Exercice 1 : Prétraitement des données

Une entreprise de e-commerce collecte des données sur ses clients. Les données brutes sont les suivantes :

ID	Âge	Revenu (en €)	Catégorie d'achat	Score de satisfaction	Avis client	Montant dépensé (€)
1	28	35000	Luxe	Très Bon	Positif	500
2	NaN	22000	Économique	Moyen	Neutre	150
3	45	NaN	Moyen	Excellent	Positif	NaN
4	30	27000	Standard	Mauvais	Négatif	300
5	52	50000	L	Très Mauvais	Négatif	800

Travail à faire :

1. Imputer les valeurs manquantes :
    - Âge par la moyenne des âges disponibles.
    - Revenu par la médiane.
    - Montant dépensé par la moyenne selon la catégorie d'achat.
2. Uniformiser la colonne "Catégorie d'achat" en trois groupes : Luxe, Standard, Économique.
3. Transformer les colonnes "Score de satisfaction" et "Avis client" en valeurs numériques.
4. Ajouter une colonne "Dépense moyenne" (Montant dépensé / Revenu).
5. Ajouter et Normaliser les colonnes "Âge" et "Revenu" avec une normalisation min-max.

Exercice 2 : Arbres de décision avec l'indice de Gini

Un hôpital souhaite prédire si un patient doit être hospitalisé ou non en fonction des données suivantes :

ID	Température	Symptômes	Hypertension	Hospitalisation
1	Élevée	Oui	Oui	Oui
2	Moyenne	Non	Non	Non
3	Basse	Oui	Non	Non
4	Moyenne	Oui	Oui	Oui
5	Élevée	Oui	Non	Oui

Travail à faire :

1. Construire un arbre de décision en utilisant l'algorithme CART (indice de Gini).
2. Montrer toutes les étapes de calcul :
    - Calculer l'indice de Gini initial.
    - Calculer l'indice de Gini pour chaque attribut et ses partitions possibles.
    - Sélectionner l'attribut offrant le plus faible indice de Gini pour diviser les données.
3. Dessiner l'arbre final en utilisant les meilleures divisions.
4. Utiliser cet arbre pour prédire si un nouveau patient (Température = Moyenne, Symptômes = Non, Hypertension = Oui) doit être hospitalisé ou non.

Exercice 3 : K-means Clustering

Les données suivantes concernent des magasins de détail. Vous souhaitez les regrouper en clusters en fonction de leur chiffre d'affaires et de leur nombre de clients.

ID	Chiffre d'affaires (€)	Nombre de clients
1	50000	150
2	70000	200
3	20000	50
4	90000	250
5	30000	100

Travail à faire :

1. Appliquer l'algorithme K-means pour regrouper les magasins en 2 clusters.
    - Utiliser des centres initiaux : (50000, 150) et (20000, 50).
2. Montrer les calculs de distance euclidienne, l'attribution des points, et les nouveaux centres.
3. Indiquer les clusters finaux et interpréter les résultats.

Exercice 4 : Classification avec K-NN

Un organisme d'assurance veut classifier les clients en deux catégories : Faible risque et Haut risque, selon leur âge et leur revenu.

ID	Âge	Revenu (€)	Risque
1	25	30000	Faible
2	45	60000	Haut
3	35	50000	Faible
4	50	80000	Haut
5	23	28000	Faible

Nouvelle observation : Âge = 40, Revenu = 55000.

Travail à faire :

1. Classifier la nouvelle observation en utilisant l'algorithme K-NN (k=3).
2. Calculer les distances euclidiennes et identifier les k voisins les plus proches.
3. Prédire la catégorie de risque pour la nouvelle observation.