Concevez une application au service de la santé publique

Santé publique France

PLE Coline

HAPPY
Nutri-Score

PROJET 3

03/06/2023

Plan de la présentation

- Rappel de l’appel à projet
- Proposition d’application
- Nettoyage du jeu de données
- Exploration du jeu de données
- Pertinence de l’application
- Discussion

RAPPEL DE L'APPEL A PROJETS

Santé publique France

Trouver une idée innovante d'application en lien avec l'alimentation

Boîte à outils

Jeu de données d'Open Food Facts

320 772 produits
162 variables

IDEE D'APPLICATION: L'HAPPY Nutri-Score

Santé publique France

Elément /100g	Points
Energie (KJ)	0-10
Sucres (g)	0-10
Acides gras saturés (g)	0-10
Sodium (g)	0-10

Elément /100g	Points
Fruits, légumes, légumineuses, fruits à coque, huiles de colza, de noix et d'olive (%)	0-5*
Fibres (g)	0-5
*Protéines (g)***	*0-5*

Composante N
0-40 points

—

Composante P
0-15 points

SCORE FINAL
N-P

-15
Meilleure qualité nutritionnelle

40
Moins bonne qualité nutritionnelle

Présence d'imperfections

HAPPY
Nutri-Score

Amélioration du calcul

Elément / 100g	Points
Additifs	0 - 10
Huile de palme	0 - 1

Produits suggérés

Produit désiré

Rubriques
Bienfaits de la micro-nutrition
Dictionnaire sur les additifs

NETTOYAGE DU JEU DE DONNEES

320772 produits
162 variables

46134 produits
16 variables

Stratégie de nettoyage

01 Variables sans donnée (16)
Variables totalement inutiles (33)

02 Analyse des variables qualitatives

03 Analyse des variables quantitatives

04 Traitement des valeurs aberrantes et anomalies 'sournoises'

05 Réduction du nombre de variables
Traitement des valeurs manquantes

Santé publique France

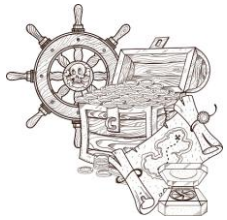
02 ANALYSE DES VARIABLES QUALITATIVES REDONDANTES

ELIMINATION DIRECTE

- **Les labels (3)**→ Contenu !!!
- **Les allergènes (2)** → Données <10% et Refonte entière
- **Les traces (3)** → Données <10%
- **Les catégories (5)** → Variables pnns beaucoup plus fiables
- **Les additifs (2/3)**

ELIMINATION APRES MODIFICATION

- **Création d'une variable 'Produits'** → Imputation du nom générique si non renseigné dans 'product name'
- **Elimination des 2 variables d'origine**
- **Elimination des produits sans nom (17643)**



CONSERVATION

- **PNNS** → Compromis pour l'exploration des données
- **'additives_fr'**

Santé publique France

02 ANALYSE DES PAYS

694 pays référencés!!!

- Création d'un fichier csv (code + 'countries_fr')
- Attribution correcte des noms de pays (147 pays)
- Création d'une variable 'Pays'
- Elimination des 3 variables initiales

Pays de vente des produits

Elimination des produits sans pays de vente (15303)

Lieu de vente	Produits	Produits catégorisés
International	287 826	64 237
France et DOM-TOM	84 062	47 224

Santé publique France

03 ANALYSE DES VARIABLES QUANTITATIVES HORS NUTRI-SCORE

- Elimination des 23 produits sans code à barres (doublons ou infos inutiles)
- Elimination de la variable ‘code’ devenue inutile
- Bonne attribution de type pour les variables

ELIMINATION DIRECTE

- Variables sans données (4)
- Variables non essentielles quasi-vides (< 1%) (12)
- Variables regroupées dans une variable unique (25)
- Variable ‘cholesterol_100g’ → Peu d’impact sur le taux sanguin

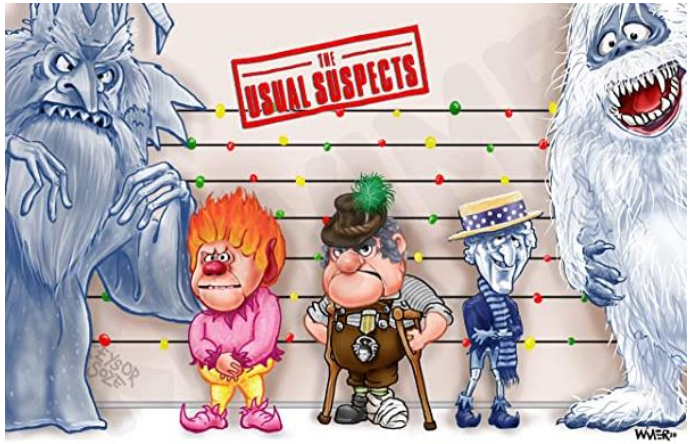
ELIMINATION APRES VERIFICATION

- **Vitamine b9** → Redondance avec **Folates** (moins de données mais plus pertinentes)
- **'ingredients_that_may_be_from_palm_oil_n’** (~doublon)

287 809 produits et 50 variables

Santé publique France

04 TRAITEMENT DES VALEURS ABERRANTES ET SOURNOISES



	Aberrance	Produits
Energie pour 100g	> 3765,6 kJ	350
	0 kJ	1402
Quantité pour 100g	> 100 g	207
	< 0 g	16

Additifs – Huile de palme – Nutri-Score

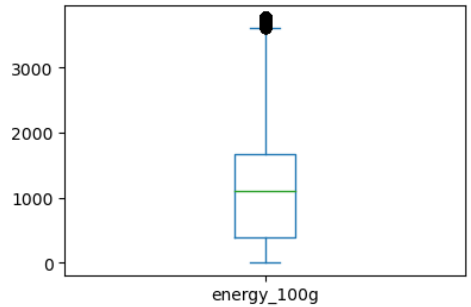
	Camouflage	Produits
Nutri-Score	Mauvais Grade	2 aliments
		226 boissons
Quantité pour 100 g	Sucres > Glucides	644
	Lipides Totaux > 'Fat'	1840
Lipides+Glucides+ Protéines + Fibres + Sel	>100 g	20544
'Faux-outliers'		62

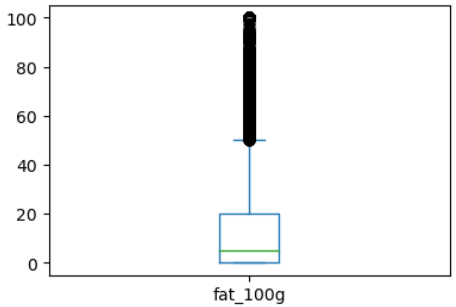
262 510 produits

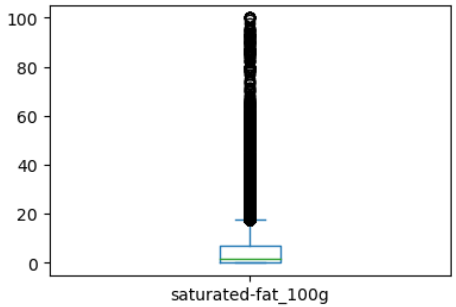
NETTOYAGE DU JEU DE DONNEES

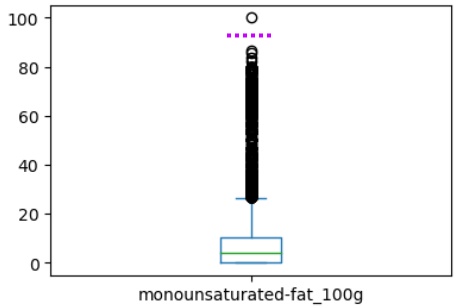
Santé publique France

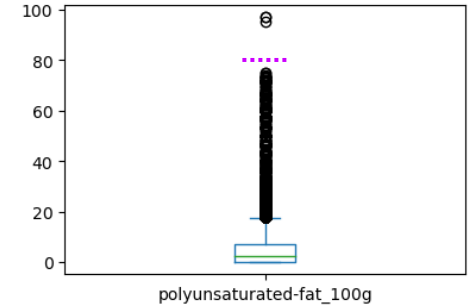
04 Les ‘faux-outliers’ des variables quantitatives

