSE pour la régression), une stratégie d'évaluation (ex: validation croisée k-fold), et l'espace des hyperparamètres à explorer.
- Initialisation: échantillonner aléatoirement des points dans l'espace des hyper-paramètres et évaluer leurs performances.
- Boucle itérative :
  - Construction d'un modèle de substitution (surrogate model): un modèle de régression (ex: krigeage ou une forêt aléatoire) est ajusté aux données déjà observées. Ce modèle prédit la performance en fonction des hyperparamètres.
  - Sélection d'un nouvel hyperparamètre: un critère basé sur le modèle de substitution sélectionne le prochain ensemble d'hyperparamètres à évaluer. Ce critère vise à explorer des régions prometteuses de l'espace des hyperparamètres qui n'ont pas encore été suffisamment explorées.
  - Évaluer les points proposés et les ajouter à l'ensemble déjà exploré: la performance du nouvel ensemble d'hyperparamètres est évaluée et ajoutée à l'ensem-

ble des données d'apprentissage du modèle de substitution afin d'orienter les recherches vers de nouveaux hyper-paramètres prometteurs.

7. Guide d'usage des forêts aléatoires

Ce guide d'entraînement des forêts aléatoires rassemble et synthétise des recommandations sur l'entraînement des forêts aléatoires disponibles dans la littérature, en particulier dans P. Probst, M. N. Wright, and A.-L. Boulesteix [20]. Ce guide comporte un certain nombre de choix méthodologiques forts, comme les implémentations recommandées ou la procédure d'entraînement proposée, et d'autres choix pertinents sont évidemment possibles. C'est pourquoi les recommandations de ce guide doivent être considérées comme un point de départ raisonnable, pas comme un ensemble de règles devant être respectées à tout prix.

7.1. Quelles implémentations utiliser?

Il existe de multiples implémentations des forêts aléatoires. Le présent document présente et recommande l'usage de deux implémentations de référence: le *package* `R ranger` et le *package* `Python scikit-learn` pour leur rigueur, leur efficacité et leur simplicité d'utilisation. Il est à noter qu'il est possible d'entraîner des forêts aléatoires avec les algorithmes `XGBoost` et `LightGBM`, mais il s'agit d'un usage avancé qui n'est pas recommandé en première approche. Cette approche est présentée dans la partie **REFERENCE A LA PARTIE USAGE AVANCE**.

7.2. Les hyperparamètres clés des forêts aléatoires

Cette section décrit en détail les principaux hyperparamètres des forêts aléatoires listés dans le tableau Table 1. Les noms des hyperparamètres utilisés sont ceux figurant dans le *package* `R ranger`, et dans le *package* `Python scikit-learn`. Il arrive qu'ils portent un nom différent dans d'autres implémentations des forêts aléatoires, mais il est généralement facile de s'y retrouver en lisant attentivement la documentation.

Table 1 : Les principaux hyperparamètres des forêts aléatoires

Hyperparamètre		Description
`ranger`	`scikit-learn`	
`num.trees`	`n_estimators`	Le nombre d'arbres
`mtry`	`max_features`	Le nombre de variables candidates à chaque noeud
`sample.fraction`	`max_samples`	Le taux d'échantillonnage des données
`replacement`		L'échantillonnage des données se fait-il avec ou sans remise?
`min.node.size`	`min_samples_leaf`	Nombre minimal d'observations nécessaire pour qu'un noeud puisse être partagé
`min.bucket`	`min_samples_split`	Nombre minimal d'observations dans les noeuds terminaux
`max.depth`	`max_depth`	Profondeur maximale des arbres
`splitrule`	`criterion`	Le critère de choix de la règle de division des noeuds intermédiaires
`oob.error`	`oob_score`	Calculer la performance de la forêt par l'erreur OOB (et choix de la métrique pour `scikit`)

