Université de Maroua
ENSPM
Master / Niveau II

Département d'Infotel
Parcours: DSC et MIF
Année : 2022 – 2023

**DATA MINING (DSC 438 – MIF 438)**
**FICHE DE TRAVAUX DIRIGES**
Proposée par : Touza Isaac

## Exercice 1 : Questions du cours

1. Définir les notions suivantes : data mining, classification, clustering.
2. Enumérer quelques tâches du data mining
3. Décrire deux objectifs du data mining
4. Décrire le processus du data Mining
5. Enumérer un exemple d'algorithme permettant de réaliser chacune des tâches suivantes :
   a. La classification
   b. La prévision
   c. Recherche des règles d'association
   d. Le clustering
6. Donner la différence entre la classification supervisée et la classification non supervisée.
7. Présenter la différence entre les méthodes descriptives et les méthodes prédictives.
8. Décrire le principe de fonctionnement de l'algorithme KNN
9. Expliquer comment doit se faire le choix de la valeur de k lors de l'exécution de l'algorithme KNN avec un jeu des données quelconque.
10. Expliquer pourquoi la valeur de k dans l'algorithme KNN doit être impair.
11. Présenter deux inconvénients de l'algorithme KNN.
12. Quels sont les différentes sources de données collectées ?
13. Quels outils sont utilisés pour collecter les données ?
14. Comment se déroule la collecte de données sur le Web ?
15. Donner quelques opérations de prétraitement des données textuelles.
16. Quelles bibliothèques Python sont utilisées pour le pré-traitement des données ?
17. Quelles sont les étapes du pré-traitement des données ?
18. Qu'est-ce qu'une transaction dans le contexte des règles d'association ?
19. Quels sont les concepts de base des règles d'association ?
20. Quels sont les algorithmes couramment utilisés pour les règles d'association ?
21. Quels sont les processus de classification ?
22. Quelles sont les méthodes de pondération ou de calcul des fréquences ?
23. Quels sont les processus de classification ?
24. Quelles sont les méthodes de pondération ou de calcul des fréquences ?

## Exercice 2 : Règles d'association

Vous disposez d'un ensemble de données représentant les symptômes de patients atteints de différentes maladies. Chaque patient est décrit par les symptômes qu'il présente. Utilisez l'algorithme Apriori pour trouver les ensembles de symptômes fréquents et générer des règles d'association basées sur l'ensemble de données donné.

Patient 1 : {Fièvre, Maux de tête, Fatigue}
Patient 2 : {Maux de tête, Toux, Frissons, Fatigue}
Patient 3 : {Fièvre, Toux, Douleurs musculaires}

Patient 4 : {Fièvre, Maux de tête, Toux, Frissons}
Patient 5 : {Toux, Frissons, Fatigue}

1. Donner une représentation binaire de ces données.
2. Utilisez l'algorithme Apriori pour trouver tous les ensembles fréquents avec un support minimum de 2 patients.
3. En utilisant les ensembles fréquents trouvés à l'étape précédente, générez toutes les règles d'association possibles avec une confiance minimale de 60%.
4. Interprétez les règles d'association intéressantes trouvées en termes de symptômes de maladies.

## Exercice 3 : knn

Le tableau suivant contient des données sur des individus d'une population décrite selon deux attributs : attribut 1 et attribut 2. La classe d'un individu peut être : C1, ou C2, ... ou C6.

Tableau des données

| N° | Attribut 1 | Attribut 2 | Classe |
|---|---|---|---|
| 1 | 1 | 2 | C1 |
| 2 | 2 | 6 | C1 |
| 3 | 2 | 5 | C2 |
| 4 | 2 | 1 | C3 |
| 5 | 4 | 2 | C5 |
| 6 | 5 | 6 | C4 |
| 7 | 6 | 5 | C3 |
| 8 | 6 | 1 | C6 |

1. Représentez sur le plan les données du tableau précédent (On prendre l'attribut 1 en abscisse et l'attribut 2 en ordonnée).
2. On veut classer un nouvel individu U ayant comme attributs (1, 4) en utilisant la méthode KNN. Quelle sera la classe de U si on choisit k=3. Justifiez. (Utiliser la distance euclidienne)
3. On utilise maintenant la variante de KNN qui utilise la distance $1/d^2$ (inverse de la distance au carré) pour calculer les voisins. Quelle sera la classe de U avec k=3 ?. Justifiez.

