Application de santé publique

Projet 3 Parcours Data Scientist
Python - Support Jupyter Notebook

Nalron Octobre 2020
OpenClassrooms - Centrale Supélec

Contexte de la mission

Élaborer une idée d'application en lien avec l'alimentation…

Mission: Identifier des arguments justifiant la faisabilité (ou non) de l'application
Donneur d'ordre / Appel à projet: L'agence "Santé publique France"
Partenaire: Open Food Facts
Données: *https://world.openfoodfacts.org/*
Variables: *https://world.openfoodfacts.org/data/data-fields.txt*

Santé publique France est l’agence nationale de santé publique.
Créée en mai 2016 par ordonnance et décret, c’est un établissement public administratif sous tutelle du ministère chargé de la Santé.

Mission : améliorer et protéger la santé des populations. Cette mission s'articule autour de trois axes majeurs : anticiper, comprendre et agir.

Open Food Facts est une base de données sur les produits alimentaires faite par tout le monde, pour tout le monde.

Elle vous permet de faire des choix plus informés, et comme les données sont ouvertes (open data), tout le monde peut les utiliser pour tout usage.

Idée d’application simplifiée

Double affichage souhaité pour le maximum de produit alimentaire.

- **Affichage d’un Nutri-score prédictif**
- **Affichage d’une classification Nova prédictive**

Comment et à partir de quoi l’application peut-elle être viable?
Identifier des arguments justifiant la faisabilité (ou non) de l’application à partir des données Open Food Facts.

Quels intérêts pour cette application?

Des milliers de produits **sans étiquette de qualité nutritionnelle.**

Simple d'utilisation avec **affichage simplifié : Nutri-score / Nova**

- Moins d'information sur l'écran du consommateur
- Rapidité du résultat avant consommation et/ou achat du produit
- Trouver des réponses aux produits sans indicateur
- Mieux manger !
- etc…

Nutri-score et Nova groupe

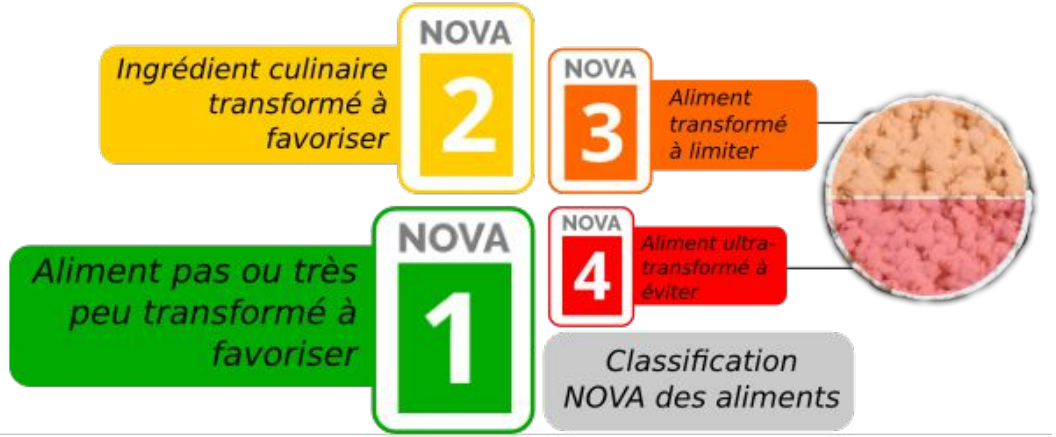
Le Nutri-Score est un logo non-obligatoire mis au point par les autorités de santé que l'on peut retrouver sur l'emballage des produits.
Il informe en un coup d'œil sur leur qualité nutritionnelle. Il vous permet de **choisir plus facilement vos produits.**

Ce score prend en compte (pour 100 g ou 100 ml de produit) :

la teneur en
« COMPOSANTES » FAVORABLES :
fibres, protéines, fruits et légumes, légumineuses, fruits à coque, huile de colza, de noix et d'olive.

la teneur en
« COMPOSANTES » DÉFAVORABLES :
énergie, acides gras saturés, sucres, sel.

Après calcul, chaque produit se voit attribuer une note associée à une couleur.



Données Open Food Fact en 4 sections

Produits alimentaires vendus dans le Monde

Informations générales: code barre, nom du produit, créateur, etc…

Tags: catégorie, pays de vente, marque, distributeur, etc…

Ingrédients: liste des ingrédients, additifs.

Données nutritionnelles: énergie, protéines, sucre, sel, graisses, etc…

Traitement et nettoyage

Chargement des données dans un Notebook, traitement des éventuels doublons, des valeurs manquantes, des valeurs outliers.

Ciblage des variables nutritionnelles avec correction éventuelle en vue des prochaines étapes de Data Mining.

Aperçu des données

Observations : 32309
Variables : 181
Type de variables : quantitative, qualitative et temporelle

Informations nutritionnelles :
Valeurs manquantes pouvant annuler totalement des variables

code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name
17	http://world-en.openfoodfacts.org/product/0000...	kiliweb	1529059080	2018-06-15T10:38:00Z	1561463718	2019-06-25T11:55:18Z	Vitória crackers
31	http://world-en.openfoodfacts.org/product/0000...	isagoofy	1539464774	2018-10-13T21:06:14Z	1539464817	2018-10-13T21:06:57Z	Cacao
0000000000003327986	http://world-en.openfoodfacts.org/product/0000...	kiliweb	1574175736	2019-11-19T15:02:16Z	1574175737	2019-11-19T15:02:17Z	Filetes de pollo empanado
100	http://world-en.openfoodfacts.org/product/0000...	del51	1444572561	2015-10-11T14:09:21Z	1444659212	2015-10-12T14:13:32Z	moutarde au moût de raisin
0000000001111111111	http://world-en.openfoodfacts.org/product/0000...	openfoodfacts-contributors	1560020173	2019-06-08T18:56:13Z	1560020173	2019-06-08T18:56:13Z	Sfiudwx

uantity	...	carbon-footprint-from-meat-or-fish_100g	nutrition-score-fr_100g	nutrition-score-uk_100g	glycemic-index_100g	water-hardness_100g	choline_100g	phylloquinone_100g	beta-glucan_100g	inositol_100g	carnitine_100g
NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
130 g	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dgesc	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
NaN	...	NaN	7.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	...	NaN	6.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Données manquantes

Restriction des observations sur la condition des 50% de données disponibles / variable.

Conservation des variables nutritionnelles fibres, protéines, énergie, acides gras saturés, sucres, sel).

