Une application Open Food Facts ?

Préparation des données: Sélection des cibles

- Base de données très grandes avec plus de 320,000 lignes et 162 colonnes descriptives
- Variables cibles: Nutrition score et Nutrition Grade

Nutri-score	Nutri-grade	Couleur
-15 à -1	A	Vert
0 à 2	B	Vert clair
3 à 10	C	Jaune
11 à 18	D	Orange
19 ou plus	E	Rouge

Préparation des données: Sélection des features

- Sélection d'environ quinze nutriments présentant un taux de valeurs manquantes inférieur à 50 %
- Les nutriments sélectionnés pour faire l'étude de faisabilité:

1. Vitamine A
2. Vitamine C,
3. Fer
4. Calcium,
5. Gras transformés,
6. Cholestérol,
7. Fibre,
8. Glucides,
9. Graisse,
10. Energie (kJ)
11. Sodium
12. Sel
13. Sucres
14. Graisses saturées
15. Protéines

Préparation des données: Gestion des outliers

Méthodes utilisées:

➔ Gestion des valeurs aberrantes cas par cas

➔ Bornage forcé [0,100]

➔ Automatisation bornage interquartile

➔Imputation par la médiane

Gestion des outliers : Cas par cas

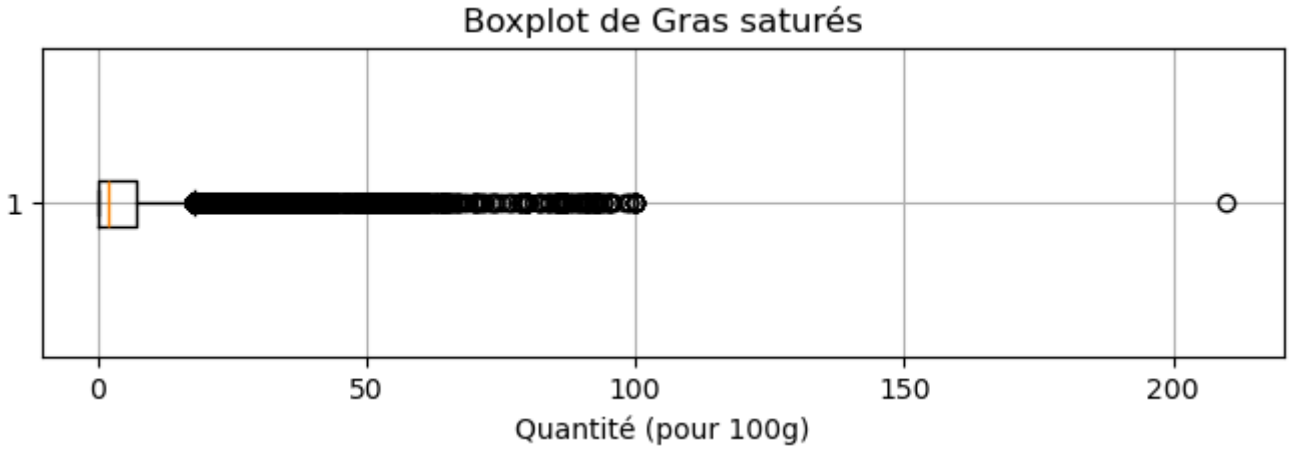
Chaque aliment a un poids maximum de 100 g.

Par conséquent, la somme des nutriments pour chaque aliment ne doit pas dépasser 100.

Pour chaque nutriment, nous analysons les lignes supérieures à 100 (valeurs non valides), les valeurs négatives (valeurs non valides) et celles égales à 100 pour voir si la composition semble normale.

Nous supprimons aussi les lignes qui ne sont pas alimentaires (DVD, livres etc.)

Par exemple, le nutriment des gras saturés : le graphique montre une valeur supérieure à 200. En sortant la ligne correspondante, nous pourrons ajuster les valeurs nutritionnelles avec les données réelles.



Gestion des outliers: Bornage [0;100]

Création de la variable « total_nutriment » et observation de la distribution du total des nutriments dans notre échantillon :

- 30 % des aliments présentent un total de nutriments supérieur ou égal à 106.
- 5% de l'échantillon a un total de nutriment supérieur ou égal à 162.

