

Module: Data Mining

**Travail Dirigé N° 1**

**Exercice 1**

Soit une matrice de confusion ci-dessous:

	X	Y	Z	E
X	156	12	0	19
Y	23	173	16	7
Z	14	0	97	9
E	2	15	12	89

1. Donner les valeurs de TP, FP, FN et TN.
2. Calculer les performances d'évaluation suivantes: accuracy, detection rate, precision and specificity.
3. Déduire le coefficient F1-Score.

**Exercice 2**

Soit le Data set ci-dessous :

Att1	Att2	Att3	Att4	Att5	class
125	12	3260	25	243	OUI
113	15	3144	31	235	NON
131	16	3055	12	239	NON
121	17	3800	16	310	OUI
109	18	3456	17	325	OUI
129	14	3250	19	267	???

1. Est-il étiqueté ou non? Justifier votre réponse.
2. Normaliser le Data Set en utilisant la technique MinMax.
3. Prédire la classe de la dernière instance en utilisant l'algorithme K-NN avec K=3.

**Exercice 3**

Soit un le Data Set :

A	B	C	D	E	F	G	H	Calss
A1	B2	C1	D2	E2	F2	G2	H1	L1
A2	B3	C2	D2	E2	F2	G2	H1	L1
A4	B3	C2	D3	E1	F1	G1	H3	L2
A2	B2	C2	D1	E1	F1	G1	H3	L3
A1	B1	C3	D1	E1	F1	G1	H3	L2
A2	B2	C1	D1	E3	F3	G3	H3	L3
A1	B3	C1	D2	E3	F3	G3	H2	L2
A3	B2	C2	D1	E3	F2	G2	H2	L1

A3	B2	C1	D3	E1	F1	G2	H2	L1
A2	B3	C2	D3	E2	F3	G1	H3	L2
A1	B3	C3	D1	E1	F3	G1	H1	L1
A3	B1	C3	D1	E2	F2	G3	H2	L2

1. En utilisant l'algorithme Naïve Bayes, calculer la probabilité a priori pour chaque classe.
2. Prédire la classe de l'instance suivante :

A2	B1	C3	D1	E2	F3	G1	H2	??????
----	----	----	----	----	----	----	----	--------

## Exercice 4

Soit l'algorithme k-means pour une segmentation de données:

	Att1	Att2	Att3
A	2,5	12,3	11
B	5,2	13	7
C	3,2	10	8,5
D	4,5	8	7,5
E	5	7,25	7
F	3,7	6	7,3
G	4	5,7	6

1. Quelles sont les qualités d'un bon clustering ?
2. Quelles sont les recommandations générales pour le choix du k ?
3. Calculer les itérations nécessaires pour mettre en place un clustering avec K=3 (B, E et G sont des centres initiaux).

## Exercice 5

En utilisant la méthode hiérarchique agglomérative pour clustering. On a les distances inter-instances suivantes:

- Dist(5, 6)=16; Dist(1, 2)=13; Dist(1, 4)=9; Dist(1, 3)=5; Dist(1, 5)=11; Dist(2, 3)=4; Dist(2, 4)=14; Dist(1, 6)=3; Dist(2, 5)=7; Dist(3, 4)=8; Dist(3, 5)=10; Dist(3, 6)=15; Dist(4, 5)=6; Dist(2, 6)=12 ; Dist(4, 6) = 17.
1. Proposer une matrice de distance adéquate.
  2. Trouver les divers clusters en élaborant les matrices obtenues dans chaque étape.
  3. Dessiner le Dendrogramme correspondant.