EXAMEN

Semestre : 1

Session : Principale

Module: **DATA MINING**

Enseignantes: **Amal Tarifa, Hiba Lahmer, Wiem TRABELSI**

Classes : **4GL**

Nombre de pages : ***3 pages***

Documents **NON** autorisés, Calculatrice autorisée, Internet **NON** autorisée.

Date : 08/01/2020 *Heure : 13h00* *Durée : 01h30mn*

Exercice 1 : Questions de cours [3pts]

1-Donner les différentes étapes du processus CRISP-DM. Expliquer brièvement la phase de modélisation.

Crisp DM :

Compréhension du métier : Cette première phase est essentielle et doit permettre de **comprendre les objectifs et les besoins métiers** afin de les intégrer dans la définition du projet DM et de décliner un plan permettant de les atteindre et les satisfaire.

Compréhension des données : Il s'agit de collecter et de **se familiariser avec les données à disposition**. Il faut également identifier le plus tôt possible les **problèmes de qualité** des données, développer les premières intuitions, détecter les premiers ensembles et hypothèses à analyser.

Préparation des données : Cette phase comprend toutes les étapes permettant de construire le jeu de données qui sera utilisé par le(s) modèle(s). Ces étapes sont souvent exécutées plusieurs fois, en fonction du modèle proposé et du retour des analyses déjà effectuées. Il s'agit entre autres d'**extraire, transformer, mettre en forme, nettoyer et de stocker de façon pertinente les données.** La préparation des données peut constituer environ **60 à 70% du travail total**.

Phase de modélisation : C'est ici qu'entrent en jeu les méthodologies de modélisation issues notamment de la statistique. Les modèles sont souvent validés et construits avec l'**aide d'analystes du côté métier et d'experts en méthodes quantitatives**. Il y a dans la plupart des cas plusieurs façons de modéliser le même problème de DM et plusieurs techniques pour arriver à ajuster au mieux un modèle aux données. La boucle de feedback vers les points précédents est fréquemment utilisée pour améliorer le modèle.

Évaluation du modèle : Une fois arrivés à cette phase, un ou plusieurs modèles sont construits. Il faut s'assurer que les **résultats** sont jugés **satisfaisants** et sont **cohérents** notamment vis-à-vis des objectifs métier.

Déploiement : Cela peut être aussi simple que de fournir une synthèse descriptive des données ou aussi complexe que de mettre en œuvre un processus complet de fouille de données pour l'utilisateur métier final.

2- Quelle est la différence entre apprentissage supervisé et apprentissage non supervisé ? [0.5pt]

L'apprentissage supervisé est distingué par des données labellisées. L'algorithme va essayer d'apprendre un modèle convenable par rapport aux données fournies et d'ajuster à chaque fois ce modèle lors de la phase d'apprentissage jusqu'à l'obtention de résultat pertinentes (phase de test).

L'apprentissage non supervisé essaye de détecter les patterns relatifs aux données afin de créer des groupes d'observations qui sont les plus similaires que possibles en s'assurant de l'éloignement des différents clusters ou groupes créés.

3- Dans le cas d'une classification, qu'elle est la métrique utilisée pour mesurer la qualité du modèle obtenu ? Expliquer par un tableau. [0.5pt]

Recall : ce qui a été prédit convenablement dans classe (i)/ le nombre d'observations que contient cette classe (i)

Précision: ce qui a été prédit convenablement dans classe (i)/ ce qui a été prédit dans la classe (i)

F1-score : 2*(précision* recall)/(precision+recall)

4- Pour le cas d'une segmentation comment peut-on mesurer la qualité des segments obtenu ? [0.5pt]

Il faut s'assurer de minimiser la distance intra-cluster et de maximiser la distance inter-cluster. Pour ce faire, il existe plusieurs méthodes :

elbow method : la méthode elbow qui calcule la distance intra-cluster, donc plus on minimise plus on optimise. Pour le graphe on retient le premier coude.

5- Nous disposons d'un dataset de petite de taille pour faire une classification. Quel algorithme choisir pour obtenir un bon résultat ; justifier votre choix. [0.5pt]

SVM car il se base uniquement sur les vecteurs de support et ignore les autres observations,

KNN car il ne nécessite pas un apprentissage, il est l'algorithme le plus paresseux.

