Université Mohammed-V de Rabat
École Nationale Supérieure d'Informatique et
d'Analyse des Systèmes
ENSIAS

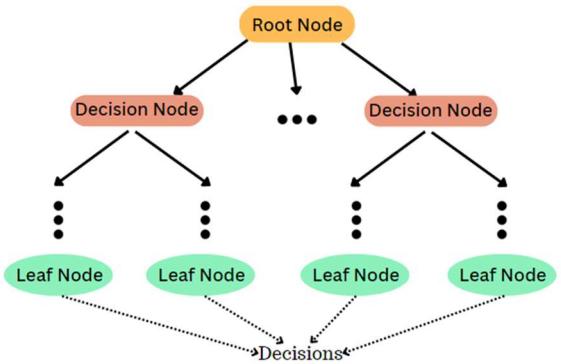
Réalisé par :
ASMAE KARMOUCHI
MOHEMMED AMINE KENDI

Supervised learning : decision tree

2A GL

Introduction :

Un arbre de décision est un algorithme d'apprentissage automatique supervisé utilisé pour des tâches de classification et de régression. C'est un modèle en forme d'arbre où un nœud interne représente une caractéristique ou un attribut, la branche représente une règle de décision, et chaque nœud feuille représente le résultat. Le nœud le plus haut dans un arbre de décision est appelé nœud racine.



Principe de fonctionement de l'algorithme d'arbre de décision:

Feature Selection/ Sélection de la caractéristique : L'algorithme sélectionne la meilleure caractéristique pour diviser l'ensemble de données à chaque nœud en se basant sur certains critères comme l'impureté de Gini, le gain d'information, ou l'entropie. Il choisit la caractéristique qui résulte en des sous-ensembles les plus homogènes.

Splitting/ Division : La caractéristique sélectionnée est utilisée pour diviser l'ensemble de données en deux ou plusieurs sous-ensembles basés sur différentes valeurs de cette caractéristique.

Recursive Partitioning/ Partitionnement récursif : Ce processus continue de façon récursive sur chacun des sous-ensembles résultants jusqu'à ce qu'un critère d'arrêt soit rencontré. Le critère d'arrêt peut être une profondeur d'arbre maximale, un nombre minimum d'échantillons dans un nœud feuille, ou d'autres conditions.

Leaf Node Creation/ Création de nœuds feuilles : Une fois que le critère d'arrêt est rencontré, les sous-ensembles finaux deviennent des nœuds feuilles, et chaque nœud feuille se voit attribuer une étiquette de classe (dans le cas de la classification) ou une valeur numérique (dans le cas de la régression).

Prediction/ Prédiction : Pour classifier une nouvelle instance, on traverse l'arbre de décision depuis le nœud racine jusqu'à un nœud feuille basé sur les valeurs de caractéristique de l'instance, et la classe majoritaire ou la valeur moyenne des instances dans ce nœud feuille est retournée comme prédiction.

Principes théoriques et les formules associées :

Entropie et Gain d'Information :

- L'entropie mesure l'incertitude dans un ensemble de données. Plus l'entropie est élevée, plus l'ensemble de données est hétérogène.
- L'entropie d'un ensemble S est définie comme :
  $H(S) = -\sum_{I} pi \log2(pi)$ Où c est le nombre de classes dans l'ensemble S, et pi est la proportion d'instances de la classe i dans S.
- Le gain d'information mesure la réduction d'entropie obtenue en divisant un ensemble de données en fonction d'un attribut particulier.
- Le gain d'information IG pour un ensemble S et un attribut A est défini comme :
  $IG(S,A) = H(S) - \sum_{v \in A} \frac{|S|}{|S_v|} H(S_v)$ Où $|S|$ est le nombre total d'instances dans S, A est l'ensemble des valeurs possibles de l'attribut A, $|Sv|$ est le nombre d'instances dans S où l'attribut A a la valeur v, et $H(Sv)$ est l'entropie de l'ensemble des instances où l'attribut A a la valeur v.
- L'entropie est particulièrement utile lorsque les classes sont équilibrées dans l'ensemble de données et qu'il n'y a pas de déséquilibre significatif entre les classes.

Choix du meilleur attribut de division :

- Pour chaque nœud de l'arbre, l'algorithme choisit l'attribut qui maximise le gain d'information ou minimise l'entropie.
- Différents critères peuvent être utilisés, comme l'indice de Gini ou le gain d'information.

