IA & Data science (LU3IN026) – Agribalyse 3.1 – Yassine Alallah

Prédiction du groupe d’aliment en fonction de différents indicateurs environnementaux via algorithmes d’apprentissage **supervisé**

Abstract

Les données sont issues des données publiques du site de l'ADEME: https://agribalyse.ademe.fr/

Pour ce projet, vous travaillerez sur les données sur les produits alimentaires dont la version originale est visible ici : https://doc.agribalyse.fr/documentation/acces-donnees

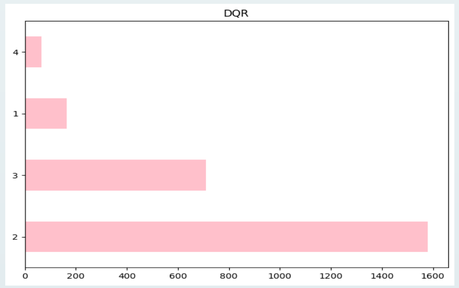
Problem

Nous allons analyser les donnée « Agribalyse v.3.1 » afin de chercher à predire le groupe d’aliments associé à differentes caractéristiques environnementales. De plus nous chercherons à prédire la qualité des données de notre dataset.

Introduction

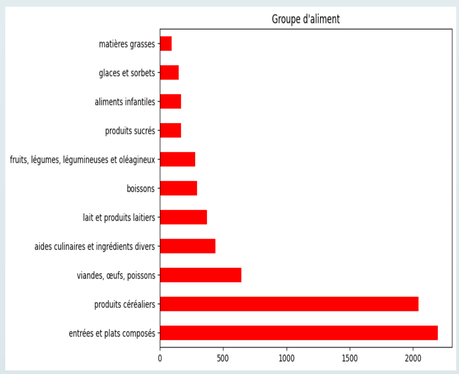
Pour réaliser nos differentes analyses. Nous utiliserons deux algorithmes d'apprentissage supervisé : kNN et Arbre de décision et un algorithme d'apprentissage non supervisé : K-means.

Nos analyses s’articuleront autour des données ci-dessous :



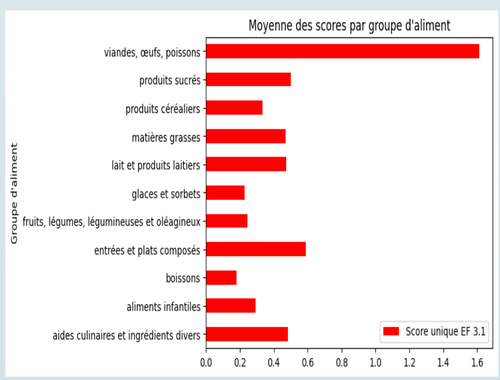
Une note de qualité - le Data Quality Ratio (DQR) - de 1, très bon, à 5, très mauvais - est associée à chaque produit agricole et alimentaire pour lequel Agribalyse fournit des inventaires de cycle de vie et des indicateurs d’impacts. Dans la base de données AGRIBALYSE, 67 % des données ont un DQR jugé bon ou très bon (1 à 3).

Se réferer au données de synthèse.



Il existe 11 groupes d’aliments dans nos datasets.

Se réferer au données sur les différents ingrédients.



Un score unique est également proposé : il s’agit du « single score EF » préconisé par la Commission Européenne , calculé avec des facteurs de pondération pour chacun des indicateurs ; la pondération prend à la fois en compte la robustesse relative de chacun de ces indicateurs et les enjeux environnementaux.

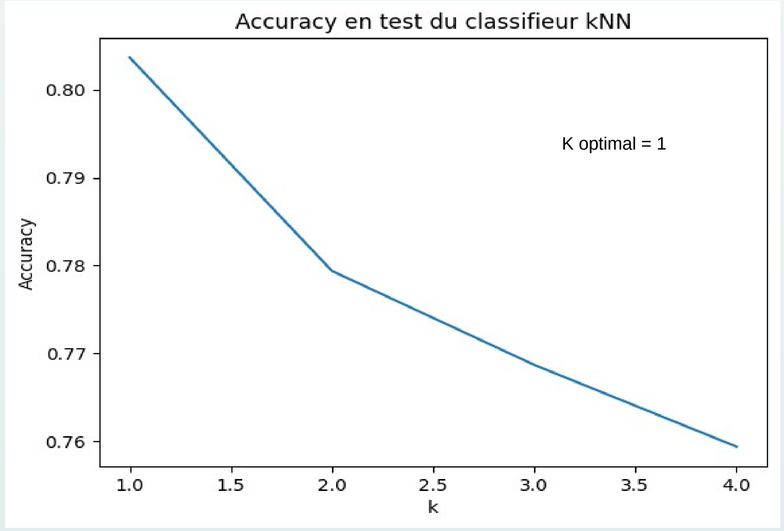
Se réferer au données de synthèse.

