Arbre de Décision

Filière : Génie Informatique et Ingénierie des Données

Réalisé par :

BOULIDAM ABDELLAH

BOUMLIK YOUNESS

HAFSI GHIZLANE

Encadré par :

Mr.GHAZDALI

Université Sultan Moulay Slimane

École Nationale des Sciences Appliquées

- Khouribga -

Année universitaire :

2024/2025

Plan

Qu'est-ce que le Machine Learning ?

Définition

- Sous-discipline de l'intelligence artificielle (IA).
- Développement d'algorithmes capables d'apprendre à partir des données.
- Identification automatique des schémas et prise de décisions.

Particularité

Contrairement aux systèmes classiques basés sur des règles, le ML s'appuie sur :

- Des techniques statistiques.
- Des méthodes mathématiques.
- L'apprentissage par l'exemple.

Principes Fondamentaux du Machine Learning

1. **Entraînement sur des données** :
   - Utilisation de données d'apprentissage.
   - Découverte de patterns.
2. **Généralisation** :
   - Application à de nouvelles données.
3. **Apprentissage itératif** :
   - Ajustement continu des paramètres.
   - Optimisation des performances.

Types d'Apprentissage Automatique

Apprentissage Supervisé

- Données étiquetées.
- Exemple : Prédiction de prix.
- Algorithmes : régression, arbres de décision.

Apprentissage Non Supervisé

- Données non étiquetées.
- Exemple : Segmentation clients.
- Algorithmes : clustering, PCA.

Apprentissage Semi-supervisé

- Mélange de données étiquetées et non étiquetées.