- Le **nombre d'arbres** par défaut varie selon les implémentations (500 dans `ranger`, 100 dans `scikit-learn`). Il s'agit d'un hyperparamètre particulier car il n'est associé à aucun arbitrage en matière de performance: la performance de la forêt aléatoire croît avec le nombre d'arbres, puis se stabilise. Le nombre optimal d'arbres est celui à partir duquel la performance de la forêt ne croît plus (ce point est détaillé plus bas) où à partir duquel l'ajout d'arbres supplémentaires génère des gains marginaux. Il est important de noter que ce nombre optimal dépend des autres hyperparamètres. Par exemple, un taux d'échantillonnage faible et un nombre faible de variables candidates à chaque noeud aboutissent à des arbres peu corrélés, mais peu performants, ce qui requiert probablement un plus grand nombre d'arbres. Dans le cas d'une classification,

l'utilisation de mesures comme le score de Brier ou la fonction de perte logarithmique est recommandée pour évaluer la convergence plutôt que la précision (métrique par défaut de `ranger` et `scikit-learn`).

- Le **nombre (ou la part) de variables candidates à chaque noeud** (souvent appelé `mtry`) est un hyperparamètre essentiel qui détermine le nombre de variables prédictives sélectionnées aléatoirement à chaque nœud lors de la construction des arbres. Ce paramètre exerce la plus forte influence sur les performances du modèle, et un compromis doit être trouvé entre puissance prédictive des arbres et corrélation entre arbres. Une faible valeur de `mtry` conduit à des arbres moins performants mais plus diversifiés et donc moins corrélés entre eux. Inversement, une valeur plus élevée améliore la précision des arbres individuels mais accroît leur corrélation (les mêmes variables ayant tendance à être sélectionnées dans tous les arbres). La valeur optimale de `mtry` dépend du nombre de variables réellement pertinentes dans les données: elle est plus faible lorsque la plupart des variables sont pertinentes, et plus élevée lorsqu'il y a peu de variables pertinentes. Par ailleurs, une valeur élevée de `mtry` est préférable si les données comprennent un grand nombre de variables binaires issues du *one-hot-encoding* des variables catégorielles (LIEN AVEC LA PARTIE PREPROCESSING). Par défaut, cette valeur est fréquemment fixée à $\sqrt{p}$ pour les problèmes de classification et à $p/3$ pour les problèmes de régression, où p représente le nombre total de variables prédictives disponibles.
- Le **taux d'échantillonnage** et le **mode de tirage** contrôlent le plan d'échantillonnage des données d'entraînement. Les valeurs par défaut varient d'une implémentation à l'autre; dans le cas de `ranger`, le taux d'échantillonnage est de 63,2% sans remise, et de 100% avec remise. L'implémentation `scikit-learn` ne propose pas le tirage sans remise. Ces hyperparamètres ont des effets sur la performance similaires à ceux du nombre de variables candidates, mais d'une moindre ampleur. Un taux d'échantillonnage plus faible aboutit à des arbres plus diversifiés et donc moins corrélés (car ils sont entraînés sur des échantillons très différents), mais ces arbres peuvent être peu performants car ils sont entraînés sur des échantillons de petite taille. Inversement, un taux d'échantillonnage élevé aboutit à des arbres plus performants mais plus corrélés. Les effets de l'échantillonnage avec ou sans remise sur la performance de la forêt aléatoire sont moins clairs et ne font pas consensus. Les travaux les plus récents semblent toutefois suggérer qu'il est préférable d'échantillonner sans remise (P. Probst, M. N. Wright, and A.-L. Boulesteix [20]).

- Le **nombre minimal d'observations dans les noeuds terminaux** contrôle la taille des noeuds terminaux. La valeur par défaut est faible dans la plupart des implémentations (entre 1 et 5). Il n'y a pas vraiment de consensus sur l'effet de cet hyperparamètre sur les performances, bien qu'une valeur plus faible augmente le risque de sur-apprentissage. En revanche, il est certain que le temps d'entraînement décroît fortement avec cet hyperparamètre: une valeur faible implique des arbres très profonds, avec un grand nombre de noeuds. Il peut donc être utile de fixer ce nombre à une valeur plus élevée pour accélérer l'entraînement, en particulier si les données sont volumineuses et si on utilise une méthode de validation croisée pour le choix des autres hyperparamètres. Cela se fait généralement sans perte significative de performance.
- Le **critère de choix de la règle de division des noeuds intermédiaires**: la plupart des implémentations des forêts aléatoires retiennent par défaut l'impureté de Gini pour la classification et la variance pour la régression, même si d'autres critères de choix ont été proposés dans la littérature (p-value dans les forêts d'inférence conditionnelle, arbres extrêmement randomisés, etc.). Chaque règle présente des avantages et des inconvénients, notamment en termes de biais de sélection des variables et de vitesse de calcul. A ce stade, aucun critère de choix ne paraît systématiquement supérieur aux autres en matière de performance. Modifier cet hyperparamètre relève d'un usage avancé des forêts aléatoires. Le lecteur intéressé pourra se référer à la discussion détaillée dans P. Probst, M. N. Wright, and A.-L. Boulesteix [20].

