Projet: Partie 1
Exploitation des données et Extraction des règles d'associations

Les données du monde réel sont généralement bruitées, d'un volume énorme et peuvent provenir de sources hétérogènes. C'est pourquoi la première étape du Data Mining consiste en l'analyse (ou étude exploratoire) des données. Cela implique d'examiner de plus près les attributs et les valeurs des données. Il existe deux types de données : statique et temporelle. Les données statiques sont des données qui ne changent pas avec le temps. Elles sont fixes et ne dépendent pas de l'instant où elles sont observées. Les données temporelles, en revanche, sont des données qui évoluent avec le temps. Dans le cadre de cette première partie du projet, on veut apprendre à analyser et à nettoyer ces deux types de données. Nous optons pour un dataset statique : dataset 1 regroupant des données des propriétés du sol pour l'analyse de la fertilité du sol, et un dataset temporelle : dataset 2, représentant l'évolution du nombre de cas du COVID-19 au fils du temps par code postale.

Une deuxième étape de cette première partie consiste à extraire les motifs fréquents et règles d'association d'un troisième dataset 3. Afin de récupérer les relations existantes entre les attributs relatifs au climat, le sol, la végétation et l'utilisation d'engrais.

1. Analyse et prétraitement des données

L'analyse des données fait référence à : l'analyse des types d'attributs qui composent vos données, des types de valeurs de chaque attribut, de la nature des attributs (discrets ou continue), de la distribution des valeurs, de fournir des résumés visuels des données, de l'existence de valeurs aberrantes, de redondance, ou de valeurs manquantes, de la mesure de similarité de certains objets par rapport à d'autres, …etc. Une fois cette analyse des données est terminée, on passe au prétraitement de données incluant : Le **nettoyage des données** pour éliminer le bruit et corriger les incohérences et valeurs aberrantes. Cela peut améliorer la précision et l'efficacité des algorithmes de data Mining.

Ainsi, dans cette première étape de cette partie du projet, il vous est demandé de d'abord vous familiariser avec vos données, d'en extraire un maximum d'informations et de les nettoyer. Le travail requiert la conception et l'implémentation d'une application Python (IHM) permettant d'analyser et de nettoyer les dataset suivants :

1.1. Données statiques

L'objectif essentiel de cette section est d'entreprendre une analyse approfondie et un processus de nettoyage du dataset 1. Cette phase est primordiale, car elle prépare ces données en vue de leur utilisation dans la deuxième étape du projet, à savoir la classification et le clustering. Le dataset 1 renferme des informations relatives aux caractéristiques du sol, et sa préparation adéquate garantit la fiabilité et la pertinence des résultats que nous obtiendrons lors de l'étape suivante.

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

1.1.1. Manipulation de dataset :
- Importer et visualiser le contenu du dataset.
- Fournir une description globale du dataset.
- Fournir une description de chaque attribut.

1.1.2. Analyse des caractéristiques des attributs du dataset :
- Pour chaque attribut :
  - Calculer les mesures de tendance centrale et en déduire les symétries.
  - Construire une boîte à moustache et afficher les données aberrantes.
  - Construire un histogramme et visualiser la distribution des données.
- Construire et afficher des diagrammes de dispersion des données et en déduire les corrélations entre les propriétés du sol.

1.1.3. Prétraitement
1. Traitement des valeurs manquantes et aberrantes :
   a. Choix de la méthode de remplacement des valeurs manquantes.
   b. Choix de la méthode de traitement des valeurs aberrantes.
2. Réduction des données (élimination des redondances) horizontales / verticales.
3. Normalisation des données :
   a. Méthode Min-Max.
   b. Méthode z-score.

