SORBONNE UNIVERSITÉ

AGRI BA LYSE®

PROJET LU3IN026

RESUMÉ PAR BOUTALEB ET OUNI

Ensemble experimentations sur data_synthese.csv :
- Partie supervisée : Perceptron, KNN, arbre numérique, arbre catégoriel, arbre catégoriel + numérique (One Hot Encoding), Bagging Tree
- Partie non-supervisée : Clustering hiéarchique, K-means.

Ensemble experimentations sur data_etape.csv :
- Partie supervisée : Arbre numérique
- Partie non-supervisée : Clustering hiéarchique, K-means.

Ensemble experimentations sur data_ingredients.csv :
- Partie supervisée seulement : *Arbre numérique, Arbre catégoriel+numérique.*

Tout les produits avec un DQR > 3 ont été retiré pendant l'expérimentation.

PROBLÉMATIQUE

Partie supervisée : Trouver les classifieurs et les attributs qui maximisent l'accuracy
(Label set = Groupe d'aliments)

Partie non-supervisée : Estimer le nombre de clusters optimal et essayer de voir la similarité entre les différents elements au sein du meme cluster.

FORME DES DATASETS (PARTIE SUPERVISÉ)

Data synthèse :
- Perceptron, KNN, arbre numérique - [tout , 12 "Score unique EF" : dernière colonne]
- arbre catégoriel - [tout , [8 "Livraison", 9 "Mode d'emballage", 10 "Mode de prépa"]]
- arbre numérique + catégoriel :
  * **version 1 :** [tout, attributs arbre catégoriel + tout les attributs d'impacte env]
  * **version 2 :** [tout, attributs arbre catégoriel + 12 "Score unique EF"]
- Bagging Tree : [tout, attributs arbre catégoriel + 12 "Score unique EF"]

Data étape :
- arbre numérique :
  * **version 1 :** [tout, tout les facteurs numériques de la meme étape] **pour chaque étape**
  * **version 2 :** [tout, un même facteur numérique de toutes les étapes] **pour chaque facteur**

Data ingrédients:
- arbre catégoriel et numérique : [tout, la colonne ingrédients : dernière col]
- arbre numérique :[tout, score unique : dernière col, sans col ingrédients]

RÉSULTATS (PARTIE SUPERVISÉ)

Data synthèse :
1 - Perceptron : résultats très faibles, pas plus de 0.4 pour accuracy, avec ou sans normalisation.
2 - KNN :
3 - Arbre numérique : Train accuracy = 0.98, Test accuracy = 0.84 | entropie = 0.2 (meilleur résultat)
4 - Arbre catégoriel : Train accuracy = 0.89, Test accuracy = 0.87 | epsilon = 0.1 (meilleur résultat)
5 - Arbre numérique + catégoriel (One Hot Encoding) :
  5.1 - version 1 : Train accuracy = 0.99, Test accuracy = 0.91 | entropie = 0.1 (meilleur résultat)
  5.2 - version 2 : Train accuracy = 0.99, Test accuracy = 0.92 | entropie = 0.1 (meilleur résultat)
6 - Bagging Tree : Train accuracy = 0.98, Test accuracy = 0.93 | B = 11 (meilleur résultat)

Data étape :
1 **- Arbre numérique :**
  1.1 - version 1 : Train accuracy = 0.93, Test accuracy = 0.87 | entropie = 0.1 , Etape = **"Emballage"** (meilleur résultat)
  1.2 - version 2 : Train accuracy = 0.99, Test accuracy = 0.94 | entropie = 0.1 , Facteur = **"Particules"** (meilleur résultat)

Data ingrédients :
  1.1 - desc_set à partir de la colonne ingrédients : Train accuracy = 0.975, Test accuracy = 0.794 | entropie = 0.2
  1.2 - éliminer la colonne ingrédients : Train accuracy = 0.974, Test accuracy = 0.785 | entropie = 0.1 (meilleur résultat)

FORME DES DATASETS (PARTIE NON- SUPERVISÉ)

Data synthèse :

- classification hiérarchique - [50 exemples, [8 "Livraison", 9 "Mode d'emballage", 10 "Mode de prépa", 12 "score unique EF"]]
.50 lignes aléatoires (En effet, cet algorithme est très gourmand en temps de calculs donc on a choisi 50 exemples pour aller plus vite et avoir un dendrogramme plus lisible.)

- Algorithme de K-moyennes - [tout , [8 "Livraison", 9 "Mode d'emballage", 10 "Mode de prépa", 12 "score unique EF"]]

Jeu de données appelé : **clustering_data (encodé en ONE HOT)**

Data étape:

- classification hiérarchique - [50, un même facteur numérique de toutes les étapes]
- Algorithme de K-moyennes - [tout , un même facteur numérique de toutes les étapes]

RESULTATS (PARTIE NON- SUPERVISÉ)

Data synthèse :

1 - classification hiérarchique : nb_clusters = 10 est le meilleur pour classifier les différents sous-groupes d'aliments.
en effet pour **nb_clusters = 6** on voit que boissons alcoolisées et fruits appartiennent au même cluster ce qui n'est pas logique et pour **nb_clusters = 20** on voit que le même type de produit alimentaire est distribué sur plusieurs clusters (viand crues par exemples).

2 - K-moyennes: Nombre de clusters optimal selon l'indice de Dunn : **27 (Valeur choisie pour plus d'évaluations)**
Nombre de clusters optimal selon l'indice de Xie-Beni : **33**

Data étape:

1 - classification hiérarchique : nb_cluster = 9 est le meilleur pour classifier les différents groupes d'aliments.
2 - K-moyennes: Valeur de K fixée à 27.
Meilleur groupe d'attributs selon l'indice de Dunn : Rayonnements ionisants (kBq U-235 eq/kg de produit)
Meilleur groupe d'attributs selon l'indice de Xie-Beni : : Eutrophisation eaux douces (E-03 kg P eq/kg de produit)

AUTRES INFORMATIONS

- **data_synthese** et **data_etape** sont remplacés par **data_synthese_fiable** et **data_etape_fiable (DQR <= 3)**
- Plusieurs colonnes **retirées** des datasets lors de l'analyse : Code AGB/CIQUAL, Nom du produit, Saisonnalité, Transport (par avion), DQR.
- La colonne **(Groupe d'aliments)** est utilisée comme **label_set** pour la partie supervisée.
- La colonne **(Groupe d'aliments)** est convertie en **numérique** pour Perceptron et KNN grace à notre méthode **label_num.**

- Nouvelles méthodes ajoutées à iads :
  - **Classifiers.py : one_hot_encoder(X)** qui rend un nouveau dataset encodé en One Hot.
  - **Clusters.py** : **clustering_info(df, results, nbr_clusters, index_col) -** qui décrit la valeur de la colonne **index_col** des exemples qui se trouvent dans les différents **clusters**, à partir de **df**.