Sorbonne Université
LU3IN026 - Science de données
2021/2022

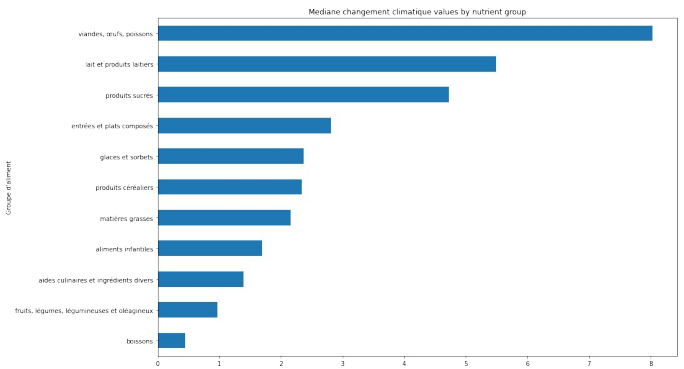
Etude de la base de données Agribalyse

Binôme :
- Samy NEHLIL
- Amel BELDJILALI

Encadrés par : Guigue & MARSALA

Problématique

Faire une étude sur la base de données Agribalyse issue du programme AGRIBALYSE, La base de données met à disposition des données de référence sur les impacts environnementaux des produits agricoles et alimentaires à travers une base de données construite selon la méthodologie des Analyses du Cycle de Vie (ACV).



Objectifs

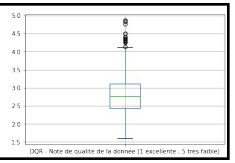
Notre étude consiste à explorer la base synthèse, cibler des attributs qui peuvent servir d'étiquettes pour notre problème de prédiction, et pour lesquels il serait intéressant de faire des prédictions.

Table	Attributs cibles (apprentissage supervisé)
Classification binaire	DQR, changement climatique
Classification multiclasse	Score unique
Classification multiclasse avec combinaison de deux attributs	Eutrophisation

Apprentissage supervisé

Construction de classes

Affichage à l'aide de histogramme des différents attributs.
Diviser les valeurs de l'attributs en utilisant les quartiles.
Ajustement des classes construites (selon l'attribut).

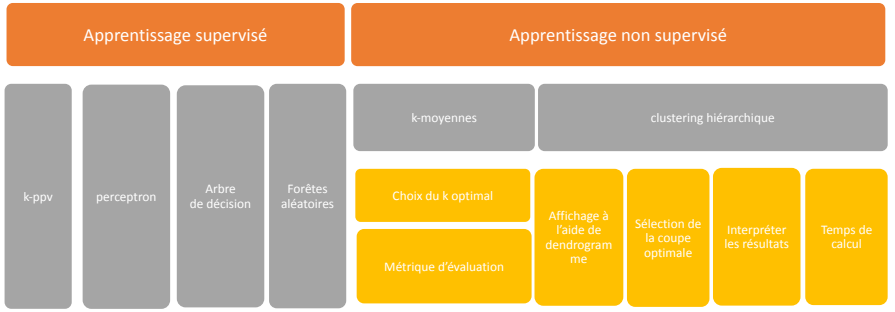


Apprentissage non supervisé

Algorithmes

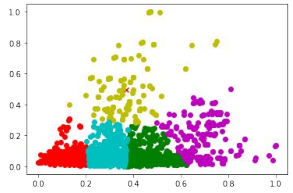
Clustering hiérarchique sur les données réduits avec PCA.
Algorithme de k-moyennes et évaluation par l'index de Dunn.

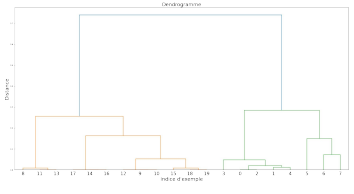
Schéma expérimental & Résultats



Approche de résolution

Exploration des données : Visualisations, types de données, dispersion des données, relations entre les attributs.
Prétraitement sur les données : Traitement des valeurs nulles, traitement des valeurs aberrantes, données catégorielles, normalisation des données (CR).
Entraînement des modèles : Sélectionner les attributs cibles, construire des classes, tester les algorithme d'apprentissage: k-ppv, perceptron, arbres de décisions, forêts aléatoires, classifieur hiérarchique, k-moyennes.
Evaluation des modèles : Utiliser la procédure de validation croisée, utiliser split en train/test, dunn index.
Synthèse : Analyse succincte des résultats et interprétation du comportement des différents algorithmes sur les données choisies.





Analyse des résultats et synthèse

Apprentissage supervisé :
- L'étude de cette base permet de prédire plusieurs attributs.
- La construction de classes pour le problème de classification supervisé.
- Les algorithmes d'apprentissage supervisé donnent de bonnes sur les étiquettes choisies (moyenne de 80% de précision).

Apprentissage non supervisé:
L'utilisation de l'algorithme de k-means sur la base synthèse avec k=5 permet d'obtenir une bon clustering confirmé par l'index de Dunn (0.04).

Références

Documentation officielle sur la base de données Agribalyse du programme AGRIBALYSE
https://doc.agribalyse.fr/documentation/

https://www.altexsoft.com/blog/datascience/how-to-organize-data-labeling-for-machine-learning-approaches-and-tools/