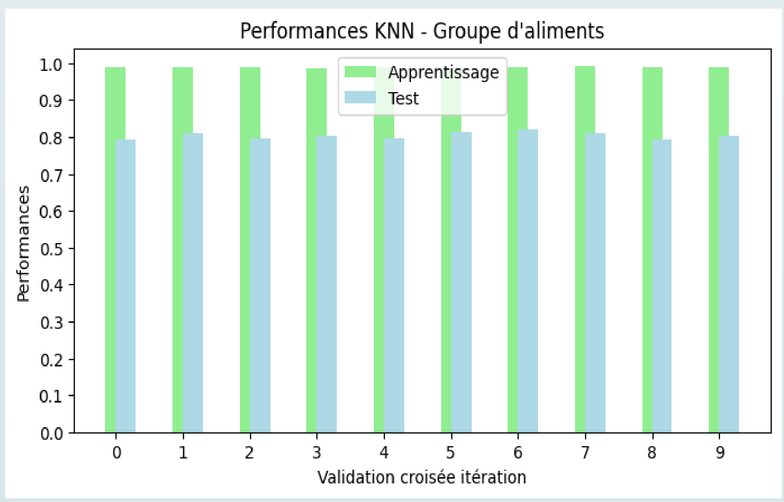
Methodology n°1

Nous avons choisis le dataset « AGRIBALYSE3-ingredients.csv » afin de réaliser notre première classification. En effet, afin d’appliquer notre algorithme kNN pour prédire le groupe d’aliments, nous avons procéder de la sorte :

1 – Récupération des seules données numérique de notre dataset.
2 – Normalisation de notre dataset.
3 – Récupération de nos labels provenant de la colonne « Groupe d’aliments » et association de chaque label à un entier.
4 – Recherche de notre k - nombre de voisin - optimal pour notre kNN.
5 – Validation croisée en 10 pour mesurer le taux de bonnes classification sur le dataset de test.

Results



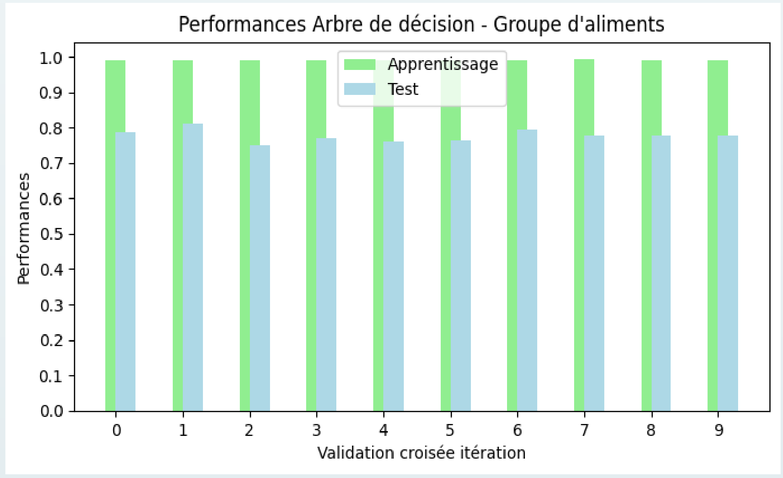


Methodology n°2

Nous avons choisis le dataset « AGRIBALYSE3-ingredients.csv » afin d’appliquer notre algorithme Arbre de décision numérique pour prédire le groupe d’aliments, pour cela nous avons procéder de la sorte :

1 – Récupération des seules données numérique de notre dataset.
2 – Normalisation de notre dataset.
3 – Récupération de nos labels provenant de la colonne « Groupe d’aliments » et association de chaque label à un entier.
4 – Validation croisée en 10 pour mesurer le taux de bonnes classification sur le dataset de test.

Results



Conclusion

Les performances de nos classifieurs sont assez bonnes, en entrainement et en test ; elles sont généralement supérieure à 75%. En d'autres termes, en connaissant différents indicateurs d'impacts environnementaux des différents ingrédients, il est possible de prédire le groupe auquel cet ingrédient appartient.

IA & Data science (LU3IN026) – Agribalyse 3.1 – Yassine Alallah

Prédiction de la qualité des données en fonction de différents indicateurs environnementauxvia algorithmes d’apprentissage **supervisé**

Abstract

Les données sont issues des données publiques du site de l'ADEME: https://agribalyse.ademe.fr/

Pour ce projet, vous travaillerez sur les données sur les produits alimentaires dont la version originale est visible ici : https://doc.agribalyse.fr/documentation/acces-donnees

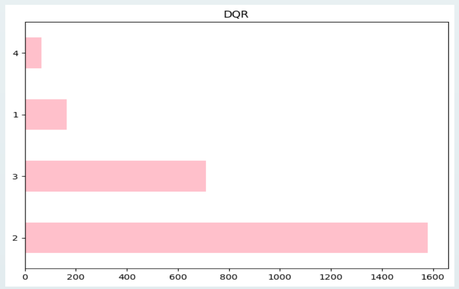
Problem

Nous allons analyser les donnée « Agribalyse v.3.1 » afin de chercher à predire le groupe d’aliments associé à differentes caractéristiques environnementales. De plus nous chercherons à prédire la qualité des données de notre dataset.

