Master 1 – Informatique Décisionnelle/Fouille de données

Toute document autorisé, Calculatrice autorisée

Durée 2h

Il y a bien longtemps, dans une galaxie lointaine, très lointaine...
La guerre civile fait rage entre l'Empire galactique et l'Alliance rebelle. Capturée par les troupes de choc de l'Empereur menées par le sombre et impitoyable Dark Vador, la princesse Leia Organa dissimule les plans de l'Etoile Noire, une station spatiale invulnérable, à son droïde R2-D2 avec pour mission de les remettre au Jedi Obi-Wan Kenobi.
Accompagné de son fidèle compagnon, le droïde de protocole C-3PO, R2-D2 échoue sur la planète Tatooine et termine sa quête chez le jeune Luke Skywalker. Rêvant de devenir pilote mais confiné aux travaux de la ferme, ce dernier se lance à la recherche de ce mystérieux Obi-Wan Kenobi, devenu ermite au coeur des montagnes désertiques de Tatooine...

Exercice 1 – L'empire à besoin de vous

Fondé par le Chancelier Palpatine, l'Empire est un régime qui se veut, à l'origine, la solution à l'injustice et la corruption régnantes.
L'empire, véritable état intergalactique, décide de se doter d'un système de ressource humaine pour gérer son armé impériale (Renseignements Impériaux, Marine Impériale, Armée Impériale, Soldats de chocs). Ce service du personnel impérial gère des personnels de différents grades du moins élevé au plus élevé :

- cadet (infanterie).
- pilote (infanterie, marine).
- enseigne (infanterie, marine).
- second lieutenant, officier pilote (infanterie, marine).
- Sous lieutenant, lieutenant, officier de vol (infanterie, marine).
- Lieutenant, capitaine, lieutenant de vol (infanterie, marine).
- Lieutenant-commandant, major, chef d'escadrille (infanterie, marine).
- Commandant, lieutenant-colonel, commandant d'escadron (infanterie, marine).
- Capitaine, colonel, capitaine de groupe (infanterie, marine).
- Capitaine de ligne, haut colonel, capitaine d'escadre (infanterie, marine).
- Commodore, général de brigade (infanterie, marine).
- Contre-amiral, général-major (infanterie, marine).
- Vice-amiral, lieutenant général (infanterie, marine).
- Amiral, général (infanterie, marine).
- Amiral de la flotte, haut général (infanterie, marine).
- Haut amiral, maréchal, maréchal de force (infanterie, marine).
- Grand amiral, grand maréchal (infanterie, marine).
- Moff (infanterie, marine).
- Grand Moff (infanterie, marine).

Chaque militaire appartient à une structure de l'armée divisée du plus petit au plus grand en :

- Escouade
- Peloton
- Compagnie
- Régiment
- Groupe de Bataille
- Corps d'Armée
- Armée sectorielle

Pour chaque employé on décide d'avoir son profil (avec au moins 100 attributs) comme la date d'embauche, le coefficient hiérarchique, le salaire, des dates d'évaluation des résultats d'évaluation, des droits aux vacances, le service, l'adresse. A tout moment, les employés sont embauchés, transférés et affectés, promus, exécutés et leurs profils sont modifiés. Il est indispensable de suivre et d'analyser les événements (transactions) sur les employés.

Question 1 : Proposez un schéma en étoile pour ce datawarehouse. Vous justifierez vos choix et préciserez les mesures.

Question 2 : Une armée sectorielle comprend 1.180.309 hommes dont 774.576 combattants et elle dispose de 66.640 véhicules à répulseurs et 13.992 blindés. L'empire comprend 20 secteurs. Un employé subit en moyenne 10 transactions de modification de profil par an. L'empire veut stocker ces informations sur 10 ans.
Estimez la taille du datawarehouse.