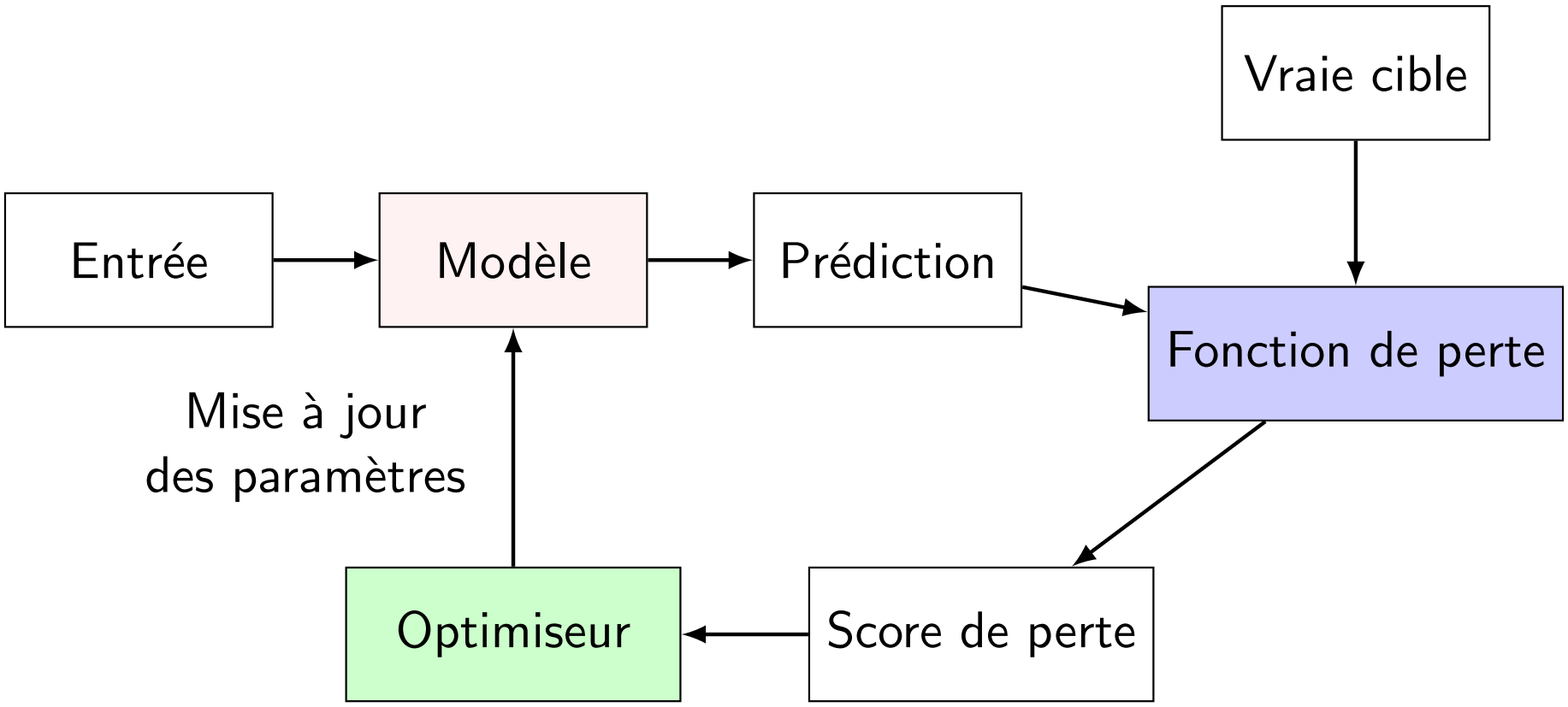
Apprentissage par Renforcement

- Apprentissage par essai-erreur.
- Exemple : IA pour les jeux.
- Algorithmes : Q-learning.

Composantes Clés du Machine Learning

- **Données** :
  - Base de l'apprentissage.
  - Qualité et quantité importantes.
- **Caractéristiques (Features)** :
  - Variables d'entrée du modèle.
- **Modèle et Optimisation** :
  - Structure algorithmique.
  - Fonction de coût.
  - Algorithmes d'optimisation (ex : descente de gradient).

Processus d'Apprentissage



Objectif d'un Arbre de Décision

- **Prédiction** : Prédire une classe (classification) ou une valeur continue (régression).
  Exemple : Prédire si un client remboursera son crédit.
- **Séparation optimale des données** : Choisir les attributs les plus pertinents pour maximiser l'information.
  Exemple : Séparer les patients en groupes à risque.
- **Interprétabilité** : Modèle facile à comprendre et à visualiser.
  Exemple : Expliquer une décision de prêt.
- **Généralisation** : Fonctionner sur des données non vues.
  Exemple : Détecter des spams.
- **Simplicité et efficacité** : Rapide à entraîner et à exécuter.
  Exemple : Classer les requêtes d'un chatbot.
- **Adaptabilité** : Gérer des données numériques ou catégorielles.
  Exemple : Prédire la satisfaction client.
- **Gestion des données manquantes** : Traiter des données incomplètes.
  Exemple : Analyse médicale avec des informations manquantes.

Structure d'un Arbre de Décision

1. **Racine** : Premier nœud, meilleure séparation initiale.
2. **Nœuds internes** : Tests sur les attributs.
3. **Branches** : Résultats des tests.
4. **Feuilles** : Décisions finales ou prédictions.

