Université de Constantine 2
Faculté NTIC, département IFA
M1 SDIA

Mini -projet du module DAMI 1
(2024-2025)

Enoncé

Le dataset **(cancer_poumon.csv)** contient les niveaux d'expression de **50 gènes** mesurés dans **150 échantillons** de poumon.
Les trois dernières colonnes fournissent des informations complémentaires : le statut de l'échantillon : (normal ou tumeur), la source du dataset : (TCGA-LUAD ou TCGA-LUSC) et la classe associée, nommée **class**.
Trois (03) sous-classes sont disponibles :

- **NTL** (non tumoral lung) : poumon non tumoral,
- **ADK** (adenocarcinoma) : adénocarcinome du poumon,
- **SQC** (squamous cell carcinoma) : carcinome épidermoïde du poumon.

Le tableau suivant résume les informations du dataset.

Dataset	**cancer_poumon.csv**
Nombre d'instances	150
Nombres d'attributs	52+1(classe)
Attribut classe	La dernière colonne
Nombre de classes	3

Il est demandé de :

1. **Lire les données.**
2. **Faire une analyse et visualisation de données :**
    - Identifier le nombre de lignes et de colonnes,
    - Examiner le type des données et la distribution des classes,
    - Afficher les boxplots des valeurs d'expression des gènes du dataset.
3. **Effectuer un prétraitement et nettoyage des données :**
    - Traitement de données manquantes,
    - Traitement des outliers,
    - Standardisation des données.
4. **Effectuer une analyse en composantes principales (ACP) sur le dataset, puis visualiser les résultats.**
5. **Tracer le dendrogramme associé au dataset.**
6. **Réaliser un clustering hiérarchique agglomératif des données puis visualiser les résultats.**

7. **Appliquer l'algorithme k-means sur les données en déterminant le nombre optimal de clusters avec la méthode du code (*El Bow method*).**
8. **Sauvegarder tous les résultats obtenus dans des fichiers.**
9. **Appliquer la validation croisée (***avec k=3***) sur les données du dataset.**
10. **Réaliser une classification des données en utilisant :**
    - KNN (*pour k= 3, 4 et 5*),
    - Les arbres de décision,
    - Le Naive Bayes

    Tout en évaluant les modèles proposés à l'aide de l'accuracy, précision, rappel et le F1-score puis sauvegarder les résultats dans un tableau .csv

C'est permis

- Travail en monôme, binôme ou Trinôme.

C'est interdit

- Travail avec plus de trois personnes,
- Plagiat,
- Falsifier les résultats.

NB. Des points bonus seront attribués pour les interfaces graphiques.