Exercice 2 : Régression linéaire multiple [6 pts]

Une école a élaboré un suivit de 116 étudiants. Ce suivi comprend 6 sessions. Pour chaque session, on récupère la note obtenue pour chaque étudiant. La figure ci-dessous illustre quelques lignes obtenues.

Student ID	Session 2	Session 3	Session 4	Session 5	Session 6
1	5.0	0.0	4.5	4.0	2.25
2	4.0	3.5	4.5	4.0	1.00
3	3.5	3.5	4.5	4.0	0.00
4	6.0	4.0	5.0	3.5	2.75
5	5.0	4.0	5.0	4.0	2.75
6	5.5	3.5	4.5	3.0	3.00
7	4.0	4.0	4.5	4.0	2.00
8	4.0	3.5	4.0	3.0	0.00
9	0.0	3.0	4.0	4.0	2.00

Tableau 1:extrait du Dataset utilsé

On souhaite prédire pour un nouvel étudiant la note qu'il aura au cours de la session 6. Pour ce faire, on élabore une régression linéaire qui donne le résultat suivant :

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.884223   0.370616  -2.386  0.01877 *
`Student Id` 0.004578   0.002811   1.628  0.10632
`Session 2`  0.046934   0.050969   0.921  0.35917
`Session 3`  0.210853   0.069771   3.022  0.00313 **
`Session 4`  0.222271   0.072598   3.062  0.00277 **
`Session 5`  0.281214   0.069391   4.053 9.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8374 on 109 degrees of freedom
Multiple R-squared:  0.5494,    Adjusted R-squared:  0.5287
F-statistic: 26.58 on 5 and 109 DF,  p-value: < 2.2e-16
```

1- Définir l'algorithme de régression linéaire utilisé et expliquer son fonctionnement [1,5pt]

Il s'agit d'utiliser l'algorithme de régression linéaire qui permet d'identifier les paramètres de la fonction linéaire (de matching entre X et Y) à travers le théorème des moindres carrées. En effet, notre but est de minimiser l'erreur e(i) qui se trouve entre chaque y(i)_réelle et chaque y(i)_prédite.

$$S = \sum_i e_i^2$$

$$avec\ e_i = Y - X\hat{a}$$

Par ailleurs, il suffit de dériver la somme des erreurs par rapport aux paramètres afin de les identifier.

2- En déduire l'équation globale qui correspond à notre analyse. [1pt]

Y=-0.884+0.0045*(studentid)+0.04*(session 2)+0.21*(session03)+0.22*(session 4)+0.28*(session 5)

3- Comment jugez-vous la pertinence du modèle obtenu ? justifiez votre réponse. [1pt]

Adjusted R squared=0.5287 , elle n'est pas bonne

4- Quelle est la différence entre R-squared et Adjusted R-squared. [0.5pt]

$$R^2 adjusted = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Autrement, le R2ajusté est écrit en fonction de R2 en pénalisant le nombre de paramètres dans l'équation de régression. Cependant, R^2 augmente avec l'ajout des variables même si elles sont non significatives.

5- Quelle est la signification de la quatrième colonne affichée dans la figure 1. [0.5pt]

Elle démontre la significativité d'une variable : plus elle s'approche de 0, plus elle est significative.pour une valeur de p_value de 0.50 la variable est non significatif d'un pourcentage de 50%.

6- En déduire le ou les attributs qui doivent être éliminés si nous souhaitons réduire l'équation obtenue. [0.5pt]

Student_id et session2

7- Donner les méthodes permettant de réduire l'équation de régression. [1pt]

Méthode ascendante avec un critère (AIC, ou BIC ou R2)

Méthode descendante

