TD 1

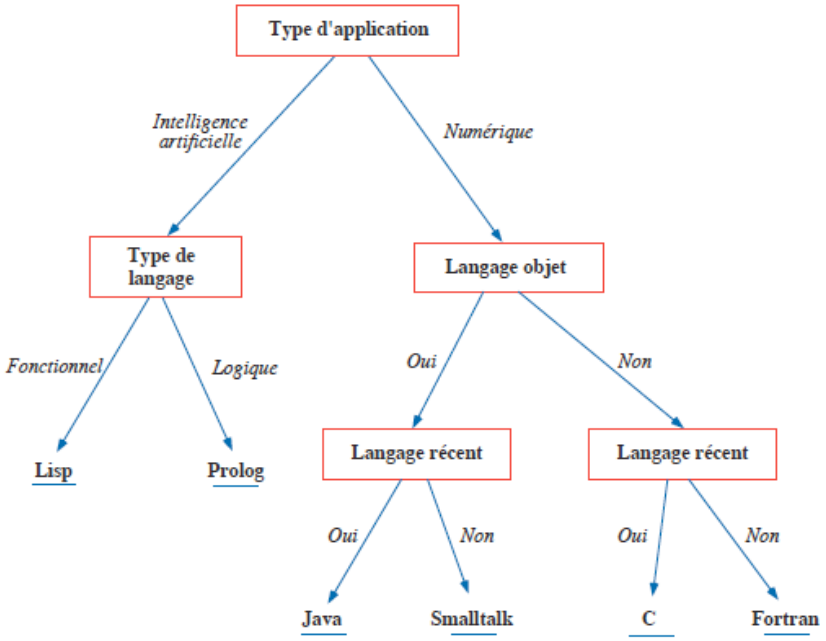
Exercice 1 :

Illustrer à travers un exemple :

1- un problème d'apprentissage supervisé
2- un problème d'apprentissage non supervisé

Exercice 2 :

Soit le schéma suivant :



1- A quoi correspond il ? Donnez une légende pour expliquer les éléments du diagramme.
2- Eliciter les règles de décision découlant de ce schéma.
3- Quel est la hauteur de ce schéma ? Quelle est la taille de schéma ?
4- Donner une structure de données permettant de stocker ce schéma en machine ?
5- Enumérer la description de deux objets (qui ne correspondent pas à des chemins dans l'arbre) et indiquer la classe de chacun d'eux.

Exercice 3 :

Tableau 1 :

	t1	t2	t3	t4	t5
C1	0	1	2	3	4
C2	6	5	4	3	2

Tableau 2 :

classes	t1	t2	t3	t4	t5
C1	0	0	1	2	3
C2	0	3	3	2	3
C3	9	6	5	5	3

Etant donné la répartition des objets dans les nœuds t1, t2, t3, t4 et t5. Pour chaque nœud dans chacun des 2 tableaux 1 et 2, appliquer les formules données en annexe, pour calculer :

1- l'indice de Gini
2- l'entropie
3- l'indice d'erreur en classification

Exercice 4 :

Soit l'exemple suivant comprenant 13 objets, chacun étiqueté dans la classe A ou B, et décrit par trois variables x1, x2 et x3 :

	x_1	x_2	x_3	Classe
1	0	V	N	A
2	1	V	I	A
3	0	F	O	B
4	1	V	N	A
5	1	V	O	A
6	1	F	N	A
7	0	F	O	B
8	0	V	I	A
9	0	F	N	B
10	1	V	I	B
11	1	F	O	A
12	1	F	I	A
13	0	V	O	B

1- Quelle est la nature des variables (attributs) ? (binaire, nominal, ordinal, ou continu, …)
2- Soit l'ensemble d'apprentissage constitué des objets **1 à 9**. On souhaite construire 2 modèles d'arbre de décision t1 et t2. Ces modèles sont aussi appelés « classifieur ».
   a. Construire l'arbre de décision correspondant au classifieur t1, en choisissant les attributs dans l'ordre x3, x2 et x1.
   b. Construire l'arbre de décision, appelé t2, en choisissant les attributs dans l'ordre x1, x2 et x3.
   c. Appliquer les 2 arbres de décision sur les objets **1 à 9**, et indiquer les erreurs commises par chacun des arbres de décision. Il s'agit de **l'erreur en apprentissage**.
   d. Construire la matrice de confusion des 2 modèles avec les objets 1 à 9.
   e. Appliquer les 2 arbres de décision sur les objets **10 à 13**, et indiquer les erreurs commises par chacun des arbres de décision. Il s'agit de **l'erreur en généralisation**.
   f. Construire la matrice de confusion des 2 modèles avec les objets 10 à 13.
3- On décide de rajouter l'objet 10 dans l'ensemble d'apprentissage : objets **1 à 10**
   a. Peut-on trouver un arbre de décision d'erreur apparente nulle durant l'apprentissage ? Pourquoi ?
      i. Si oui, proposer cet arbre ?
      ii. Si non, comment peut-on résoudre le problème mentionné ?
4- Soit l'ensemble d'apprentissage constitué des objets **1 à 9**, et l'ensemble de test constitué des objets **11 à 13**. Soit l'arbre de décision t3 qui ne comporte aucun test et attribue la classe de plus forte probabilité a priori estimée sur l'échantillon. Soit t4 l'arbre de décision constitué d'un seul test sur x1 et de deux feuilles. Ces deux feuilles portent l'étiquette de la classe majoritaire dans le sous-ensemble qu'elles définissent.
   a. Calculer l'erreur apparente *sur l'ensemble d'apprentissage* pour t3 et t4.
   b. Calculer l'erreur apparente *sur l'ensemble test* pour t3 et t4.
   c. Calculer l'erreur apparente sur *l'échantillon complet* pour t1, t2, t3 et t4.
5- On souhaite utiliser les mesures de qualité pour choisir les variables. Trois mesures ont été vues en cours : l'indice de Gini, l'entropie et l'indice d'erreur en classification, dont les formules sont ici rappelées (voir annexe).
   a. On souhaite déterminer la racine de l'arbre de décision.
      i. Calculer l'indice de Gini sur l'ensemble d'apprentissage
      ii. Calculer l'indice d'entropie sur l'ensemble d'apprentissage
      iii. Calculer l'indice d'erreur en classification sur l'ensemble d'apprentissage.
      iv. Pour chacune des variables, calculer le gain avec chacune des mesures :