Exemple de Structure d'un Arbre de Décision

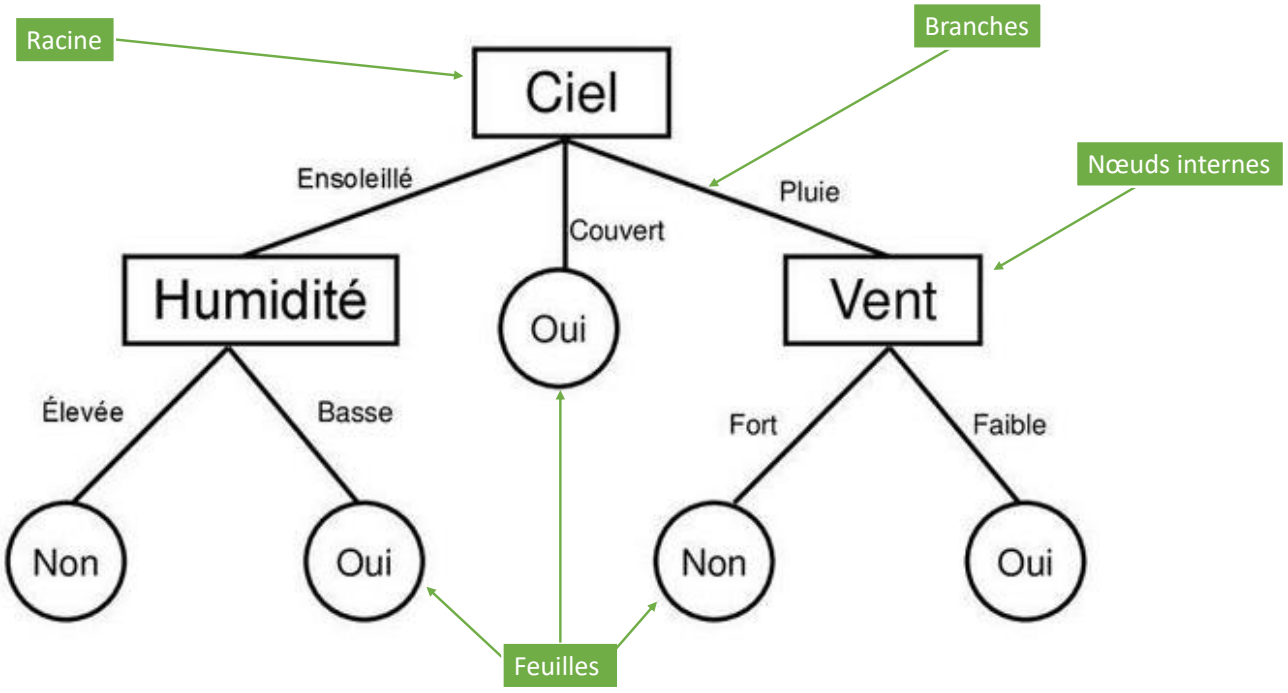


Figure: Exemple d'un arbre de décision illustrant les concepts clés : la **racine**, les **nœuds internes**, les **branches** et les **feuilles**.

Construction d'un Arbre de Décision

- **Processus itératif** : Diviser les données en sous-groupes homogènes.
- **Critères de sélection** : Utiliser des mesures comme l'entropie ou l'indice de Gini.
- **Apprentissage supervisé** : Basé sur des données étiquetées.

Principe Général de Construction

Processus Récursif

1. Sélectionner le meilleur attribut pour diviser les données.
2. Créer un nœud et établir des branches.
3. Répéter pour chaque sous-groupe.

Conditions d'Arrêt

- Profondeur maximale atteinte.
- Absence d'amélioration significative.
- Niveau de pureté suffisant dans les feuilles.

Gestion des Variables Continues dans un Arbre de Décision

Problème :

Remarque

Lorsqu'une variable explicative est continue (ex: taille, âge, température), un arbre de décision doit choisir un **seuil** pour la transformer en une variable binaire. Ce choix influence fortement la performance du modèle.

Méthode :

- Trier les valeurs possibles de la variable.
- Tester plusieurs seuils possibles.
- Choisir le seuil maximisant le **gain d'information** ou minimisant l'**indice de Gini**.

Exemple :

- Variable : $\hat{A}ge$ d'un client en années.
- Seuils candidats : 25, 30, 35, 40...
- On choisit le seuil $\theta = 30$ si **âge** ≤ 30 optimise la séparation des classes.

Choix du Critère de Division

Critères Disponibles

- Entropie et gain d'information.
- Indice de Gini.
- Variance (pour la régression).

Impact

- Structure de l'arbre.
- Performance du modèle.

