Projet : Partie 2

Apprentissage supervisé et non supervisé

Suite à l'étape préliminaire de prétraitement du **dataset 1**, qui consolide les données relatives aux propriétés du sol dans la partie 1, nous entamons désormais une phase plus avancée de notre étude. Cette nouvelle étape revêt une importance cruciale, s'articulant autour de deux composantes majeures : d'abord, l'analyse approfondie de la fertilité du sol à travers des techniques de classification, visant à attribuer des catégories spécifiques aux différents types de sols. En parallèle, nous mettons en œuvre une approche de clustering afin de regrouper les propriétés du sol qui présentent des similarités, facilitant ainsi la détection de tendances et de structures intrinsèques au sein de notre ensemble de données. Cette double approche analytique nous permettra d'obtenir une vision holistique et nuancée des caractéristiques du sol étudié, ouvrant la voie à des conclusions plus éclairées et à des recommandations pertinentes dans le contexte d'agriculture.

1. Analyse supervisée

La classification est une forme d'analyse de données qui extrait des modèles décrivant des classes de données importantes permettant d'étiqueter de nouvelles données en se basant sur des exemples déjà vus. De nombreuses méthodes ont été proposées afin de développer des techniques de classification et de prédiction capables de gérer de grandes quantités de données. Chacune de ces méthodes comporte des avantages et des inconvénients la rendant plus ou moins adaptée à un type de dataset. Ainsi, dans cette deuxième partie du projet, il vous est demandé de construire un classificateur permettant de prédire si une instance représentant un échantillon du sol est faiblement (low), moyennement (medium), ou fortement (high) fertile. Soit l'attribut *"Fertility"* du **dataset 1**. Il vous est alors demandé de :

A. Application des algorithmes de classification :
   a. Séparer le dataset en données d'apprentissages et données de tests (80% par classe / 20% par classe, respectivement).
   b. Programmer les deux algorithmes de classification "**KNN "**, **"Decision Trees"** et **"Random Forest"**.
   c. Appliquer les trois algorithmes sur les instances du dataset.
   d. Illustrer par des exemples.
   e. Donner la Matrice de confusion.
   f. Évaluer et Comparer les modèles de classification en calculant les mesures : EXACTITUDE, SPÉCIFICITÉ, PRÉCISION, RAPPEL, F-SCORE pour chaque classe & globale en plus du temps moyen d'exécution.

B. Options avancées de l'IHM :
   a. Choix de la méthode et de ses paramètres à exécuter.
   b. Insertion des données sur un échantillon du sol et en déduire sa classe.

Belkadi Widad Hassina (G1-G2): belkadi.wh.usthb@gmail.com
Khelfa Celia (G3-G4): khelfacelia99@gmail.com

2. Analyse non supervisée

Le clustering est le processus de partitionnement d'un ensemble d'objets de données (ou d'instances) en sous-ensembles. Chaque sous-ensemble est un cluster, de sorte que les objets d'un cluster sont similaires les uns aux autres, mais différents des objets des autres clusters. Dans ce contexte, différentes méthodes de clustering peuvent générer différents clusterings sur le même ensemble de données. De nombreux algorithmes de clustering ont été proposés afin de découvrir des regroupements utiles dans de larges banques de données. Ainsi, dans cette deuxième partie du projet, il vous est demandé d'implémenter, tester et comparer des algorithmes de clustering permettant de trouver des partitionnements intéressants à partir du **dataset 1** après **pré-traitement** et en éliminant la classe (l'attribut *"Fertility")*. Il vous est alors demandé de :

A. Application d'algorithme de clustering basé partitionnement :
   a. Programmer l'algorithme de clustering **"k-means"**
   b. Expérimentation en variant les paramètres de k-means sur les instances du dataset.
   c. Évaluer, comparer et analyser les résultats de k-means.
   d. Illustrer par des exemples et graphes.

B. Application d'algorithme de clustering basé densité :
   a. Programmer l'algorithme de clustering **"DBSCAN"**
   b. Expérimentation des paramètres de DBSCAN sur les instances du dataset.
   c. Évaluer, comparer et analyser les résultats de DBSCAN.
   d. Illustrer par des exemples et graphes.

C. Comparer les deux algorithmes de clustering k-means et DBSCAN en employant les métriques adéquates tel que la distance inter-cluster et intra-cluster, la silhouette….
D. Intégrer l'exécution de ces algorithmes à l'IHM.

Notes :

- Chaque binôme est tenu d'envoyer la version électronique du rapport et le code source au plus tard : **Mardi 2 janvier 2024 à 23h59**.
- Chaque binôme devra présenter son interface et code source en Janvier [TBD].

Bon courage !

Belkadi Widad Hassina (G1-G2): belkadi.wh.usthb@gmail.com
Khelfa Celia (G3-G4): khelfacelia99@gmail.com