Méthode step-wise

Exercice 2 : Segmentation k-means [3pts]

On considère l'ensemble E des entiers suivants : E= {2, 5, 8, 10, 11, 18, 20 }

Nous voulons répartir les données de E en trois clusters, en utilisant l'algorithme K-means. La distance *d* entre deux nombres *a* et *b* est calculée comme suit : ***d(a , b) =*** **|a - b|** (la valeur absolue de a moins b)

1- On applique l'algorithme K-means en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11.
   a- Montrer toutes les étapes de calcul ; [1.5]

	C1 : 8	C2 : 10	C3 : 11	cluster
2	6	8	9	C1

5	3	5	6	C1
8	0	2	3	C1
11	3	1	0	C3
10	2	0	1	C2
18	10	8	7	C3
20	12	10	9	C3

Nouveaux centres :

C1'=(2+5+8)/3=5 || C2'=10|| C3'=(11+18+20)/3=16

	C1' : 5	C2' : 10	C3' : 16	cluster
2	3	8	14	C1
5	0	5	11	C1
8	3	2	8	C2
10	5	0	6	C2
11	6	1	5	C2
18	13	8	2	C3
20	15	10	4	C3

Nouveaux centres :

C1''=(2+5)/2=3.5 || C2''=(8+11+10)/2=9.66|| C3''=(18+20)/2=14

	C1'' : 3.5	C2' : 9.66	C3' : 14	cluster
2	1.5	7.66	12	C1
5	1.5	4.66	9	C1
8	4.5	1.66	6	C2
10	6.5	0.34	4	C2
11	7.5	1.34	3	C2
18	14.5	8.34	4	C3
20	17.5	11.34	6	C3

Une stabilisation des centres est très évidente, l'algorithme est arrêté à ce niveau.

b- Donner le résultat final et préciser le nombre d'itérations qui ont été nécessaires. [0.5]

Résultat final : C1''=(2+5)/2=3.5 || C2''=(8+11+10)/2=9.66|| C3''=(18+20)/2=14

3 itérations

2- Pouvons-nous avoir un nombre d'itérations inférieur pour ce problème ? Discutez. [1]

Dans le cas général, on peut minimiser le nombre d'itérations car le choix de centres initiaux est laissé au hasard.

Si on choisit bien les centres de sorte qu'ils soient proches des centres trouvés on aura forcément moins d'itérations.

Exercice 3 : Analyse en Composantes Principales [5pts]

Nous désirons étudier la composition organique de plusieurs verres. Chaque verre est représenté par 10 variables. Nous les énumérons comme suit :

V1	RI : refractive index	**V2**	Na :Sodium	**V3**	Mg: Magnesium	**V4**	Al: Aluminum
V5	Si: Silicon	**V6**	K: Potassium	**V7**	Ca: Calcium	**V8**	Ba: Barium
V9	Fe: Iron	**V10**	Type : le type de verre (de 1 à 7)				

Nous présentons dans la figure suivante quelques lignes de nos données.

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
1	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0	0.00	1
2	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0	0.00	1
3	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0	0.00	1
4	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0	0.00	1
5	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0	0.00	1
6	1.51596	12.79	3.61	1.62	72.97	0.64	8.07	0	0.26	1

Figure 2: extrait du dataset pour ACP

Nous effectuons une analyse en composantes principales pour pouvoir comprendre mieux nos données. Le résultat est comme suit :

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.055597415	30.55597415	30.55597
comp 2	2.291337939	22.91337939	53.46935
comp 3	1.409424034	14.09424034	67.56359
comp 4	1.166325743	11.66325743	79.22685
comp 5	0.914084695	9.14084695	88.36770
comp 6	0.547422471	5.47422471	93.84192
comp 7	0.369434911	3.69434911	97.53627