Construction récursive de l'arbre :

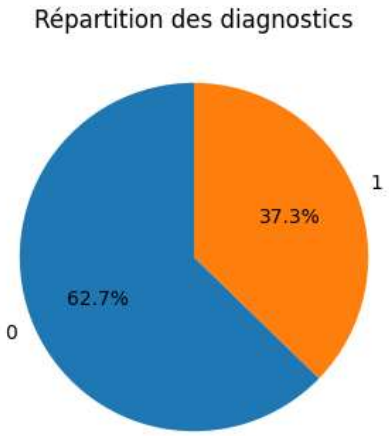
- Une fois l'attribut de division choisi, l'ensemble de données est divisé en sous-ensembles en fonction des valeurs de cet attribut.
- Cette division est répétée récursivement pour chaque sous-ensemble jusqu'à ce qu'un critère d'arrêt soit atteint, tel que la profondeur maximale de l'arbre ou le nombre minimum d'instances dans un nœud feuille.

Élagage (Pruning) :

- Après la construction de l'arbre, il peut être trop complexe et souffrir d'un surajustement (overfitting). L'élagage est une technique pour réduire la complexité de l'arbre en supprimant les branches qui n'apportent pas d'amélioration significative à la performance du modèle sur un ensemble de validation

Jeu de données utilisé:

Le jeu de données **Breast Cancer Wisconsin** (Diagnostic) contient des informations sur des biopsies de tumeurs mammaires. Il comprend 569 instances avec 30 caractéristiques mesurées à partir d'images de cellules biopsiées. L'objectif est de prédire si une tumeur est maligne (cancer) ou bénigne (non-cancéreuse). Les attributs comprennent des mesures de taille, de forme et de texture des noyaux cellulaires. Ce jeu de données est souvent utilisé pour développer des modèles de classification en apprentissage automatique pour le diagnostic du cancer du sein.



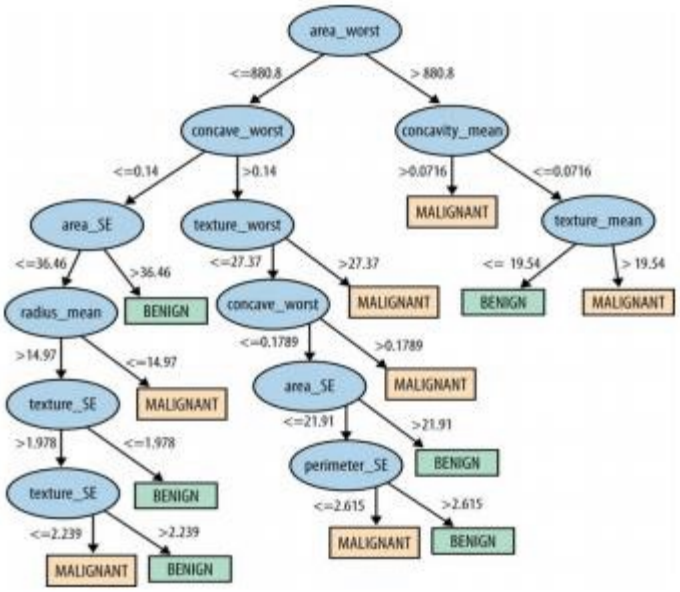
Accuracy : L'accuracy est définie comme le ratio des prédictions correctes sur le nombre total d'instances. Mathématiquement, elle est calculée comme suit :

$$\textbf{Accuracy=Nombre de pre´dictions correctes/Nombre total d'instancesAccuracy}$$

Précision : La précision est définie comme le ratio des vrais positifs (instances positives correctement prédites) sur le nombre total d'instances prédites comme positives. Mathématiquement, elle est calculée comme suit :

$$\textbf{Precision=Vrais positifs/Vrais positifs+Faux positifsPrecision}$$

> Une précision de 0.6228 signifie que notre modèle d'arbre de décision est correct dans environ 62,28% des cas pour prédire si une tumeur mammaire est maligne ou bénigne.


area_worst
<=880.8
>880.8
concave_worst
concavity_mean
<=0.14
>0.14
>0.0716
<=0.0716
area_SE
texture_worst
MALIGNANT
texture_mean
<=36.46
>36.46
<=27.37
>27.37
<= 19.54
> 19.54
radius_mean
BENIGN
concave_worst
MALIGNANT
BENIGN
MALIGNANT
<=0.1789
>0.1789
>14.97
<=14.97
area_SE
MALIGNANT
texture_SE
MALIGNANT
<=21.91
>21.91
>1.978
<=1.978
perimeter_SE
BENIGN
texture_SE
BENIGN
<=2.615
>2.615
<=2.239
>2.239
MALIGNANT
BENIGN
MALIGNANT
BENIGN