Algorithmes d'Apprentissage

ID3

- Utilise l'entropie.
- Gain d'information.
- Choix optimal des attributs.

C4.5

- Amélioration d'ID3.
- Gestion des valeurs continues.
- Traitement des données manquantes.

CART

- Indice de Gini.
- Classification et régression.

Entropie

- L'entropie mesure l'incertitude ou l'impureté d'un ensemble de données.
- Formule de l'entropie :

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

- n : nombre total de classes dans l'ensemble S.
- p_i : proportion d'exemples appartenant à la classe i dans S.

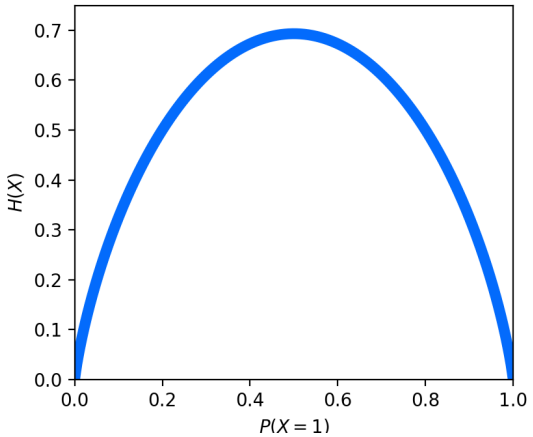


Figure: Graphe de l'entropie en fonction de la probabilité

Données Utilisées

Forme des oreilles (x_1)	**Forme du visage (x_2)**	**Moustaches (x_3)**	**Classe (y)**
Pointues	Rond	Présentes	1
Ovales	Non rond	Présentes	1
Ovales	Rond	Absentes	0
Pointues	Non rond	Présentes	0
Ovales	Rond	Présentes	1
Flasques	Non rond	Absentes	0
Ovales	Rond	Absentes	1
Flasques	Rond	Absentes	0
Flasques	Rond	Absentes	0

Table: Exemple de données

- **Forme des oreilles (x_1)** : Pointues, Ovales, Flasques.
- **Forme du visage (x_2)** : Rond, Non rond.
- **Moustaches (x_3)** : Présentes, Absentes.
- **Classe (y)** : 1 (chat), 0 (non-chat).

Ces données seront utilisées pour illustrer les calculs d'entropie et d'indice de Gini.

Exemple de Calcul de l'Entropie

- Ensemble S de 9 exemples basé sur les données du tableau.
- Classes possibles (y) : 1 (chat) et 0 (non-chat).
- Nombre d'exemples dans chaque classe :
  - 4 exemples dans la classe 1.
  - 5 exemples dans la classe 0.
- Probabilités associées à chaque classe :

$$p_1 = \frac{4}{9}, \quad p_0 = \frac{5}{9}.$$

Calcul de l'Entropie (Étapes)

- Appliquons la formule de l'entropie :

$$H(S) = -\left(\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}\right).$$

- Calcul des logarithmes :

$$\log_2\frac{4}{9} \approx -1.169, \quad \log_2\frac{5}{9} \approx -0.847.$$

- Substitution dans la formule :

$$H(S) = -\left(\frac{4}{9}\cdot(-1.169) + \frac{5}{9}\cdot(-0.847)\right).$$

- Calcul final :

$$H(S) \approx -\left(-0.519 - 0.472\right) = 0.991 \text{ bits}.$$

Résultat du Calcul de l'Entropie

- L'entropie de l'ensemble S est :

$$H(S) \approx 0.991 \text{ bits.}$$

- Interprétation :
  - Une entropie proche de 1 indique une distribution relativement équilibrée des classes.
  - Cela signifie que l'ensemble S est assez hétérogène.