Échantillon de taille : 20166 x 52

countries_tags	100.0
product_name	100.0
categories_tags	100.0
countries	100.0
energy-kcal_100g	99.0
ingredients_text	97.0
additives_n	97.0
ingredients_that_may_be_from_palm_oil_n	97.0
ingredients_from_palm_oil_n	97.0
serving_size	96.0
serving_quantity	96.0
fiber_100g	94.0
nova_group	94.0
brand_owner	92.0
cholesterol_100g	92.0
iron_100g	91.0
trans-fat_100g	91.0
calcium_100g	90.0
vitamin-c_100g	78.0
vitamin-a_100g	77.0
additives_tags	67.0
additives_en	67.0
brands	59.0
brands_tags	59.0
potassium_100g	27.0
allergens	26.0
polyunsaturated-fat_100g	17.0
monounsaturated-fat_100g	17.0
stores	10.0
image_url	10.0
image_small_url	10.0
vitamin-b1_100g	9.0
vitamin-b2_100g	8.0
vitamin-pp_100g	8.0

Observations Variables doublon

Code/Produit en doublon

Variables redondantes :

- Modalités multilingues
- Valeurs équivalentes

Échantillon de taille : 20164 x 38

```
#Comparaison de quelques variables en apparence similaires
def variable_duplicate(data, var1, var2):
    return data[var1].nunique() == data[var2].nunique()

display(variable_duplicate(openfoodfacts, 'states', 'states_tags'))
display(variable_duplicate(openfoodfacts, 'additives_en', 'additives_tags'))
display(variable_duplicate(openfoodfacts, 'countries_en', 'countries_tags'))
display(variable_duplicate(openfoodfacts, 'categories_en', 'categories_tags'))
```

```
countries / countries_en
categories / categories_en
states / states_en
```

Valeurs outliers

Conditions des valeurs outliers :

- Nutriment pour 100g de produit > 100g ou négatif
- Val énergie pour 100g de produit > 2000kcal (8373,6kJ)

Après affichage et analyse de ces valeurs, les produits sont maintenus dans l'analyse…

```
fat_100g : 0
saturated-fat_100g : 0
trans-fat_100g : 0
cholesterol_100g : 1
carbohydrates_100g : 1
sugars_100g : 0
fiber_100g : 0
proteins_100g : 1
salt_100g : 3
sodium_100g : 1
vitamin-a_100g : 0
vitamin-c_100g : 0
calcium_100g : 0
iron_100g : 0
nutrition-score-fr_100g : 0
energy_100g : 1
```

Seulement 8 observations concernées

Valeurs manquantes restantes

Beaucoup moins de valeurs manquantes sur notre corpus de variables nutritionnelles.

Échantillon de taille : 20164 x 38

nutrition-score-fr_100g	100.0
energy_100g	100.0
url	100.0
creator	100.0
created_t	100.0
last_modified_t	100.0
product_name	100.0
categories	100.0
countries_en	100.0
nutriscore_grade	100.0
pnns_groups_1	100.0
pnns_groups_2	100.0
main_category	100.0
main_category_en	100.0
states	100.0
fat_100g	100.0
sugars_100g	100.0
sodium_100g	100.0
salt_100g	100.0
proteins_100g	100.0
saturated-fat_100g	100.0
code	100.0
carbohydrates_100g	100.0
ingredients_that_may_be_from_palm_oil_n	97.0
ingredients_from_palm_oil_n	97.0
additives_n	97.0
ingredients_text	97.0
nova_group	94.0
fiber_100g	94.0
cholesterol_100g	92.0
brand_owner	92.0
iron_100g	91.0
trans-fat_100g	91.0
calcium_100g	90.0
vitamin-c_100g	78.0
vitamin-a_100g	77.0
additives_en	67.0
brands	59.0

Echantillon de travail…

Suppression des observations dont les données nutritionnelles ne sont pas à 100% disponibles.

Échantillon de taille : 10521 x 38

nutrition-score-fr_100g	100.0
iron_100g	100.0
pnns_groups_1	100.0
nova_group	100.0
nutriscore_grade	100.0
ingredients_that_may_be_from_palm_oil_n	100.0
ingredients_from_palm_oil_n	100.0
additives_en	100.0
additives_n	100.0
ingredients_text	100.0
countries_en	100.0
categories	100.0
product_name	100.0
last_modified_t	100.0
created_t	100.0
creator	100.0
url	100.0
pnns_groups_2	100.0
states	100.0
sugars_100g	100.0
main_category	100.0
calcium_100g	100.0
vitamin-c_100g	100.0
vitamin-a_100g	100.0
sodium_100g	100.0
salt_100g	100.0
proteins_100g	100.0
fiber_100g	100.0
code	100.0
carbohydrates_100g	100.0
cholesterol_100g	100.0
trans-fat_100g	100.0
saturated-fat_100g	100.0
fat_100g	100.0
energy_100g	100.0
main_category_en	100.0
brand_owner	99.0
brands	64.0

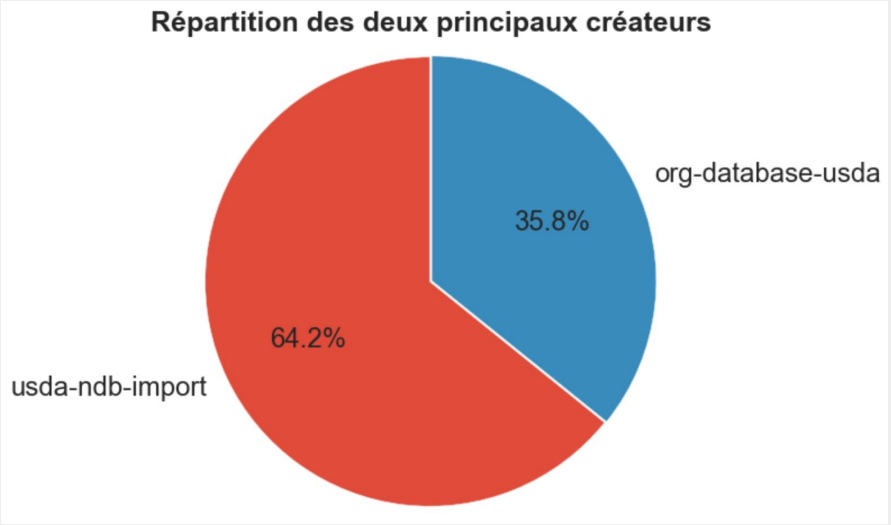
Analyse des variables clés

Description et analyse des principales variables explicatives. Visualisation afin de mieux comprendre leur comportement.