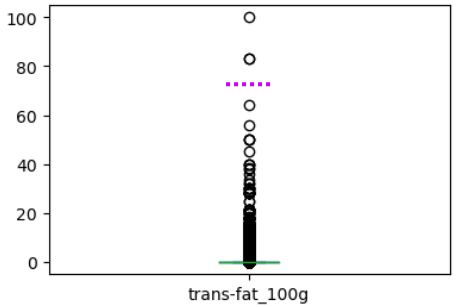


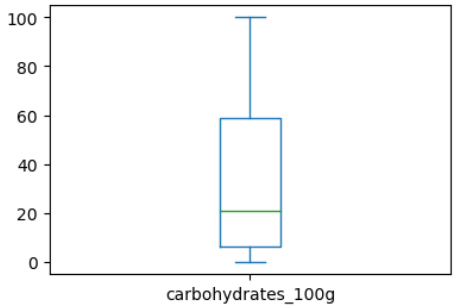


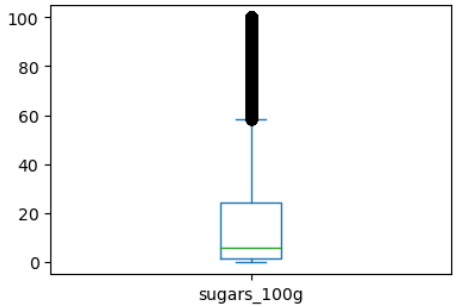


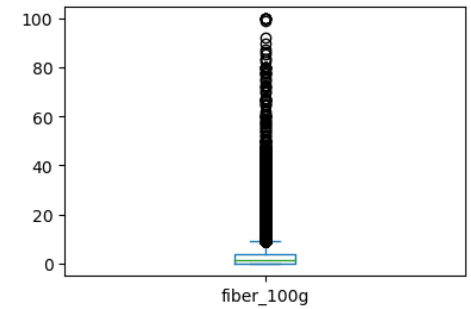


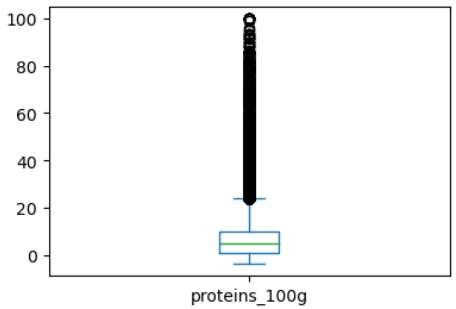


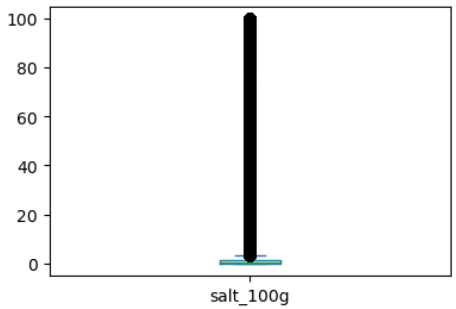












62 produits

7 classiques
36 vitamines
19 minéraux/alcool

Seuil

NETTOYAGE DU JEU DE DONNEES

ETAPES AVANT TRAITEMENT DES VALEURS MANQUANTES

Choix des produits conservés

Lieu de vente	Produits catégorisés	Perte due au nettoyage
International	63 195	1042 (1,65%)
France et DOM-TOM	46 134	605 (1,28%

Réduction du nombre de variables

- Données < 55%
- Variable ‘Pays’
- Variable ‘Sodium’ (Sodium = Sel / 2,5)

Santé publique France

05 TRAITEMENT DES VALEURS MANQUANTES

Variables quantitatives

Imputation par la médiane en fonction de la catégorie de produits

Le Nutri-Score : KNN Imputer vs Imputation par la médiane

Produit	KNN Imputer	Médiane
Les lentilles	8 (Grade C)	-4 (Grade A)

Médiane

Variables qualitatitives

- **Additifs** : Remplacement des NaN par ‘Non renseigné’ (Seul choix possible)
- **Nutri-Grade** : Imputation en fonction du Nutri-Score
- **pnns_groups_1** : Pas d’imputation car non nécessaire

ETAPES PREALABLES AVANT EXPLORATION

Santé publique France

01 VERIFICATION DE LA QUALITE DU JEU DE DONNEES

	Anomalies	Produits
Lipides +Glucides + Protéines + Fibres + Sel	> 100 g	284
Quantité pour 100 g	AG saturés > Lipides	511
	Sucres > Glucides	760

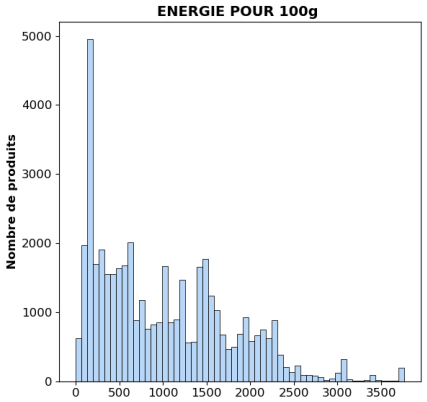
44579 produits

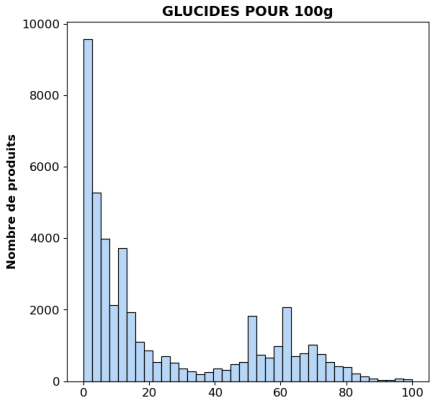
02 CREATION DE CATEGORIES DE PRODUITS

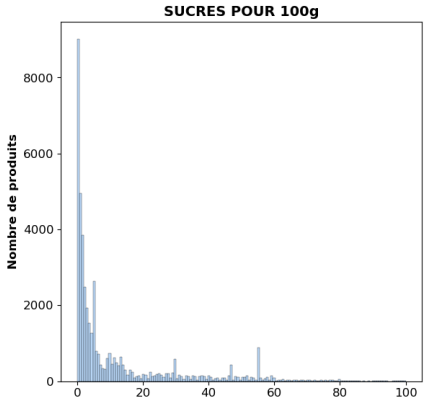
Aliments	Boissons
Fruits et Légumes	Boissons non sucrées
Fruits secs	Boissons sucrées
Céréales et Légumineuses	Jus et Nectars de fruits
Produits marins, viandes et œufs	Boissons alcoolisées
Produits sucrés	
Produits salés	
Produits composés	
Sauces et condiments	
Produits laitiers	

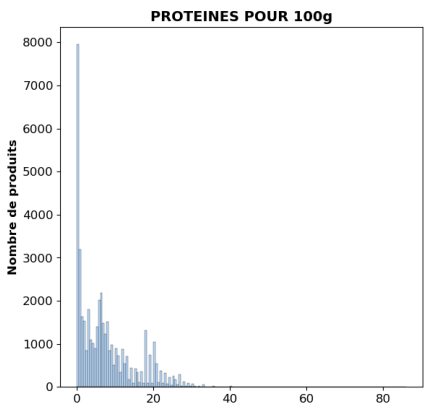
Elimination d'un produit (mauvais grade) : 44578

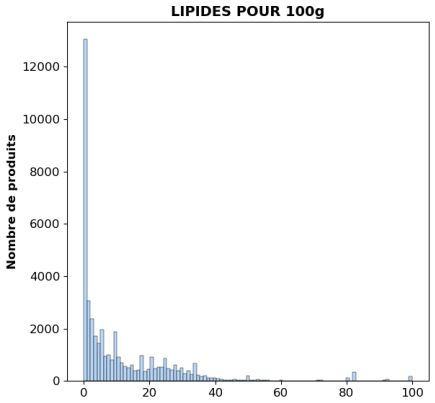
ANALYSES UNIVARIEES: Variables quantitatives

