esprit Se former autrement	**EXAMEN** Semestre : 1 ☐ 2 ■ Session : Principale ■ Rattrapage ☐

Module : Data Mining
Enseignantes : Amal Tarifa, Dorra Trabelsi, Wiem Trabelsi
Classes : ERP-BI
Documents autorisés : OUI ☐ NON ■ Nombre de pages : 4 + Annexe
Calculatrice autorisée : OUI ■ NON ☐ Internet autorisée : OUI ☐ NON ■
Date : 18/05/2019 Heure: 10h30 Durée : 1h30

Exercice 1 : Analyse en Composantes Principales [5pts]

On donne sur **l'annexe 1** un échantillon de données pour quelques villes. Ces dernières sont décrites par :

le nombre de personnes atteintes de cancer	le nombre de mortalités du au cancer
le revenu moyen de la ville	le nombre de population
le pourcentage de pauvreté	l'âge moyen de la population
l'âge moyen des femmes	le pourcentage des mariés
le pourcentage de la population atteinte du cancer dont l'âge est entre 18 et 24, etc	

Les statistiques descriptives sont données par la **figure 1 de l'annexe1.**
On se propose de réaliser une analyse en composantes principales afin de comprendre les données. Les résultats sont illustrés dans les graphes de **l'annexe 2**.

1- Définir l'analyse en composantes principales et préciser son utilité. [1 pt]

2- En se référant au tableau des valeurs propres donné par la figure 4, comment choisir les axes factoriels les plus adéquats. Quel est le critère utilisé ? [1 pt]

3- Pour des raisons de visualisation, on a choisi de représenter nos variables sur les axes **Dim1** et **Dim2**. Comment jugez-vous ce repère de projection. Interpréter les corrélations **variables/ variables** et **variables /dimensions.** [1 pt]

4- En se référant à la carte des individus représentée par la *figure 3* préciser la liste des individus mal représentés sur les axes Dim1 et Dim 2. [0.5 pt]

5- On donne la *figure 6*, carte des individus selon le $\cos^2$. En se référant à cette dernière, vérifier les résultats obtenus dans la question précédente. Expliquer et réordonner les individus par ordre décroissant selon leur contribution dans l'axe Dim1 et Dim2. [1 pt]

6- A l'issue de cette étude, proposer un profiling. [0.5]

Exercice 2 : Règles d'associations [2 pts]

On considère un ensemble de 6 tablettes {T1, T2, T3, T4, T5, T6}. Chaque tablette est décrite par une liste de caractéristiques:

T1	GPS , LCD, 3G , Bluetooth (B), Wi-Fi (WF), USB, HDMI
T2	Wi-Fi (WF), USB
T3	LCD, Bluetooth (B)
T4	Bluetooth (B), Wi-Fi(WF)
T5	GPS, LCD, Bluetooth (B), Wi-Fi
T6	GPS, 5G , Bluetooth (B), Wi-Fi (WF), USB, HDMI

1- Appliquer l'algorithme APRIORI pour rechercher les itemsets fréquents ayant un support supérieur à 35%. [1 pt]
2- Donner les règles d'association les plus pertinentes, avec une confiance minimum de 50%. [1 pt]

Exercice 3 : Classification Ascendante Hiérarchique [3 pts]

On dispose d'un tableau de données décrivant 8 individus en fonction de 2 variables ainsi que le tableau de distances (euclidiennes) entre eux.

Individu	*V1*	*V2*
1	1	3
2	2	4
3	1	5
4	5	5
5	5	7
6	4	9
7	2	8
8	3	10

	1	2	3	4	5	6	7	8
1	0	1.41	2.00	4.47	5.66	6.71	5.10	7.28
2		0	1.41	3.16	4.24	5.39	4.00	6.08
3			0	4.00	4.47	5.00	3.16	5.39
4				0	2.00	4.12	4.24	5.39
5					0	2.24	3.16	3.61
6						0	2.24	1.41
7							0	2.24
8								0

1- Appliquer l'algorithme **CAH** aux données précédentes en précisant à chaque étape la nouvelle table de distances sachant que le critère d'agrégation est le saut Minimal. [2 pts]
2- Dessiner un dendrogramme associé à cette segmentation. [1 pt]

Exercice 4 : méthodes supervisées (arbres de décisions, Régression SVM) [10 pts]

Les maladies cardiovasculaires constituent un ensemble de troubles affectant le cœur et les vaisseaux sanguins. Ces malaises sont la première cause de la mortalité dans le monde. Un groupe de médecins chercheurs se proposent d'étudier les facteurs majeurs de risque cardiovasculaire en se basant sur une liste de paramètres, à savoir : hypertension, diabète, angine, dépression, maladie déjà installé, etc.
Un aperçu sur les données est donné par *l'annexe 3*.
I- Une première étude consiste à réaliser un modèle **à base de règles logiques** en utilisant la commande suivante :

Model1=rpart(cœur ~ . , dataC)