Nous appliquons une méthode d'ajustement automatique en contraignant les valeurs >100 ou <0 pour qu'elles respectent le bornage [0;100].

En effectuant cette méthode, il reste encore 74,602 lignes dont le total des nutriments dépasse 100.

```
df_final['total_nutriments'].quantile(q=[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.95,0.96,0.97,0.98,0.99])
```

```
0.10     15.847290
0.20     25.045276
0.30     34.735626
0.40     50.194480
0.50     68.336569
0.60     88.313375
0.70    106.026399
0.80    123.205592
0.90    146.630000
0.95    162.552704
0.96    166.000000
0.97    170.070003
0.98    174.707473
0.99    182.451462
Name: total_nutriments, dtype: float64
```

Gestion des outliers: Bornage interquartile

Méthode interquartile: Pour chaque nutriment, à l'exception des graisses transformées, nous appliquons la méthode de bornage interquartile. Les valeurs sont limitées à un maximum correspondant à la somme du troisième quartile et de 1,5 fois l'IQR.

IQR=Q3-Q1
Max = Q3 + (1,5*IQR)
Min = Q1-(1,5*IQR)
Q1 = 1er quartile (25%)
Q3 = 3ème quartile (75%)

Avec cette méthode, plus de 6000 lignes ne sont plus considérées comme aberrantes (74,602 à 68,694 lignes).

Gestion des outliers: Imputation par la médiane

La dernière étape de gestion des valeurs aberrantes consiste à remplacer les données aberrantes par la médiane des nutriments concernés. Lorsque la valeur d'un nutriment dépasse les seuils suivants, elle est automatiquement remplacée par la médiane. Plus de 37,000 lignes ont été ajustées avec cette méthode.

Glucides = 60 g/ 100 g.
Réaliste pour des aliments comme le riz ou les pâtes, peut être légèrement bas pour des aliments très transformés.

Gras transformés = 10g / 100g.
Réaliste mais relativement élevé. Seuls les aliments comme la margarine, les fritures ou les aliments ultra-transformés ont ce niveau.

Graisse totale = 50g /100g.
Réaliste pour beaucoup d'aliments très gras, peut être légèrement faible pour des produits comme le beurre ou l'huile.

Sucre = 50g /100g.
Réaliste pour des aliments très sucrés comme les sodas ou les confiseries.

Gestion des valeurs NaN

Nous maintenons une marge d'erreur en conservant les lignes avec un total de nutriments strictement inférieur à 111.

Seulement 4% de l'échantillon ont un total de nutriment > 100.

```
df_final2['total_nutriments']
```

```
0.10     14.534545
0.20     22.893701
0.30     29.273500
0.40     37.359054
0.50     49.607770
0.60     61.158619
0.70     70.507580
0.80     80.078974
0.90     94.226772
0.95     99.927874
0.96    102.087402
0.97    104.546053
0.98    107.064850
0.99    109.251350
1.00    110.998640
```

Méthode utilisée:

➔ Suppression des lignes entières de NaN (191)