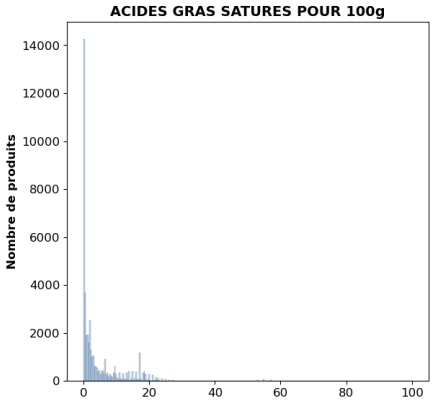


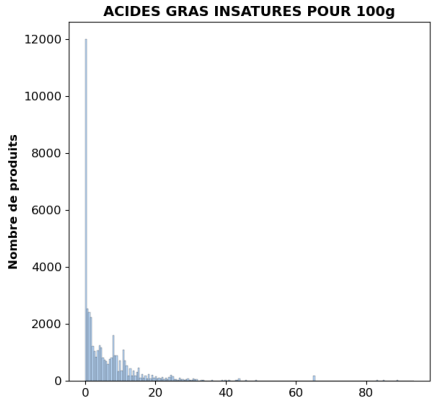


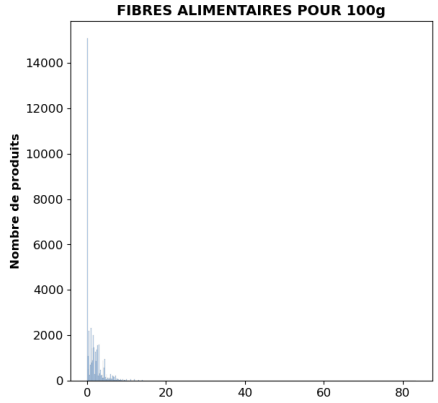


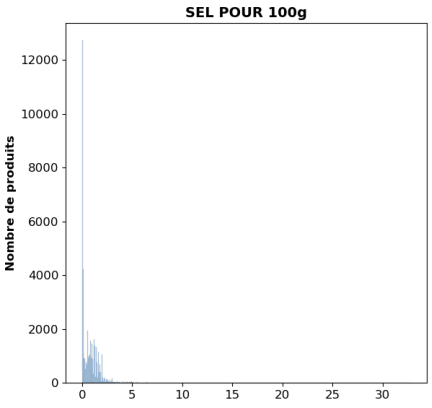


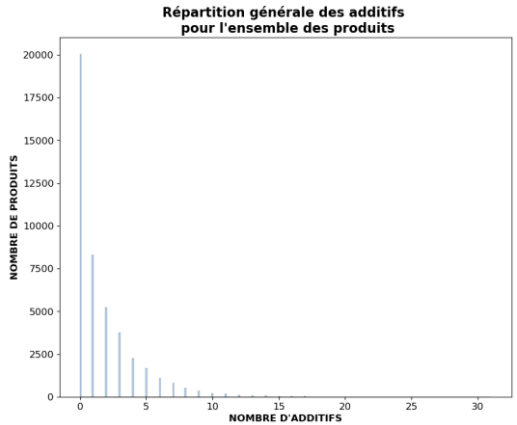








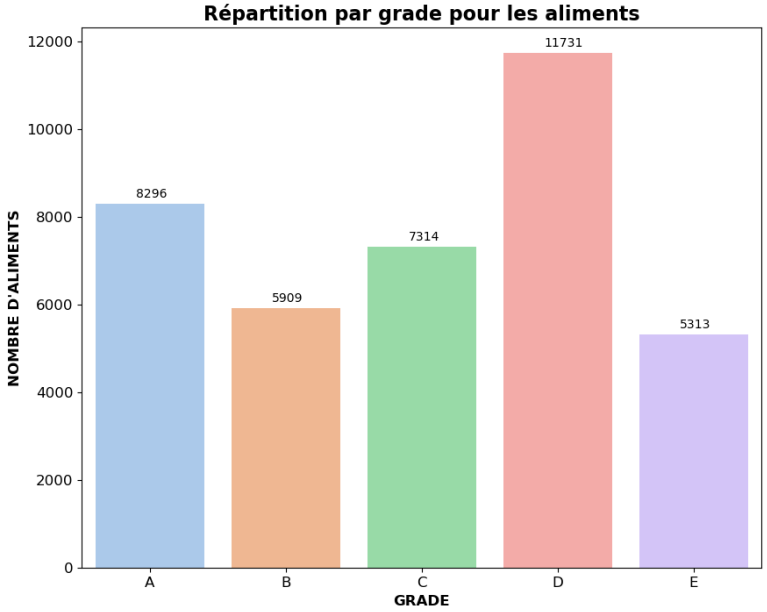


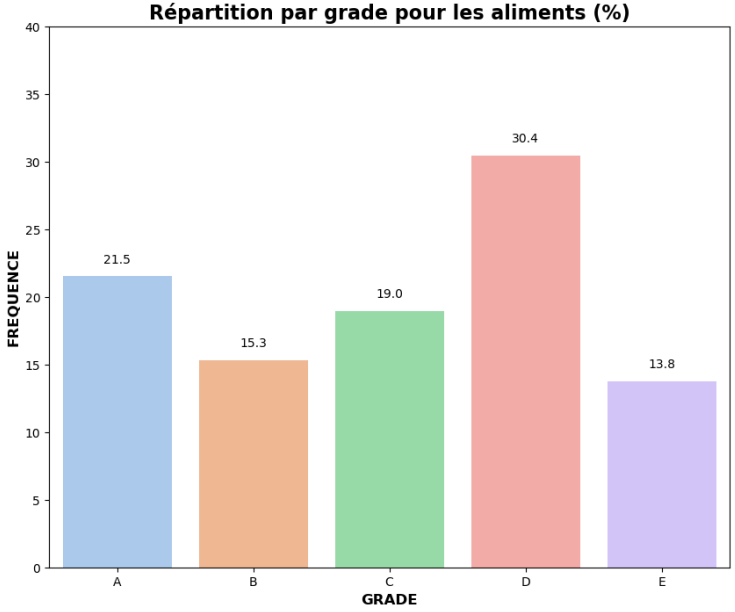


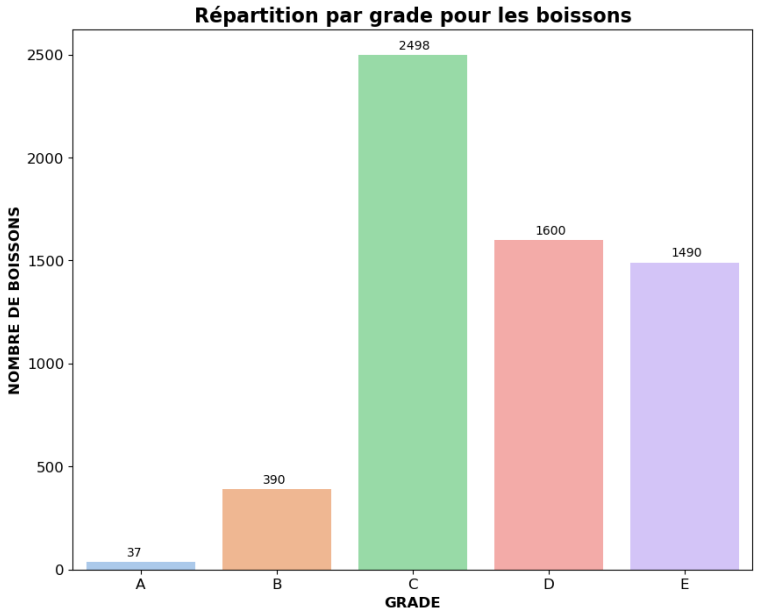
Distribution unimodale/asymétrique pour toutes les variables