Gain d'Information

- Le gain d'information mesure la réduction de l'incertitude (ou de l'entropie) après avoir effectué une division des données selon un attribut donné.
- Il est utilisé pour déterminer quel attribut doit être choisi comme racine ou à chaque nœud d'un arbre de décision.
- Formule du gain d'information :

$$\text{Gain}(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)$$

- Objectif : Maximiser le gain d'information pour réduire l'incertitude de manière optimale.

Explication des termes de la formule

- $\text{Gain}(S, A)$: Gain d'information de l'attribut A pour l'ensemble de données S.
- $H(S)$: Entropie de l'ensemble S, mesurant l'incertitude ou le désordre dans les données avant la division.
- $\mathit{Values}(A)$: Ensemble des valeurs possibles que peut prendre l'attribut A.
- S_v : Sous-ensemble des données où l'attribut A prend la valeur v.
- $|S|$: Nombre total d'éléments dans l'ensemble S.
- $|S_v|$: Nombre d'éléments dans le sous-ensemble S_v.
- $H(S_v)$: Entropie du sous-ensemble S_v, mesurant l'incertitude après la division.
- $\frac{|S_v|}{|S|}$: Proportion des éléments dans S_v par rapport à l'ensemble S.

Exemple de Calcul du Gain d'Information

Entropie initiale de l'ensemble S :

- 5 exemples de classe 0 et 4 exemples de classe 1.
- Entropie initiale :

$$H(S) = -\left(\frac{5}{9}\log_2\left(\frac{5}{9}\right) + \frac{4}{9}\log_2\left(\frac{4}{9}\right)\right) \approx 0.991 \text{ bits.}$$

Cas 1 : Division selon l'attribut x_1 (Forme des oreilles)

- Valeurs possibles : {Pointues, Ovales, Flasques}.
- Pour $x_1 =$ Pointues (2 exemples : 1 classe 0, 1 classe 1) :

$$H(S_{\text{Pointues}}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1.$$

- Pour $x_1 =$ Ovales (4 exemples : 1 classe 0, 3 classe 1) :

$$H(S_{\text{Ovales}}) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) \approx 0.811.$$

- Pour $x_1 =$ Flasques (3 exemples : 3 classe 0, 0 classe 1) :

$$H(S_{\text{Flasques}}) = 0 \text{ (ensemble homogène).}$$

Exemple de Calcul du Gain d'Information (Suite)

Cas 2 : Division selon l'attribut x_2 (Forme du visage)

- Valeurs possibles : {Rond, Non rond}.
- Pour $x_2 =$ Rond (6 exemples : 3 classe 0, 3 classe 1) :

$$H(S_{\text{Rond}}) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1.$$

- Pour $x_2 =$ Non rond (3 exemples : 2 classe 0, 1 classe 1) :

$$H(S_{\text{Non rond}}) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \approx 0.918.$$

- Gain d'information pour x_2 :

$$\text{Gain}(S, x_2) = 0.991 - \left(\frac{6}{9}\cdot 1 + \frac{3}{9}\cdot 0.918\right) \approx 0.073.$$

Exemple de Calcul du Gain d'Information (Suite)

Cas 3 : Division selon l'attribut x_3 (Moustaches)

- Valeurs possibles : {Présentes, Absentes}.
- Pour $x_3 =$ Présentes (4 exemples : 1 classe 0, 3 classe 1) :

$$H(S_{\text{Présentes}}) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) \approx 0.811.$$

- Pour $x_3 =$ Absentes (5 exemples : 4 classe 0, 1 classe 1) :

$$H(S_{\text{Absentes}}) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) \approx 0.721.$$

- Gain d'information pour x_3 :

$$\text{Gain}(S, x_3) = 0.991 - \left(\frac{4}{9}\cdot 0.811 + \frac{5}{9}\cdot 0.721\right) \approx 0.230.$$

Conclusion du Calcul du Gain d'Information

Gains d'information pour chaque attribut :

$$\text{Gain}(S, x_1) \approx 0.408, \quad \text{Gain}(S, x_2) \approx 0.073, \quad \text{Gain}(S, x_3) \approx 0.230.$$

Conclusion :

- L'attribut x_1 (Forme des oreilles) présente le gain d'information le plus élevé (0.408).
- Il réduit le plus l'entropie de l'ensemble S.
- Par conséquent, x_1 est le meilleur choix pour la racine de l'arbre de décision.

