Analyse du nutriscore

Damien et Amaury

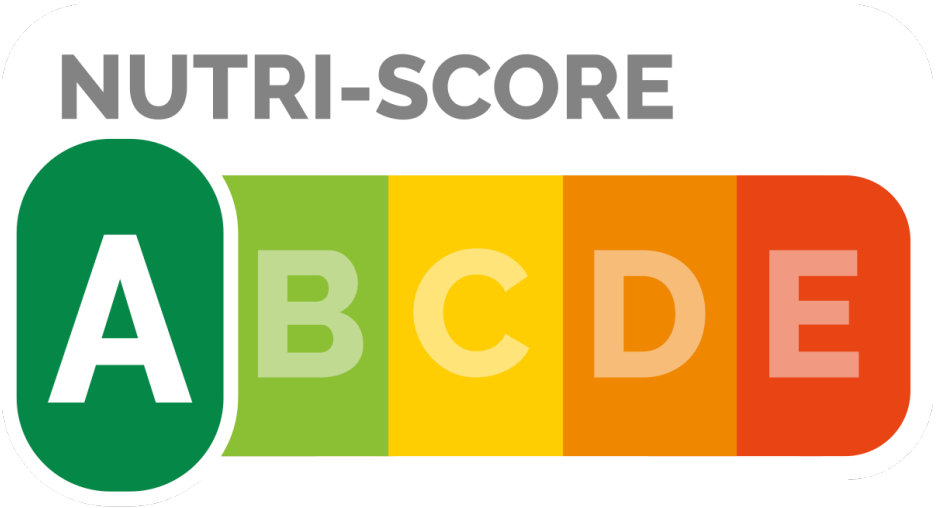
Objectif

L'objectif du projet est de prédire le nutriscore grâce aux données quantitatives
De la base de donnée https://fr.openfoodfacts.org/ avec des algorithmes de machine learning.

Sommaire

- Présentation des données du nutriscore
- Méthodologie
- Analyse exploratoire des données du nutriscore
- Nettoyage des données
- Visualisation des données
- ACP
- Modèle prédictif
  - Random forest
  - KNN
  - Test des modèles
  - Conclusion

Présentation du nutriscore

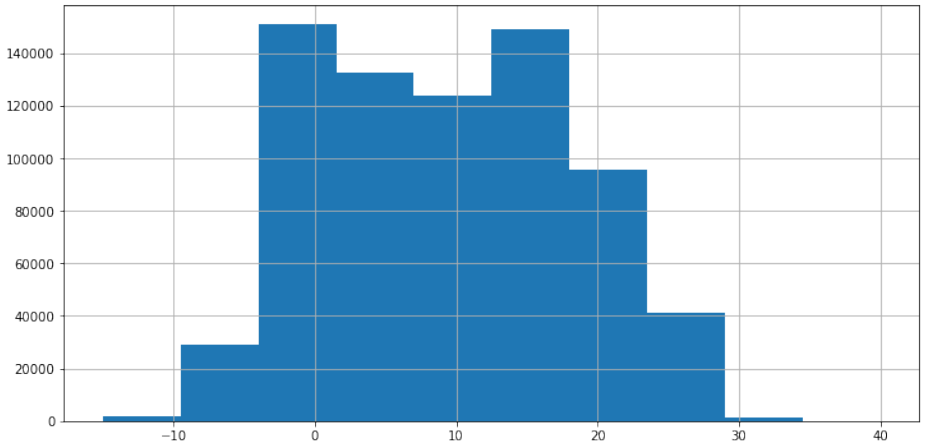


- Un logo apposé en face avant des emballages qui informe sur la qualité nutritionnelle des produits sous une forme simplifiée et complémentaire à la déclaration nutritionnelle obligatoire (fixée par la réglementation européenne)
- Basé sur une échelle de 5 couleurs : du vert foncé au orange foncé
- Associé à des lettres allant de A à E pour optimiser son accessibilité et sa compréhension par le consommateur

Méthodologie

- Chargement des données avec pandas
- Visualisation des données
- Sélection des colonnes utiles dans le cadre de notre projet
- Vérification des types
- Traitement des outliers
- Visualisation des données
- ACP
- Machine learning
- Dashboard