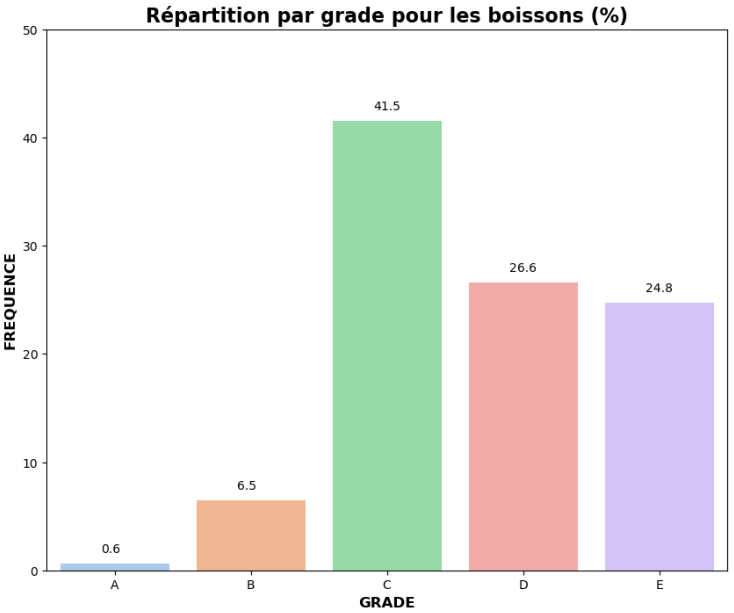
ANALYSES UNIVARIEES: Distribution par grade

Santé publique France



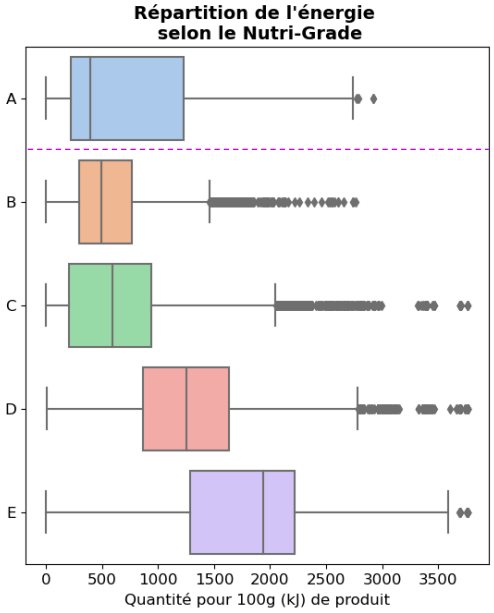


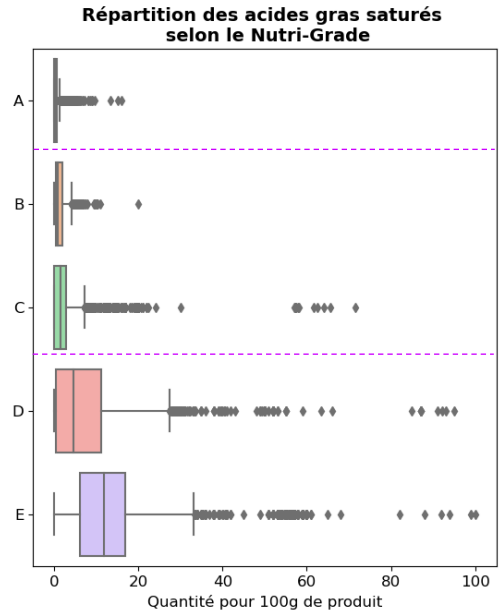


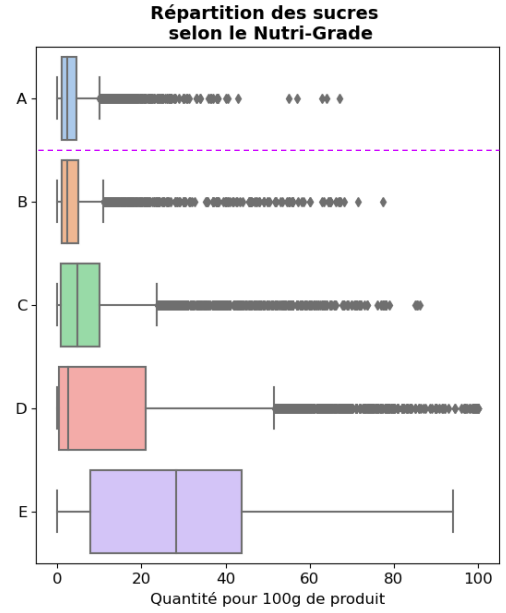


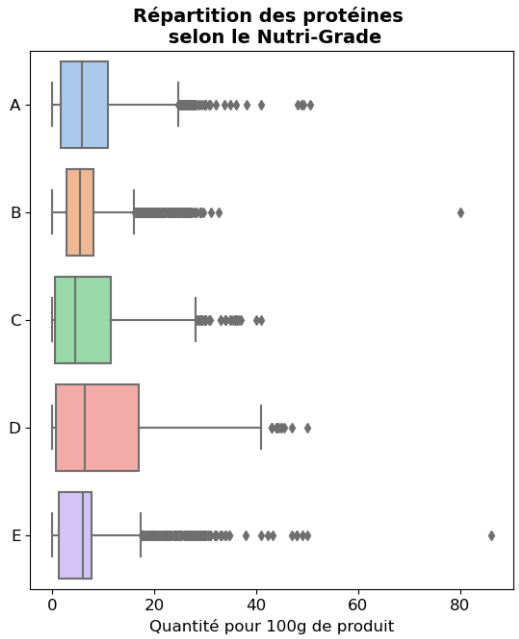
ANALYSES BIVARIEES: Répartition de variables quantitatives selon le Nutri-Grade