Contributeurs identifiés

+99% des contributeurs se résument à deux principales entités gérées par le **département de l'Agriculture Américaine.**

Source de qualité

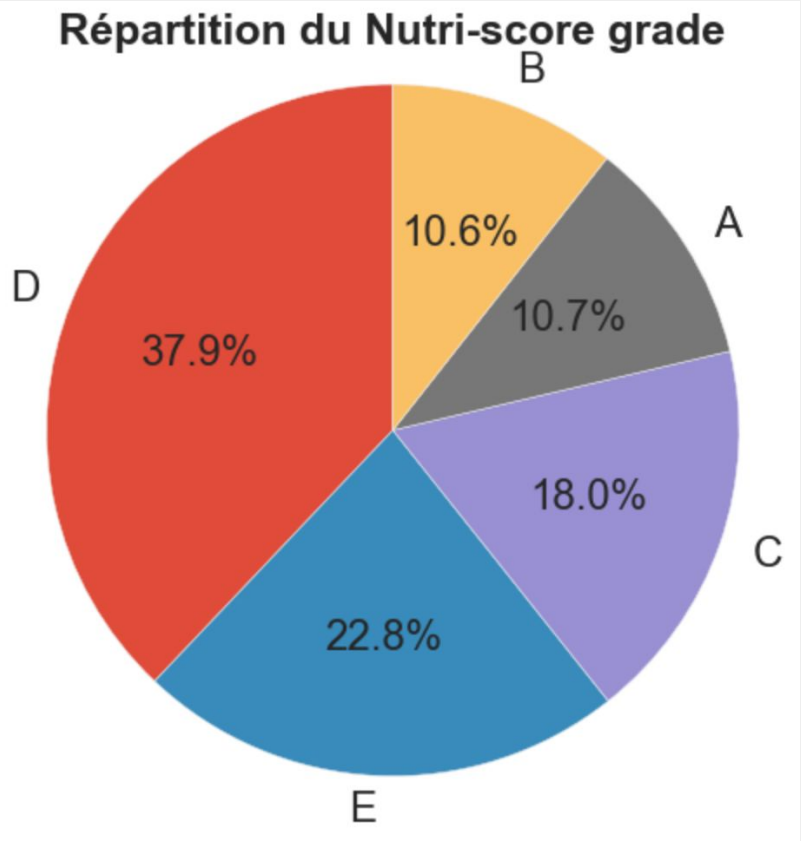


	creator	percentage
0	usda-ndb-import	61.39
1	org-database-usda	34.23
2	openfoodfacts-contributors	1.44
3	kiliweb	1.16
4	foodvisor	0.31
5	tacinte	0.23
6	waistline-app	0.19
7	stephane	0.15
8	bdwyer	0.12
9	veganeamos	0.11
10	date-limite-app	0.07

Nutri-score grade

+60% des produits ont une étiquette D et E

Échantillon de produits alimentaires plus représentatif des produits à la **qualité nutritionnelles Médiocre**.



Nutri score vs grade

Correspondances entre le Nutri-score et son grade.

Vérification des données avec le référentiel de base…

Utile dans le cadre d'une modélisation statistique.

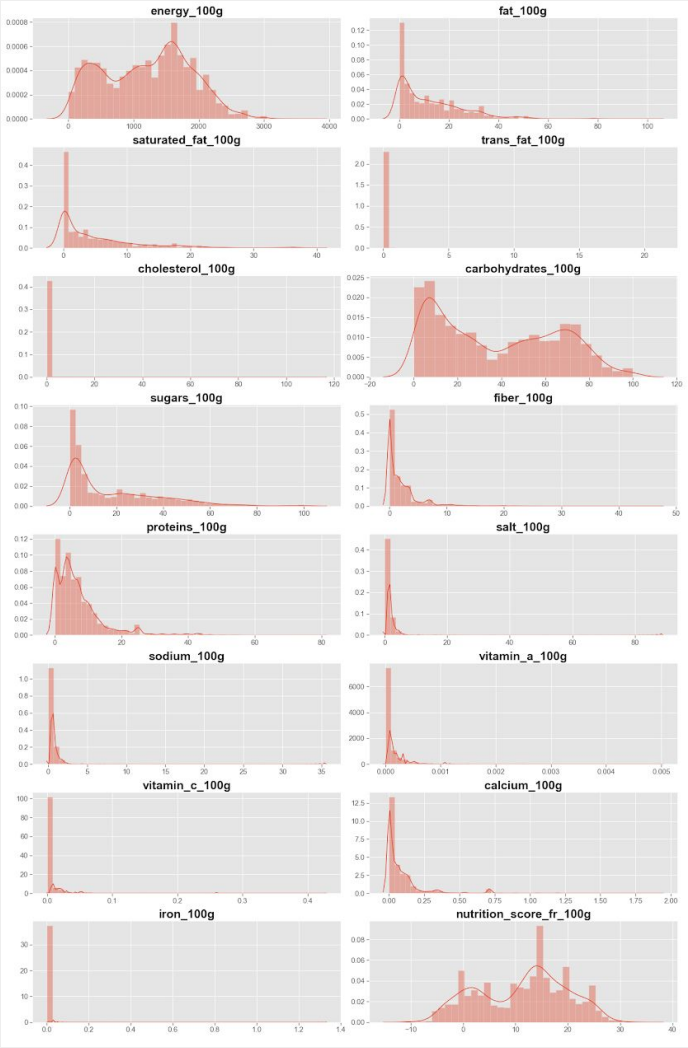
	nutriscore_grade	nutrition_score_fr_100g
0	a	[-2.0, -6.0, -4.0, -3.0, -1.0, -5.0, -9.0, -8.0, -7.0, -10.0, -11.0]
1	b	[2.0, 0.0, 1.0, -3.0]
2	c	[9.0, 8.0, 10.0, 3.0, 7.0, 5.0, 4.0, 6.0, 2.0]
3	d	[17.0, 11.0, 12.0, 13.0, 15.0, 16.0, 14.0, 18.0, 9.0, 6.0, 7.0, 8.0]
4	e	[25.0, 24.0, 27.0, 21.0, 23.0, 20.0, 19.0, 22.0, 26.0, 28.0, 30.0, 29.0, 17.0, 13.0, 16.0, 14.0,...

Rating	Score range	Colour
A	min. to -1	Dark green
B	0 to 2	Light green
C	3 to 10	Light orange
D	11 to 18	Medium orange
E	19 to max.	Dark orange

Distribution des variables nutritionnelles

Distribution non gaussienne confirmée également par un test de Kolmogorov-Smirnov.