➔ Calcul de la matrice de corrélation entre nutriments

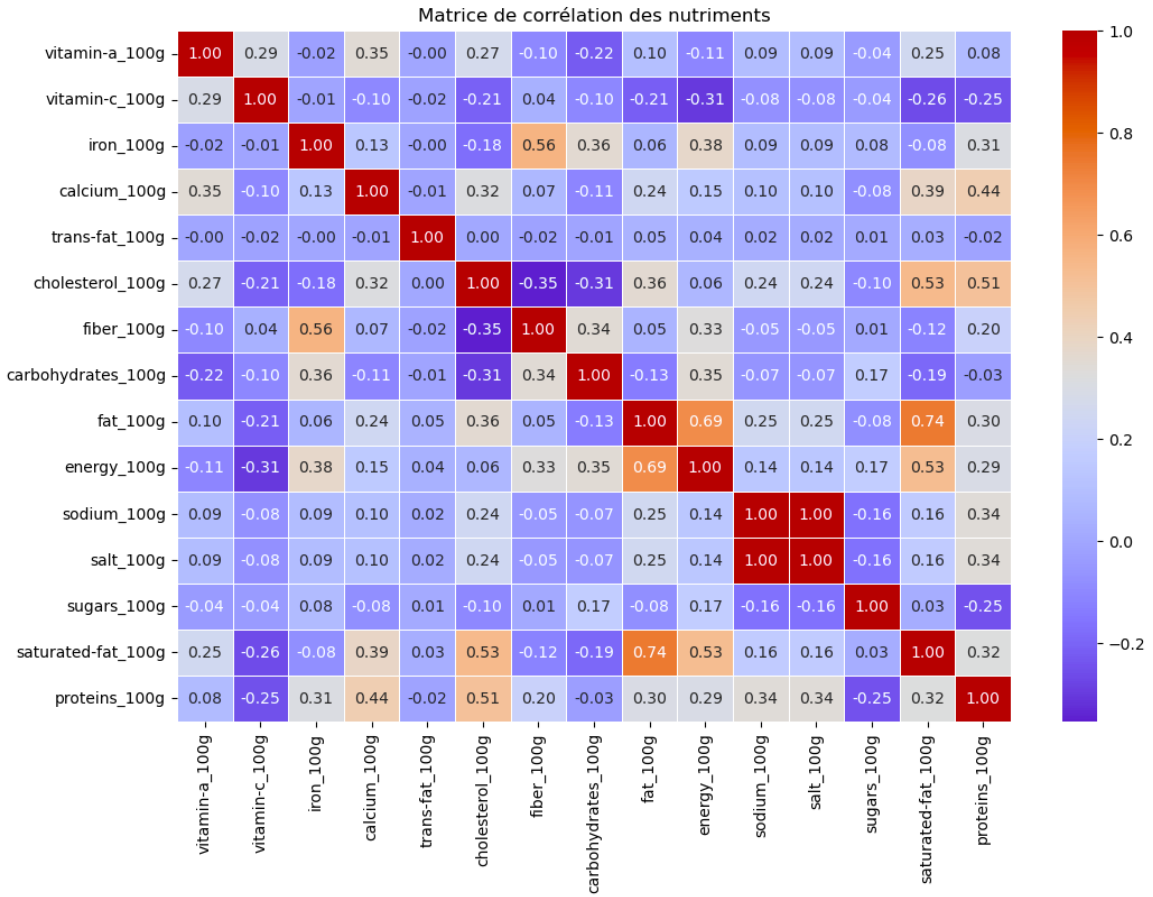
➔ Imputation des données manquantes via la régression linéaire

➔ Imputation des données manquantes via la médiane pour le nutriment graisse transformée

Gestion des valeurs NaN: Matrice de corrélation

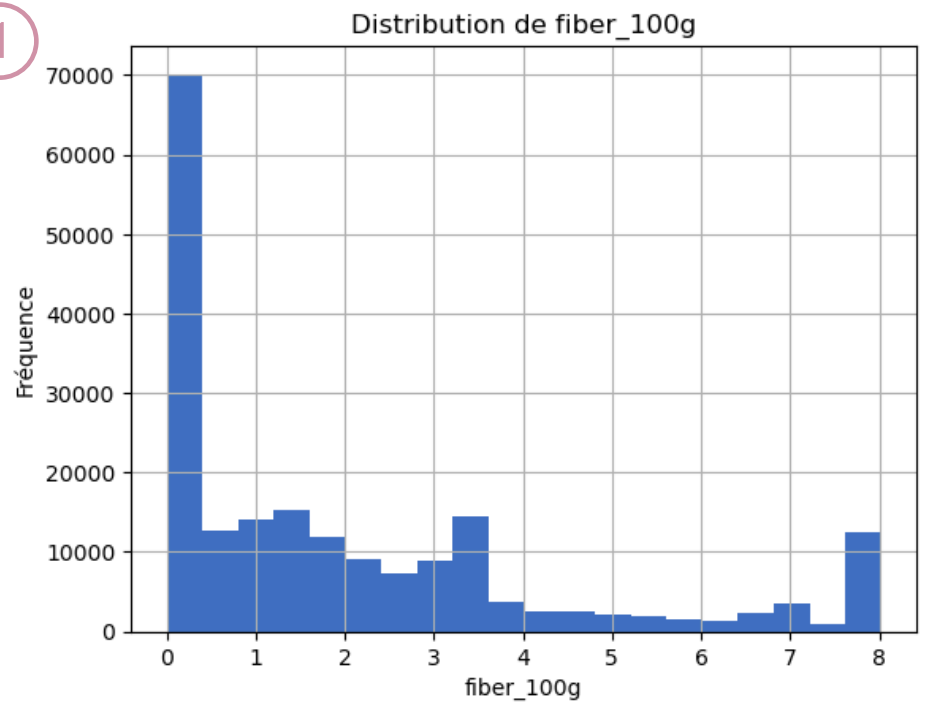
Cette matrice montre les relations entre nutriments. Voici l'ordre d'imputation par régression linéaire des valeurs manquantes suivi:

