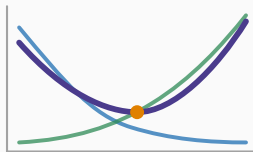


Apprentissage statistique

Chapitre 8 : Les arbres de décision



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Un modèle qui découpe

La construction

L'élagage

Forces et faiblesses

Ce qu'il faut retenir

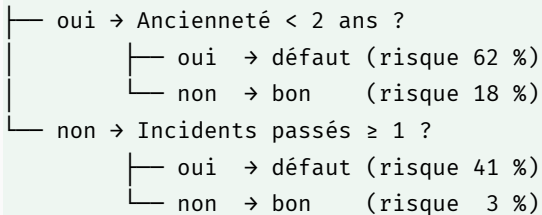
Un modèle qui découpe

Arbre de décision

Un **arbre** découpe l'espace des variables en régions rectangulaires par une suite de questions binaires, et prédit dans chaque région la moyenne — ou la classe majoritaire — des observations qui s'y trouvent.

Un arbre sur le portefeuille de crédit

Endettement $> 0,45$?



Quatre règles, lisibles par un chargé de clientèle qui n'a jamais vu une équation.

Intuition

Il capte les **non-linéarités** et les **interactions** sans qu'on les spécifie.

L'effet de l'ancienneté n'y est pas le même selon le niveau d'endettement — c'est une interaction, que le modèle linéaire du chapitre 4 aurait fallu écrire à la main.

La construction

Un choix glouton, à chaque nœud

L'algorithme

À chaque nœud, essayer **toutes** les variables et **tous** les seuils, retenir le découpage qui réduit le plus l'impureté, puis recommencer sur chaque branche.

Mesurer l'impureté

Cible	Critère
Quantitative	somme des carrés des écarts
Qualitative	indice de Gini ou entropie

Une région pure ne contient qu'une classe : son impureté est nulle.

L'indice de Gini, sur deux nœuds

$$\text{Gini} = 1 - \sum_k p_k^2$$

L'élagage

Un arbre complet est inutilisable

Le problème

Rien n'empêche de découper jusqu'à ce que chaque feuille contienne une seule observation. L'erreur d'apprentissage est alors **nulle**, et l'erreur de test désastreuse.

C'est le sur-ajustement du chapitre 2, dans sa forme la plus pure.

L'élagage par coût-complexité

On fait pousser un grand arbre, puis on le **taille** en pénalisant le nombre de feuilles :

$$\text{critère} = \text{erreur} + \alpha \times \text{nombre de feuilles}$$

α se choisit par **validation croisée** — exactement comme λ au chapitre 4.

Intuition

S'arrêter dès qu'un découpage n'apporte rien serait myope, pour la raison déjà vue : un découpage sans intérêt immédiat peut ouvrir la voie à deux découpages excellents.

On laisse donc pousser, puis on taille — ce qui permet de juger chaque branche sur ce qu'elle a réellement produit.

Forces et faiblesses

Ce que l'arbre a pour lui

1. **Interprétable** — un schéma que n'importe qui lit.
2. Aucun besoin de **standardiser**.
3. Gère naturellement les variables **qualitatives**.
4. Capte **interactions et non-linéarités** sans spécification.
5. Robuste aux valeurs **extrêmes**.

Son défaut, qui est grave

La variance est très élevée. Changer quelques observations peut produire un arbre entièrement différent — d'autres variables, d'autres seuils, une autre structure.

Un modèle qui change de forme quand les données changent à peine n'inspire pas confiance, et prévoit mal.

Application en gestion

	Biais	Variance	Interprétabilité
Régression linéaire	élevé si non-linéarité	faible	bonne
Arbre unique	faible	très élevée	excellente

L'arbre est donc rarement utilisé seul en prévision. Il est en revanche **irremplaçable pour expliquer** une décision à un client ou à un régulateur — ce que les chapitres 9 et 10 vont perdre.

Intuition

Le chapitre 2 l'a établi : moyenner des modèles **indépendants** divise leur variance.

L'arbre a précisément beaucoup de variance et peu de biais — c'est le candidat idéal pour cette opération. Il suffit d'en construire beaucoup, sur des échantillons différents, et de les agréger.

Ce qu'il faut retenir

Cinq points

1. Un **arbre** découpe l'espace en régions rectangulaires par questions binaires.
2. La construction est **gloutonne** : meilleur découpage immédiat, sans anticipation.
3. On mesure l'impureté par **Gini** ou l'**entropie**; Gini vaut 0,18 à la racine ici.
4. Un arbre complet sur-ajuste; on **élague** par coût-complexité, avec α choisi par validation croisée.
5. Interprétabilité excellente, **variance très élevée** — d'où la suite.

La suite

Le chapitre 9 exploite exactement cette faiblesse : beaucoup d'arbres instables, moyennés, donnent un modèle stable.