Introduction

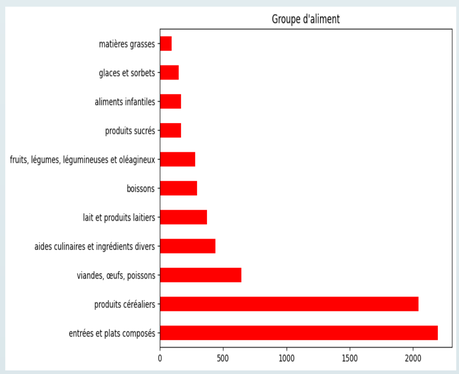
Pour réaliser nos differentes analyses. Nous utiliserons deux algorithmes d'apprentissage supervisé : kNN et Arbre de décision et un algorithme d'apprentissage non supervisé : K-means.

Nos analyses s’articuleront autour des données ci-dessous :



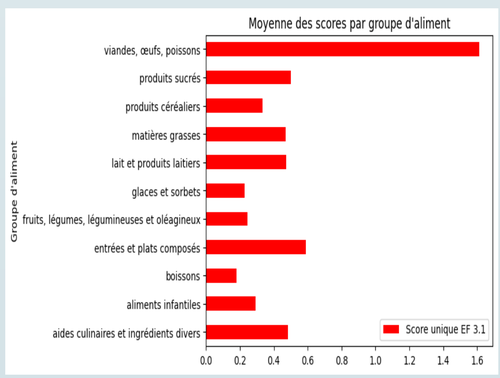
Une note de qualité - le Data Quality Ratio (DQR) - de 1, très bon, à 5, très mauvais - est associée à chaque produit agricole et alimentaire pour lequel Agribalyse fournit des inventaires de cycle de vie et des indicateurs d'impacts. Dans la base de données AGRIBALYSE, 67 % des données ont un DQR jugé bon ou très bon (1 à 3).

Se réferer au données de synthèse.



Il existe 11 groupes d’aliments dans nos datasets.

Se réferer au données sur les différents ingrédients.



Un score unique est également proposé : il s’agit du « single score EF » préconisé par la Commission Européenne , calculé avec des facteurs de pondération pour chacun des indicateurs ; la pondération prend à la fois en compte la robustesse relative de chacun de ces indicateurs et les enjeux environnementaux.

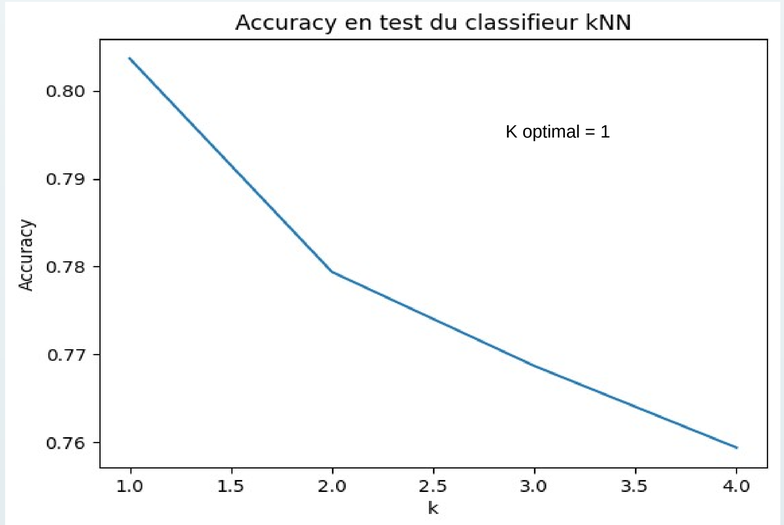
Se réferer au données de synthèse.