1. L'indice de Gini
2. L'entropie
3. L'indice d'erreur en classification

v. Quel attribut choisira-t-on avec chacune de ses mesures ?

b. Lorsqu'un nœud-fils de l'arbre n'est pas pur, on décide de poursuivre son partitionnement en utilisant **les attributs non choisis**.
   i. Reprendre les questions 5.a) sur les nœuds-fils obtenus après de le choix de la racine de l'arbre.
   ii. Itérer le processus sur chacun des nœuds-fils obtenus de nouveau et qui ne sont pas purs, pour construire un arbre de décision avec chacune des mesures.

c. Soit t5, t6 et t7, les trois arbres obtenus respectivement avec l'indice de Gini, l'entropie et l'indice d'erreur en classification. Pour chacun d'eux :
   i. Calculer l'erreur apparente sur l'ensemble d'apprentissage
   ii. Calculer l'erreur apparente sur l'ensemble de test
   iii. Calculer l'erreur apparente sur l'échantillon complet
   iv. Proposer leur matrice de confusion avec l'ensemble de test
   v. Calculer le taux de précision, le rappel, la f-mesure pour chacun des classifieurs t5 à t7.
   vi. Proposer les règles de décision associées à chacun des modèles obtenus.

Exercice 5 : Arbres non binaires

On s'intéresse aux arbres de décision non binaires sur des variables réelles. D'un nœud partent donc q branches, avec $q \geq 2$, portant des tests du type : $(x_i \leq a1)$, $(a1 \leq x_i \leq a2)$, … , $(x_i \geq a_{q-1})$

1- Montrer que pour tout arbre de décision de ce type, avec éventuellement un nombre de branches différent à chaque nœud, il existe un arbre de décision binaire réalisant la même classification.
2- Soit un arbre de décision non binaire ayant un seul nœud (la racine) et q feuilles, avec $q > 2$. Sa hauteur est définie comme la longueur maximale en nombre de nœuds entre la racine et une feuille. Quelle est la hauteur la plus grande d'un arbre binaire réalisant la même classification ? La hauteur minimale ?
3- Quel est le nombre minimal et maximal de nœuds de cet arbre binaire ?

Exercice 6 : Arbres avec attributs numériques

		A1	A2	A3		Classe
1		1.	1.	F		+
2		2.	2.	F		+
3		3.	0.	F		-
4		4.	1.	F		-
5		3.	3.	V		+
6		2.	4.	F		+
7		4.	2.	F		-
8		5.	3.	F		-
9		4.	4.	V		+

Soit l'ensemble d'apprentissage ci-dessus composé d'exemples (5) et de contre-exemples (4). Les attributs A1 et A2 sont numériques et l'attribut A3 est binaire.
Construire un arbre de décision sur ces données d'apprentissage.

Annexes :

p(j / t) est la fréquence relative de la classe j au nœud t.

1.1- Indice de Gini pour le nœud t :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

1.2- Indice de Gini pour l'attribut *split*

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

1.2- Gain d'information avec l'indice de Gini pour l'attribut *split*

$$Gain_{split} = Gini(r) - Gini_{split}$$

Le nœud parent **r** a n objets, et est divisé en k partitions. La partition i possède n_i objets.

2.1- Entropie du nœud t:

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

2.2- Gain d'information avec l'entropie pour l'attribut *split*:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Le nœud parent **p** a n objets, et est partitionné en k partitions. La partition i possède n_i objets.

3.1- Indice d'Erreur en classification au nœud t

$$Error(t) = 1 - \max_i P(i \mid t)$$

3.2- Gain d'information avec l'indice d'erreur en classification

$$Gain_{split} = Error(r) - Error_{split}$$

Le nœud parent **r** a n objets, et est partitionné en k partitions. La partition i possède n_i objets.