Analyse exploratoire des données du nutriscore

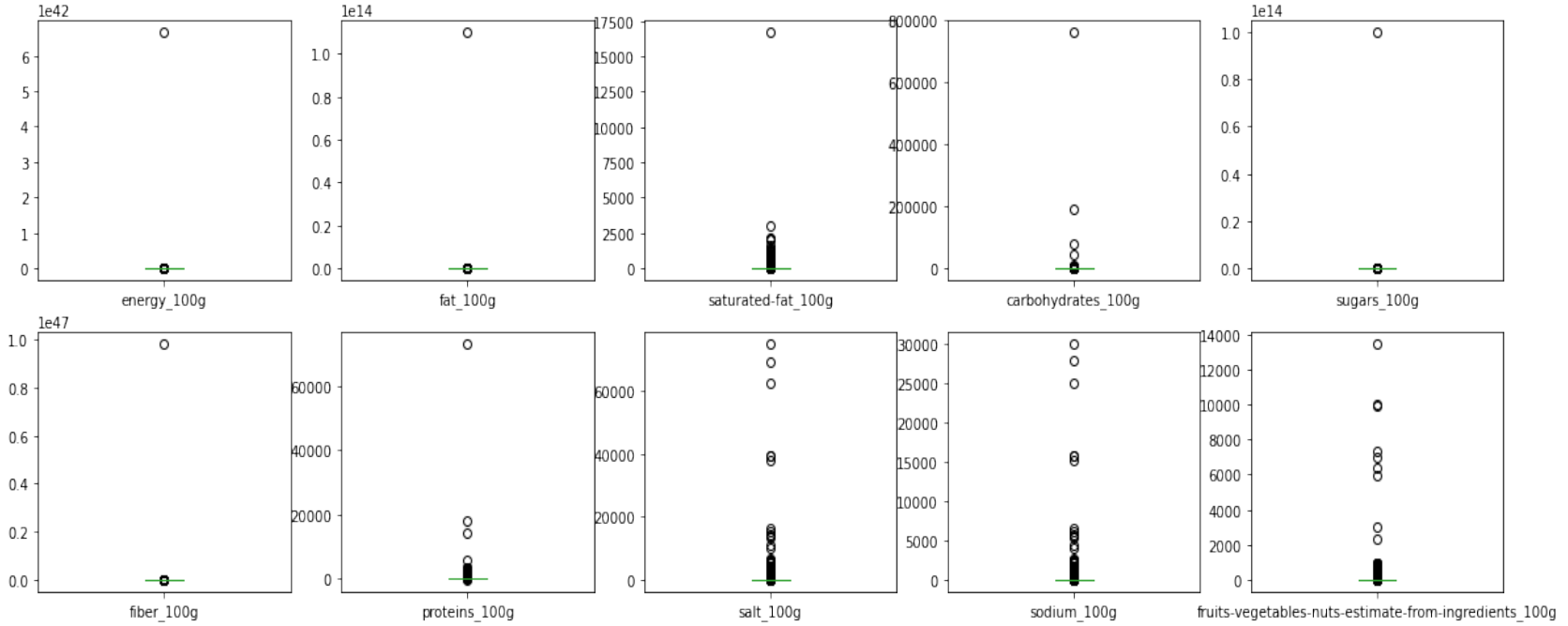


Répartition du nutriscore

On remarque que le nutriscore oscille du -10 à + 30

Analyse exploratoire des données du nutriscore

On remarque que les données possèdent de nombreux outliers, donc nous allons les nettoyer.



Nettoyage des données

Données avant nettoyage : (2033614, 187)

Après la suppressions des NAN : (2033356, 18)

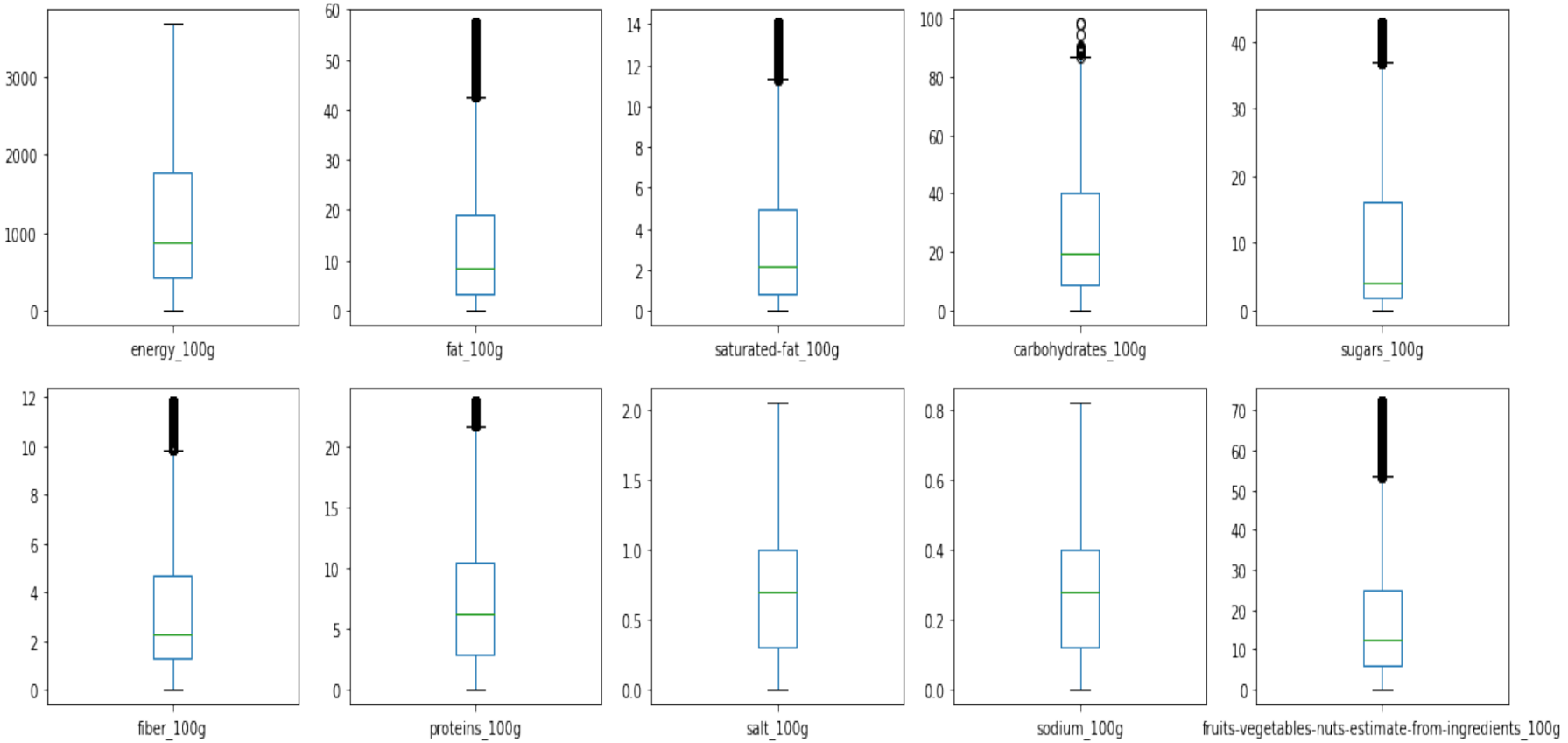
Après la suppressions des doublons et des colonnes inutiles pour la prédiction du nutriscore : (2033356, 18)