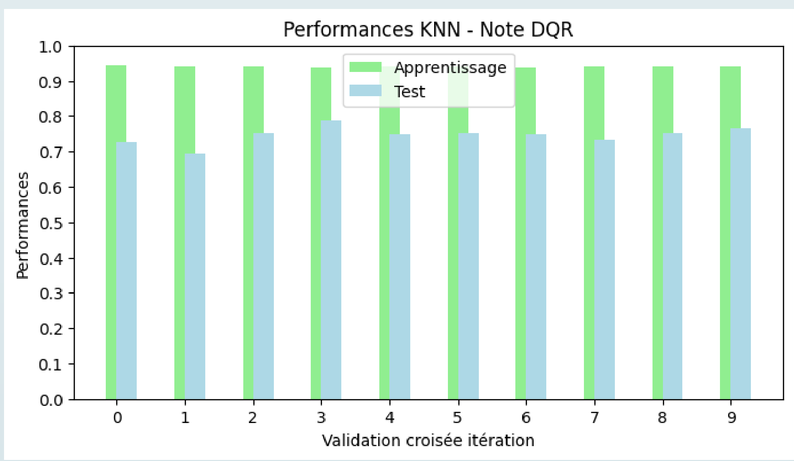
Methodology n°1

Nous avons choisis le dataset « AGRIBALYSE3-synthese.csv » afin de réaliser notre seconde classification. En effet, afin d’appliquer notre algorithme kNN pour prédire la qualité des données de notre dataset, nous avons procéder de la sorte :

1 – Récupération des seules données numérique de notre dataset.
2 – Normalisation de notre dataset.
3 – Récupération de nos labels provenant de la colonne « DQR » et découpage de nos classes en 5 familles afin de rendre entier nos labels.
4 – Recherche de notre k - nombre de voisin - optimal pour notre kNN.
5 – Validation croisée en 10 pour mesurer le taux de bonnes classification sur le dataset de test.

Results



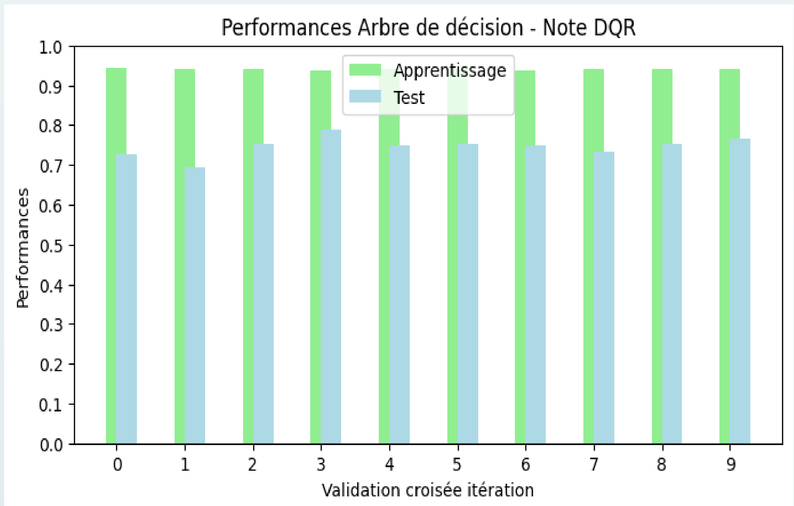


Methodology n°2

Nous avons choisis le dataset « AGRIBALYSE3-synthese.csv » afin d’appliquer notre algorithme Arbre de décision numérique pour prédire la qualité des données, pour cela nous avons procéder de la sorte :

1 – Récupération des seules données numérique de notre dataset.
2 – Normalisation de notre dataset.
3 – Récupération de nos labels provenant de la colonne « DQR » et découpage de nos classes en 5 familles afin de rendre entier nos labels.
4 – Validation croisée en 10 pour mesurer le taux de bonnes classification sur le dataset de test.

Results



Conclusion

Les performances de nos classifieurs sont assez bonnes, en entrainement et en test ; elles sont généralement supérieure à 75%. En d'autres termes, en connaissant différents indicateurs d'impacts environnementaux des différents ingrédients, il est possible de prédire la note DQR associée à ces indicateurs.

IA & Data science (LU3IN026) – Agribalyse 3.1 – Yassine Alallah

Prédiction du groupe d’aliment en fonction de différents indicateurs environnementaux via algorithmes d’apprentissage **non-supervisé**

Abstract

Les données sont issues des données publiques du site de l'ADEME: https://agribalyse.ademe.fr/

Pour ce projet, vous travaillerez sur les données sur les produits alimentaires dont la version originale est visible ici : https://doc.agribalyse.fr/documentation/acces-donnees

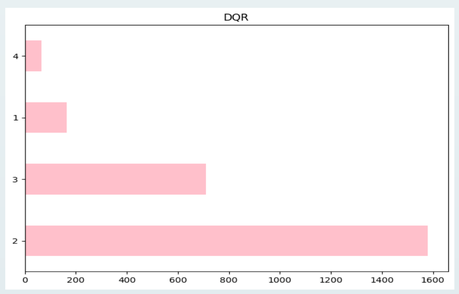
Problem

Nous allons analyser les donnée « Agribalyse v.3.1 » afin de chercher à predire le groupe d’aliments associé à differentes caractéristiques environnementales. De plus nous chercherons à prédire la qualité des données de notre dataset.