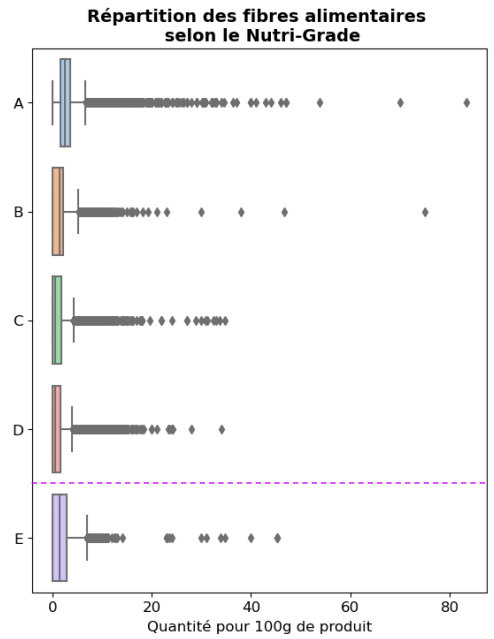
Santé publique France

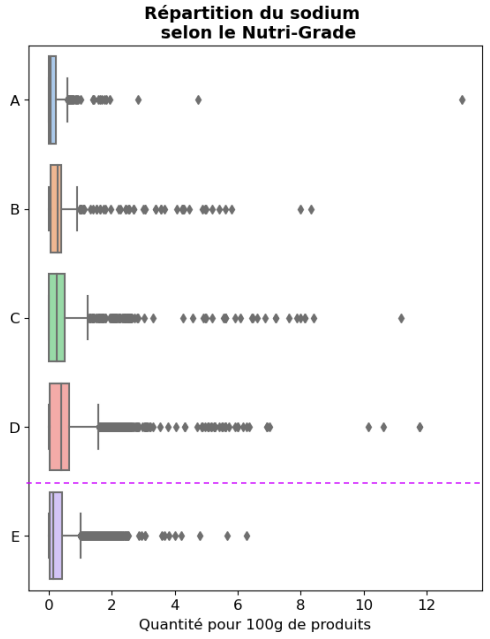












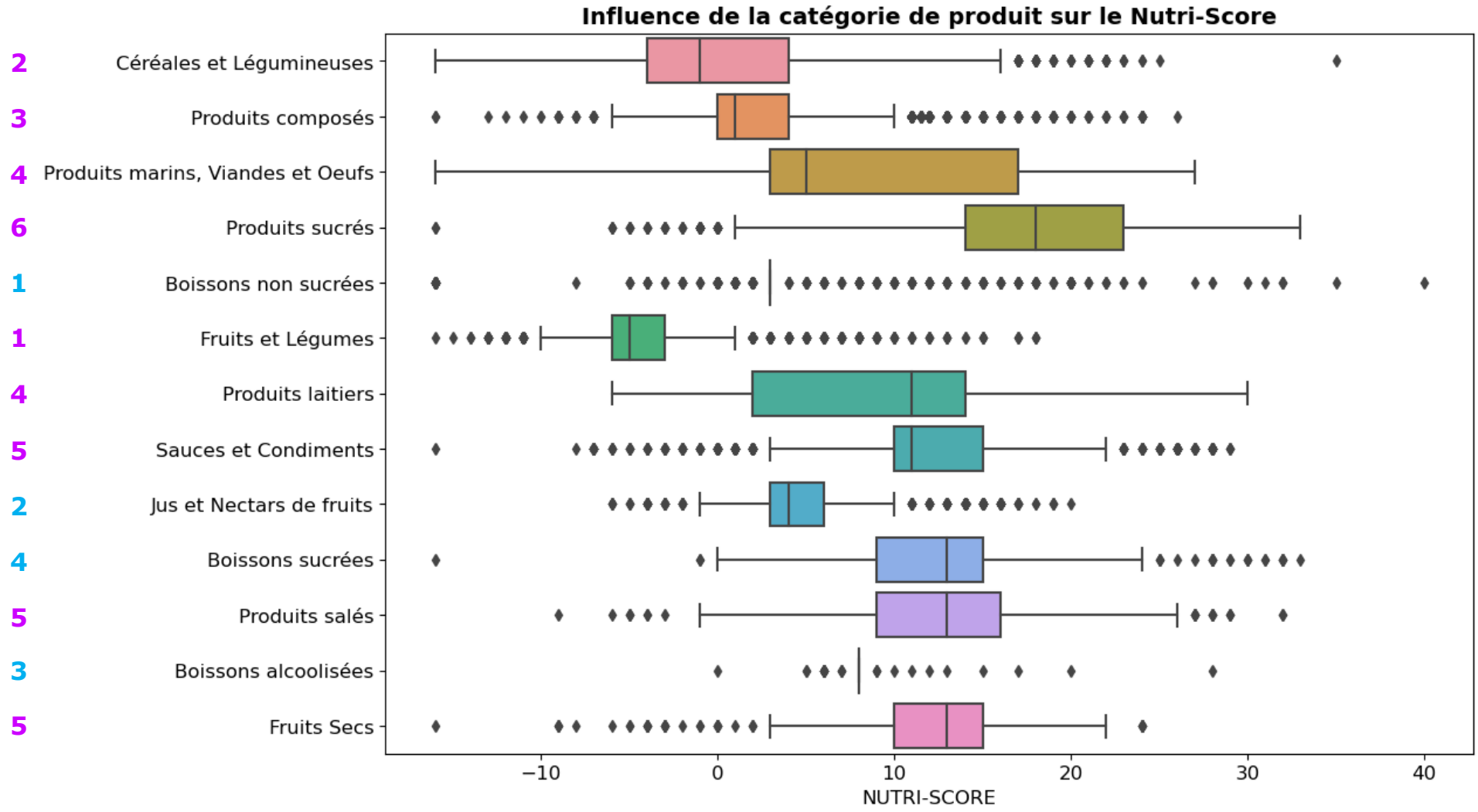
Seuil de significativité de corrélation linéaire

ANALYSES BIVARIEES: Répartition du Nutri-Score selon la catégorie de produits

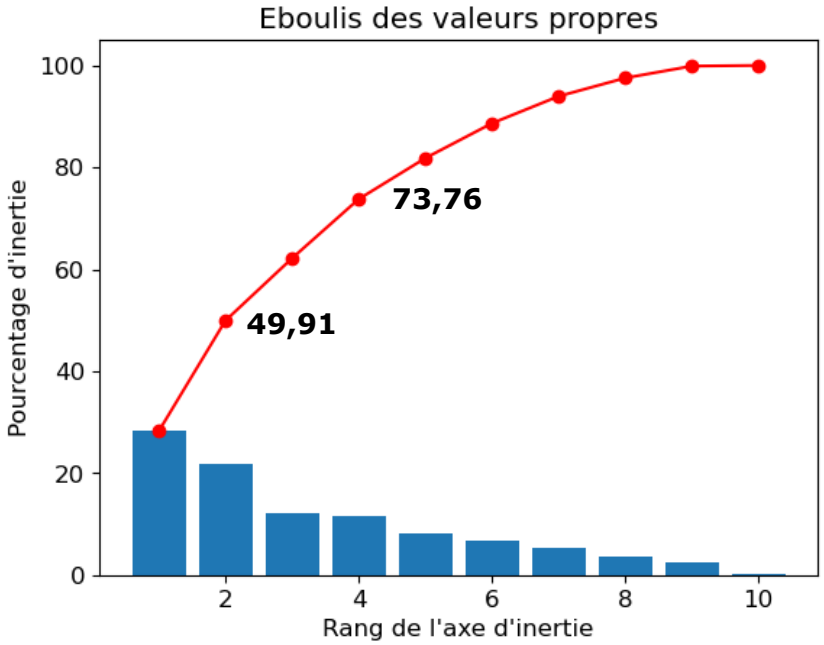
Santé publique France

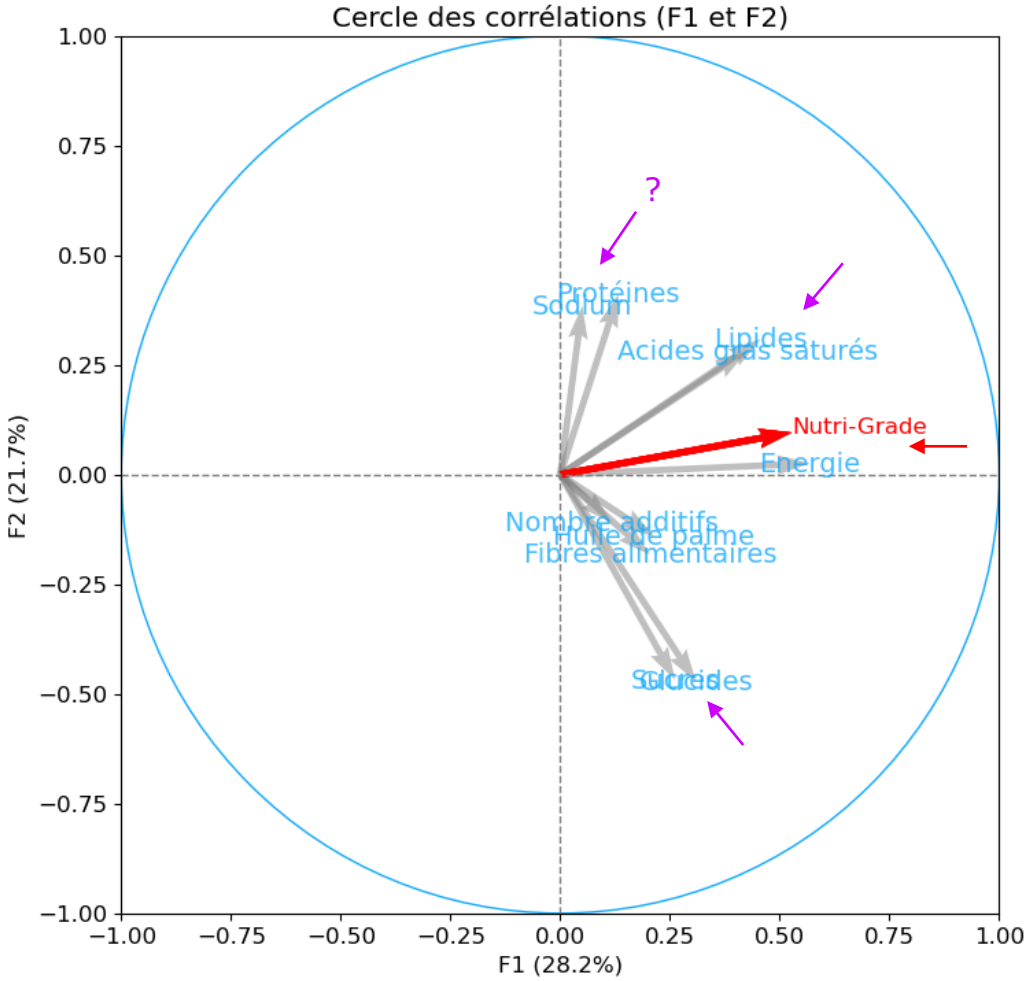
Aliments

Boissons



ANALYSES MULTIVARIEES: L'analyse en composantes principales (ACP)