Après la suppressions des valeurs aberrantes : (59254, 10)

Après la suppressions des données atypique : (42597, 10)

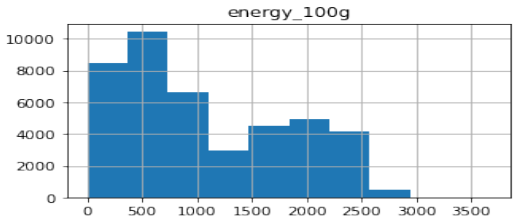
Nettoyage des données

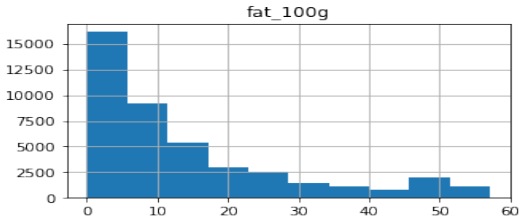
Les données après traitement

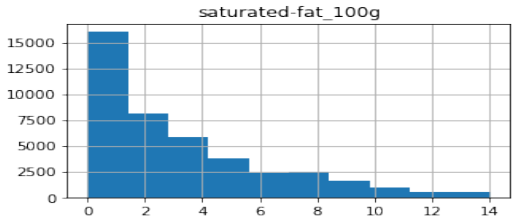


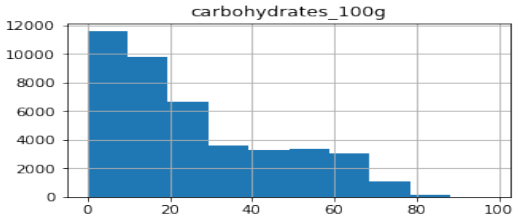
Visualisation des données

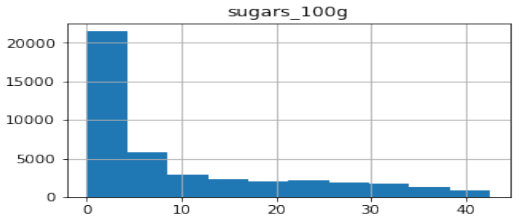
Répartition des données dans les différentes colonnes.

