— Problématique 1 :
Sur quels critères se baser afin de mesurer la qualité d'un ensemble de donnée destiné à la création d'un modèle d'arbre de décision? peut-on définir des caractéristiques d'un modèle parfait?

— Réponse :
A ce sujet, nous devons d'abord préciser que les arbres de décision font partie des techniques statistiques *non paramétrique*, c-a-d qu'aucune restriction sur les paramètres ni sur la structure n'est faite au préalable.
Ce qui rend impossible la définition d'un ensemble parfait, vu le nombre infini de possibilité que l'on peut avoir, en variant les paramètres (nombre de ligne, nombre d'attribut, nombre de catégorie de classes, ... etc).
Néanmoins, il existe une méthode, appelée *Validation croisée*, qui permet d'évaluer un ensemble de données (en terme de précision), donnant ainsi la possibilité à une comparaison entre ensembles, afin de déterminer lequel serait le meilleur pour l'algorithme X.

— Validation croisée
La technique est très utilisée dans le domaine d'apprentissage automatique, elle tien sa force du fait qu'elle permet d'utiliser tout l'ensemble de données, à la fois en apprentissage et en teste. Elle peut être utilisée à deux fin :

1. Vérifier les performances d'erreur du modèle
2. Valider l'ensemble de données d'apprentissage

— Déroulement

1. Découper l'ensemble de données en K parties

FIGURE 1 – Validation croisée : découpage de l'ensemble en K parties

2. Faire K itérations, tel que :
Chacune des K parties est utilisée comme données de teste, afin de tester le modèle crée par les K-1 parties restantes. A la fin, à partir des K arbres de décision crées. la moyenne de leurs taux de précision est calculée afin d'évaluer la taux de précision de notre jeu de données avec l'algorithme utilisé.
Ainsi, comparer deux ensemble en terme de qualité pour un algorithme donné, revient à appliquer la validation croisée afin de comparer leurs résultats.

— Problématique 2 :
A partir d'un ensemble de données, comment peut on choisir le critère de division adéquat afin de construire le meilleur arbre de décision possible ?

— Réponse :
Étant donnée que la formule de calcul diffère d'un critère à un autre, ça se répercute naturellement sur leurs performances, nous pouvons donc déterminer dans quels circonstances il est préférable d'utiliser tel critère et non pas un autre.

1. GAIN D'INFORMATION

   — Utilisé dans l'algorithme ID3
   — Ce critère de division est mieux adapté au ensemble de données équilibré en terme de valeur d'attribut, car il privilégie ceux qui on en beaucoup, ce qui se répercute mal sur l'arbre de décision produit (en terme de précision et de complexité). Exemple : Soit l'ensemble de données suivant :

Attribut 1	Attribut 2	Classe
A	V	C1
B	V	C1
C	V	C1
D	T	C2
E	T	C2

   — En appliquant le Gain d'information, l'attribut 1 sera choisi comme attribut de division, un choix qui donnera en résultat 5 nœuds, alors que l'attribut 2 aurait donné un résultat meilleur (2 nœuds).

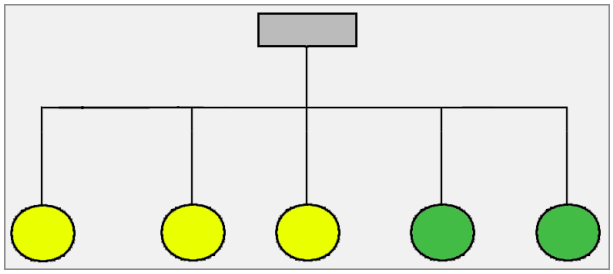

FIGURE 2 – Arbre conçu en utilisant le Gain d'information

   — Les logarithmes utilisés par la formule du gain font améliorer la complexité de ses calculs.

2. Rapport de gain

   — Utilisé dans l'algorithme C4.5
   — Cette mesure est une amélioration du gain d'information, l'idée est de prendre en compte le nombre de valeur des attributs lors du calcul, afin de normaliser le gain d'information.
   Elle est donc meilleure lorsque le nombre de valeur des attribut n'est pas équilibré.
   Avec le même exemple vu précédemment, le rapport de gain choisira l'attribut 2 comme attribut de division, ce qui est meilleur en terme de précision de prédiction ainsi qu'en terme de complexité.
      — Problématique
      Lorsque le nombre de modalité des attributs est équilibré, l'application du rapport de Gain ferait augmenter la complexité du calcul tout en donnant le même résultat que le Gain d'information.