7.3. Comment entraîner une forêt aléatoire?

Les forêts aléatoires nécessitent généralement moins d'optimisation que d'autres modèles de *machine learning*, car leurs performances varient relativement peu en fonction des hyperparamètres. Les valeurs par défaut fournissent souvent des résultats satisfaisants, ce qui réduit le besoin d'optimisation intensive. Cependant, un ajustement précis des hyperparamètres peut apporter des gains de performance, notamment sur des jeux de données complexes.

Comme indiqué dans la partie Section 5.3.3, la performance prédictive d'une forêt aléatoire varie en fonction de deux critères essentiels: elle croît avec le pouvoir prédictif des arbres, et décroît avec la corrélation des arbres entre eux. L'optimisation des hyperparamètres d'une forêt aléatoire vise donc à trouver un équilibre optimal où les arbres sont suffisamment puissants pour être prédictifs, tout en étant suffisamment diversifiés pour que leurs erreurs ne soient pas trop corrélées.

La littérature propose de multiples approches pour optimiser simultanément plusieurs hyperparamètres: la recherche par grille (*grid search*), la recherche aléatoire (*random search*) et

l'optimisation basée sur modèle séquentiel (SMBO), et il peut être difficile de savoir quelle approche adopter. Ce guide propose donc une première approche délibérément simple, avant de présenter les approches plus avancées.

7.3.1. Approche simple

Voici une procédure simple pour entraîner une forêt aléatoire. Elle ne garantit pas l'obtention d'un modèle optimal, mais elle est lisible et permet d'obtenir rapidement un modèle raisonnablement performant.

- **Entraîner une forêt aléatoire avec les valeurs des hyperparamètres par défaut**. Ce premier modèle servira de point de comparaison pour la suite.
- **Ajuster le nombre d'arbres**: entraîner une forêt aléatoire avec les hyperparamètres par défaut en augmentant progressivement le nombre d'arbres, puis déterminer à partir de quel nombre d'arbres la performance se stabilise (en mesurant la performance avec l'erreur OOB avec pour métrique le

  score de Brier

). Fixer le nombre d'arbres à cette valeur par la suite.
- **Ajuster le nombre de variables candidates et le taux d'échantillonnage**: optimiser ces deux hyperparamètres grâce à une méthode de *grid search* évaluée par une approche de validation-croisée, ou par une approche reposant sur l'erreur OOB.
- **Ajuster le nombre minimal d'observations dans les noeuds terminaux**: optimiser cet hyperparamètre grâce à une méthode de *grid search* évaluée par une approche de validation-croisée, ou par une approche reposant sur l'erreur OOB. Ce n'est pas l'hyperparamètre le plus important, mais s'il est possible de le fixer à une valeur plus élevée que la valeur par défaut sans perte de performance, cela permet d'accélérer le reste de la procédure.
- **Entraîner du modèle final**: entraîner une forêt aléatoire avec les hyperparamètres optimisés déduits des étapes précédentes.
- **Évaluer du modèle final**: mesurer la performance du modèle final soit avec l'approche *out-of-bag* (OOB), soit avec un ensemble de test. Il est souvent instructif de comparer les performances du modèle final et du modèle entraîné avec les valeurs des hyperparamètres par défaut (parfois pour se rendre compte que ce dernier était déjà suffisamment performant…).

7.3.2. Approches plus avancées

Lorsque l'espace des hyperparamètres est large ou que les performances initiales sont insuffisantes, adopter des méthodes avancées comme l'ptimisation basée sur un modèle séquentiel (SMBO). En `R`, il existe plusieurs implémentations d'appuyant sur cette méthode: `tuneRF` (limité à l'optimisation de `mtry`), tuneRanger (optimise simultanément `mtry`, node size, et sample size). La méthode SMBO est généralement la plus performante, mais demande un temps de calcul plus important.