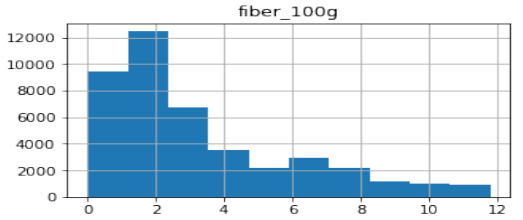


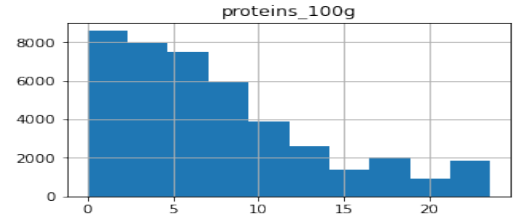


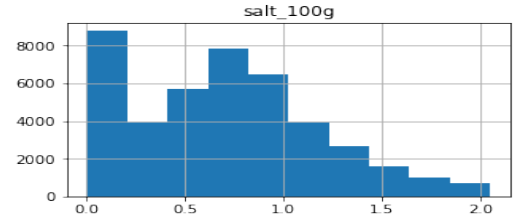


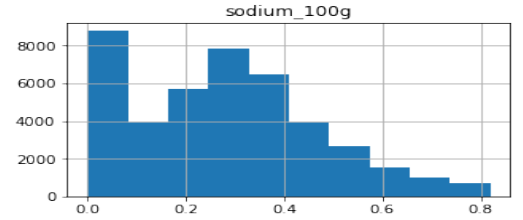


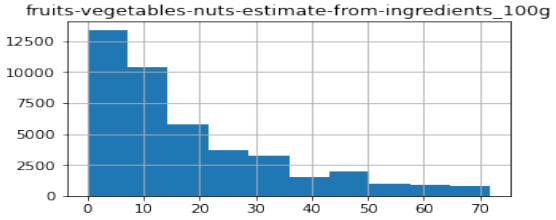




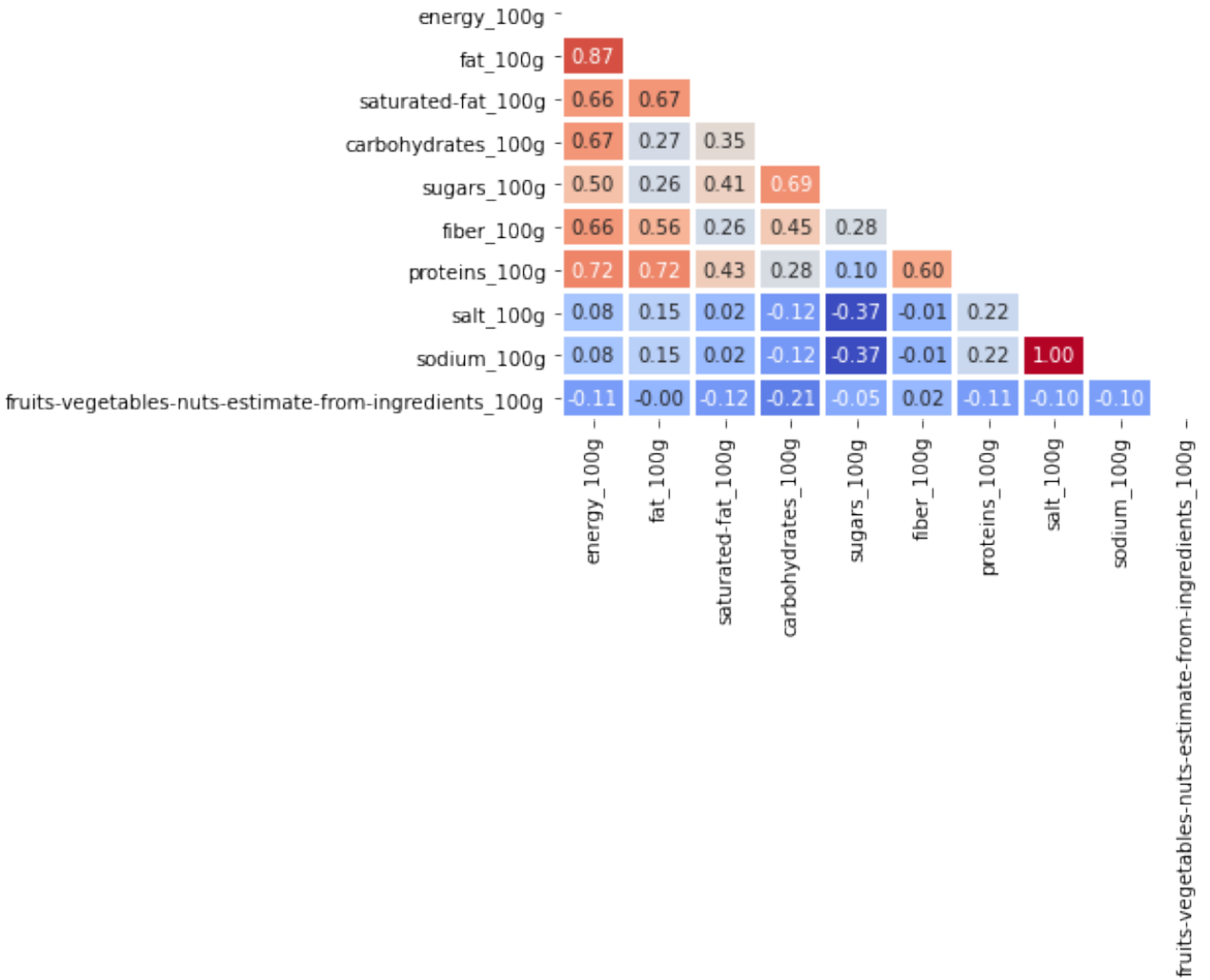