Introduction

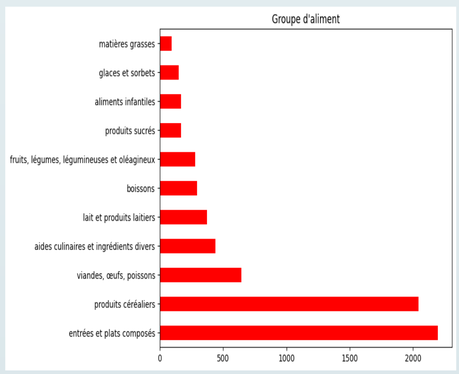
Pour réaliser nos differentes analyses. Nous utiliserons deux algorithmes d'apprentissage supervisé : kNN et Arbre de décision et un algorithme d'apprentissage non supervisé : K-means.

Nos analyses s’articuleront autour des données ci-dessous :



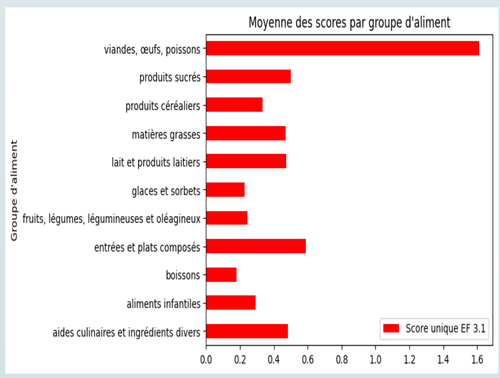
Une note de qualité - le Data Quality Ratio (DQR) - de 1, très bon, à 5, très mauvais - est associée à chaque produit agricole et alimentaire pour lequel Agribalyse fournit des inventaires de cycle de vie et des indicateurs d’impacts. Dans la base de données AGRIBALYSE, 67 % des données ont un DQR jugé bon ou très bon (1 à 3).

Se réferer au données de synthèse.



Il existe 11 groupes d’aliments dans nos datasets.

Se réferer au données sur les différents ingrédients.



Un score unique est également proposé : il s’agit du « single score EF » préconisé par la Commission Européenne , calculé avec des facteurs de pondération pour chacun des indicateurs ; la pondération prend à la fois en compte la robustesse relative de chacun de ces indicateurs et les enjeux environnementaux.

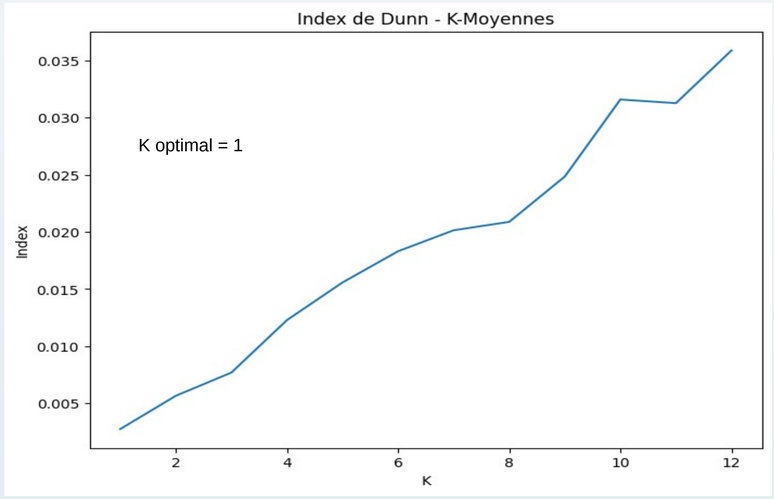
Se réferer au données de synthèse.

Methodology n°1

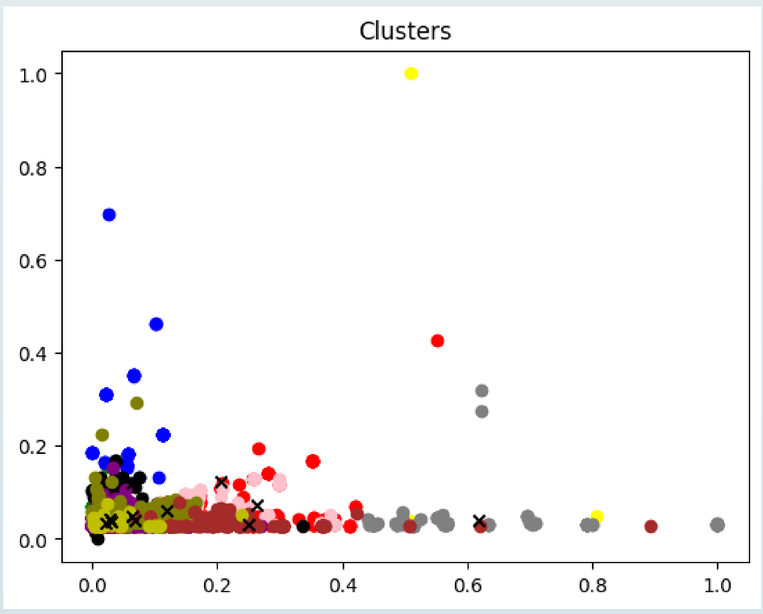
Nous avons choisis le dataset « AGRIBALYSE3-etape.csv » afin de réaliser notre clustering. En effet, afin d’appliquer notre algorithme K-means pour observer s’il existe une correlation entre les N clusters et les N groupes d’aliments, nous avons donc procéder de la sorte :

