Université Mohamed Boudiaf M'sila **M2 SIGL**
Faculté MI, Dépt Informatique **M'sila, Le 15/02/2022**

Examen Final

Data Mining et Recherche d'Information

Exercice 1 (4 points)

Soit la grandeur suivante : $E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$

a) Dites quel algorithme utilise cette grandeur ?
b) Que signifient k, C_i, p et m_i ?
c) Est-ce que l'algorithme utilisant cette grandeur tend à la réduire ou l'augmenter ?

Exercice 2 (8 points)

Supposons qu'il y a 10 objets numérotés de 1 à 10 et, 10 paniers numérotés également de 1 à 10. L'objet *i* est dans le panier *b* si et seulement si *i divise b* (division entière). D'où, l'objet 1 doit être dans tous les paniers, l'objet 2 est dans la moitié des paniers ayant un numéro pair, …etc. Le panier 10 contient les objets {1, 2, 5, 10} c.-à-d. tous les entiers qui divisent le nombre 10.

(a) Compléter le tableau suivant qui contient les objets de tous les paniers.

Panier	Objets
1	1
2	1, 2
…	…
10	1, 2, 5, 10

(b) Si le support est égal à 3, donner tous les objets fréquents
(c) Appliquer l'algorithme A-Priori pour trouver tous les k-itemsets fréquents.

Exercice 3 (8 points)

Les données ci-contre sont utilisées pour l'apprentissage d'un arbre de décision afin de prédire si un plat est délicieux ou non selon sa texture, sa couleur et son odeur.

(a) Quel attribut peut-on choisir comme racine de l'arbre si on utilise le Gain en information.

(b) Construire l'arbre de décision complet à partir de ces données.

(c) Soit l'ensemble de test suivant. Déterminer le taux d'erreur de classification de l'arbre.

Texture	Couleur	Odeur	Délicieux
C	B	2	No
D	B	2	Yes
C	W	2	Yes

Texture	Couleur	Odeur	Délicieux
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

BONNE REUSSITE

Dr. Tahar Mehenni

Université Mohamed Boudiaf M'sila **M2 SIGL**
Faculté MI, Dépt Informatique **M'sila, Le 15/02/2022**

Examen Final (Correction)

Data Mining et Recherche d'Information

Exercice 1 (4 points)

Soit la grandeur suivante : $E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$

a) L'algorithme Kmeans **(1 pt)**
b) k : nombre de clusters, C_i : un cluster i, p : objet du cluster C_i et m_i : centre du cluster C_i **(0.5 pt x 4)**
c) E détermine le total des distances qui existent entre les différents objets p des clusters et leur centre m_i. Kmeans tend à réduire E pour que le clustering soit efficace. **(1 pt)**

Exercice 2 (8 points)

2 pts

1	1
2	1, 2
3	1, 3
4	1, 2, 4
5	1, 5
6	1, 2, 3, 6
7	1, 7
8	1, 2, 4, 8
9	1, 3, 9
10	1, 2, 5, 10

2 pts

1	**10**
2	**5**
3	**3**
4	2
5	2
6	1
7	1
8	1
9	1
10	1

	1	**2**	**3**
1			
2	**5**		
3	**3**	1	

1.5 pts

	21	**31**
21		
31	1	

1.5 pts

Les k-itemsets fréquents sont : (1), (2), (3), (1,2), (1,3) **1 pt**

Exercice 3 (8 points)

Texture	Couleur	Odeur	Délicieux
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

n= 11 E= 0.994030211 **0.5 pt**
C1= 5
C2= 6

Texture	Yes	No	Somme	ni/n		Entropie	Somme	Gain
D	3	3	6	0.55	1.00000	0.545455	0.9868	0.00723
C	2	3	5	0.45	0.97095	0.441341		

0.5 pt

Couleur	Yes	No	Somme	ni/n		Entropie	Somme	Gain
B	3	2	5	0.45	0.97095	0.441341	0.80498	0.18905
W	2	2	4	0.36	1.00000	0.363636		
G	0	1	1	0.09	0.00000	0		
U	0	1	1	0.09	0.00000	0		

0.5 pt

Odeur	Yes	No	Somme	ni/n		Entropie	Somme	Gain
1	3	0	3	0.27	0.00000	0	0.44134	0.55269
2	2	3	5	0.45	0.97095	0.441341		
3	0	3	3	0.27	0.00000	0		

0.5 pt

Attribut choisi comme racine:Odeur E(Odeur)= 0.552689

Pour Odeur = 1 → Délicieux = Yes
Pour Odeur = 3 → Délicieux = No
Pour Odeur = 2 → ??

0.5 pt

Texture	Couleur	Odeur	Délicieux
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No

n= 5
C1= 2
C2= 3
E= 0.97095

0.5 pt

Texture	Yes	No	Somme	ni/n		Entropie	Somme	Gain
D	1	2	3	0.60	0.91830	0.550978	0.95098	0.01997
C	1	1	2	0.40	1.00000	0.4		

0.5 pt

Couleur	Yes	No	Somme	ni/n		Entropie	Somme	Gain
B	1	1	2	0.18	1.00000	0.181818	0.18182	0.81221
W	1	0	1	0.09	0.00000	0		
G	0	1	1	0.09	0.00000	0		
U	0	1	1	0.09	0.00000	0		

0.5 pt

Attribut choisi pour le branchement: **Couleur**
Couleur= W → Délicieux=Yes
Couleur= G → Déliciciex=No
Couleur= U → Déliciciex=No
Couleur= B → ??

0.5 pt

Texture	Couleur	Odeur	Délicieux
C	B	2	Yes
D	B	2	No

n= 2
C1= 1
C2= 1
E= 1

0.25

Texture	Yes	No	Somme	ni/n		Entropie	Somme	Gain
C	1	0	1	0.20	0.00000	0	0	1
D	0	1	1	0.20	0.00000	0		

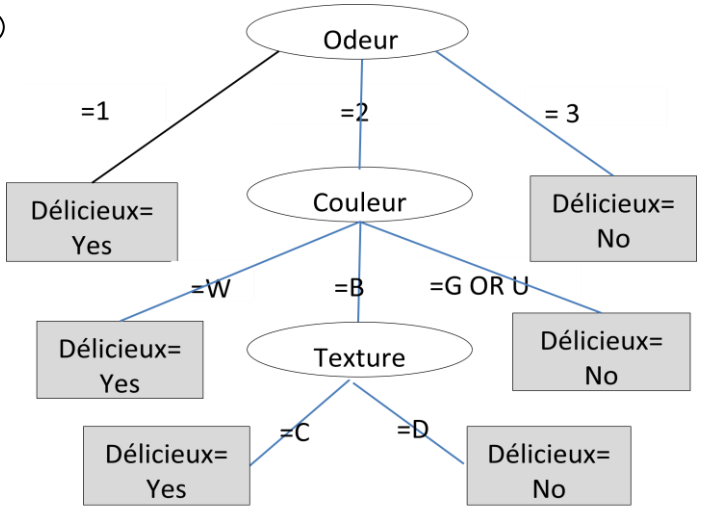
0.25

Attribut choisi pour le branchement: **Texture**

Texture= C→ Délicieux=Yes
Texture= D → Délicieux=No

0.25

Arbre de décision **1 pt**



Texture	Couleur	Odeur	Délicieux	Délicieux (Prédiction)
C	B	2	No	Yes
D	B	2	Yes	No
C	W	2	Yes	Yes

0.75

Taux d'erreur=(Faux Négatifs+Faux positifs)/Total= (1+1)/3=0.66 (66%) **1 pt**