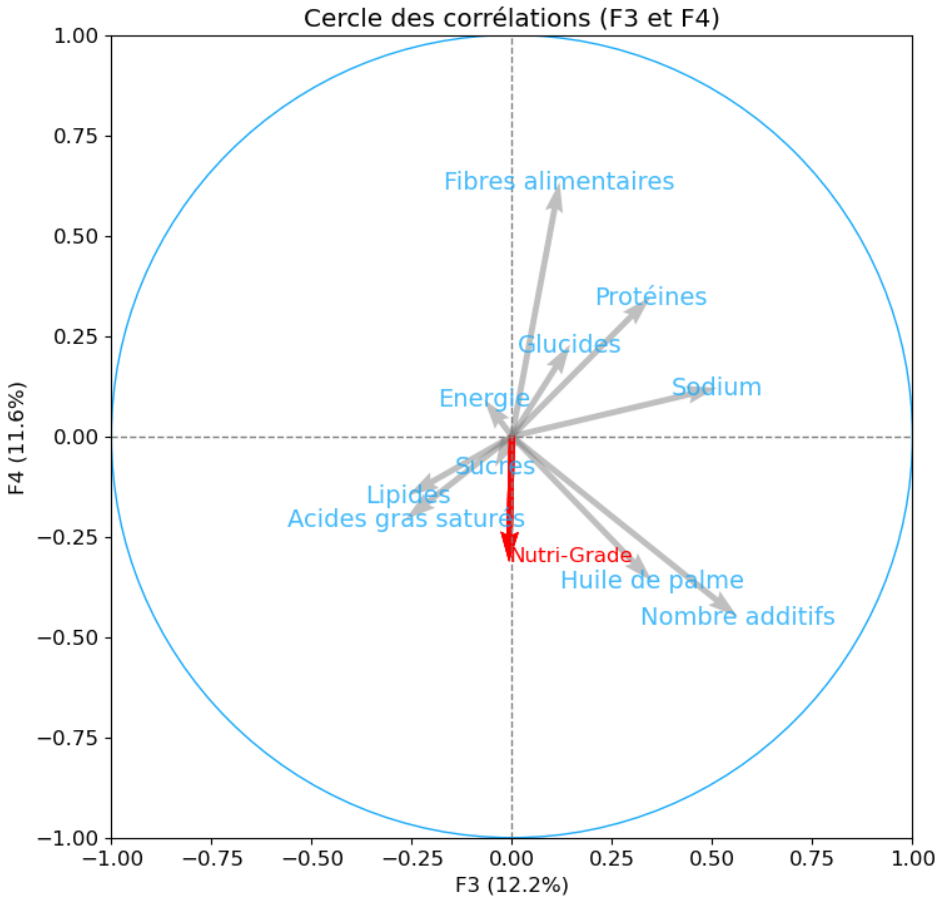


Premier plan factoriel (F1 et F2)

- Non associées à une variable particulière
- Couverture d'~50% de l'inertie → Trop de perte d'informations

→ **2ème plan factoriel (F3 et F4)**

ANALYSES MULTIVARIEES: L'analyse en composantes principales (ACP)



2éme plan factoriel (F3 et F4)

- F3 plutôt associée aux Additifs
- F4 associée aux Fibres alimentaires

- Additifs absents du Nutri-Score
- Corrélation inverse Sucres/Glucides!!!

Pertinence de l'ACP dans ce projet?

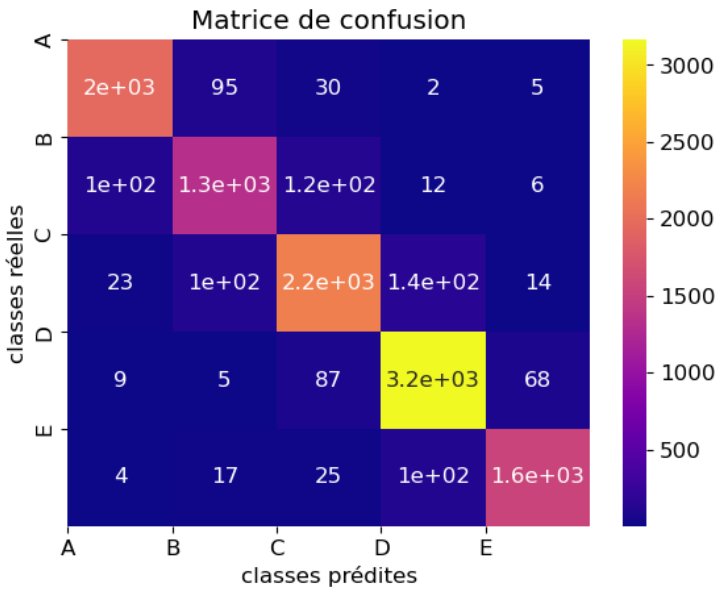
ACP vs Machine Learning pour l'importance des variables

ANALYSES MULTIVARIEES: Importance des variables pour la prédiction du Nutri-Grade