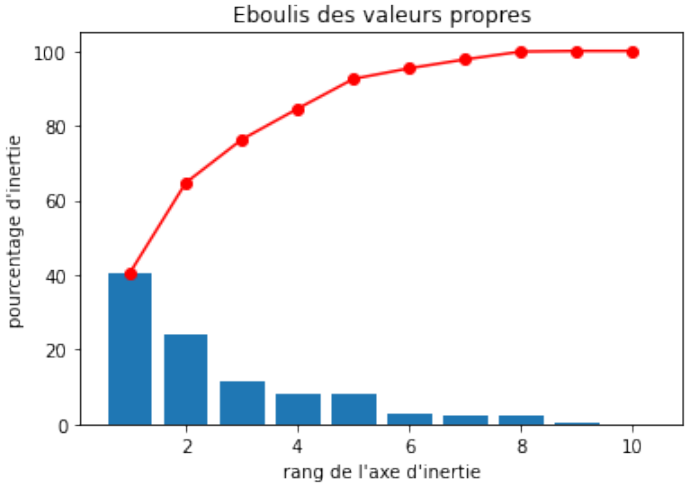




Visualisation des corrélations



Analyse en Composantes Principales

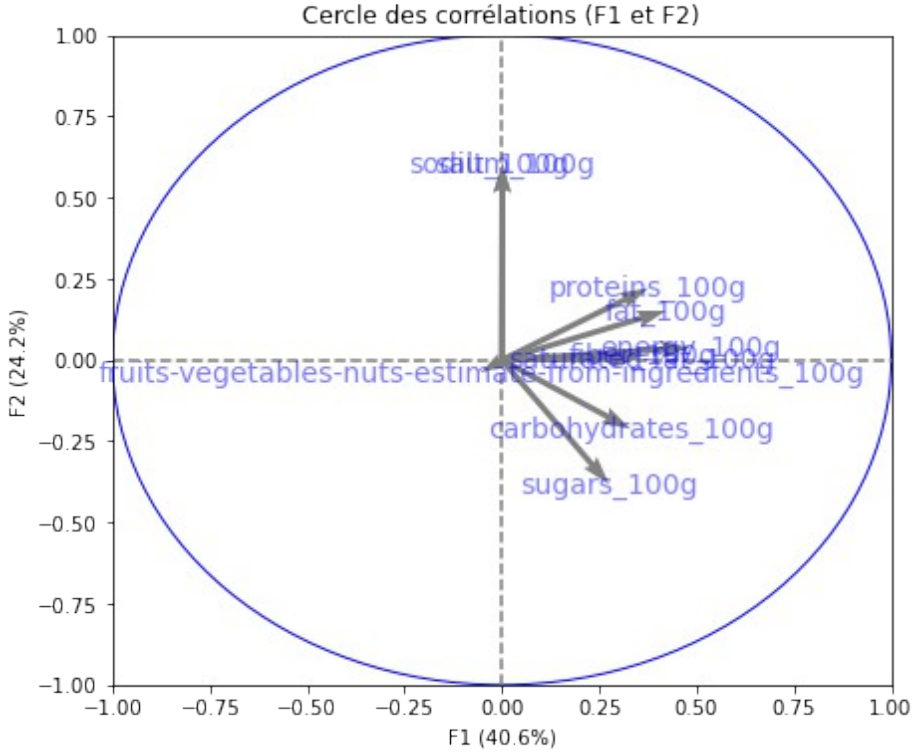


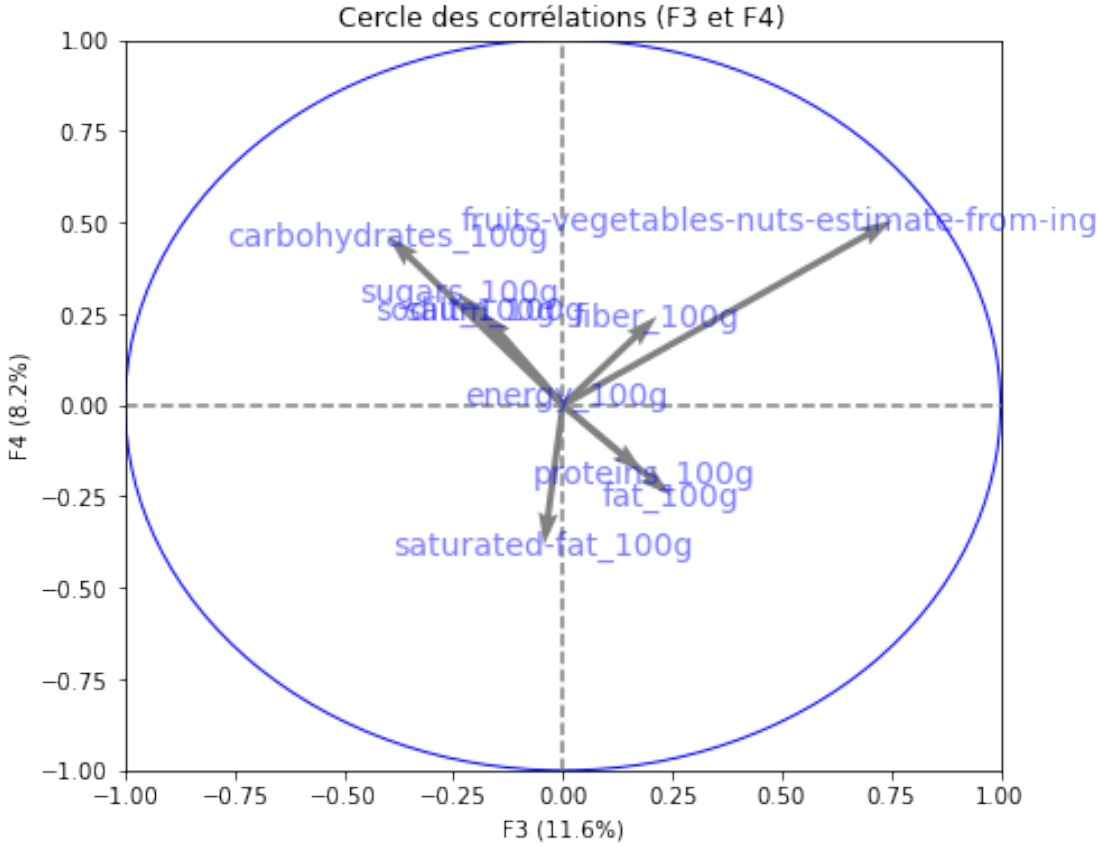
Grâce aux éboulis on remarque que les 4 premières composantes nous permet de décrire 80 des données