1.2. Données temporelles

Le but de cette partie est d'apprendre à analyser, nettoyer et visualiser des données temporelles. En utilisant le dataset 2, on désire extraire des conclusions sur la propagation du covid 19 de 2019 à 2023 aux états unis. La dataset 2 offre une vue d'ensemble complète des cas, des tests et des taux de positivité par code postal (ZIPCODE) au fil du temps. Pour ce faire on passe par les étapes suivantes :

1.2.1. Prétraitement : On procède par les prétraitements nécessaires, notamment :
- Traitement des valeurs manquantes
- Traitement des données aberrantes

1.2.2. Visualisation : Pour obtenir les conclusions nécessaires, vous devriez répondre aux questions suivantes en utilisant les **visualisations (graphes)** appropriées :
- La distribution du nombre total des cas confirmés et tests positifs par zones (**Indication :** Tree Map/Bar chart)
- Comment les tests COVID-19, les tests positifs et le nombre de cas évolue au fil du temps (hebdomadaire, mensuel et annuel) pour une zone choisit ? (**Indication** : Line chart)
- Comment les cas covid positifs sont distribués par zone et par année ? (**Indication :** Stacked Bar chart)
- Comment peut-on efficacement graphiquement représenter le rapport entre la population et le nombre de tests effectués ?
- Quelles sont les 5 zones les plus fortement impactées par le coronavirus ?

- Quel est le rapport entre les cas confirmés, les tests effectués et les tests positifs au fil du temps pour chaque zone ? (La période doit être choisie)

2. Extraction de motifs fréquents, règles d'associations et corrélations

L'exploration des motifs fréquents conduit à la découverte d'associations et de corrélations entre les éléments d'un dataset transactionnel ou relationnel. La découverte de relations de corrélation intéressantes entre d'énormes quantités d'enregistrements de transactions commerciales peut aider dans de nombreux processus décisionnels commerciaux tels que la conception de catalogues, le marketing croisé et l'analyse du comportement d'achat ou de consommation des clients. Cependant, ces applications ne se limitent pas au domaine du marketing, car l'extraction de motifs fréquents et de règles d'association trouve également sa place dans des secteurs tels que l'environnement et l'agriculture.

Dans le cadre de ce projet spécifique, notre objectif est d'analyser et d'extraire les motifs fréquents, les règles d'association et les corrélations à partir du dataset 3. Nous visons à mettre en lumière les relations existantes entre les attributs relatifs au climat (Température, Humidité, Précipitation), le sol, la végétation et l'utilisation d'engrais. Cette analyse nous permettra de dégager des informations essentielles pour prendre des décisions éclairées dans le contexte de la gestion des ressources environnementales et agricoles.

1. Choisissez un des attributs représentant le climat : 'Temperature' ou bien 'Rainfall' ou bien 'Humidity', puis discrétiser ces données continues de cet attribut, en utilisant ces 2 méthodes :
   - En classes d'effectifs égaux (equal-frequency).
   - En classes d'amplitudes égales (equal-width).
2. Extraction des motifs fréquents puis les règles d'association en utilisant l'algorithme ***Apriori.*** Effectuer des expérimentations en variant les valeurs de Min_Supp et Min_Conf.
3. Extraction des fortes règles d'associations (en utilisant les mesures de corrélation (lift, confidence, cosine...)).
4. Options avancées de l'IHM :
   a. Exécution de la méthode de DataMining à appliquer (Apriori).
   b. Choix du Min_Supp et Min_Conf.
   c. Insertion de nouvelles observations et en déduire une recommandation.

Notes :

- Le rapport doit être bien structuré et comporter **au moins :** une section analyse et pré-traitement de chacun des dataset 1 et 2, une section Extraction des motifs fréquents décrivant en détail le fonctionnement de l'algorithme Apriori, une section Extraction des règles d'associations et corrélations expliquant leur intérêt, equation, etc ainsi que les résultats de l'expérimentation des différentes valeurs de Min_Supp et Min_Conf.
- Chaque binôme est tenu d'envoyer la version électronique du rapport au plus tard : **30 Jeudi novembre 2023 à 23h59**.
- Chaque binôme devra présenter son interface et code source le **Lundi 06 et Mardi 07 Decembre 2023** durant la séance de TP.

Bon courage !

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