Participants

Ilyes DJERFAF

Anyes TAFOUGHALT

Projet Fin Semestre : Etude sur la Base de Données Agribalyse

SCIENCES SORBONNE UNIVERSITÉ

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

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.
- Etude sur les emballages dans l'industrie agroalimentaire et la toxicité non cancérogène : Apprentissage non supervisé.

Jeu de Données : Agribalyse (v3.1)

- Qu'est-ce qu'Agribalyse ?

Agribalyse est une base de données de référence des indicateurs d'impacts environnementaux des produits agricoles produits en France et des produits alimentaires consommés en France.

- Que trouve-t-on dans ces données ?

Cette base de donnée contient 3 fichiers .csv : AGRIBALYSE3-synthese.csv, AGRIBALYSE3-etapes.csv et AGRIBALYSE3-ingredients.

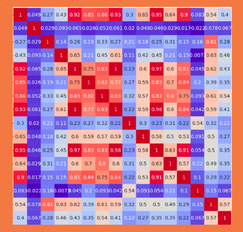
Fichiers	Lignes	Colonnes
AGRIBALYSE3-synthese.csv	2517	29
AGRIBALYSE3-etapes.csv	2517	132
AGRIBALYSE3-ingredients.csv	2517	27

Dans ce projet, nous avons faire nos études sur le fichier AGRIBALYSE3-synthese.csv.

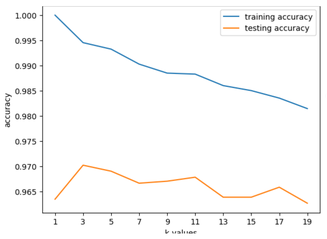
Méthodologie

Notre démarche est la suivante :

- Pré-traitement des données/ Nettoyage de Données : Nous avons géré les valeurs manquantes, encodé les variables catégorielles, normalisé les données pour les préparer à l'analyse et enfin supprimé les colonnes non pertinentes (en utilisant la matrices de corrélation).
- Visualisation des données prétraitées
- Application des algorithmes d'apprentissages :
  - Pour le problème supervisé, nous avons appliqué KNN, Perceptron et Perceptron Biais
  - Pour le problème non supervisé, nous avons appliqué K-moyennes
- Recherche des paramètres optimaux pour chaque algorithmes et comparer les résultats.
- Visualisation des résultats.

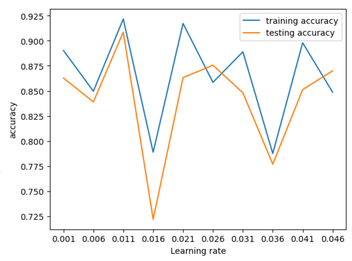


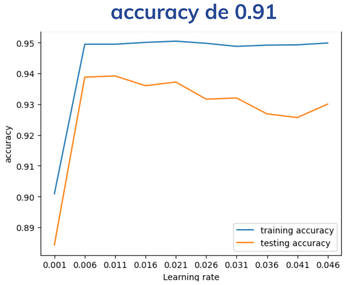
Résultats des Expérimentations sur le problème supervisé : Comment prédire l'impact environnemental des produits alimentaires ?



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.

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





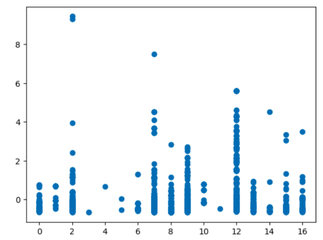
Enfin, pour le perceptron biais, le taux d'apprentissage optimal a été identifié avec un learning rate égale à 0,006, atteignant une accuracy de 0,94.

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.

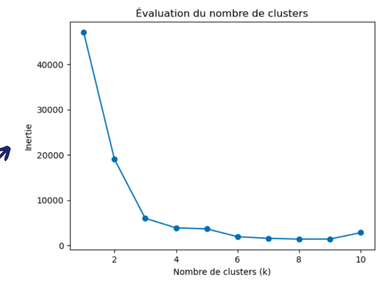
Algorithme de classification	Accuracy
KNN	0.97
Perceptron	0.91
Perceptron Biais	0.94

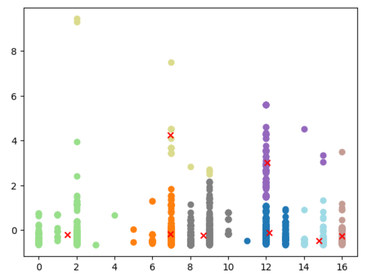
Résultats des Expérimentations sur le problème non supervisé : Etude sur les emballages dans l'industrie agroalimentaire et la toxicité non cancérogène



Un plot qui représente les matériaux d'emballage (encodé) sur l'abscisse des X, et l'Effets toxicologiques sur la santé humaine : substances non-cancérogènes (normalisé) sur l'abscisse des Y

Application de l'algorithme K-moyennes et mesure des inerties globales à chaque exécution, Conclusion k optimal = 8





Résultat final de l'application de l'algorithme K-moyennes avec k=k_otimal=8

Conclusion

- Suite à nos expérimentations faites sur bases de données, on a conclu que pour avoir des bons résultats à la sortie des algorithmes (supervisé et non supervisé), le prétraitement des données est l'étape la plus importante.
- Le choix de la normalisation varie selon les problèmes supervisés et non supervisés
- La matrice de corrélation est un moyen qui permet de détecter les relations cachées entre les attributs