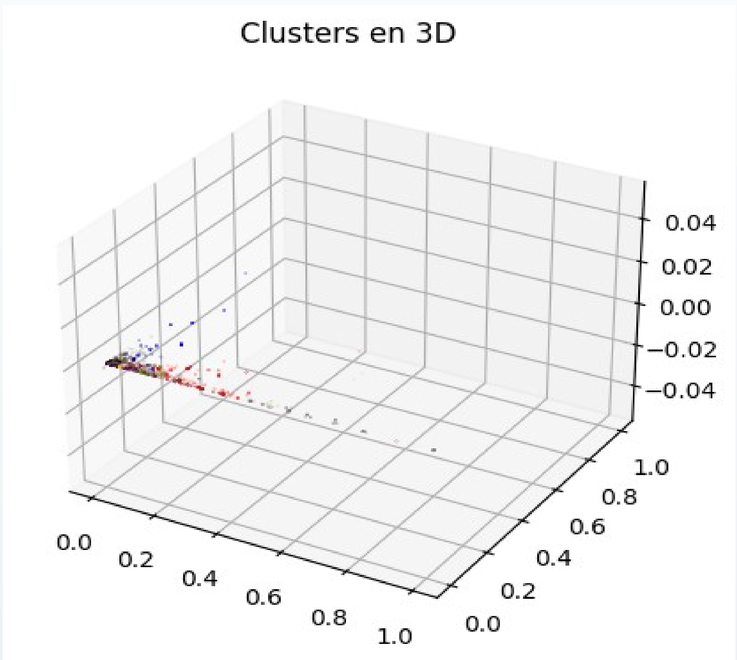
1 – Récupération des seules données numérique de notre dataset.
2 – Normalisation de notre dataset.
3 – Calcul du K optimal afin de faire une premiere observation quant à la correlation entre clusters et groupe d’aliments.
4 – Aplication de l’algorithme K-moyennes pour N clusters.
5 – Analyse de la classe majoritaire dans chaque cluster

Results



Il existe 11 groupes d’aliments, nous allons donc choisir K = 11.






```
for (centre, indices) in U.items():
    print("Cluster numéro :",centre)
    count = np.unique(np.array(data_etapes.iloc[indices,[2]]), return_counts=True)
    print(count[0][np.argmax(count[1])])
```

```
Cluster numéro : 0
viandes. œufs. poissons
Cluster numéro : 1
viandes. œufs. poissons
Cluster numéro : 2
viandes. œufs. poissons
Cluster numéro : 3
fruits. légumes. légumineuses et oléagineux
Cluster numéro : 4
viandes. œufs. poissons
Cluster numéro : 5
lait et produits laitiers
Cluster numéro : 6
lait et produits laitiers
Cluster numéro : 7
produits céréaliers
Cluster numéro : 8
viandes. œufs. poissons
Cluster numéro : 9
entrées et plats composés
Cluster numéro : 10
boissons
```

Conclusion

Le clustering en 11 clusters ne permet pas de faire ressortir un groupe d'aliment unique different pour chacun de nos clusters.