Indice de Gini

- L'indice de Gini mesure l'impureté d'un ensemble de données.
- Formule de l'indice de Gini :

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2$$

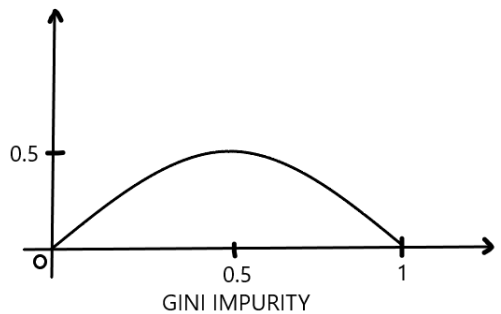


Figure: Graphe de l'indice de Gini en fonction de la probabilité

Introduction

- Ensemble S de 9 exemples basé sur les données du tableau.
- Classes possibles (y) :
  - 1 (chat) : 4 exemples.
  - 0 (non-chat) : 5 exemples.
- Probabilités associées :

$$p_1 = \frac{4}{9}, \quad p_0 = \frac{5}{9}$$

- Objectif : Calculer l'indice de Gini pour S et le gain de Gini pour chaque attribut.

Calcul de l'Indice de Gini pour S

- Formule de l'indice de Gini :

$$Gini(S) = 1 - (p_1^2 + p_0^2)$$

- Application numérique :

$$Gini(S) = 1 - \left(\left(\frac{4}{9} \right)^2 + \left(\frac{5}{9} \right)^2 \right)$$

$$Gini(S) = 1 - (0.1975 + 0.3086) = 0.4939$$

- L'indice de Gini de S est 0.4939.

Gain de Gini pour l'Attribut x_1

- Attribut x_1 : Forme des Oreilles.
- Distribution :
  - Pointues : 2 exemples (1 classe 1, 1 classe 0).
  - Ovales : 4 exemples (3 classe 1, 1 classe 0).
  - Flasques : 3 exemples (0 classe 1, 3 classe 0).
- Calcul de $Gini(D_{x_1})$:

$$Gini(D_{x_1}) = \frac{2}{9} \times 0.5 + \frac{4}{9} \times 0.375 + \frac{3}{9} \times 0 = 0.2778$$

- Gain de Gini :

$$Gain_{Gini}(x_1) = Gini(S) - Gini(D_{x_1}) = 0.4939 - 0.2778 = 0.2161$$

Gain de Gini pour l'Attribut x_2

- Attribut x_2 : Forme du Visage.
- Distribution :
  - Rond : 6 exemples (3 classe 1, 3 classe 0).
  - Non rond : 3 exemples (1 classe 1, 2 classe 0).
- Calcul de $Gini(D_{x_2})$:

$$Gini(D_{x_2}) = \frac{6}{9} \times 0.5 + \frac{3}{9} \times 0.4444 = 0.4814$$

- Gain de Gini :

$$Gain_{Gini}(x_2) = Gini(S) - Gini(D_{x_2}) = 0.4939 - 0.4814 = 0.0125$$

Gain de Gini pour l'Attribut x_3

- Attribut x_3 : Moustaches.
- Distribution :
  - Présentes : 4 exemples (3 classe 1, 1 classe 0).
  - Absentes : 5 exemples (1 classe 1, 4 classe 0).
- Calcul de $Gini(D_{x_3})$:

$$Gini(D_{x_3}) = \frac{4}{9} \times 0.375 + \frac{5}{9} \times 0.32 = 0.3445$$

- Gain de Gini :

$$Gain_{Gini}(x_3) = Gini(S) - Gini(D_{x_3}) = 0.4939 - 0.3445 = 0.1494$$

Conclusion

- Comparaison des gains de Gini :
  - $Gain_{Gini}(x_1) = 0.2161$.
  - $Gain_{Gini}(x_2) = 0.0125$.
  - $Gain_{Gini}(x_3) = 0.1494$.
- L'attribut x_1 (Forme des Oreilles) a le plus grand gain de Gini (0.2161).
- Conclusion : x_1 est le meilleur choix pour la division.

