USTHB
Faculté d'Informatique
Département d'Intelligence Artificielle
Et des sciences de données
Master S2I, 2ème année
Module : Data mining

4 octobre 2022

TD 1-2

Exercice 1

Considérer les données suivantes de l'attribut *âge* écrites par ordre croissant :

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

1) Quelle est la *moyenne* de ces données ? Quelle est la *médiane* ?
2) Quel est le *mode* des données ? Commenter la modalité des données (i.e., bimodales, trimodales, etc.).
3) Pouvez-vous déterminer (approximativement) le premier quartile ($Q1$) et le troisième quartile ($Q3$) des données ?
4) Donner le *résumé des cinq nombres* des données.
5) Dessiner un *boxplot* des données.
6) quelle est la différence entre un *quantile-quantile plot* et un *quantile plot* ?

Exercice 2

Supposer que dans un hôpital, on ait effectué des tests sur le taux des graisses de 18 adultes sélectionnés au hasard et dont les résultats accompagnés de l'âge sont comme suit :

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculer la moyenne, la médiane et l'écart-type de l'*âge* et du taux de graisse *%fat*.
(b) Dessiner les boxplots pour *âge* and *%fat*.
(c) Dessiner un *scatter plot*.
(d) Normaliser les deux variables avec la technique du *z-score*.

(e) Calculer le *coefficient de corrélation* (Pearson's product moment coefficient). Ces deux variables sont-elles corrélées positivement, négativement ou pas du tout?

Exercice 3

Considérer les notes suivantes d'un groupe de 10 étudiants pours les cours respectifs de 'datamining' et de 'méta-heuristiques' :

	Datamining				**Méta-heuristiques**			
	TP1/10	**Test1/10**	**EMD1/20**	**Final1/20**	**TP2/10**	**Test2/10**	**EMD2/20**	**Final2/20**
1	7,75	5	13,5	13,13	7,25	6,50	16	14,88
2	6	6	11	11,50	7,75	5,50	11,5	12,38
3	6	4	7,5	8,75	7,87	6,25	10	12,06
4	6,17	4	5,5	7,84	7,75	4,00	11,5	11,44
5	8	5	11	12,00	6,62	6,50	13	13,63
6	7,75	2	10,5	10,13	7,50	7,50	12,5	13,31
7	6	4	5,5	7,75	7,37	3,50	12,5	11,75
8	6,75	5	9	10,38	7,62	7,50	5,5	10,19
9	5,67	6	6	8,84	8,37	4,00	8,5	10,06
10	6,5	6	9	10,75	7,37	5,50	17	15,44

La note 'Final' est obtenue comme suit :
Final = (TP+Test+EMD)/2

1) A-t-on besoin de normaliser les notes de TP, Test et EMD pour calculer la note 'Final' ? Pourquoi ? Si oui, procéder à la normalisation des données.
2) Que représente concrètement 'Final' par rapport aux notes de TP, Test et EMD ?
3) Pour chacune des notes Final1 et Final2, Calculer :
    a. La moyenne
    b. La médiane
    c. Le mode

   Que peut-on en conclure?
    d. Calculer approximativement le premier et le troisième quartiles
    e. Donner le résumé des cinq nombres pour ces données.
    f. Dessiner une *boxplot* pour ces données.
4) Dessiner un *q-q plot* pour les deux variables Final1 et Final2. Que peut-on en conclure ?
5) Calculer le *coefficient de corrélation* (Pearson's product moment coefficient) pour les deux variables. Que peut-on en conclure?

Exercice 4

L'analyse de données et le data mining sont des technologies transversales très utiles ces dernières années à cause de la présence des données dans quasiment toute entreprise.

1) Expliquer clairement la différence entre l'analyse de données et le datamining.
2) Quelles sont les fonctionalités du datamining ? Citer au moins un algorithme de datamining pour chacune des fonctionalités.
3) Enumérer les différentes étapes d'un système de data mining orienté intelligence économique ou business intelligence.
4) Considérer le domaine de la finance et plus particulièrement l'analyse et la gestion financière.

a. Imaginer au moins trois sources de données pour une telle application. A-t-on besoin d'analyse de données ou de datamining pour traiter les données ? pourquoi ?
b. Imaginer trois types de fonctionalités pour le traitement des données de l'application et donner une technique pour chacune d'elles.
c. Considérer la requête suivante de datamining exprimée dans le langage DMQL (Data Mining Query Langage) :

use database AllElectronics-db
use hierarchy location-hierarchy for T.branch, age-hierarchy for C.age
mine classification as promising-customers
in relevance to C.age, C.income, I.type, I.place-made, T.branch
from customer C, item I, transaction T
where I.item-ID = T.item-ID and C.cust-ID = T.cust-ID
and C.income >= 40,000 and I.price >= 100
group by T.cust-ID
having sum(I.price)>= 1,000
display as rules

Quels sont les resultats que la requête doit rendre ? expliquer.

Exercice 5

Supposer qu'un groupe de 18 étudiants ont obtenu les notes suivantes dans deux examens de cours différents :

Note1	2	2	4	4	4	7	8	8	10
Note2	4	5	6	7	7	9	10	19	11
Note1	10	12	12	12	13	14	14	15	17
Note2	8	11	15	15	16	16	16	19	19

1) Calculer la moyenne, la médiane et l'écart-type pour les deux notes.
2) Dessiner les boites à moustache pour *les deux notes. Que pouvez-vous conclure ?*
3) Dessiner un *histogramme* pour représenter ces données.
4) Calculer le *coefficient de corrélation* de Pearson. Ces deux variables sont-elles corrélées ?
5) Comment peut-on réduire ces données ? Donner le résultat de la réduction.