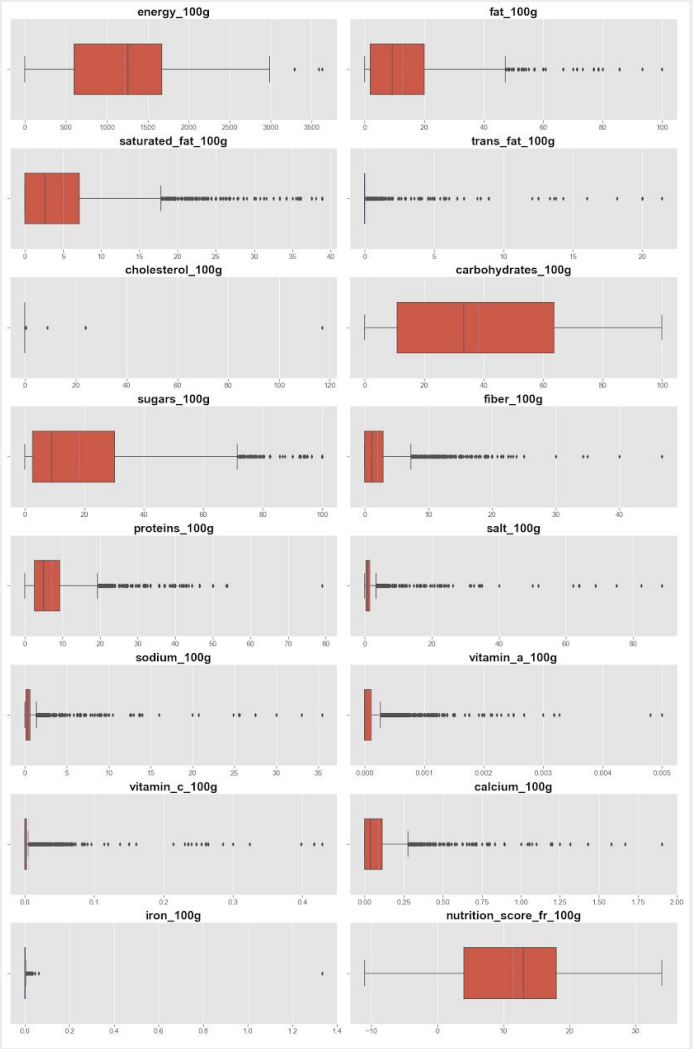
Forme souvent **asymétrique** avec étalement à droite, ou binomiale pour la variable cible *nutrition_score_fr_100g*.



Mesures de dispersion Boxplot

Le profilage de nos séries statistiques nutritionnelles permet une première approche.

Mais quel est l'impact du Nutri-score grade?



Analyse des corrélations

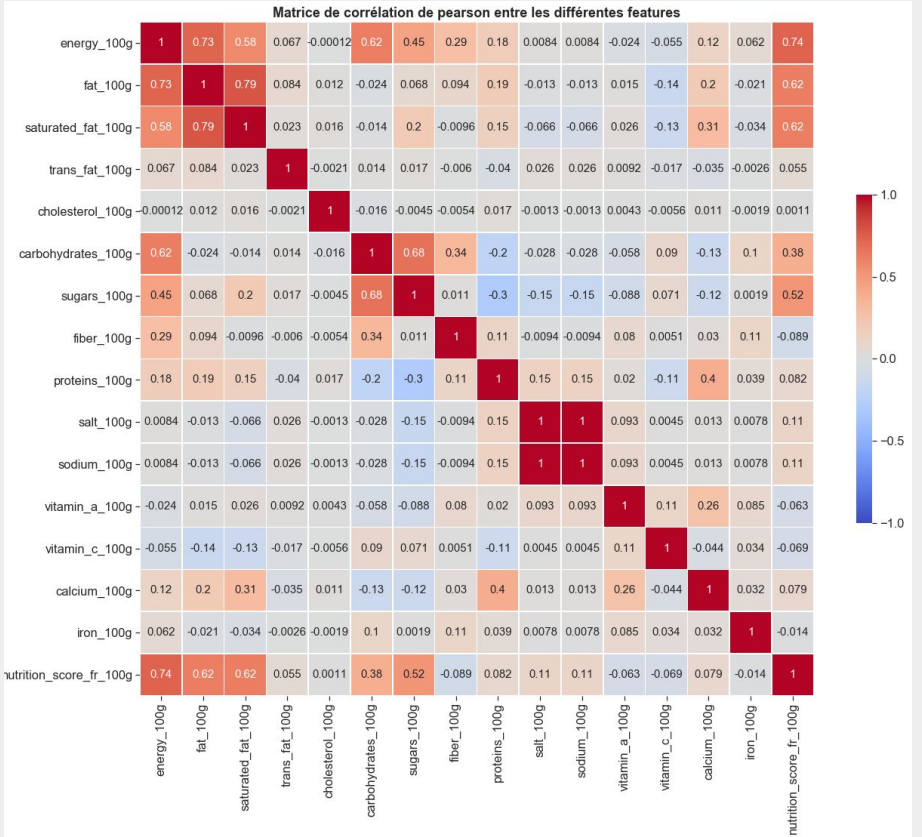
L’analyse multivariée permettra d’avoir une approche statistique tenant compte des liens entre nos variables nutritionnelles. Les corrélations seront traitées dans le contexte métier de l’application souhaitée.

Matrice des corrélations

Forte corrélation du Nutri-score:

- Graisses et graisses saturées
- Sucre
- énergie

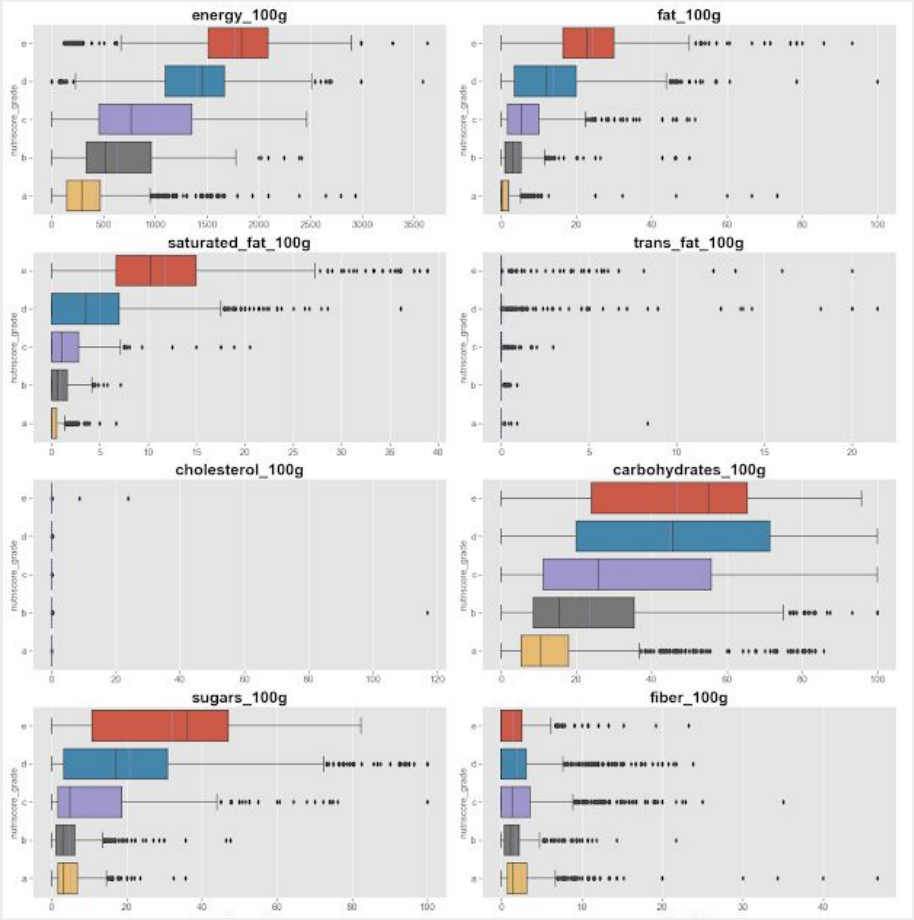
L'ANOVA permettra d'en apprendre plus sur les liens du Nutri-score grade / variables…