1. Graisse => Energie, gras saturés et protéines
2. Cholestérol => Protéines, gras saturés et graisse
3. Calcium => Protéines, gras saturés et cholestérol
4. Vitamine A => Gras saturés, calcium et cholestérol
5. Glucides => Energie, cholestérol et vitamine A
6. Fibres => Cholestérol, énergie et glucides
7. Fer => Fibres, glucides et énergie
8. Vitamine C => Energie, vitamine A et gras saturés
9. Trans-fat => Les corrélations sont trop faibles. Nous imputerons les valeurs manquantes via la médiane.

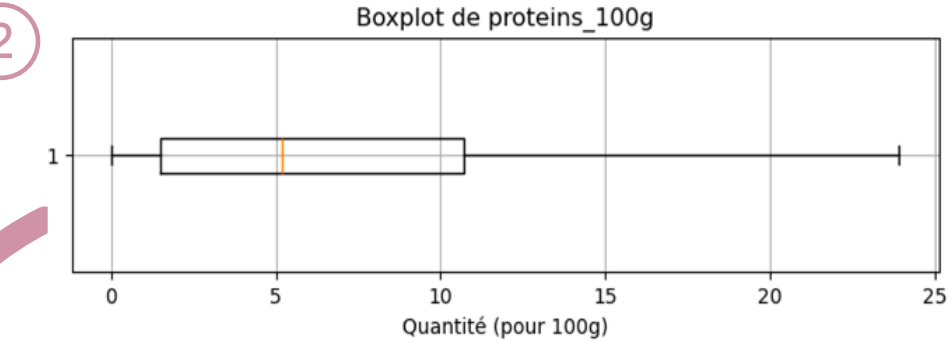


Analyse univariée

1



2



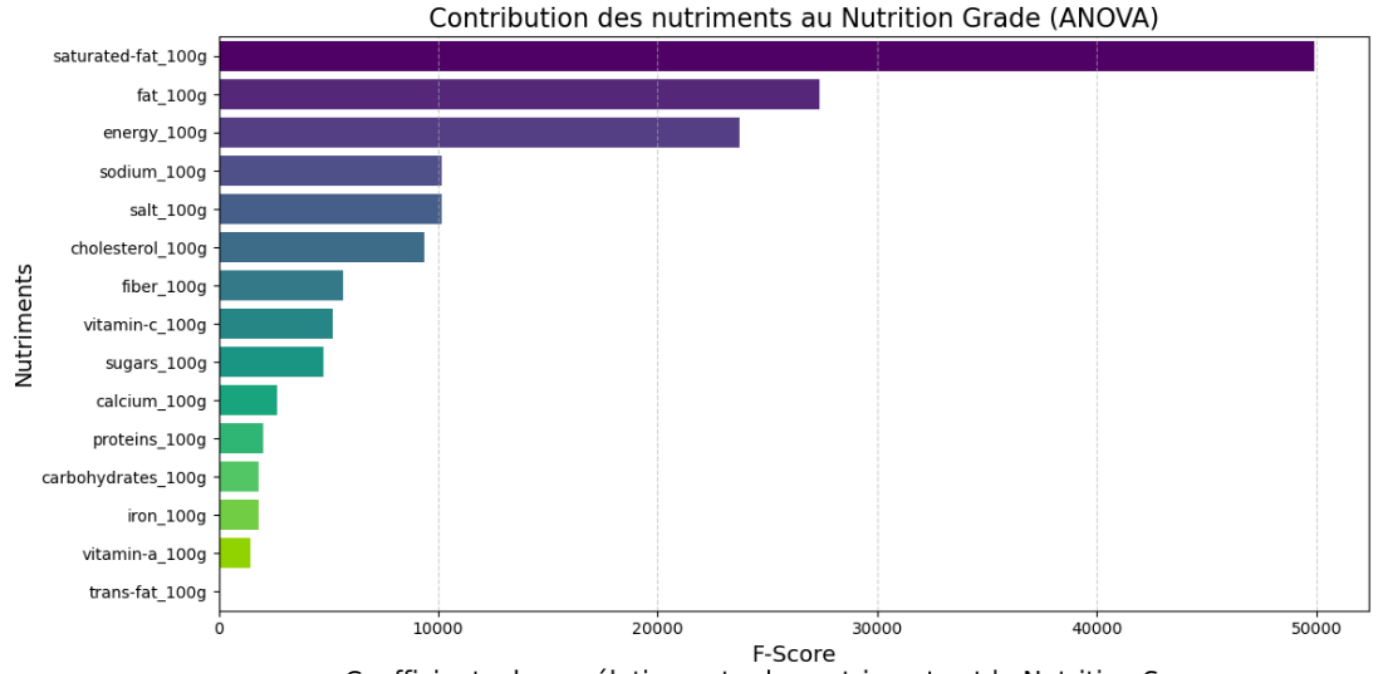
3

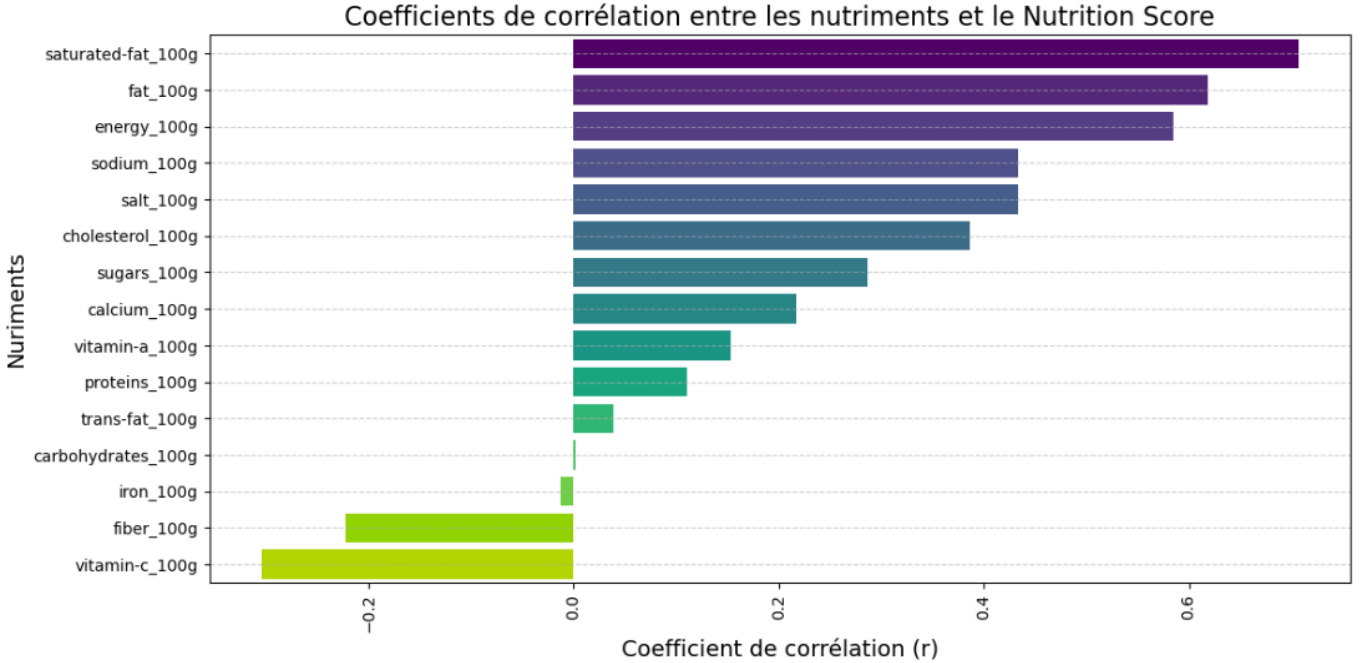
```
nutrition_grade_fr                nutrition_grade_fr
a    17.911823                    a    35577
b    17.103256                    b    33971
c    22.092104                    c    43880
d    27.664973                    d    54949
e    15.227844                    e    30246
Name: count, dtype: float64       Name: count, dtype: int64
```