Exercice 3
L'empire cherche à regrouper les planètes et les étoiles selon le nombre d'espèces qu'elles comportent, le nombre de bases rebelles, le diamètre de la planète et la présence de végétation. Les agents de l'empire ont relevé les informations suivantes :

	Nb Espèces	*Nb Base*	*Diamètre*	*Végétation*
P1	*11*	*1*	*120536*	*oui*
P2	*5*	*2*	*6700*	*non*
P3	*10*	*4*	*12756*	*oui*
P4	*7*	*2*	*12100*	*non*
P5	*2*	*0*	*4880*	*non*
P6	*0*	*0*	*2300*	*non*
P7	*8*	*2*	*12100*	*oui*

Question 1. Définissez formellement une distance permettant de considérer tous les attributs. Donnez la distance de P1 à P2 avec la distance précédemment définie.

Question 2. Les généraux de l'empire proposent la matrice de distance suivante (ne comparez pas ces distances avec votre réponse à Q1, on ne peut pas faire confiance à l'empire) :

	P1	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*
P1	*0*	*0.462*	*0.666*	*0.538*	*0.478*	*0.334*	*0.666*
P2		*0*	*0.428*	*0.466*	*0.334*	*0.52*	*0.786*
P3			*0*	*0.5*	*0.44*	*0.652*	*0.846*
P4				*0*	*0.482*	*0.52*	*0.572*
P5					*0*	*0.636*	*0.76*
P6						*0*	*0.487*
P7							*0*

Proposez un regroupement via la méthode single-link. Vous détaillerez les calculs et donnerez le dendogramme obtenu.

Exercice 4

Arrivé dans la cantina de la planète tatooine, Han Solo décide de donner des indications à Luke pour qu'il ne provoque pas les extraterrestes belliqueux. Il repère quelques caractéristiques et vous demande de l'aider à fournir des éléments à Luke pour ne pas créer de problèmes.

Couleur	*Taille*	*Poids*	*Yeux par pair ?*	*Belliqueux*
Jaune	*moyenne*	*léger*	*non*	*Non*
Jaune	*grande*	*moyen*	*oui*	*Oui*
Vert	*petite*	*moyen*	*oui*	*Oui*
Jaune	*petite*	*moyen*	*non*	*Non*
Rouge	*moyenne*	*lourd*	*non*	*Non*
Vert	*grande*	*lourd*	*non*	*Oui*
Vert	*moyenne*	*lourd*	*non*	*Oui*
Jaune	*petite*	*léger*	*oui*	*Oui*

Question 1 : Construire l'arbre de décision en utilisant l'indice de Gini. Vous détaillerez les calculs.

Question 2 : R2D2 propose d'utiliser Weka pour aider Han Solo. Il obtient la sortie suivante

```
1.  === Classifier model (full training set) ===
2.  Id3
3.  Couleur = Jaune
4.  |  Yeux par pair ? = non: Non
5.  |  Yeux par pair ? = oui: Oui
6.  Couleur = Vert: Oui
7.  Couleur = Rouge: Non
8.  Time taken to build model: 0 seconds
9.  === Evaluation on training set ===
10. === Summary ===
11. Correctly Classified Instances           8               100      %
12. Incorrectly Classified Instances         0                 0      %
13. Kappa statistic                          1
14. Mean absolute error                      0
15. Root mean squared error                  0
16. Relative absolute error                  0      %
17. Root relative squared error              0      %
18. Total Number of Instances                8

19. === Detailed Accuracy By Class ===

20. TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
21. 1         0          1         1          1          1        Non
22. 1         0          1         1          1          1        Oui
23. Weighted Avg.    1          0          1          1          1          1

24. === Confusion Matrix ===

25. a b   <-- classified as
26. 3 0 | a = Non
27. 0 5 | b = Oui
```

Q2.1 Commentez la sortie de Weka

Q2.2 Redessiner l'arbre obtenu.

Question 3 : Calculez la performance de chaque algorithme sur l'ensemble de test T. Qu'allez vous donner comme modèle à Han Solo ?

Couleur	*Taille*	*Poids*	*Yeux par pair ?*	*Bélliqueux*
Jaune	*grande*	*léger*	*non*	*Non*
Jaune	*moyenne*	*moyen*	*oui*	*Oui*
Vert	*petite*	*moyen*	*non*	*Oui*