Algorithmes d'arbres de décision

- Les arbres de décision sont des modèles de machine learning utilisés pour la classification et la régression.
- Plusieurs algorithmes existent pour construire des arbres de décision :
  - ID3 (Iterative Dichotomiser 3).
  - C4.5 (Successeur de ID3).
  - CART (Classification and Regression Trees).
  - SPRINT (Scalable Parallelizable Induction of Decision Trees).
- Chaque algorithme a ses propres caractéristiques, critères de division et domaines d'application.

ID3 (Iterative Dichotomiser 3)

- **Origine** : Développé par Ross Quinlan en 1986.
- **Critère de division** : Utilise le **gain d'information** (basé sur l'entropie).
- **Types de problèmes** : Uniquement pour la **classification**.
- **Caractéristiques** :
  - Ne supporte pas les attributs numériques (uniquement catégoriels).
  - Ne gère pas les valeurs manquantes.
  - Génère des arbres **non binaires**.
- **Fonctionnement** :
  1. Calcule l'entropie de l'ensemble de données.
  2. Pour chaque attribut, calcule le gain d'information.
  3. Choisit l'attribut avec le gain d'information maximal.
  4. Répète le processus récursivement.
- **Limitations** :
  - Tendance au surajustement (overfitting).
  - Ne supporte pas la régression.

Exemple pour ID3

- **Problème** : Prédire le temps (Ensoleillé, Pluvieux, Nuageux) en fonction de l'humidité et de la température.
- **Attributs** : "Humidité" (Élevée, Normale), "Température" (Chaude, Douce, Froide).
- **Étapes** :
  1. Calcule l'entropie de l'ensemble de données.
  2. Calcule le gain d'information pour "Humidité" et "Température".
  3. Choisit l'attribut avec le gain d'information maximal (par exemple, "Humidité").
  4. Divise les données en sous-ensembles et répète le processus.

C4.5 (Successeur de ID3)

- **Origine** : Développé par Ross Quinlan comme amélioration de ID3.
- **Critère de division** : Utilise le **gain ratio** (normalisation du gain d'information).
- **Types de problèmes** : Principalement pour la **classification**, mais adaptable à la régression.
- **Caractéristiques** :
  - Supporte les attributs **numériques** et **catégoriels**.
  - Gère les **valeurs manquantes**.
  - Génère des arbres **non binaires**.
  - Inclut un mécanisme d'élagage pour réduire le surajustement.
- **Fonctionnement** :
  1. Calcule le gain ratio pour chaque attribut.
  2. Choisit l'attribut avec le gain ratio maximal.
  3. Pour les attributs numériques, détermine un seuil optimal.
  4. Répète le processus récursivement.
- **Avantages** :
  - Plus robuste que ID3.
  - Moins sujet au surajustement.

Exemple pour C4.5

- **Problème** : Prédire le risque de crédit en fonction de l'âge, du revenu, et de l'historique de crédit.
- **Attributs** : "Âge" (numérique), "Revenu" (numérique), "Historique de crédit" (catégoriel).
- **Étapes** :
  1. Calcule le gain ratio pour chaque attribut.
  2. Choisit l'attribut avec le gain ratio maximal (par exemple, "Âge").
  3. Détermine un seuil optimal pour "Âge" (par exemple, "Âge $\leq$ 30").
  4. Divise les données et répète le processus.