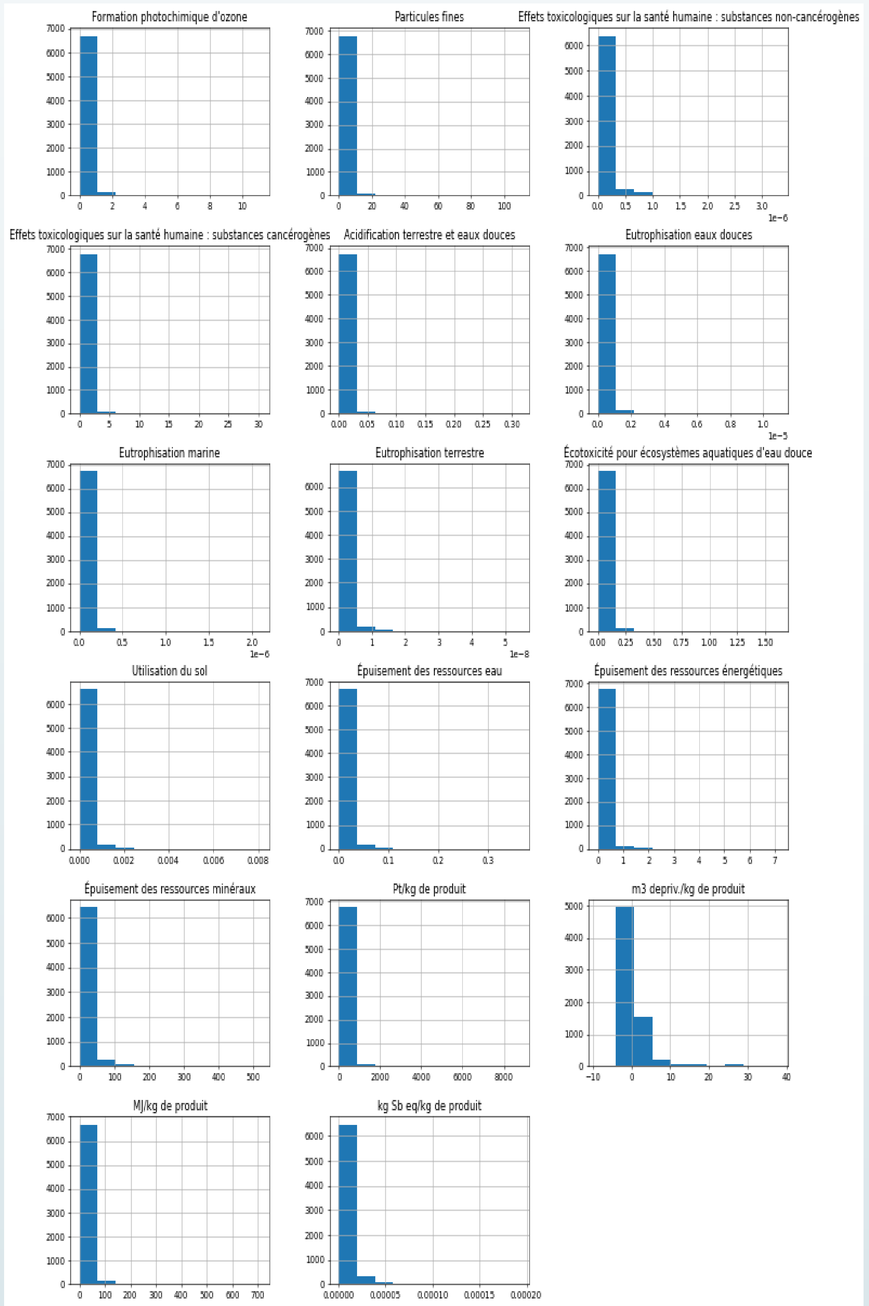
IA & Data science (LU3IN026) – Agribalyse 3.1 – Yassine Alallah

