Projet Fin Semestre :
Etude sur la Base de Données Agribalyse

IA & Data Science (LU3IN0226) -- 2022-2023

Participants

Ilyes DJERFAF

Anyes TAFOUGHALT

Introduction

Dans le cadre de notre projet de fin de semestre du module IA & Data Science (LU3IN266), nous avons exploré les données de la base AGRIBALYSE en appliquant les algorithmes d'apprentissage vus tout au long du semestre. Notre objectif était de mettre en évidence des résultats intéressants et de relever deux types de problèmes : l'apprentissage supervisé et l'apprentissage non supervisé.

Problématique

Lors de cette étude nous allons traiter deux problématique différentes avec deux formes d'apprentissage différents.

- Prédiction de l'impact environnemental des produits alimentaires : Apprentissage supervisé. --> Problème 1
- Analyse des Méthodes de Normalisation - Min-Max et StandardScalar - sur la Performance des Algorithmes Non Supervisé pour voir l'Impact Environnemental des types d'Emballages dans l'Industrie Agroalimentaire et les Effets toxicologiques sur la santé humaine : substances non cancérogènes --> Problème 2

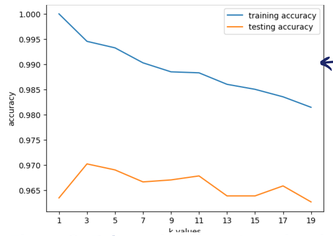
Problème 1 : Prédiction de l'impact environnemental des produits alimentaires : Apprentissage supervisé

Méthodologie :

- Nettoyage de Données
- Suppression des données non pertinentes grâce à la matrice de corrélation
- Visualisation des données pré-traitées
- Application des algorithmes d'apprentissages :
  - KNN, Perceptron et Perceptron Biais
- Recherche des paramètres optimaux pour chaque algorithmes et comparer les résultats.
- Visualisation des résultats.

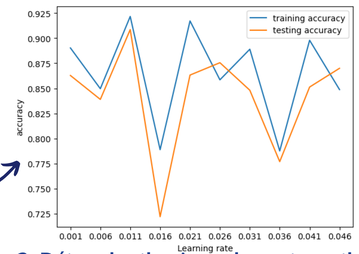
Figure1. Matrice de corrélation avant et après suppression des variables fortement corrélé (seuil = 0.8)

Résultats des Expérimentations sur le problème supervisé



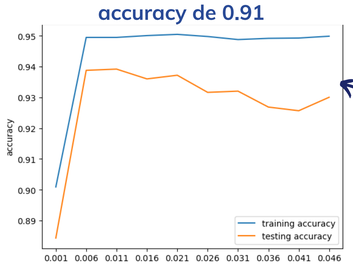
Après avoir analysé les résultats d'accuracy en fonction des différents K utilisés dans l'algorithme KNN, On a constaté que le K optimal est de 3, avec une précision de 0,97.

Figure2 Détermination K-optimal pour Knn



Concernant le perceptron, nous avons trouvé que le learning rate le plus optimal est égale à 0.011, avec une accuracy de 0.91

Figure3 Détermination Learning rate-optimal pour Perceptron



Enfin, pour le perceptron biais, le taux d'apprentissage optimal a été identifié avec un learning rate égale à 0,006, atteignant une accuracy de 0,94.

Figure4 Détermination Learning rate-optimal pour Perceptron Biais

Conclusion sur les classifieurs supervisés

D'aprés ce tableau, on peut conclure que la classifieur KNN avec k = 3 est le meilleur classifieur pour traiter notre problématique qui s'agit de trouver l'impact envirennmental d'un produit alimentaire.

Algorithme de classification	Accuracy
KNN	0.97
Perceptron	0.91
Perceptron Biais	0.94

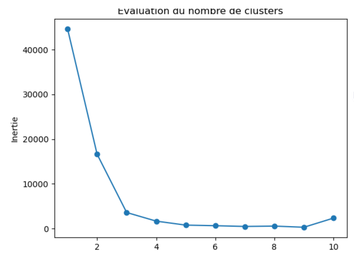
Tableau1. Table récapitulative des résultats problème supervisé

Problème 2 : Analyse des Méthodes de Normalisation - Min-Max et StandardScalar - sur la Performance des Algorithmes Non Supervisé

Méthodologie :

- Nettoyage de Données
- Suppression des données non pertinentes grâce à la matrice de corrélation
- Encodage des variables catégorielle (Label Encoding)
- Génération des deux datasets selon la méthode de normalisation utilisé
  - Min-Max : X_norm = (X - X.min()) / (X.max() - X.min())
  - StandardScalar : X_scaled = (X - X.mean()) / X.std()
- Application de K-moyennes pour comparer les deux méthodes
- Comparer les résultats de notre algorithmes avec l'algorithme KMEANS de SKlearn

Résultats des Expérimentations sur le problème non supervisé



On peut voir que le nombre optimal de clusters est : 9 (avec une inertie minimale de 303.78

Figure5. Evaluation du nombre du cluster selon l'inertie globale - Normalisation Min-Max

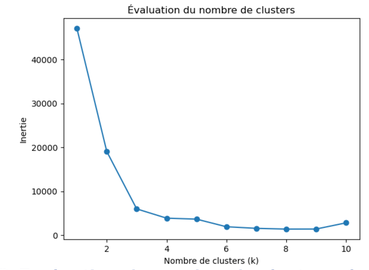
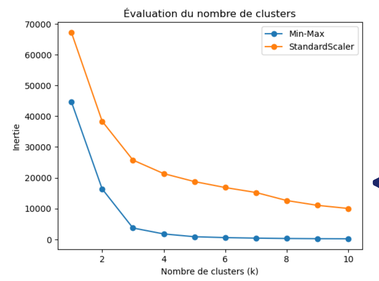


Figure6. Evaluation du nombre du cluster selon l'inertie globale - Normalisation StandardScalar

on peut voir que le nombre optimal de clusters est : 8 (avec une inertie minimale de 1403.32



La normalisation avec min-max est meilleure dans notre cas d'étude non supervisé que la normalisation standard-Scalar

Figure7. Résultat d'application de Kmeans de Sclearn avec deux normalisations

Conclusion sur le problème non supervisé

Normalisation \ Algorithme	Knn-Implémenté	KMEANS-Sklearn
Min-Max	303.78 (9 cluters)	175.37 (10 cluster)
Standard_Scalar	1403.32 (8 cluster)	10031.13 (10 cluster)

Tableau2. Table récapitulative des résultats problème non supervisé

la normalisation Min-Max s'est avérée plus adaptée à notre problème non supervisé utilisant l'algorithme K-means. Cependant, il est important de noter que le choix de la méthode de normalisation dépendra des caractéristiques spécifiques du jeu de données et des objectifs de l'analyse.

Références

- Cours IA & Data Science LU3IN0226 - 2022/2023
- Label-Encoding
- KMEANS
- Standard Scalar Normalization
- scikit-learn
- Correlation Matrix