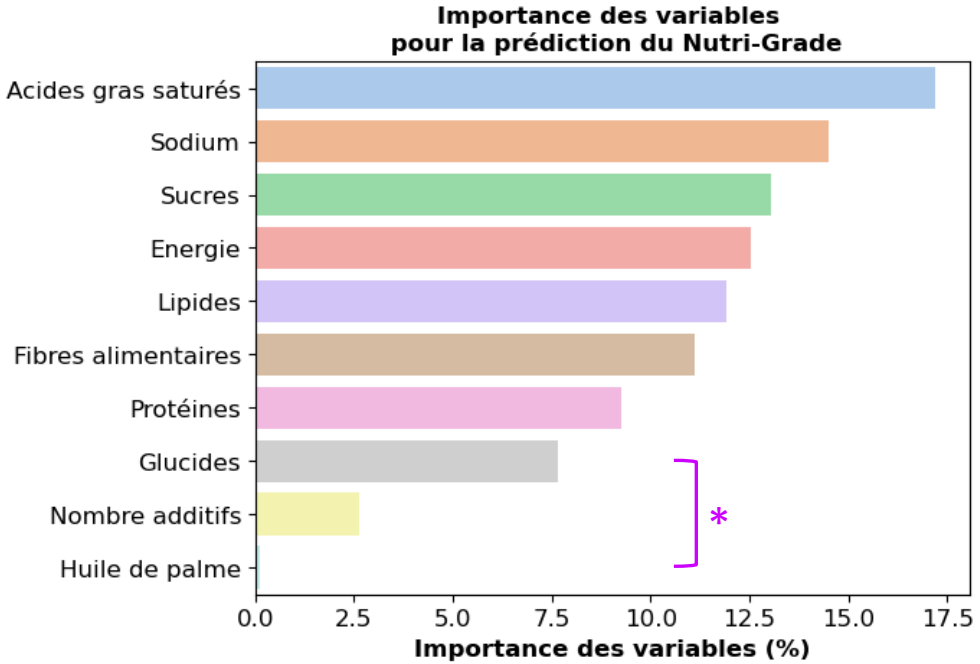
Santé publique France

Modèle: Random Forest Classifier

- Performance: 91,30%
- Pourcentage d'erreurs: 0,15%



Taux d'erreurs les plus importants entre classes voisines



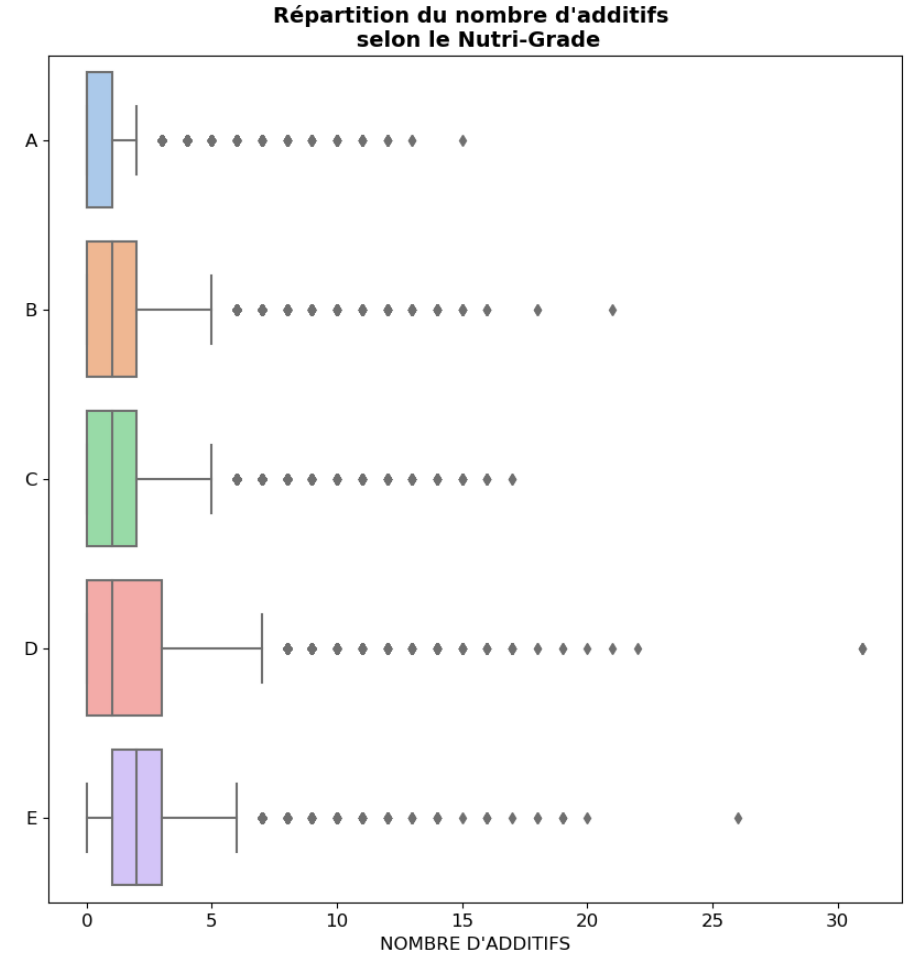
Adéquation avec le calcul du Nutri-Grade

* Non pris en compte

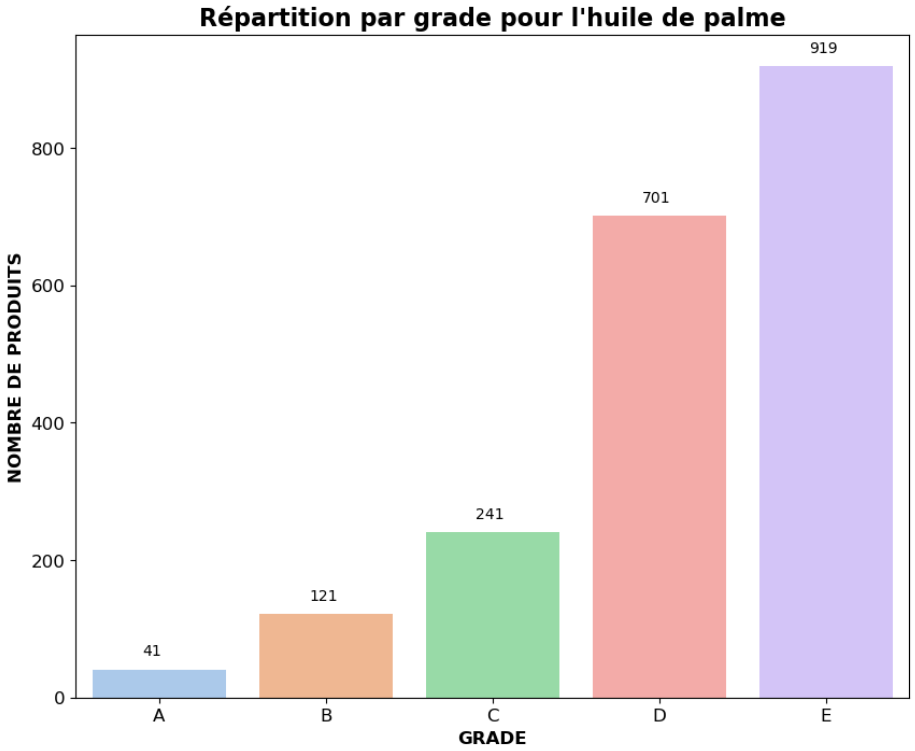
PERTINENCE DE L'HAPPY Nutri-Score

Santé publique France

Corrélation linéaire significative
Nombre d'additifs / Nutri-Grade

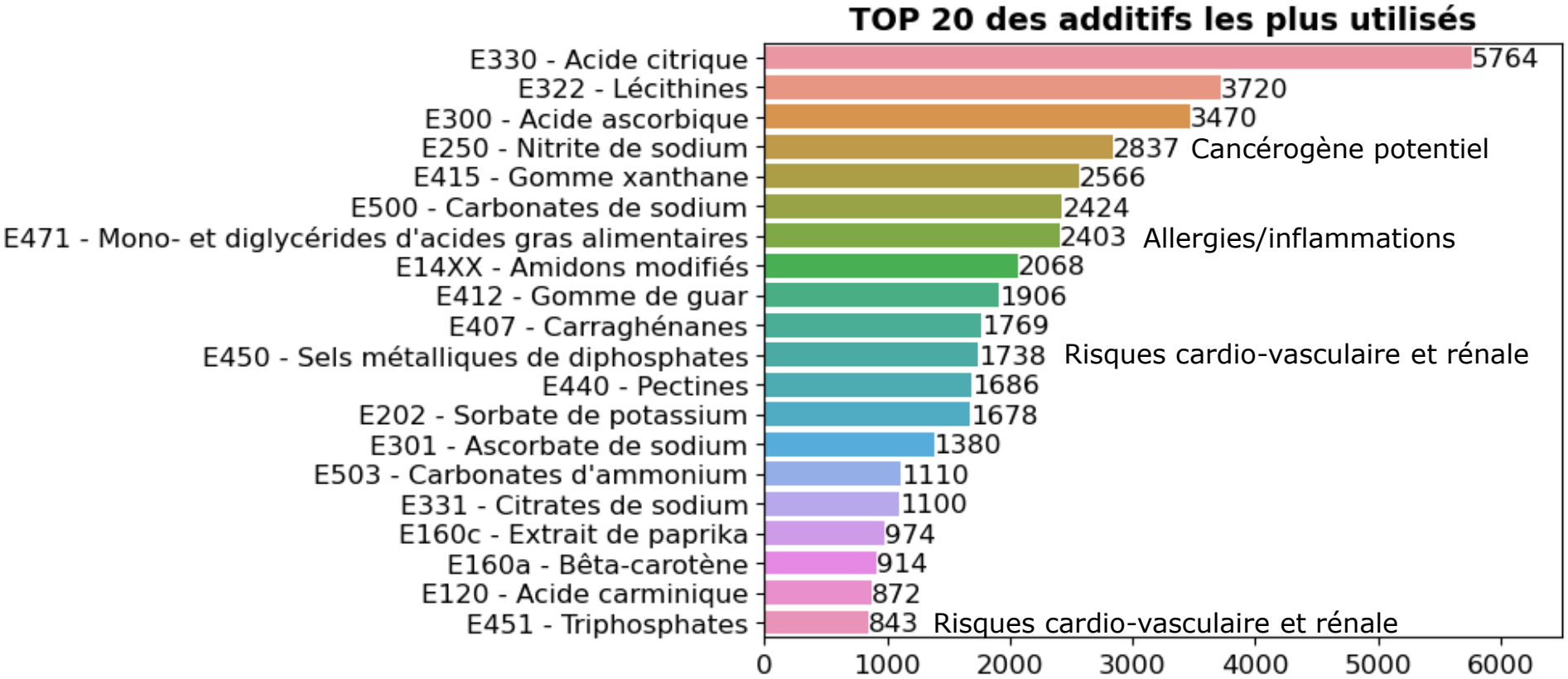


Corrélation linéaire NS
Présence Huile de palme/ Nutri-Grade



p-value = 0,166 (Two-side)

PERTINENCE DE L'HAPPY Nutri-Score



- ✓ Pertinente
- ✓ Réalisable
- ✓ Création d'une database sql avec implémentation possible de nouveaux produits (Clé primaire = Code à barres)

