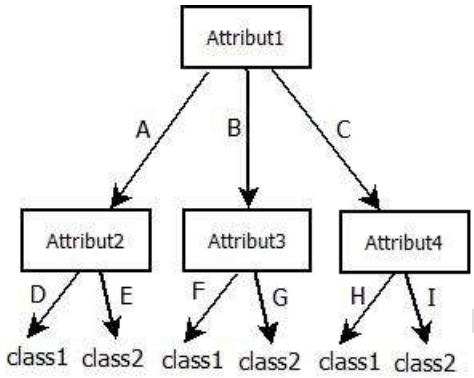
Modules "Datamining" " Fouille & Extraction de données " (M2-IL-ISIA-RFIA)

Fiche n° 3 : Weka & "Arbres de Décisions"

Exercice 1 : Rappelez l'objectif général du modèle des Arbres de Décisions (AD).

Exercice 2 : Quelle différence y-a-t-il entre un AD de *classification* et un AD de *régression* ? .

Exercice 3 : On considère l'AD de classification suivant. Les classes sont (class1 et class2).



1/ Traduire l'arbre sous la forme d'un ensemble de règles.

2/ Transformer l'arbre en un AD binaire.

Exercice 4 : On considère l'AD présenté en cours (diagnostic médical). On dispose d'un échantillon de 200 patients. Dans cet échantillon, 100 sont sains et 100 sont malades. La répartition entre les deux classes H (Healthy) et S (Sick) est donnée par le tableau suivant :

	gorge irritée	*gorge non irritée*
température <= 37	*(0 H, 45 S)*	*(100 H, 0 S)*

1/ Calculez pour chaque position p de l'arbre :

- N(p) : le cardinal de l'ensemble des exemples associé à *p*
- N(k/p) : le cardinal de l'ensemble des exemples associé à *p* qui sont de classe *k*,
- P(k/p) = *N*(*k*/*p*)/*N*(*p*) la proportion d'éléments de classe *k* à la position *p*.

2/ Calculer l'entropie pour chaque position de l'arbre. Que constatez-vous ?

Exercice 5 : On considère les données suivantes représentant les clients d'une banque.

Montant (M)	*Age (A)*	*Résidence (R)*	*Etudes (E)*	*I*
moyen	Adulte	village	oui	oui
élevé	Adulte	campagne	non	non
faible	Agé	campagne	non	non
faible	Adulte	campagne	oui	oui
moyen	Jeune	ville	oui	oui
élevé	Agé	ville	oui	non
moyen	Agé	ville	oui	non
faible	Adulte	village	non	non

M est le montant que possède le client à la banque. A est son âge, et R son lieu de résidence. E est une variable booléenne qui précise si oui ou non le client a poursuivi des études supérieures. I est la classification qu'on veut obtenir : elle est égale à "oui" si le client utilise internet pour accéder à son compte bancaire, et "non" sinon.

1/ Utilisez l'algorithme ID3 pour construire un AD à partir de ces données. Montrez toutes les étapes de calcul.

2/ Vérifiez que l'arbre obtenu donne les bonnes prédictions des données du dataset.

3/ Quelle sera la prédiction de l'arbre pour une nouvelle donnée (M= « moyen », « A= « Jeune », R= « Campagne », E= « Non ») ?.

Exercice 7 : On reprend le dataset de l'exercice 6. Utilisez le programme de construction des AD J48 de Weka. Décrivez l'expérience avec des captures.

Exercice 8 : On reprend le dataset "weather" (de Quinlan, voire Fiche n°1), et on souhaite en construire un AD **binaire**. Utiliser pour cela l'algorithme J48 disponible sous Weka. Décrivez l'expérience avec des captures. D'après le résultat obtenu, peut-on dire qu'i y'a des attributs qui influent, plus que les autres, sur la décision des joueurs de tennis.