Exercice 6

Considérer les attributs *longueur du sépale en cm* et *longueur de la pétale en cm* du dataset Iris. La table ci-dessous exhibe 12 instances du dataset pour les deux attributs apparaissant respectivement en deuxième et troisième colonnes.

Instance	**Sépale**	**Pétale**
1	4.9	1.4
2	5.0	1.4
3	5.4	1.7
4	4.6	1.4

5	5.5	4.0
6	5.1	3.0
7	5.7	4.5
8	5.0	3.3
9	4.9	4.5
10	5.7	5.0
11	5.8	5.1
12	5.6	4.9

1) Dessiner les boites à moustaches pour chacun des attributs Sépale et Pétale. Que pouvez-vous conclure ?
2) Appliquer l'algorithme *Chimerge* ci-dessous pour discrétiser l'attribut Sépale.

Algorithme ChiMerge
1. trier les valeurs de l'attribut par ordre croissant.
2. considérer chaque valeur dans un intervalle distinct.
3. calculer la valeur de χ^2 pour tous les intervalles adjacents.
4. fusionner les paires d'intervalles qui ont la plus petite valeur de χ^2.
5. arrêter le processus quand le nombre d'intervalles est égal à 4 sinon aller à (3).

La formule du χ^2 est donnée comme suit:
$\chi^2 = \sum_{i=1}^{i=m} \frac{(R_i - E)^2}{E}$

où:
m est le nombre d'intervalles à comparer (2 dans ce cas),
R_i est le nombre de valeurs de l'intervalle *i*,
E est la fréquence moyenne calculée comme: $E = n/MaxIntervalles$,
n est le nombre total de valeurs,
MaxIntervalles est le nombre maximum d'intervalles.

Exercice 7
Considérer les attributs *longueur du sépal en cm* et *longueur de la pétale en cm* du dataset Iris. La table ci-dessous exhibe les 20 premières entrées du dataset pour les deux attributs apparaissant respectivement en deuxième et troisième colonnes.

Instance	Sépal	Pétale
1	5.1	1.4
2	4.9	1.4
3	4.7	1.3
4	4.6	1.5
5	5.0	1.4
6	5.4	1.7
7	4.6	1.4
8	5.0	1.5
9	4.4	1.4
10	4.9	1.5
11	5.4	1.5
12	4.8	1.6
13	4.8	1.4
14	4.3	1.1
15	5.8	1.2
16	5.7	1.5
17	5.4	1.3
18	5.1	1.4
19	5.7	1.7
20	5.1	1.5

1) Calculer la moyenne, la médiane et le mode pour chacun des deux attributs. Que pouvez-vous conclure ?
2) Dessiner les boites à moustaches pour les deux attributs. Que pouvez-vous conclure ?
3) Dessiner un histogramme pour représenter ces données. Qu'observez-vous ?
4) Donner le diagramme de dispersion des deux attributs (q-q plot et scatter plot). Que pouvez-vous conclure ?
5) Calculer le coefficient de corrélation pour les deux variables. Que pouvez-vous conclure ?
6) Les 20 instances se divisent en 2 catégories :
   - ✓ Celles qui dépassent la moyenne pour la longueur du sépal.
   - ✓ Celles qui dépassent la moyenne pour la longueur de la pétale.

Calculer le χ^2 pour ces deux sous-groupes d'instances. Que pouvez-vous conclure ?

7) Comment peut-on réduire ces données ? Donner le résultat de la réduction.

Exercice 8

Considérer les 10 premières instances suivantes du dataset Heart.

Age	sex	chest pain	blood pressure	cholestoral	FBS > 120 mg/dl	ECG (0,1,2)	heart rate	Exercise angina	oldpeak	ST	vessels (0-3)	thal	
70	1	4	130	322	0	2	109	0	2.4	2	3	3	present
67	0	3	115	564	0	2	160	0	1.6	2	0	7	absent
57	1	2	124	261	0	0	141	0	0.3	1	0	7	present
64	1	4	128	263	0	0	105	1	0.2	2	1	7	absent
74	0	2	120	269	0	2	121	1	0.2	1	1	3	absent
65	1	4	120	177	0	0	140	0	0.4	1	0	7	absent
56	1	3	130	256	1	2	142	1	0.6	2	1	6	present
59	1	4	110	239	0	2	142	1	1.2	2	1	7	present
60	1	4	140	293	0	2	170	0	1.2	2	2	7	present
63	0	4	150	407	0	2	154	0	4	2	3	7	present

1) Calculer le coefficient de corrélation entre l'attribut ***sex*** (2è colonne) et l'attribut ***fasting blood sugar* or FBS** (colonne 6). Que peut-on en déduire ?
2) Calculer le coefficient de corrélation entre l'attribut âge (1ère colonne) et l'attribut ***serum cholestoral*** (5ème colonne).
3) Dessiner le scatter plot de l'attribut ***âge*** et de l'attribut ***serum Cholestoral***. Que peut-on en conclure ?
4) Dessiner la boite à moustache de l'attribut ***serum cholestoral***. Que peut-on en conclure ?