Université de Maroua
ENSPM
Département d'INFOTEL

Parcours : DSC et MIF
Niveau : 4
Année : 2022 – 2023

EXAMEN DE DATA MINING

Code UE : DSC 438 - MIF 438

(EPREUVE THEORIQUE)

Proposée par : Touza Isaac

Durée : 3h

Exercice 1 : Questions du cours — 5pts

1. Définir Data Mining — **0,5pt**
2. Enumérer deux tâches du Data Mining — **0,5pt**
3. Expliquer pourquoi la valeur de k dans l'algorithme KNN doit être impair. — **0,5pt**
4. Présenter quatre opérations de prétraitement des données textuelles. — **1pt**
5. Enumérer deux bibliothèque python utiliser pour récupérer des données à partir d'un site web — **0,5pt**
6. Donner la différence entre la classification supervisée et la classification non supervisée. — **1pt**
7. Ecrire le code python permettant de charger dans la variable **data** le contenu d'un fichier csv nommé **data_science.csv** — **1pt**

Exercice 2 : les règles d'association — 5pts

On vous demander d'analyser les transactions de vente dans la boutique de monsieur WARDA. Pour ce faire ce dernier met à votre disposition les données relatives aux achats des produits effectués par cinq (05) clients.

Transactions	Articles
1	{ Pain, Lait, Fromage}
2	{ Pain, Lait, Beurre }
3	{ Pain, Lait, Beurre, Fromage }
4	{ Pain, Lait, Beurre }
5	{ Pain, Lait, Fromage }

1. Définir les termes suivants : règles, transaction — **0,5pt**
2. Déterminer le nombre d'itemsets qu'on peut former avec l'ensemble des données de ce tableau. — **0,5pt**
3. Donner la représentation horizontale de ce tableau — **1pt**
4. Utiliser l'algorithme Apriori pour extraire les ensembles des produits fréquents avec un support minimum de 0,6. — **1,5pt**
5. Extraire les règles d'associations à partir de l'ensemble des produits fréquents avec une confiance minimale de 80% — **1pt**
6. Donner une interprétation d'une règle de votre choix. — **0,5pt**

Exercice 3 : Le clustering — 5pts

Soit l'ensemble D des entiers suivants : D= { 2, 5, 8, 10, 11, 18, 20 }

On veut répartir les données de D en trois (3) clusters, en utilisant l'algorithme K-means. La distance d entre deux nombres a et b est calculée ainsi :

d(a , b) = |a - b| (la valeur absolue de a moins b)

1. Appliquez l'algorithme K-means pour regrouper ces données en trois cluster en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11. Montrez toutes les étapes de votre calcul. **3pts**
2. Donnez le résultat final des clusters et précisez le nombre d'itérations qui ont été nécessaires. **1pt**
3. Peut-on avoir un nombre d'itérations inférieur pour ce problème ? Discutez. **1pt**

Exercice 4 : La classification avec les arbres de décision 5pts

Considérons un ensemble de données contenant des informations sur des voitures d'occasion. Chaque voiture est décrite par les caractéristiques suivantes :

- Marque (Toyota, Honda, Ford)
- Année de fabrication (2005, 2010, 2015)
- Kilométrage (en milliers de kilomètres)
- État (Excellent, Bon, Mauvais)
- Prix de vente (en euros)

Nous souhaitons construire un arbre de décision pour prédire si une voiture sera vendue à un prix élevé ou basé sur ses caractéristiques.
Voici les données d'entraînement :

Marque	Année	Kilométrage	État	Prix
Toyota	2010	50	Excellent	Élevé
Honda	2005	120	Bon	Bas
Ford	2015	30	Mauvais	Bas
Toyota	2015	80	Bon	Bas
Honda	2010	75	Excellent	Élevé
Ford	2005	100	Mauvais	Bas
Toyota	2005	85	Mauvais	Bas
Honda	2015	40	Excellent	Élevé

1. Construisez un arbre de décision pour prédire le prix de vente des voitures. **3pts**
2. Utilisez cet arbre pour prédire le prix de vente d'une voiture de marque Honda, fabriquée en 2010, avec un kilométrage de 60 et en bon état. **1pt**
3. Discutez des avantages et des limites des arbres de décision pour ce problème. **1pt**

NB : Pour la caractéristique "Kilométrage", subdiviser les valeurs en deux groupes (Kilométrage ≤ 50 et Kilométrage 50)