Exercice 1 : Arbres de décision : Gain entropique / indice de Gini

Problème :

Dans le cadre d'une étude sur les comportements d'achat, une entreprise cherche à comprendre quels facteurs influencent la décision d'un client d'acheter un produit spécifique. Pour cela, elle a collecté des données sur un échantillon de clients, enregistrant plusieurs attributs pertinents tels que l'âge, le revenu, le statut d'étudiant, et la qualité du crédit. Chaque client est également associé à une variable cible indiquant s'il a acheté le produit ("Oui") ou non ("Non").

L'objectif principal de cette analyse est de calculer deux métriques clés utilisées dans la construction des arbres de décision :

- ✓ **Gain Entropique** : Mesure la réduction d'incertitude (entropie) apportée par chaque attribut.
- ✓ **Indice de Gini** : Évalue l'impureté des nœuds pour chaque attribut.

Ces métriques permettent de sélectionner les attributs les plus informatifs pour effectuer des splits dans un arbre de décision. Bien que cette étude se concentre uniquement sur le calcul du gain entropique et de l'indice de Gini, il est possible d'étendre le problème à d'autres aspects, tels que :

1. La construction complète de l'arbre de décision.
2. L'évaluation des performances du modèle (précision, rappel, etc.).
3. L'interprétation des résultats pour fournir des recommandations commerciales.

Description des Données

Les données sont structurées sous forme de tableau (DataFrame) avec les colonnes suivantes :

-**Âge :** Représente la tranche d'âge du client (Jeune, Adulte, Senior).
-**Revenu :** Indique le niveau de revenu du client (Faible, Moyen, Elevé).
-**Étudiant :** Statut d'étudiant du client (Oui, Non).
-**Crédit :** Qualité du crédit du client (Bon, Excellent).
-**Acheter :** Variable cible indiquant si le client a acheté le produit (Oui, Non).

Ces données permettent d'analyser comment les différents attributs influencent la décision d'achat et de sélectionner les attributs les plus informatifs pour les calculs.

	Age	Revenu	Etudiant	Crédit	Acheter
0	Jeune	Moyen	Oui	Bon	Oui
1	Jeune	Moyen	Oui	Excellent	Oui
2	Jeune	Faible	Non	Bon	Non
3	Jeune	Elevé	Non	Excellent	Non
4	Adulte	Elevé	Non	Excellent	Non
5	Adulte	Faible	Non	Bon	Non
6	Adulte	Moyen	Oui	Bon	Oui
7	Senior	Moyen	Oui	Bon	Oui
8	Senior	Elevé	Non	Excellent	Oui
9	Senior	Moyen	Non	Excellent	Oui

Calcul de l'entropie initiale I(S)

L'entropie initiale mesure l'incertitude globale dans le jeu de données avant tous split.

Formule de l'entropie :

$$I(S) = -\sum_{i=1}^{k} p_i \cdot log_2(p_i)$$

Le jeu de données contient 10 exemples.

Parmi ces exemples :

- Oui : 6 exemples (60%)
- Non : 4 exemples (40%)

Calcul de I(S) :

$$I(S) = -\big(0{,}6\ log_2(0{,}6) + 0{,}4\ log_2(0{,}4)\big) = 0{,}971$$

Calcul du Gain Entropique : Gain(S,A) et Indice de Gini : Gini(S,A)

Le gain entropique mesure la réduction d'entropie après un split sur un attribut spécifique.

Formule du gain entropique :

$$Gain(S, A) = I(S) - \sum_{v \in valeurs(A)} \frac{|S_v|}{S} \cdot I(S_v)$$

Calculer le gain entropique et l'indice de Gini pour chaque attribut :

a) Attribut "Âge"

L'attribut "Âge" a trois valeurs possibles : Jeune, Adulte et Senior.

1. Jeune :
   - Nombre d'exemples : 4.
   - Répartition des classes : 2 "Oui", 2 "Non".
   - Entropie : $I\big(S_{Jeune}\big) = -\big(0{,}5\ log_2(0{,}5) + 0{,}5\ log_2(0{,}5)\big) = 1$
   - Gini : $G\big(S_{Jeune}\big) = 1 - (0{,}5^2 + 0{,}5^2) = 0{,}5$
2. Adulte :
   - Nombre d'exemples : 3.
   - Répartition des classes : 1 "Oui", 2 "Non".
   - Entropie :$I(S_{Adulte}) = -\big(0333\ log_2(0{,}333) + 0{,}667\ log_2(0{,}667)\big) = 0{,}918$
   - Gini : $G(S_{Adulte}) = 1 - (0{,}333^2 + 0{,}667^2) = 0{,}444$

3. Senior :
    - Nombre d'exemples : 3.
    - Répartition des classes : 3 "Oui", 0 "Non".
    - Entropie : $I(S_{Senior}) = -\big(1{,}0\ log_2(1{,}0) + 0\ log_2(0)\big) = 0$
    - Gini : $G(S_{Senior}) = 1 - (1^2 + 0^2) = 0$
4. Gain Entropique pour "Âge" :

$$Gain(S, Age) = I(S) - \left(\frac{4}{10} \times 1 + \frac{3}{10} \times 0{,}918 + \frac{3}{10} \times 0\right)$$

$$Gain(S, Age) = 0{,}971 - (0{,}4 + 0{,}275 + 0) = 0{,}296$$

5. Indice de Gini pour "Âge" :

$$Gini(S, Age) = \frac{4}{10} \times 0{,}5 + \frac{3}{10} \times 0{,}444 + \frac{3}{10} \times 0 = 0{,}333$$

b) Attribut "Revenu"

L'attribut "Revenu" a trois valeurs possibles : Moyen, Faible, et Elevé.

1. Moyen:
    - Nombre d'exemples : 5.
    - Répartition des classes : 5 "Oui", 0 "Non".
    - Entropie : $I\big(S_{Moyen}\big) = -\big(1\ log_2(1) + 0\ log_2(0)\big) = 0$
    - Gini : $G\big(S_{Moyen}\big) = 1 - (1^2 + 0^2) = 0$
2. Faible:
    - Nombre d'exemples : 2.
    - Répartition des classes : 0 "Oui", 2 "Non".
    - Entropie :$I(S_{Faible}) = -\big(0\ log_2(0) + 1\ log_2(1)\big) = 0$
    - Gini : $G(S_{Faible}) = 1 - (0^2 + 1^2) = 0$
3. Elevé :
    - Nombre d'exemples : 3.
    - Répartition des classes : 1 "Oui", 2 "Non".
    - Entropie : $I(S_{Elevé}) = -\big(0{,}333\ log_2(0{,}333) + 0{,}667\ log_2(0{,}667)\big) = 0{,}918$
    - Gini : $G(S_{Elevé}) = 1 - (0{,}333^2 + 0{,}667^2) = 0{,}444$

4. Gain Entropique pour "Revenu" :

$$Gain(S, Revenu) = I(S) - \left(\frac{5}{10} \times 0 + \frac{2}{10} \times 0 + \frac{3}{10} \times 0{,}918\right)$$

$$Gain(S, Revenu) = 0{,}971 - (0 + 0 + 0{,}276) = 0{,}696$$

5. Indice de Gini pour "Revenu" :

$$Gini(S, Revenu) = \frac{5}{10} \times 0 + \frac{2}{10} \times 0 + \frac{3}{10} \times 0{,}444 = 0{,}133$$