CART (Classification and Regression Trees)

- **Origine** : Développé par Breiman et al. en 1984.
- **Critère de division** : Utilise l'**indice de Gini** pour la classification et la **variance réduite** pour la régression.
- **Types de problèmes** : **Classification** et **régression**.
- **Caractéristiques** :
  - Génère des arbres **binaires**.
  - Supporte les attributs numériques et catégoriels.
  - Ne gère pas directement les valeurs manquantes (nécessite un prétraitement).
  - Inclut un mécanisme d'élagage.
- **Fonctionnement** :
  1. Calcule l'indice de Gini pour chaque attribut.
  2. Choisit l'attribut et le seuil qui minimisent l'impureté.
  3. Divise le nœud en deux sous-ensembles.
  4. Répète le processus récursivement.

Exemple pour CART

- **Problème** : Prédire le prix d'une maison en fonction de sa taille, de son emplacement, et de son âge.
- **Attributs** : "Taille" (numérique), "Emplacement" (catégoriel), "Âge" (numérique).
- **Étapes** :
  1. Calcule l'indice de Gini pour chaque attribut.
  2. Choisit l'attribut et le seuil qui minimisent l'impureté (par exemple, "Taille ≤ 1000").
  3. Divise les données en deux sous-ensembles et répète le processus.

SPRINT (Scalable Parallelizable Induction of Decision Trees)

- **Origine** : Conçu pour les grands ensembles de données.
- **Critère de division** : Utilise l'indice de Gini ou l'entropie.
- **Types de problèmes** : Principalement pour la **classification**.
- **Caractéristiques** :
  - Conçu pour les **grands ensembles de données**.
  - Utilise des techniques de **parallélisation**.
  - Supporte les attributs numériques et catégoriels.
  - Gère les valeurs manquantes.
- **Fonctionnement** :
  1. Divise les données en plusieurs partitions.
  2. Applique l'algorithme en parallèle sur chaque partition.
  3. Combine les résultats pour former l'arbre final.
- **Avantages** :
  - Très efficace pour les données massives.
  - Adapté aux environnements distribués (Hadoop, Spark).

Exemple pour SPRINT

- **Problème** : Détecter des fraudes dans des millions de transactions.
- **Attributs** : "Montant" (numérique), "Type de transaction" (catégoriel), "Localisation" (catégoriel).
- **Étapes** :
  1. Divise les données en partitions.
  2. Applique l'algorithme en parallèle sur chaque partition.
  3. Combine les résultats pour former l'arbre final.

Comparaison des algorithmes

Algorithme	**Critère**	**Problèmes**	**Valeurs manquantes**
ID3	Gain info.	Classification	Non
C4.5	Gain ratio	Classification	Oui
CART	Gini / Variance	Class., Régression	Non (prétraitement)
SPRINT	Gini / Entropie	Classification	Oui

Table: Comparaison des algorithmes d'arbres de décision

Conclusion

- Les arbres de décision sont des modèles **puissants et intuitifs** pour la classification et la régression, capables de gérer des données complexes tout en restant interprétables.
- Les critères de division, tels que l'**entropie**, le **gain d'information**, l'**indice de Gini** et le **gain ratio**, jouent un rôle central dans la construction des arbres en maximisant la pureté des nœuds.
- Le choix de l'attribut de division est une étape clé : il détermine l'efficacité et la performance de l'arbre.
- Les algorithmes comme **ID3**, **C4.5**, **CART** et **SPRINT** offrent des approches variées, adaptées à différents types de problèmes et de données (catégorielles, numériques, massives).
- Les exemples concrets (prédiction du temps, risque de crédit, prix des maisons, détection de fraudes) illustrent l'utilité des arbres de décision dans des domaines variés.
- Enfin, les arbres de décision constituent une base solide pour des techniques plus avancées comme les forêts aléatoires (Random Forests) et le boosting (XGBoost, LightGBM).