ANOVA à un facteur Nutri-score grade

Variance plus marquée pour l'énergie, les graisses, les graisses saturées, les glucides et le sucres.

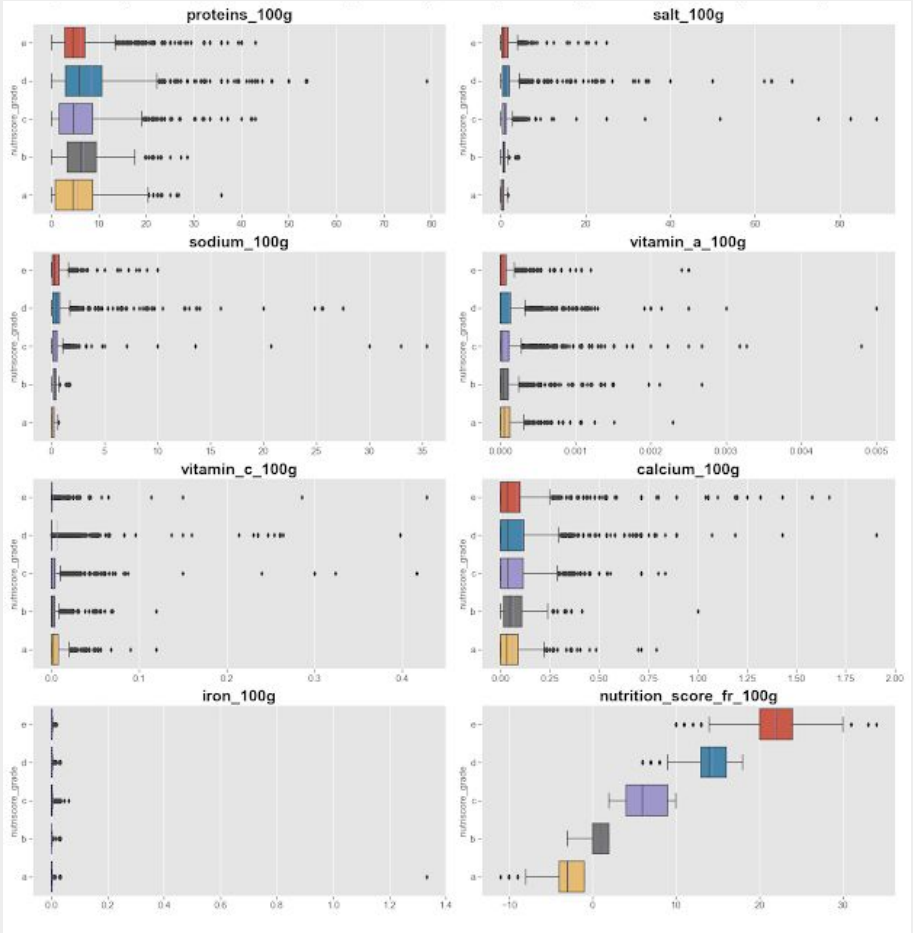
Écarts moins significatifs pour les fibres.



ANOVA
à un facteur
Nutri-score grade

Aucune variance propre au Nutri-score grade pour le fer, le cholestérol, les graisses transformées, les vitamines A et C, le sel et le sodium.

Vérification par un test de Fisher.



Test de fisher

H0 est rejetée sur un seuil 5% selon laquelle : A=B=C=D=E=0
(sauf pour le fer et le cholestérol)

La variance du Nutri-score grade a donc bien un effet sur les valeurs nutritionnelles.

L'intuition de départ via les visualisations graphiques se confirme.

```
p-value des variables…
energy_100g: 0.0
fat_100g: 0.0
saturated_fat_100g: 0.0
trans_fat_100g: 5.245462326685858e-07
cholesterol_100g: 0.12760525770236267
carbohydrates_100g: 0.0
sugars_100g: 0.0
fiber_100g: 5.70353364630666e-52
proteins_100g: 8.114763332698311e-60
salt_100g: 9.006230762005503e-50
sodium_100g: 9.012396532573376e-50
vitamin_a_100g: 5.7370139641103375e-25
vitamin_c_100g: 1.1340972298515993e-17
calcium_100g: 3.0171152570648807e-31
iron_100g: 0.13208734871364727
nutrition_score_fr_100g: 0.0
```

Scoring features indépendance?

scoring_features = ["nutriscore_grade", "nova_group"]

Test du khi-2

- Indice élevé
- p-value < 5%
- H0 (abs de relation) rejetée

Il existe un lien entre le Nutri-score grade et le Nova groupe.

nutriscore_grade / nova_group	a	b	c	d	e	Total
1.0	118	5	26	7	3	159
2.0	0	1	0	0	0	1
3.0	225	106	177	223	137	868
4.0	786	1003	1689	3760	2255	9493
Total	1129	1115	1892	3990	2395	10521