Il est également possible de remplacer les critères classiques (le taux d'erreur pour une classification par exemple) par d'autres critères de performance, comme le score de Brier ou la fonction de perte logarithmique (P. Probst and A.-L. Boulesteix [21]).

Pour gérer la contrainte computationnelle, il est possible de commencer par utiliser des échantillons réduits pour les étapes exploratoires, puis d'augmenter la taille de l'échantillon pour les tests finaux.

7.4. Mesurer l'importance des variables

Les méthodes classiques d'évaluation de l'importance des variables, telles que l'indice de Gini (Mean Decrease in Impurity - MDI) et l'importance par permutation (Mean Decrease Accuracy - MDA), peuvent produire des résultats biaisés dans certaines situations (C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn [11], C. Bénard, S. Da Veiga, and E. Scornet [9], C. Bénard, G. Biau, S. Da Veiga, and E. Scornet [10]). Notamment, lorsque les variables prédictives sont fortement corrélées, présentent des échelles de mesure différentes ou possèdent un nombre variable de catégories, ces méthodes peuvent surestimer l'importance de certaines variables. Par exemple, les variables avec un grand nombre de catégories ou des échelles continues étendues peuvent être artificiellement privilégiées, même si leur contribution réelle à la prédiction est limitée.

En pratique, il est recommandé d'utiliser des méthodes d'importance des variables moins sensibles aux biais, comme les CIF ou la Sobol-MDA. Les valeurs de Shapley, issues de la théorie des jeux, sont également une alternative intéressante. Elles attribuent à chaque variable une contribution proportionnelle à son impact sur la prédiction. Cependant, leur calcul est souvent complexe et coûteux en ressources computationnelles, surtout en présence de nombreuses variables. Des méthodes comme SHAFF (SHApley eFfects via random Forests) ont été développées pour estimer efficacement ces valeurs, même en présence de dépendances entre variables.

On conseille l'utilisation de trois implémentations pour comparer l'importances des variables d'une forêt aléatoire:

- Pour la MDI: l'algorithme CIF proposé par C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn [11] et implémenté en `R`

- Pour la MDA: l'algorithme Sobol-MDA proposé par C. Bénard, S. Da Veiga, and E. Scornet [9] et implémenté en R
- Pour les valeurs de Shapley : l'alogrithme SHAFF proposé par C. Bénard, G. Biau, S. Da Veiga, and E. Scornet [10] et implémenté en R

Enfin, nous recommandons de combiner plusieurs méthodes pour une analyse plus robuste et de tenir compte des prétraitements des données afin de minimiser les biais potentiels.

References

[1] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," *Advances in neural information processing systems*, vol. 35, pp. 507–520, 2022.

[2] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[3] D. McElfresh *et al.*, "When do neural nets outperform boosted trees on tabular data?," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Cart," *Classification and regression trees*, 1984.

[5] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.

[6] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[7] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.

[8] G. Biau, "Analysis of a random forests model," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1063–1095, 2012.

[9] C. Bénard, S. Da Veiga, and E. Scornet, "Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA," *Biometrika*, vol. 109, no. 4, pp. 881–900, 2022, doi:

10.1093/biomet/asac017

.

[10] C. Bénard, G. Biau, S. Da Veiga, and E. Scornet, "SHAFF: Fast and consistent SHApley eFfect estimates via random Forests," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 5563–5582.

[11] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, pp. 1–21, 2007.

[12] R. Shapire, "The strength of weak learning," *Machine Learning*, vol. 5, no. 2, 1990.

[13] L. Breiman, "Rejoinder: arcing classifiers," *The Annals of Statistics*, vol. 26, no. 3, pp. 841–849, 1998.

[14] A. J. Grove and D. Schuurmans, "Boosting in the limit: Maximizing the margin of learned ensembles," in *AAAI/IAAI*, 1998, pp. 692–699.

[15] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[16] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[18] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[19] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[20] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, vol. 9, no. 3, p. e1301, 2019.

[21] P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest," *Journal of Machine Learning Research*, vol. 18, no. 181, pp. 1–18, 2018.