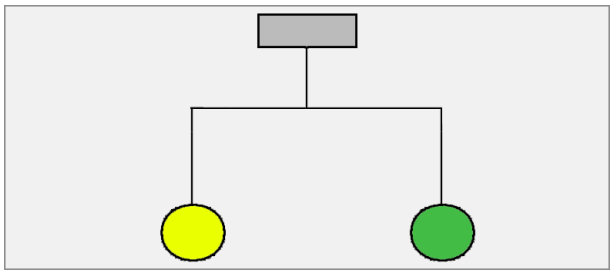

FIGURE 3 – Arbre conçu en utilisant le Rapport du gain d'information

— Solution (une piste pour l'approche à proposer)
Il faudra trouver un rapport $\frac{Nb\ de\ modalité}{Nbdeligne}$, qui, a partir duquel, nous pourrons (de manière approximative) décider lequel des deux critères serait meilleur à appliquer a un ensemble donnée.

3. Indice de Gini
   — Utilisé dans l'algorithme CART
   — La formule de ce critère de division est beaucoup plus légère en terme de complexité de calcul, comparé aux deux précédents, étant donné qu'elle n'utilise pas les logarithmes.

   — Avec ses divisions binaire, la qualité de l'arbre (en terme d'interprétation) est touchée.
   — Ce critère fait la division par rapport a une valeur d'attribut, et non pas par rapport à l'attribut, ce qui revient à calculer l'indice de Gini de toutes les combinaisons possible (VALEUR D'ATTRIBUT, CATÉGORIE DE CLASSE), ce qui augmentera sa complexité lorsque le nombre de modalité ainsi que le nombre de catégorie de classe est grand.
   — Il serai approprié pour un jeu de donnée à attributs binaires.

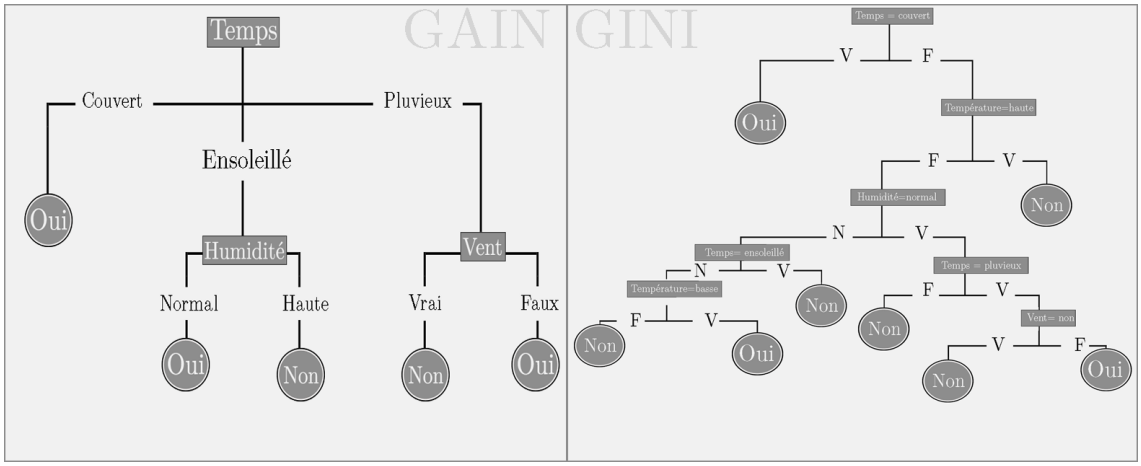

FIGURE 4 – Comparaison entre le Gain et l'indice de GINI en terme de compléxité

La figure précédente montre clairement la différence en terme de facilité d'interpréta-tion, entre un arbre construit en utilisant le Gain d'information, et un autre construit en utilisant l'indice de Gini.

4. Chi-square
   — Utilisé dans l'algorithme CHAID
   — L'utilisation de ce critère est très appropriée pour :

   — De très grands ensembles de données, car le calcul repose sur des estimations, qui, plus l'ensemble est grand, plus elle seront meilleures.

    - Un nombre de modalité d'attribut ainsi que de catégorie de classe réduit. Rappelons que la taille des deux tables sur les quelles base le calcul du X^2 sont de taille $N * M$ ou :
        - N : Le nombre de modalité d'un attribut.
        - M : Le nombre de catégorie de classe.

- Conclusion
  Cette petite étude comparative nous a permis de confirmer que chacun critère de division est meilleur dans une situation particulier, d'où l'importance de prendre les caractéristiques des données d'apprentissage, ainsi que le type de traitement voulu très au sérieux, afin de choisir le critère adéquat.