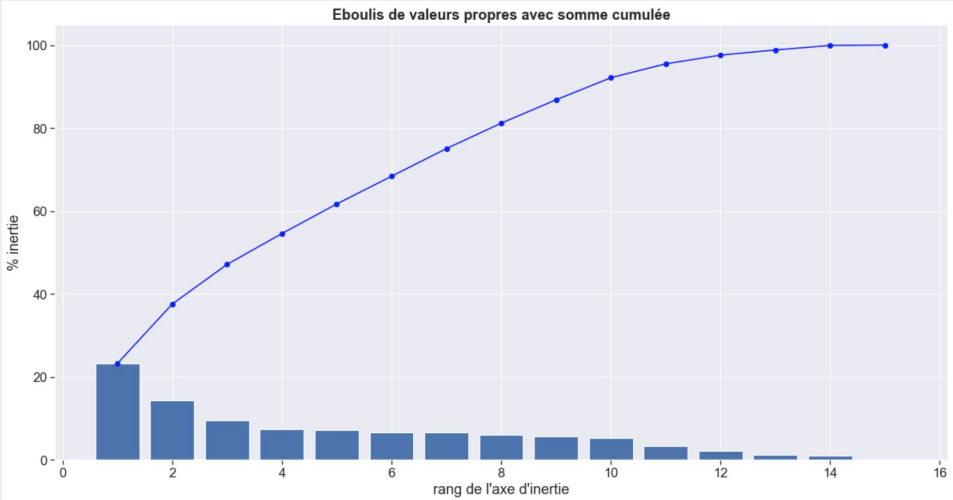
```
#Calcul du khi-2 et de la p-value
import scipy.stats as st
chi2, pvalue, degrees, expected = st.chi2_contingency(cont)
chi2, degrees, pvalue
```

```
(1000.6579215306631, 20, 2.8268196589863586e-199)
```

Structure sous-jacente

Pour choisir le nombre de composantes à utiliser, on regarde la proportion de la variance totale expliquée par k composantes.

Une analyse en composantes principales est réalisable.



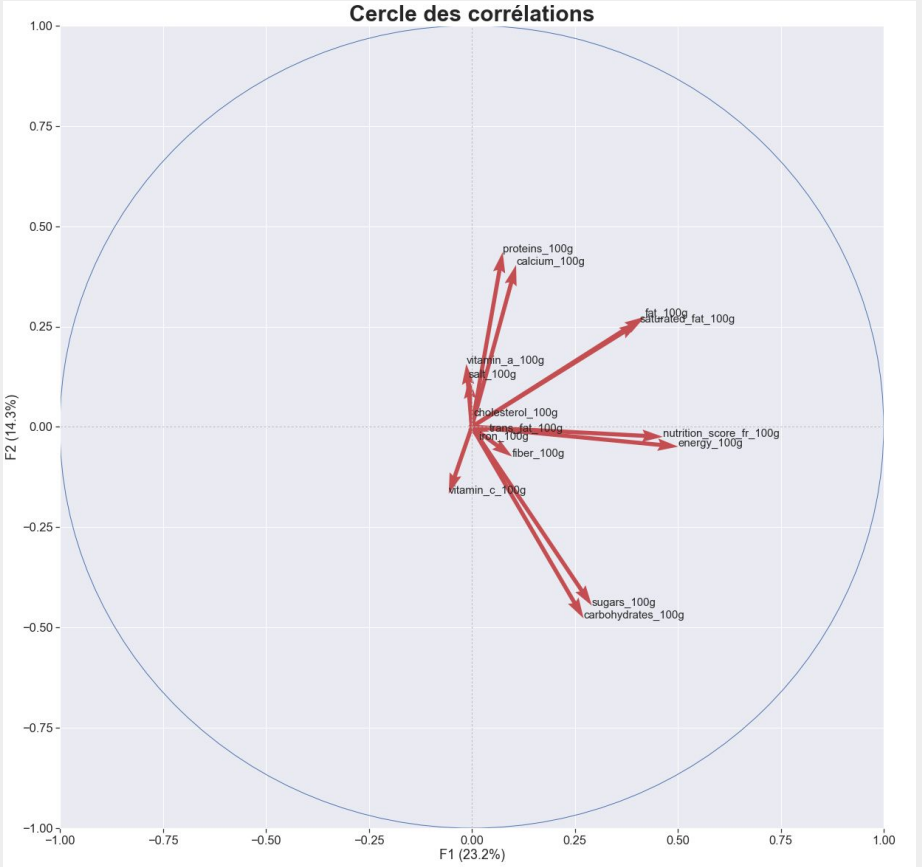
	Axe	Valeur propre	% variance expliquée	% cum. var. expliquée
0	F1	3.485848	23.0	23.0
1	F2	2.147842	14.0	38.0
2	F3	1.435725	10.0	47.0
3	F4	1.115005	7.0	55.0
4	F5	1.070521	7.0	62.0
5	F6	1.002798	7.0	68.0
6	F7	0.998658	7.0	75.0
7	F8	0.920597	6.0	81.0
8	F9	0.844932	6.0	87.0
9	F10	0.798537	5.0	92.0
10	F11	0.501386	3.0	95.0
11	F12	0.316722	2.0	98.0
12	F13	0.185798	1.0	99.0
13	F14	0.164629	1.0	100.0
14	F15	0.012428	0.0	100.0

ACP et cercle des corrélations

1er Plan factoriel permet d'expliquer **38% de la variance.**

Forte corrélation :

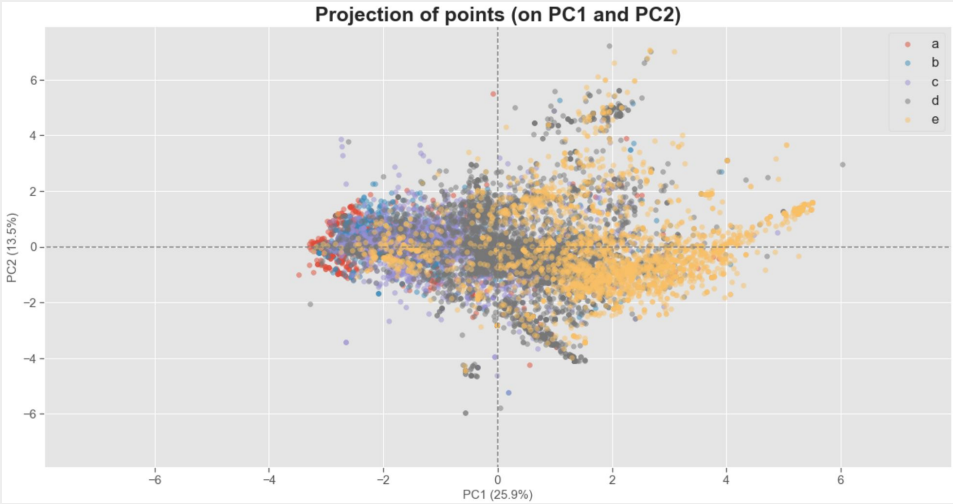
- nutri-score / énergie
- sucre / glucides
- graisses / graisses saturées
- protéines / calcium