comp 8	0.182674245	1.82674245	99.36301
comp 9	0.062090749	0.62090749	99.98392
comp 10	0.001607797	0.01607797	100.00000

Figure 3: résultat de la PCA

1. Expliquer l'utilité de la diagonalisation dans une Analyse en Composantes Principales. [1pt]

La diagonalisation nous permet d'obtenir la matrice des vecteurs propres ainsi que le vecteur des valeurs propres. Nous nous basons alors sur le vecteur des valeurs propres pour choisir les composantes principales à retenir et qui fournissent le maximum d'inertie : prendre les CP correspondantes aux valeurs propres >1 (critère de kaiser).
Puis nous faisons la projection des individus sur les CP retenues et qu'on retrouve à partir de la matrice des vecteurs propres.

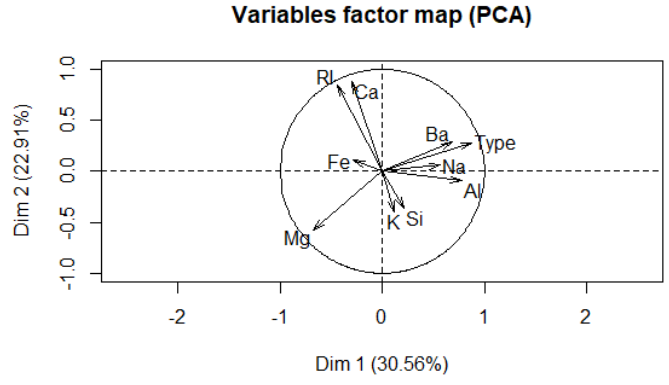
2. Comment calculer l'inertie portée par chaque composante principale. [0.5]

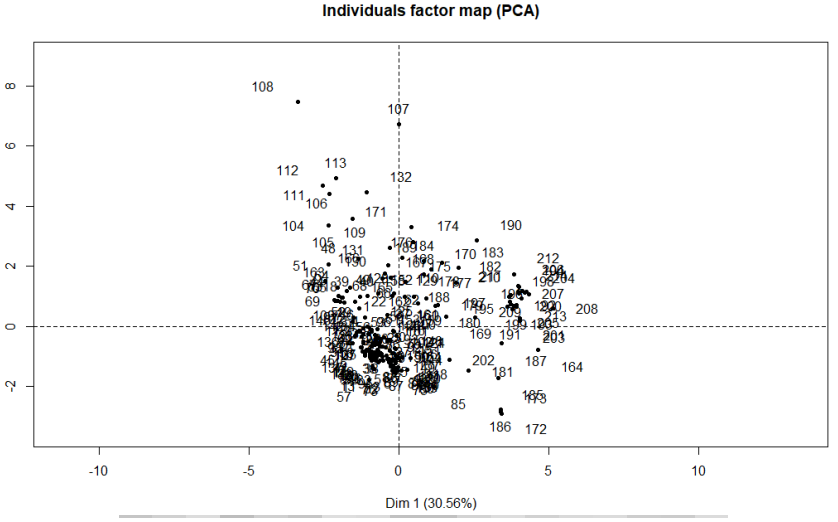
Valeur propre(i)/somme des valeurs propres

3. Comment choisir le nombre de composantes à retenir. Existe-t-il une corrélation entre les différentes composantes retenues. [0.5]

Selon le critère de kaiser on retient les CP correspondantes aux valeurs propres >1

Pour analyser les corrélations entre les différentes variables, nous projetons nos observations sur deux composantes principales. Nous obtenons les figures suivantes :





4. Interpréter les corrélations existantes entre (K et Ba), (Ba et type), (RI et Si). [1.5]
   K et Ba : pas de corrélation
   Ba et type : corrélation positive
   RI et Si : corrélation négative
5. Lister les variables corrélées positivement avec la première composante principale. [0.5]

Type, Na, Ba, Al

6. Selon La figure 5, répondre aux questions suivantes :
   a- l'individu 108 est corrélé positivement avec quelles variables ? [0.25]
      RI et CA
   c- l'individu 190 est corrélé positivement avec quelles variables ?[0.25]
      Ba et Type
   d- quelles sont les individus (3 individus) qui sont bien représentés sur le plan factoriel. [0.5]
      172, 108, 107

Exercice 4 : Arbre de décisions [3pts]

1- Donner un pseudo code du fonctionnement de l'algorithme des arbres de décisions. [1pt]

```
Procédure construire-arbre(X)
SI tous les individus I appartiennent à la même modalité de la variable décisionnelle ALORS
        créer un nœud feuille portant le nom de cette classe : Décision
SINON
        choisir le meilleur attribut pour créer un nœud // l'attribut qui sépare le mieux, le test associé à ce
        nœud sépare X en des branches
        construire-arbre(Xd), ...,construire-arbre(Xg)
FIN
```

2- En utilisant l'algorithme de Cart avec le gain en information basé sur l'indice de Gini, construire un arbre de décision sur les données suivantes : [2 pts]

Age	Sexe	Propriétaire	Intéressé
20	H	N	N
25	F	N	N
32	H	O	O
34	H	O	O
37	H	N	O
41	F	O	N
45	H	O	O
45	F	O	N
52	H	O	N
60	F	O	N

Trois attributs descriptifs sont à votre disposition :

- ✔ L'âge en deux tranches : [18; 35] et [36 et plus]
- ✔ Le sexe H : Homme ou F : Femme
- ✔ Propriétaire O : oui ou N : non

L'attribut cible qui prend deux valeurs : O (intéressé) et N (pas intéressé).

Correction dans le PowerPoint