## Exercice 5 : kNN

Soit les points de coordonnées suivantes : A(1, 6), B(2, 6), C(3, 1), D(4, 2), E(6, 0), F(7, 5), G(7, 3), H(10, 3) En utilisant la distance euclidienne, quels sont les deux plus proches voisins du point P(5, 5) ?

## Exercice 5 : K-means

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 :

$$d(a, b) = |a - b| \text{ (la valeur absolue de a moins b)}$$

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

## Exercice 6 : arbre de décision

Le tableau suivant contient des données sur les résultats obtenus par des étudiants de Tronc Commun (première année à l'Université). Chaque étudiant est décrit par 3 attributs : Est-il doublant ou non, la série du Baccalauréat obtenu et la mention. Les étudiants sont répartis en deux classes : Admis et Non Admis.

On veut construire un arbre de décision à partir des données du tableau, pour rendre compte des éléments qui influent sur les résultats des étudiants en Tronc Commun. Les lignes de 1 à 12 sont utilisées comme données d'apprentissage. Les lignes restantes ( de 13 à 16) sont utilisées comme données de tests.

| | **Doublant** | **Série** | **Mention** | **Classe** |
|---|---|---|---|---|
| 1 | Non | Maths | ABien | Admis |
| 2 | Non | Techniques | ABien | Admis |
| 3 | Oui | Sciences | ABien | Non Admis |
| 4 | Oui | Sciences | Bien | Admis |
| 5 | Non | Maths | Bien | Admis |
| 6 | Non | Techniques | Bien | Admis |
| 7 | Oui | Sciences | Passable | Non Admis |
| 8 | Oui | Maths | Passable | Non Admis |
| 9 | Oui | Techniques | Passable | Non Admis |
| 10 | Oui | Maths | TBien | Admis |
| 11 | Oui | Techniques | TBien | Admis |
| 12 | Non | Sciences | TBien | Admis |
| 13 | Oui | Maths | Bien | Admis |
| 14 | Non | Sciences | ABien | Non Admis |
| 15 | Non | Maths | TBien | Admis |
| 16 | Non | Maths | Passable | Non Admis |

1. Définir arbre de décision
2. On souhaite utiliser l'algorithme ID3 pour construire l'arbre de décision en utilisant les données d'apprentissage de ce tableau.
   2.1. Montrez toutes les étapes de calcul pour la construction de deux arbres de décisions différents puis Dessinez-les.
   2.2. Quels sont les résultats de test de l'arbre obtenu sur les données des lignes de 13 à 16 ? Ces résultats correspondent-ils à ceux du tableau ci-dessus ?
   2.3. Déduire le taux d'erreur et de succès de chacun d'arbre de décision construit.
   2.4. Déduire l'arbre à choisir puis justifiez.
3. Présenter deux limites de l'algorithme ID3.
4. Pour améliorer l'algorithme ID3, l'on a mis sur pied l'algorithme C4.5.
   4.1. Présenter deux extensions de l'algorithme ID3 introduit par C4.5
   4.2. Donner l'intérêt d'élaguer un arbre de décision.

## Exercice 7 : Prétraitement de texte

Vous avez un ensemble de documents texte que vous souhaitez prétraiter avant de les utiliser dans une tâche de classification de texte. Appliquez les étapes de prétraitement des données textuelles suivantes à un document donné :

**Document : "Le chat noir se promène dans le jardin. Il saute sur les arbres et attrape des oiseaux. Le chien aboie et le chat miaule."**

1. Donner la représentation à 4 grammes de l'ensemble du document.
2. Diviser le document en tokens individuels.
3. Supprimer les mots vides (stop words) français suivants : "le", "et".
4. Normaliser les tokens obtenus.
5. Appliquer l'algorithme de stemming de Porter pour réduire les mots à leur racine.
6. Calculez la fréquence de chaque terme dans le document.

NB : Montrez toutes les étapes de prétraitement des données textuelles et les résultats obtenus.