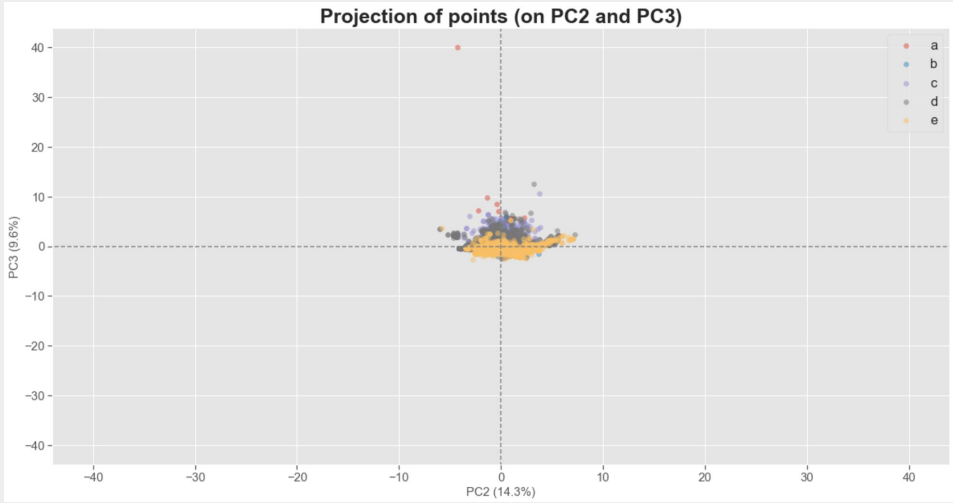


Projection des individus

Représentation optimale des grades du Nutri-score :

- **38% de la variance**
- **Forte présence du D et E**
- **1er plan factoriel plus représentatif des grades**





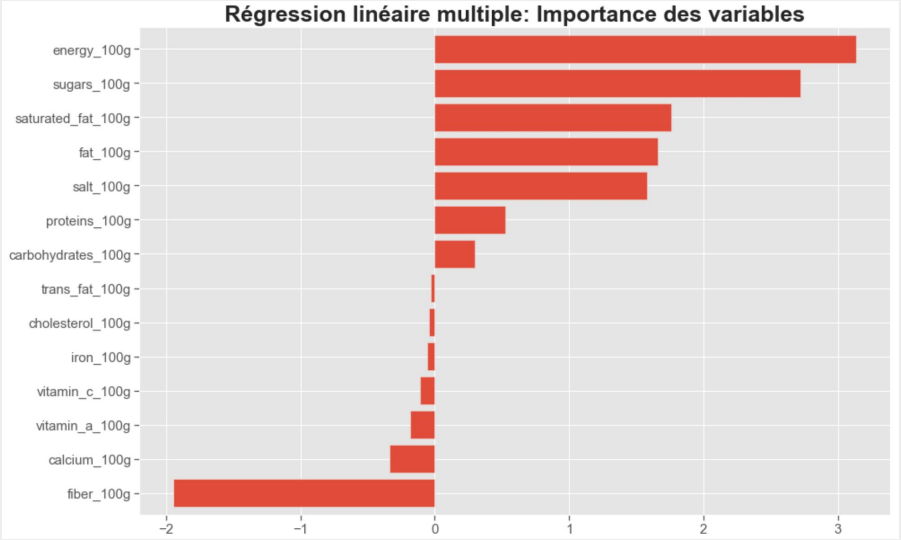
Modélisation

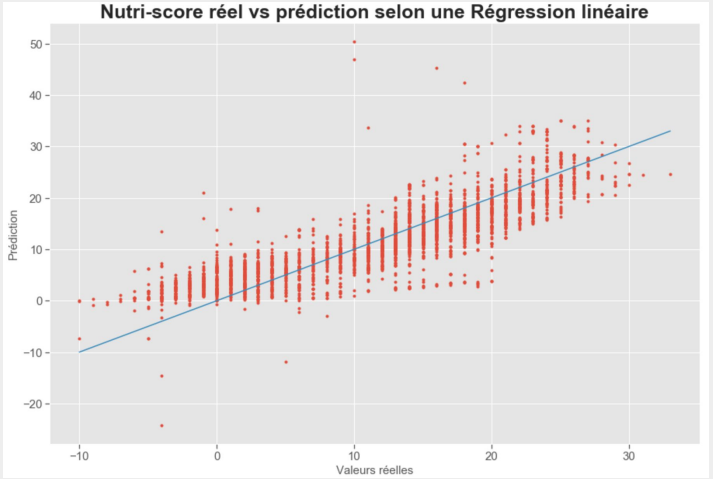
Régression linéaire multiple, régularisation par la régression ridge et lasso, forêts aléatoires. Comparaison des premiers résultats.

Régression linéaire multiple

Coefficient de détermination R^2 :
≃ 0.75

Mesure d'erreur RMSE :
≃ 4.21



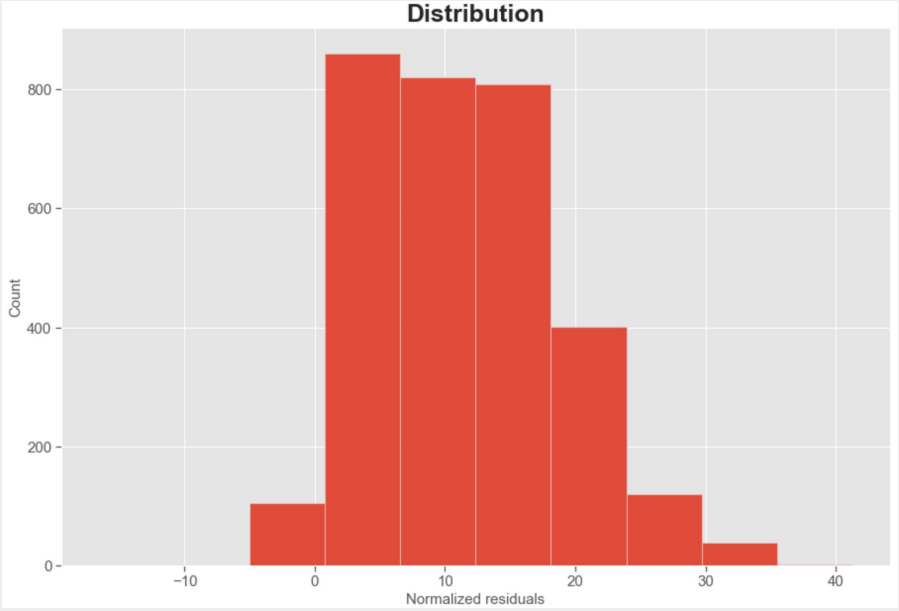


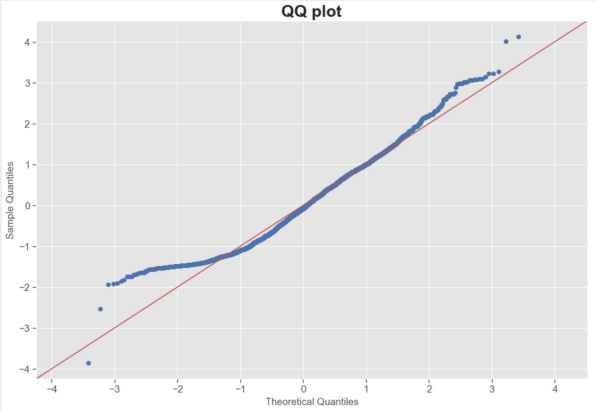
Contrôle du modèle : résidus

Test de normalité des résidus

- Résidus alignés avec une distribution théorique
- Symétrique assez proche d'une forme gaussienne

Modélisation par régression linéaire non absurde .



