

Chapitre 11 : Arbres de Décision

26 février 2026

Plan du chapitre

Structure et Intuition

Algorithme CART

Critères de Division

Élagage (Pruning)

Le principe des Arbres de Décision

Les arbres de décision sont des modèles non-paramétriques très intuitifs qui partitionnent récursivement l'espace des caractéristiques en régions rectangulaires.

Leur force est l'explicabilité.

La structure "Si... Alors..." reproduit le raisonnement humain.

- **Nœud racine** : L'ensemble initial des données.
- **Nœuds internes** : Un test binaire sur une caractéristique ($x_j \leq \theta$).
- **Feuilles (Nœuds terminaux)** : La prédiction finale (classe majoritaire ou valeur moyenne).

Structure d'un Arbre (Illustration)

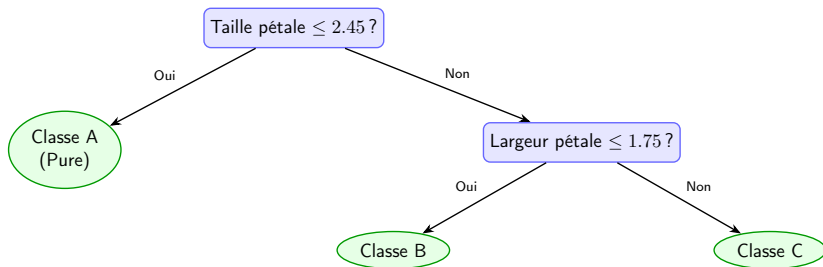


Figure – Arbre de décision typique pour la classification.

L'Algorithme CART

L'algorithme **CART** (Classification And Regression Trees) construit l'arbre de haut en bas selon une approche **gloutonne (greedy)**.

Étapes de construction

1. Partir du nœud racine avec toutes les données.
2. Chercher la "meilleure" division (caractéristique j et seuil θ) qui maximise la **pureté** des enfants.
3. Diviser les données en un sous-ensemble Gauche et Droit.
4. Répéter récursivement sur les enfants jusqu'à un critère d'arrêt.

Étant glouton, CART trouve un optimum local à chaque nœud, ce qui ne garantit pas l'arbre global le plus optimal possible.

Critères d'arrêt de la croissance

Si l'on ne fixe aucune limite, l'algorithme isolera chaque donnée d'entraînement dans une feuille unique (surapprentissage).

On arrête la division d'une branche si :

- Le nœud est **pur** (100% d'une seule classe).
- La **profondeur maximale** (`max_depth`) est atteinte.
- Le nœud contient trop peu d'exemples (`min_samples_split`).
- Le gain de pureté est inférieur à un seuil critique.

Division en Classification : Gini

En classification, on cherche à séparer les classes. Soit $p(k|t)$ la proportion de la classe k dans le nœud t .

Impureté de Gini (choix par défaut dans CART)

Mesure la probabilité de mal classer un exemple tiré au hasard :

$$\text{Gini}(t) = 1 - \sum_{k=1}^K [p(k|t)]^2$$

- Nœud pur $\rightarrow \text{Gini} = 0$
- Nœud mélangé uniformément (2 classes) $\rightarrow \text{Gini} = 0.5$

Le Gain : On choisit la division s maximisant la baisse d'impureté :

$$\text{Gain}_{\text{Gini}}(s) = \text{Gini}(t_{\text{parent}}) - \left(\frac{n_g}{n_p} \text{Gini}(t_{\text{gauche}}) + \frac{n_d}{n_p} \text{Gini}(t_{\text{droit}}) \right)$$

Division en Classification : Entropie

Gain d'Information (Entropie)

Issu de la théorie de l'information (utilisé par ID3 / C4.5), il mesure la quantité de désordre ou d'incertitude :

$$H(t) = - \sum_{k=1}^K p(k|t) \log_2 p(k|t)$$

Gini ou Entropie ?

En pratique, ils produisent des arbres très similaires.

Cependant, Gini est privilégié car il est plus rapide à calculer informatiquement (pas de logarithme).

Et pour la Régression ?

Les arbres gèrent aussi les variables continues ($\mathcal{Y} = \mathbb{R}$). La prédiction d'une feuille est simplement la moyenne des cibles des données qui s'y trouvent.

Critère de division (Erreur Quadratique)

On ne cherche plus la "pureté" des classes, mais à réduire la **Variance** (ou la MSE) dans les nœuds enfants :

$$\text{MSE}(t) = \frac{1}{n_t} \sum_{i \in t} (y_i - \bar{y}_t)^2$$

L'algorithme choisit le seuil θ qui minimise la somme des MSE pondérées des enfants gauche et droit.

Le Surapprentissage

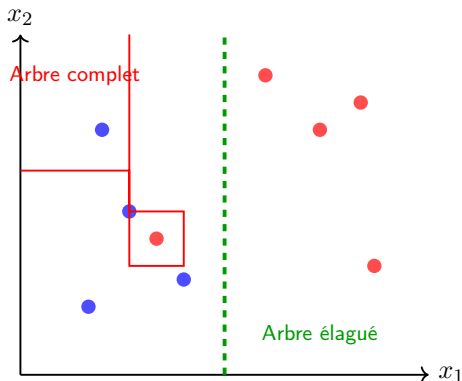


Figure – Un arbre non contraint crée des frontières absurdes pour isoler le moindre point de bruit (rouge).

Solutions : Pré et Post-Élagage

1. Pré-élagage (Hyperparamétrage)

Arrêt précoce à la construction (profondeur max, min samples).

Avantage : Rapide.

Inconvénient : Trop myope (bloque des branches potentiellement utiles plus bas).

2. Post-élagage (Cost-Complexity Pruning)

On laisse l'arbre grandir à son maximum, puis on élague.

On minimise le risque pénalisé :

$$R_{\alpha}(T) = R(T) + \alpha|T|$$

- $R(T)$: Taux d'erreur de l'arbre.
- $|T|$: Nombre de feuilles (complexité).
- α : Paramètre de pénalité (estimé par validation croisée).