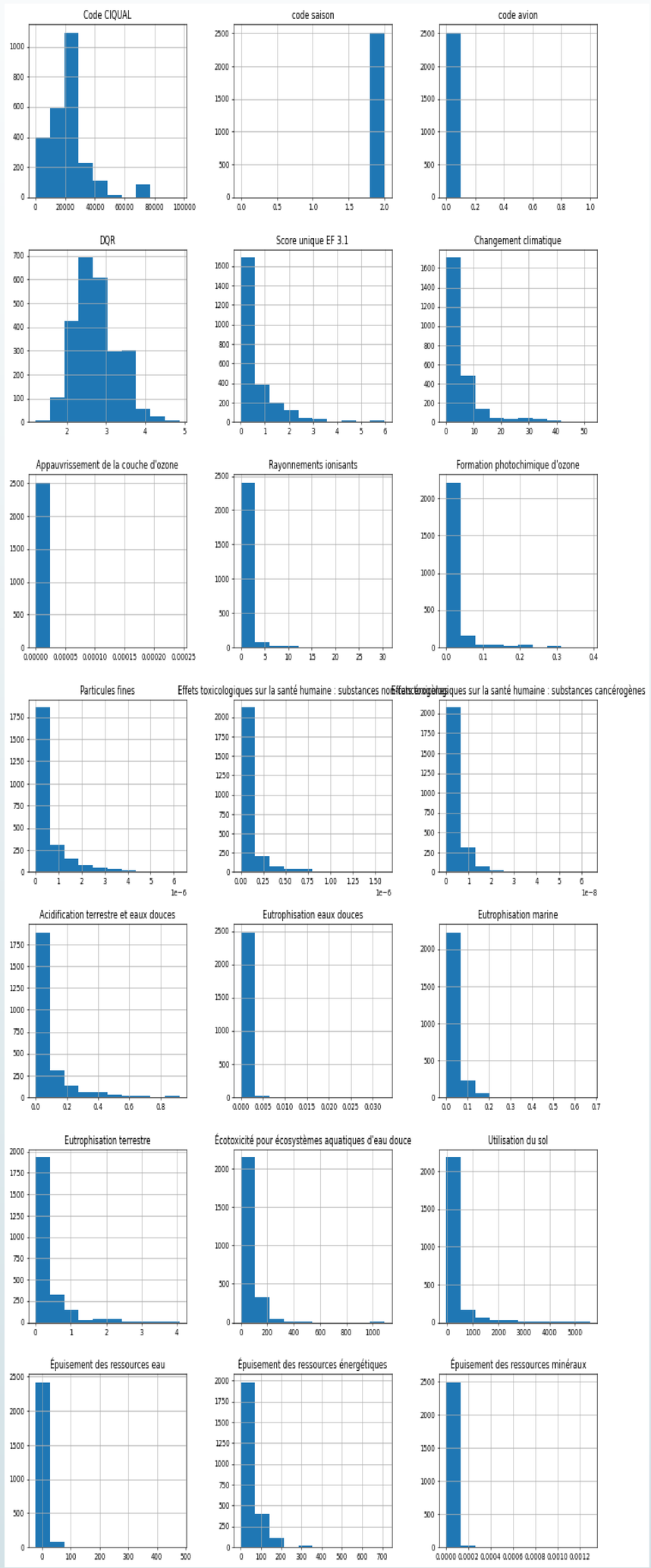
Prise en main des données

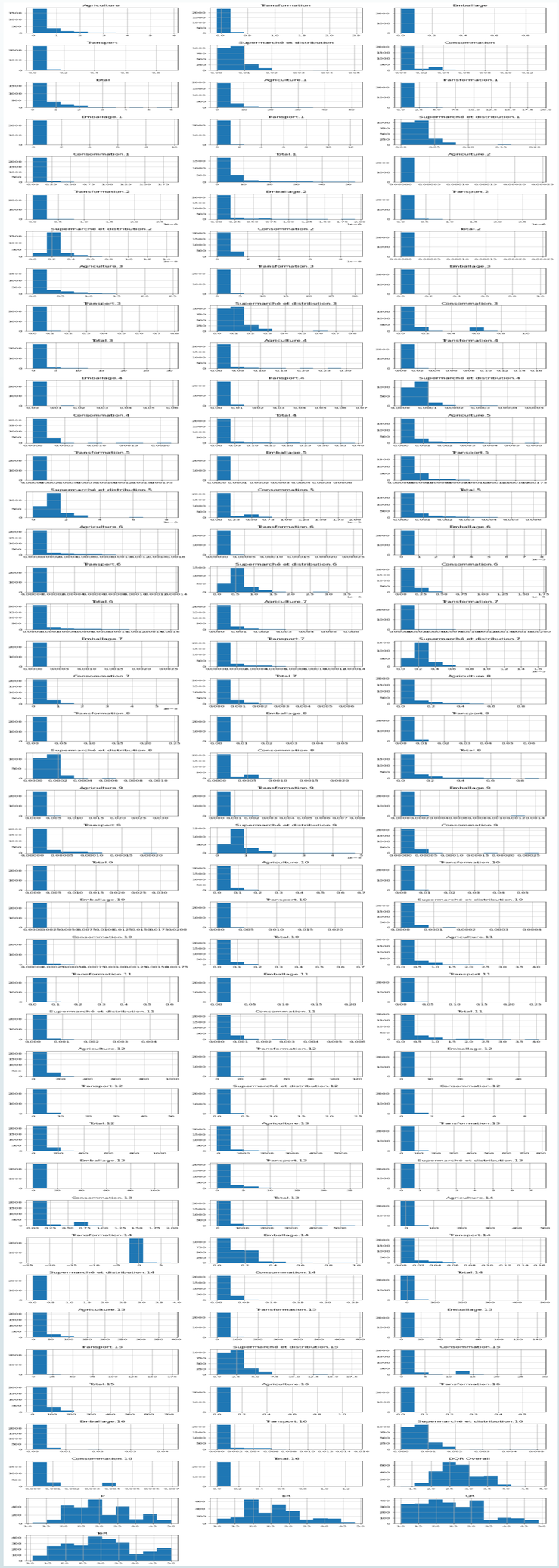
Abstract

Les données sont issues des données publiques du site de l'ADEME: https://agribalyse.ademe.fr/

Pour ce projet, vous travaillerez sur les données sur les produits alimentaires dont la version originale est visible ici : https://doc.agribalyse.fr/documentation/acces-donnees







Conclusion

Les données ne sont pas uniformes.