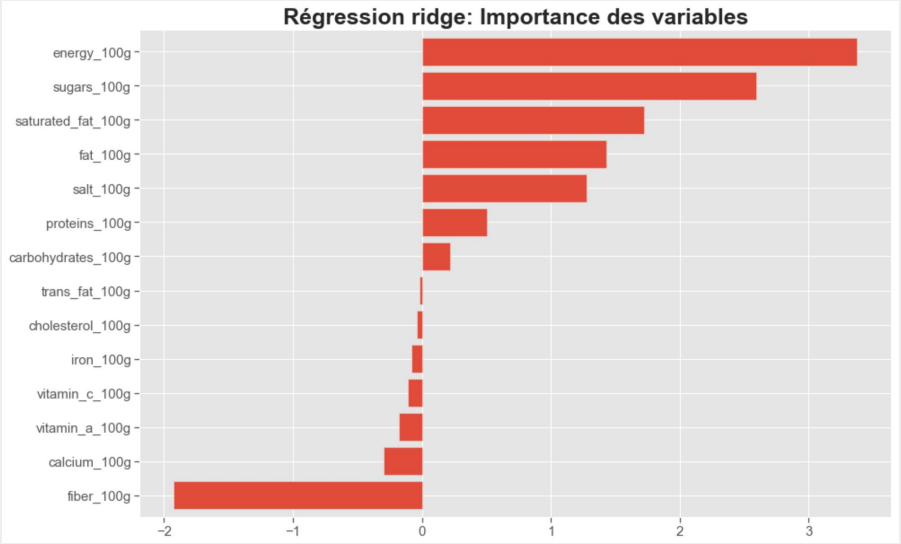


Régularisation régression ridge

Coefficient de détermination R^2 :
≃ 0.74

Mesure d'erreur RMSE :
≃ 4.21

Réduction du poids des variables explicatives.

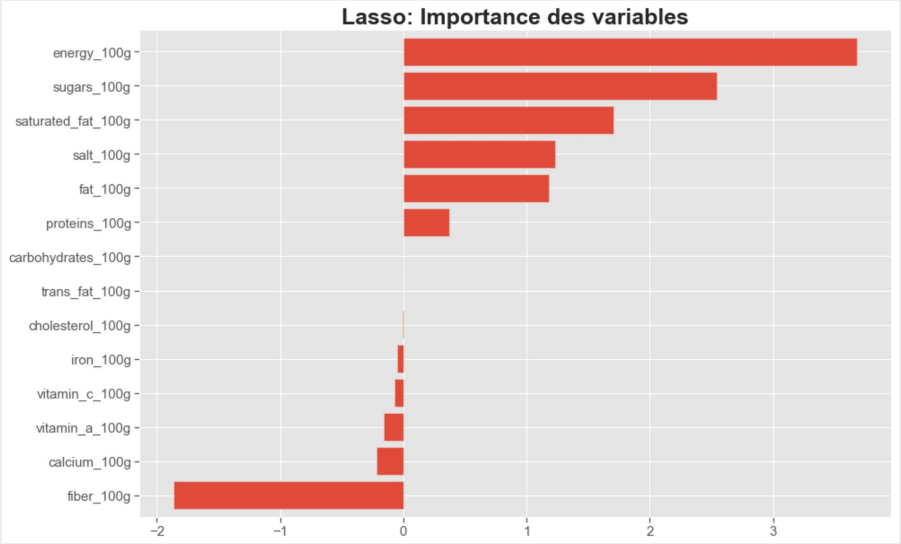


Application du lasso

Coefficient de détermination R^2 :
≃ 0.74

Mesure d'erreur RMSE :
≃ 4.34

Suppression de certaines variables, sous-ensemble permettant la généralisation.

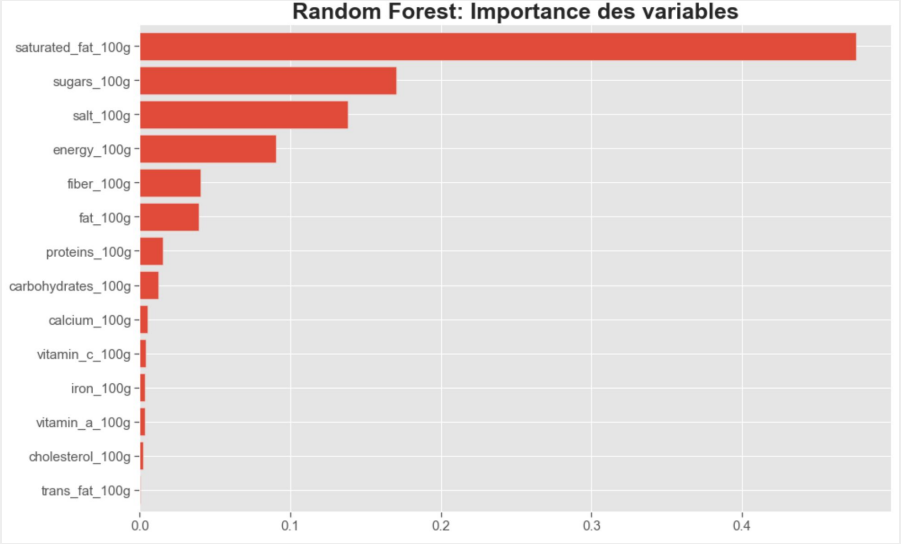


Forêts aléatoires

Coefficient de détermination R^2 :
≃ 0.90

Mesure d'erreur RMSE :
≃ 1.45

Réduction de la corrélation entre nos apprenants faibles.



Conclusion

L’application souhaitée est-elle réalisable? L’échantillon est-il en adéquation avec les besoins nécessaires?

Pertinence et faisabilité de l’application

Identification de variables nutritionnelles essentielles

Corrélation identifiée avec le Nutri-score

Affichage prédictif possible d’un Nutri-score

Affichage prédictif de la classification Nova

Premiers essais de modélisation…