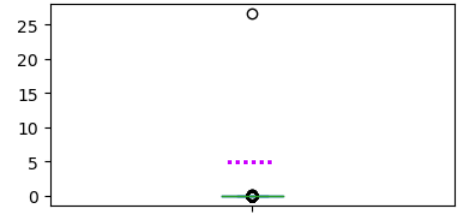
Santé publique France

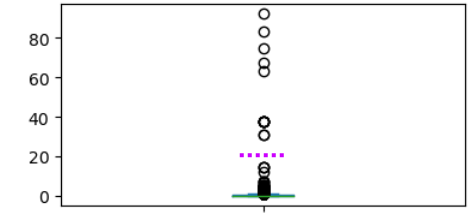
- Débat / Réflexion -

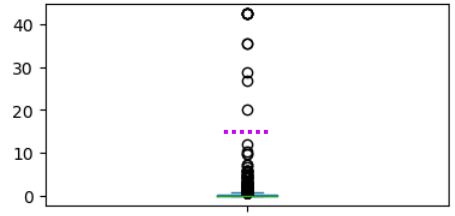
Informations supplémentaires

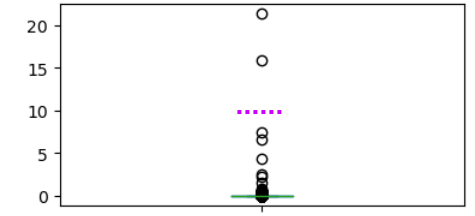
Santé publique France

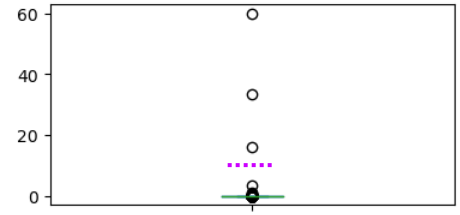
Distribution des vitamines

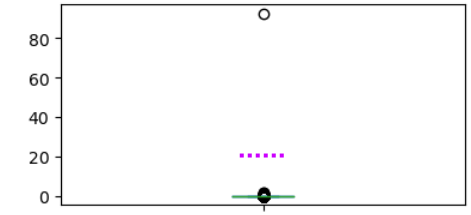


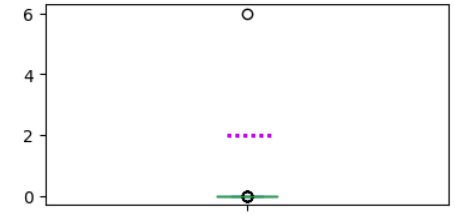


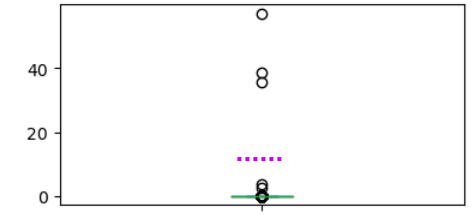


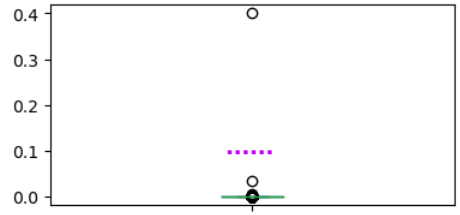


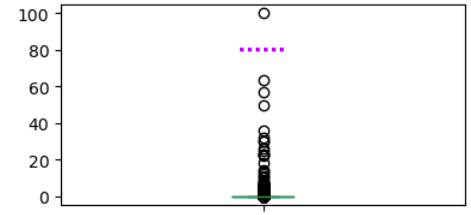


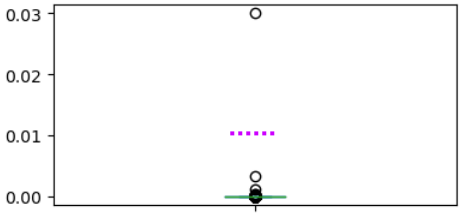


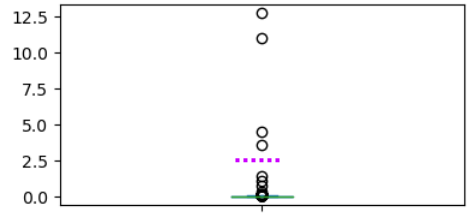


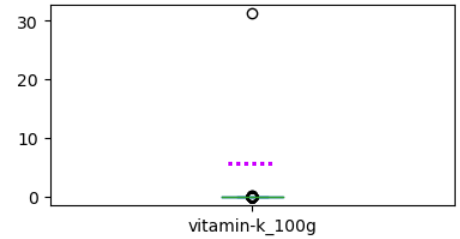










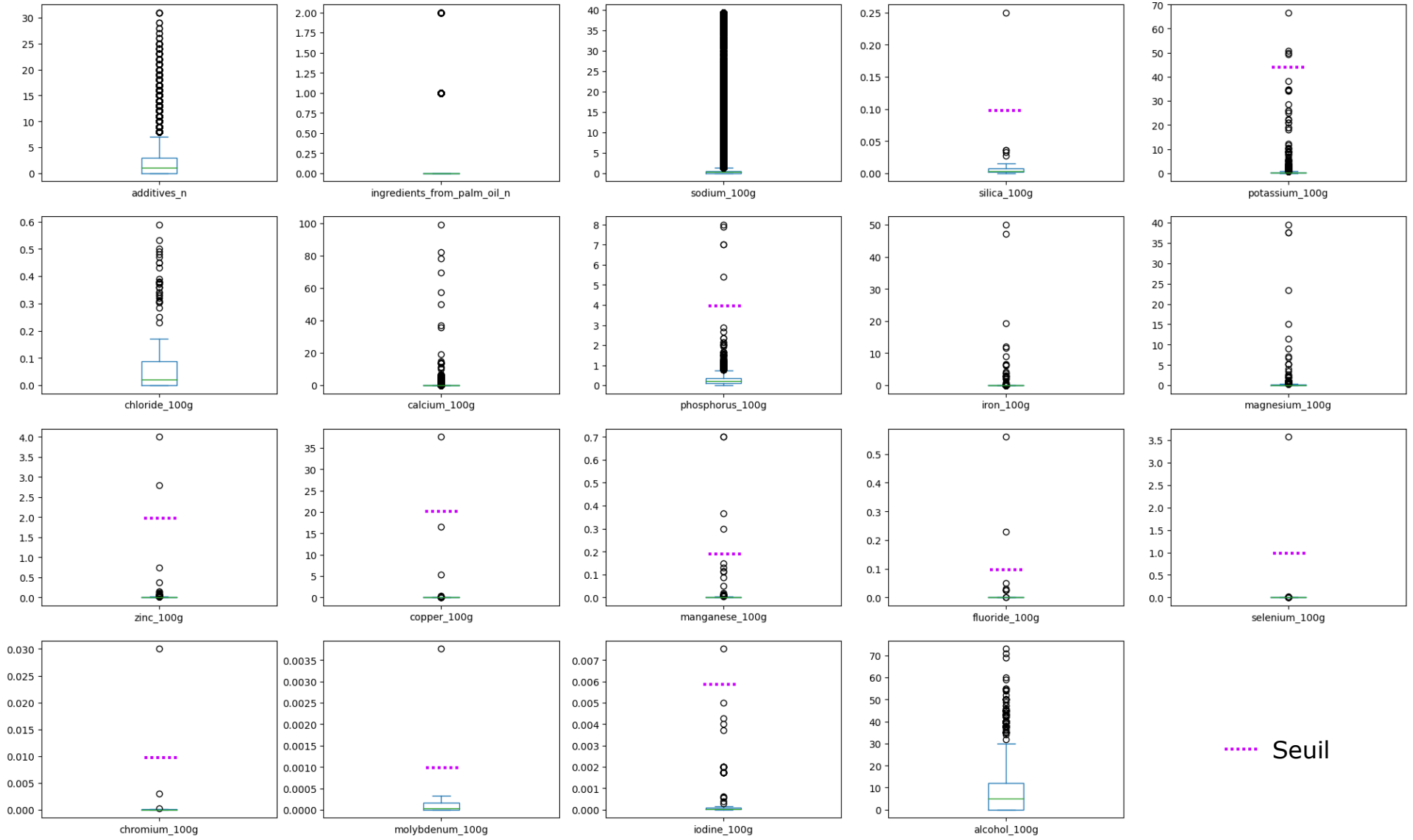


Seuil

Informations supplémentaires

Santé publique France

Distribution des minéraux/alcool/additifs/Huile de palme



Informations supplémentaires