1- S'agit-il d'un arbre de régression ou de classification ? justifier la réponse. [0.5 pt]
2- Expliciter les différents paramètres de la commande précédente. [0.5 pt]
3- Le résultat obtenu donne la figure suivante :

```
n= 200

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 200 89 absence (0.5550000 0.4450000)
   2) typedouleur=A,B,C 100 21 absence (0.7900000 0.2100000)
     4) depression< 19.5 87 12 absence (0.8620690 0.1379310) *
     5) depression>=19.5 13  4 presence (0.3076923 0.6923077) *
   3) typedouleur=D 100 32 presence (0.3200000 0.6800000)
     6) depression< 7 41 18 absence (0.5609756 0.4390244)
      12) sexe=feminin 13  3 absence (0.7692308 0.2307692) *
      13) sexe=masculin 28 13 presence (0.4642857 0.5357143)
        26) tauxmax>=167 7  2 absence (0.7142857 0.2857143) *
        27) tauxmax< 167 21  8 presence (0.3809524 0.6190476)
          54) tauxmax< 153.5 11  5 absence (0.5454545 0.4545455) *
          55) tauxmax>=153.5 10  2 presence (0.2000000 0.8000000) *
     7) depression>=7 59  9 presence (0.1525424 0.8474576) *
```

a) Donner la variable la plus séparatrice. Quel est le critère mathématique permettant son obtention. Expliciter sa formule. [1 pt]

b) Expliciter les règles décisionnelles extraites à partir de la figure précédente. [1 pt]

4- On applique le **Model1** sur un échantillon de test, ainsi on obtient la table de contingence suivante :

a) Calculer le taux de bonne classification. [0.25 pt]

b) Calculer les taux de mauvaise classification. [0.5 pt]

c) Comment jugez-vous les performances du modèle pour chacune des classes établies. [0.5 pt]

	Absence	**Présence**
absence	35	9
présence	4	22

II- On souhaite maintenant réaliser un second modèle **à base de régression**.

1- Quel est le type de régression adéquat pour cette modélisation. [0.25]

2- Donner le principe algorithmique du modèle choisi ? Vous pouvez vous appuyer sur un schéma. [1 pt]

3- On donne les commandes suivantes :

Model2=glm(cœur ~ . , family=binomial, dataC)

Summary(Model2)

```
glm(formula = coeur ~ ., family = binomial, data = dataC)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4508  -0.6627  -0.2381   0.5808   2.3058

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.434514   2.611679  -1.315 0.188490
age           0.021972   0.024123   0.911 0.362393
sexemasculin  1.711204   0.451088   3.794 0.000149 ***
typedouleurB  1.302234   0.930747   1.399 0.161775
typedouleurC  1.141397   0.792333   1.441 0.149711
typedouleurD  2.772973   0.773214   3.586 0.000335 ***
sucreB        0.534369   0.537995   0.993 0.320583
tauxmax      -0.014883   0.009818  -1.516 0.129530
angineoui     0.743791   0.431113   1.725 0.084477 .
depression    0.079096   0.021957   3.602 0.000315 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 274.83  on 199  degrees of freedom
Residual deviance: 171.63  on 190  degrees of freedom

AIC: 191.63

Number of Fisher Scoring iterations: 5
```

a) Ecrire le modèle global obtenu. [0.5 pt]

b) Quelles sont les variables que vous retiendrez pour écrire un modèle réduit. Justifier votre réponse. [0.5 pt]

4- On applique le **Model2** sur un échantillon de test, on obtient la table de contingence suivante :

a) Calculer le taux de bonne classification. [0.25 pt]

b) Calculer les taux de mauvaise classification. [0.5 pt]

c) Comment jugez-vous les performances du modèle pour chacune des classes établies. [0.5 pt]

	Absence	**Présence**
absence	32	7
présence	7	24

III- On souhaite maintenant réaliser un troisième **modèle à base de séparateurs à vaste marge.** L'exécution de la commande **: Model3= svm(formula=cœur ~ . , dataC, kernel= ''radial '')** donne le résultat suivant :

```
Call:
svm(formula = coeur ~ ., data = dataC, kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1

Number of Support Vectors:  121
```

1- Donner les différents types de noyaux dans un contexte SVM. [0.75 pt]
2- Quand est ce que l'utilisation d'un noyau est nécessaire. [0.25 pt]
3- Comparer les performances des deux méthodes SVM et régression sur un petit entrepôt de données. [0.5 pt]
4- Donner les deux points spécifiques de la population sur lesquels se base l'SVM pour le calcul de marge. [0.25 pt]
5- Enoncer mathématiquement la maximisation de la marge [0.5 pt]
6- Donner le nombre de vecteurs supports calculés. [0.25 pt]
7- On applique le **Model3** sur un échantillon de test, on obtient la table de contingence suivante :

a) Calculer le taux de bonne classification. [0.25 pt]
b) Calculer les taux de mauvaise classification. [0.5 pt]
c) Comment jugez-vous les performances du modèle 3 pour chacune des classes établies. [0.5 pt]

	Absence	**Présence**
absence	34	9
présence	5	22

IV- Evaluer les résultats obtenus pour les trois modèles précédents. Lequel choisirez-vous. [0.5 pt]