Ce qui nous est confirmé par le calcule des variances cumulées

Composantes	Variance	Variance Cumulées
composante1	40.547615	40.547615
composante2	24.136232	64.683847
composante3	11.575085	76.258932
composante4	8.262596	84.521528
composante5	7.996309	92.517837
composante6	2.840320	95.358157
composante7	2.383891	97.742048
composante8	2.093524	99.835572
composante9	0.160623	99.996195
composante10	0.003805	100.000000

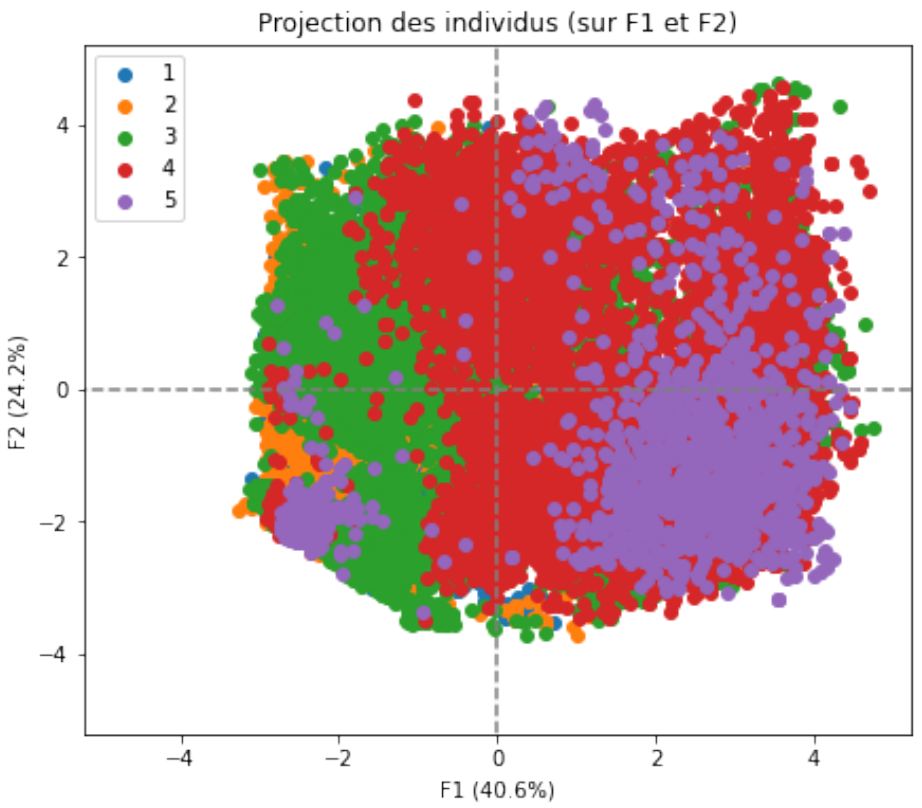
Analyse en Composantes Principales

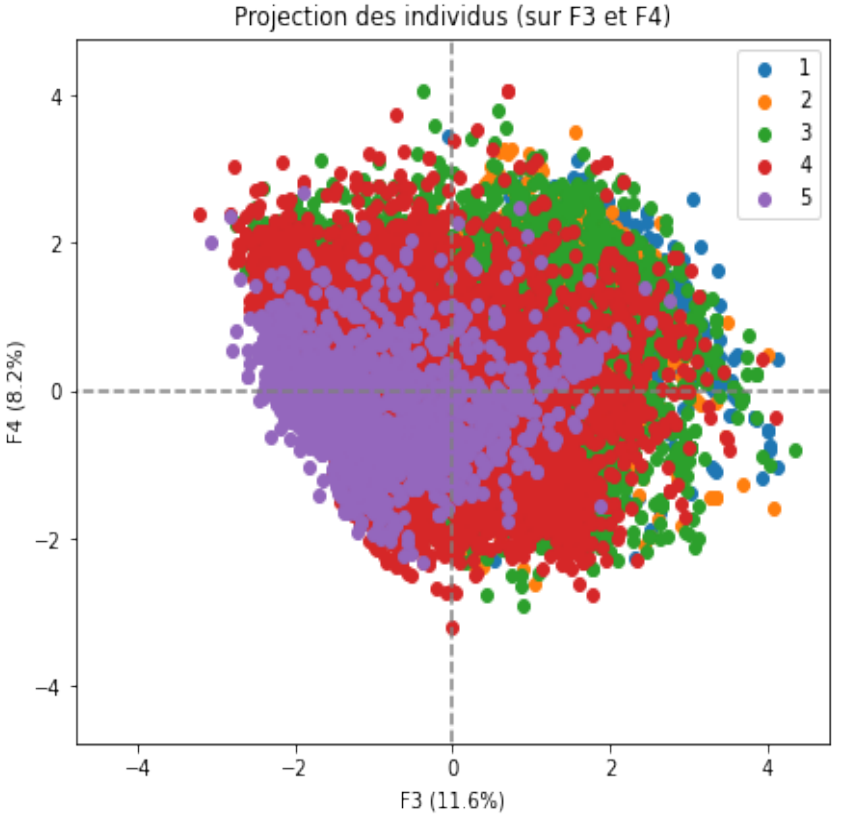




Analyse en Composantes Principales

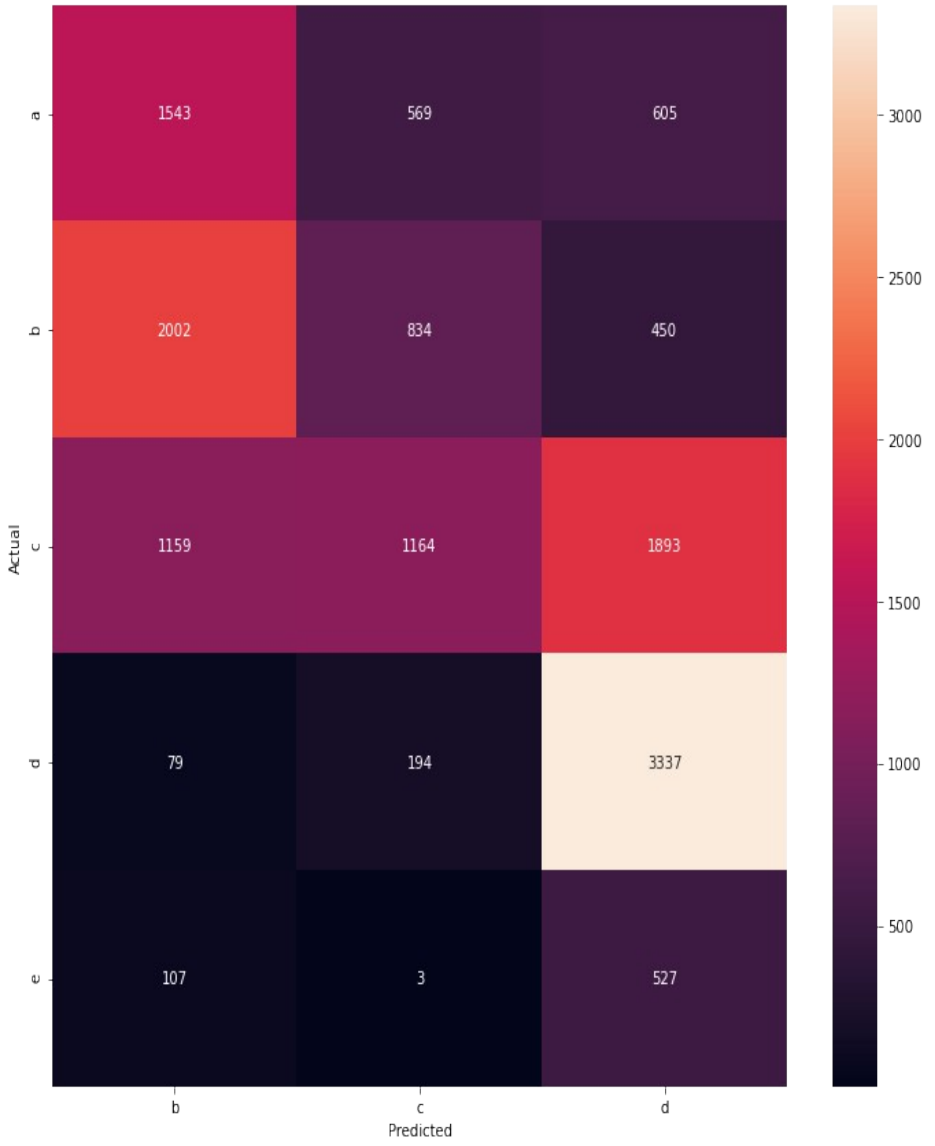
nous pouvons voir la répartition des données projetés sur les axes





Modèle prédictif : random forest

Nous observons grâce à la matrice de confusion que A et E ne sont pas prédit et les autres sont mal prédit.



