Évaluation 3 - Examen final

Code du cours : 420-A51-BB
Date : 7 juillet 2025 de 18h00 à 19h30
Durée : 1h30 minutes
Pondération : 40%

Vous devez remettre via Léa un seul document (en format pdf) contenant aussi bien toutes les lignes de code que toutes les réponses aux questions. Il est très important d'indiquer clairement le numéro de la question à laquelle vous répondez.

Note : Avant la remise, vérifiez la qualité visuelle et l'entièreté du document. Respectez l'heure limite de remise, les documents remis en retard, ne seront pas acceptés.

Partie I : Questions de compréhension (15 points)

Répondez à chacune de ces questions de façon concise (en 2 phrases).

Q1. Quelle technique d'apprentissage machine utilisiez-vous pour segmenter les clients d'un commerce en plusieurs groupes ?

Q2. Nous disposons d'une très large base de données. Quelles partitions des données et stratégies adopterez-vous pour développer un système de classification ? Justifiez votre réponse.

Q3. Nous disposons d'une petite base de données. Quelle stratégie de répartition des données adopterez-vous pour développer un système de classification ? Justifiez votre réponse.

Q4. Pourquoi le regroupement est considéré comme une méthode d'apprentissage machine non-supervisé ?

Q5. Classeriez-vous le problème de détection de spam parmi les problèmes de classification supervisée ou non-supervisée ?

Q6. Pourquoi stratifier les données ?

Partie II : Classification de données (25 points)

L'objectif de ce travail est de développer des modèles de classification sur un ensemble de données. Plus spécifiquement, nous désirons prédire la variable

cible quality (qualité de vin) en fonction de ses caractéristiques chimiques décrit dans la base de données Wine Quality Data Set (winequality-red.xlsx).

En procédant selon les étapes suivantes, vous devez réaliser un arbre de décision.

1. À partir du fichier `winequality-red.xlsx` (disponible sur Léa), téléchargez le contenu de la base de données et affichez son contenue. Enregistez ces données dans une structure `df`
2. Identifier les différentes variables et leur type.
3. Répartissez les données en deux ensembles : 70% pour l'entrainement (train) et 30% pour le test (test) (random_state =10). Vérifiez le nombre d'individus dans chaque classe de l'ensemble train.
4. Entrainez un arbre de décision qui permet de prédire la variable cible quality à partir de l'ensemble des caractéristiques. Considérez une profondeur de l'arbre de 2.
5. Représentez (afficher) l'arbre de décision.
6. Identifiez les variables retenues. Quelle est la variable la plus discriminante (importante).
7. Représentez la matrice de confusion de l'ensemble de données de test.
8. Développez un autre arbre de décision qui permet de prédire la variable cible quality à partir de l'ensemble des caractéristiques de profondeur 3. Représentez la matrice de confusion de l'ensemble de données de test et commentez les résultats.