Algorithme d'Arbre de Decision

Devoir 9 - INF5099

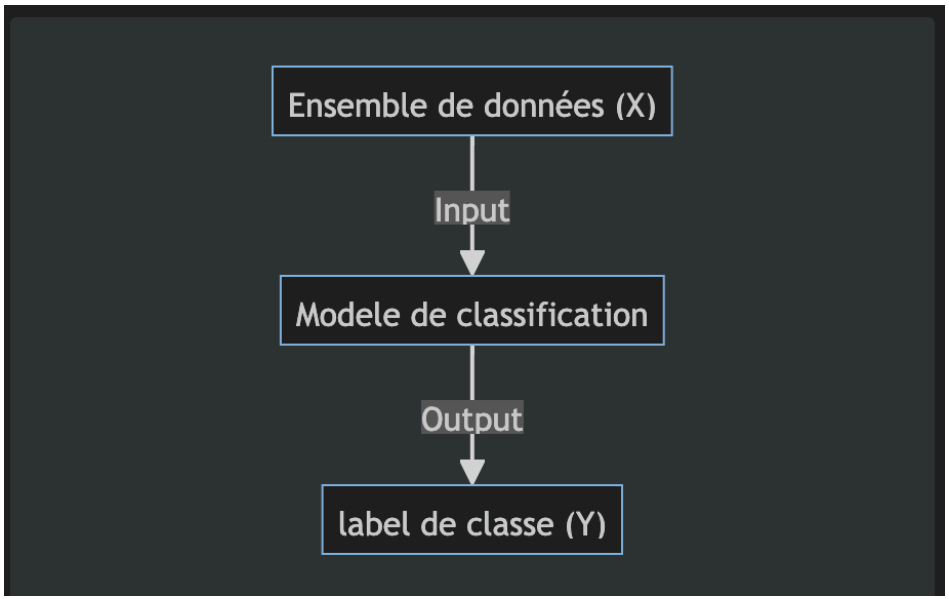
DJIEMBOU TIENTCHEU VICTOR NICO[1]

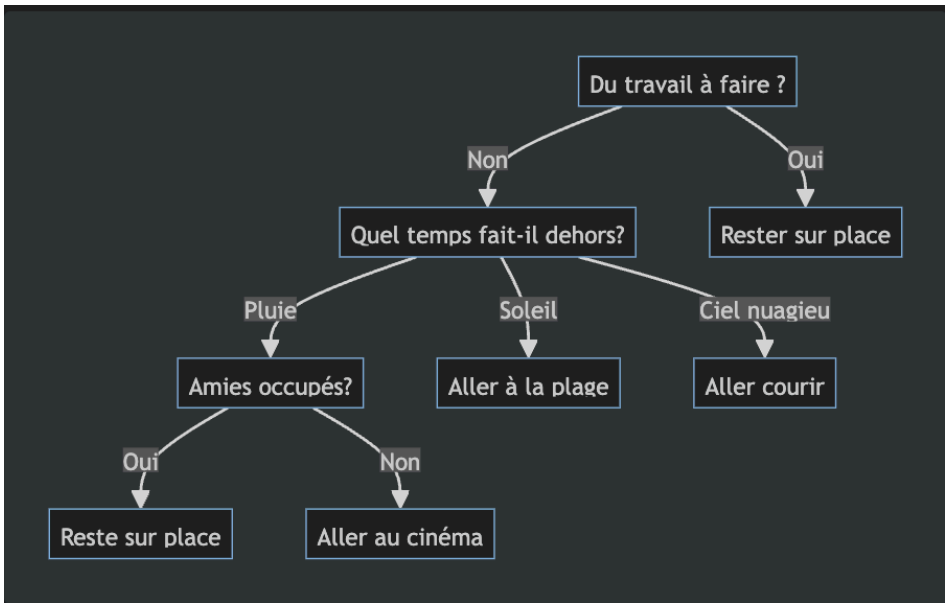
[1] *Université de Yaoundé I,*
Faculté des Sciences,
Département d'Informatique,
Étudiant Master SD

17 mai 2024

Outline

Generalites





Version existantes

Définition

Un arbre de décision est un algorithme qui permet de former un modèle qui se base sur les arbres de décision pour associer une classe ou un catégorie a une entrée.

Noms des versions

- CHAID
- ID3
- C4.5
- CART
- C5

Idée de solution

Tableau caractérisation des arbres de décision

Algorithme d'arbre de décision	Type de données	Méthode de fractionnement des données numériques	Outils possibles
CHAID (CHi-square Automatic Interaction Detector)	Catégorielle	Indéfini	SPSS answer tree
ID3 (Iterative Dichotomiser 3)	Catégorielle	Pas de restriction	WEKA
C4.5	Catégorielle et Numérique	Pas de restriction	WEKA
CART (Classification and Regression Tree)	Catégorielle et Numérique	Séparation binaire	CART 5.0

Idée de solution

Tableau de decription des arbres de décision

Nom de l'algorithme	**Classification**	**Description**
CHAID (CHi-square Automatic Interaction Detector)	Antérieur à l'implémentation originale de l'ID3	Ce type d'arbre de décision est utilisé pour une variable nominale à échelle. La technique détecte la variable dépendante à partir des variables catégorisées d'un ensemble de données
ID3 (Iterative Dichotomiser 3)	Utilise la fonction d'entropie et le gain d'information comme mesures	La seule préoccupation concerne les valeurs discrètes. Par conséquent, l'ensemble de données continues doit être classé dans l'ensemble de données discrètes
C4.5	La version améliorée de l'ID 3	Traite à la fois des données discrètes et continues. Il peut également gérer les données incomplètes
CART (Classification and Regression Tree)	Utilise l'indice de Gini comme mesure	En appliquant la division numérique, nous pouvons construire l'arbre basé sur CART

Outline

Fonctions de l'algorithmes ID3

Opération clefs

- Calcul de l'impurété
  - Entropie : $E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$
  - Indice de Gini : $Gini = 1 - \sum_{i=1}^{n} (P_i)^2$
- Calcul du gain d'information
  - Gain de classification : $Information\ Gain_{Classification} = E(d) \sum \frac{|s|}{|d|} E(s)$
  - Gain de regression :
    $Information\ Gain_{Regression} = Variance(d) \sum \frac{|s|}{|d|} Variance(s)$

Outline