Modèle prédictif : KNN

On remarque que la prédiction est de meilleur qualité.

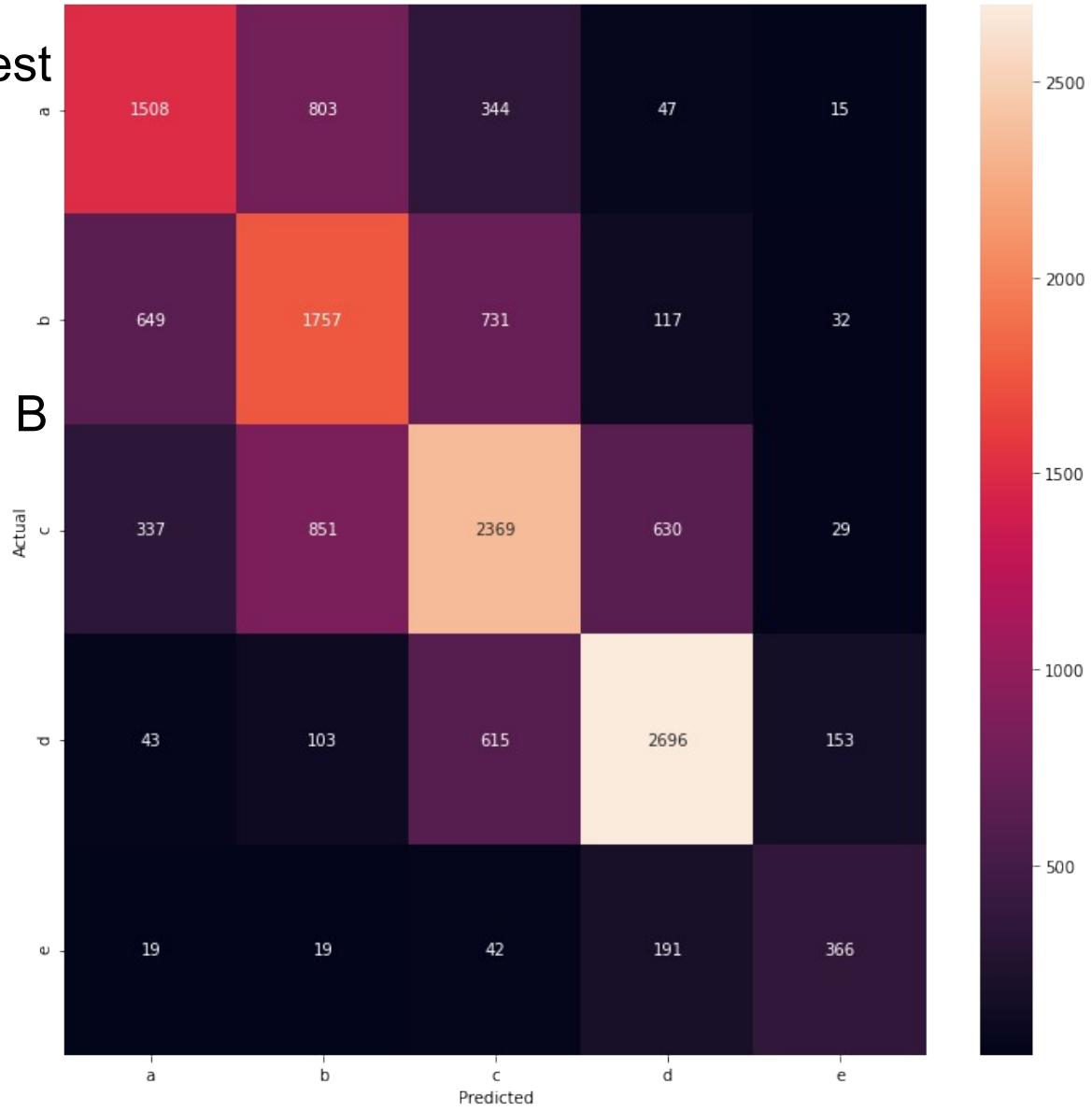
Par exemple pour le score D :

Score D à 2696 prédictions exactes sur 3610. 153 à été prédites en E, 615 en C, 103 en B et 43 en A .

Score de notre modèle :

Test set Accuracy:
0.6011336927969031

F1 Score: 0.6



Test du modèle

- Score C : Prédiction C
- Score C : Prédiction B
- Score D : Prédiction C

Régression

On remarque que les données sont bien réparties