1. **Distribution du nutriment fibres**
   Ce graphique illustre la distribution des teneurs en fibres dans les aliments. On constate que la majorité des aliments ne contiennent pas de fibres. Environ 15 000 aliments présentent des quantités entre 1 et 2 grammes de fibres, tandis que près de 3 000 aliments en contiennent environ 4 grammes.
2. **Distribution des protéines**
   Le boxplot montre la distribution des teneurs en protéines. Les valeurs varient entre 0 et 25 grammes, avec une médiane autour de 5 grammes de protéines.
3. **Répartition du Nutrition Grade**
   Ces tableaux présentent la distribution du Nutrition Grade des aliments. On observe que 18 % des produits sont classés en A, 17 % en B, 22 % en C, 27 % en D, et 15 % en E. La répartition est relativement équilibrée, bien qu'une proportion plus importante d'aliments se situe dans la catégorie D.

Analyse bivariée

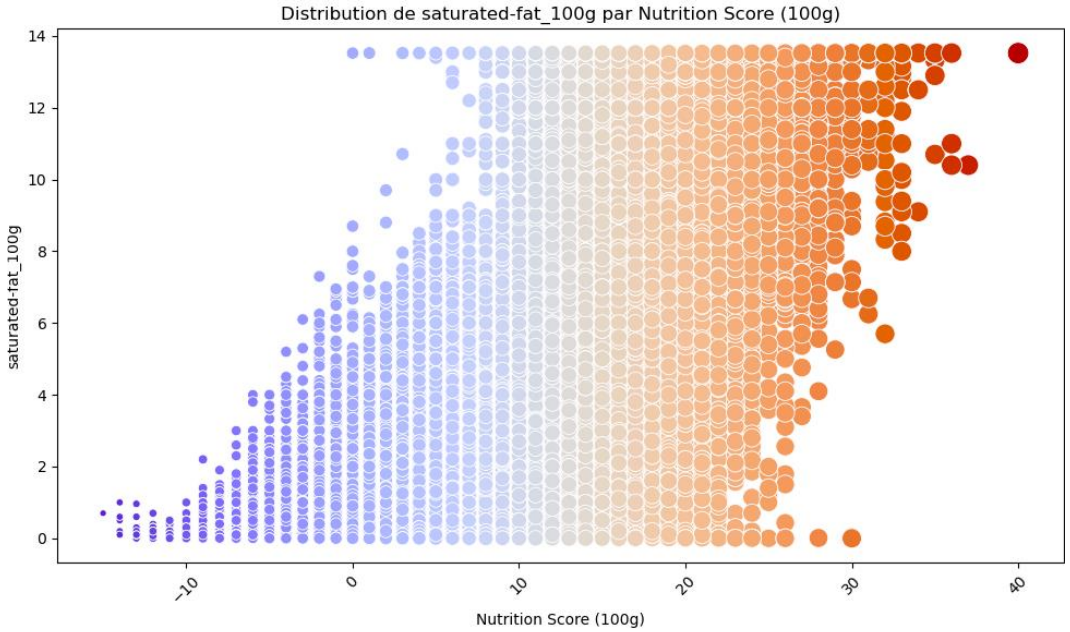
- Corrélations fortes: Gras saturés, graisses, Energie, sodium et sel pour le nutrition grade et score. Le cholestérol a aussi une corrélation assez forte pour le nutrition score.
- Les fibres et la vitamine C ont des corrélations négatives sur le nutrition score.
- Ces nutriments sont essentiels pour la prédiction du nutrition grade et nutrition score.
- Les relations claires entre certains nutriments et nos variables cibles montrent qu'il est possible de construire un modèle prédictif, et ainsi une application capable de prévoir ces variables.





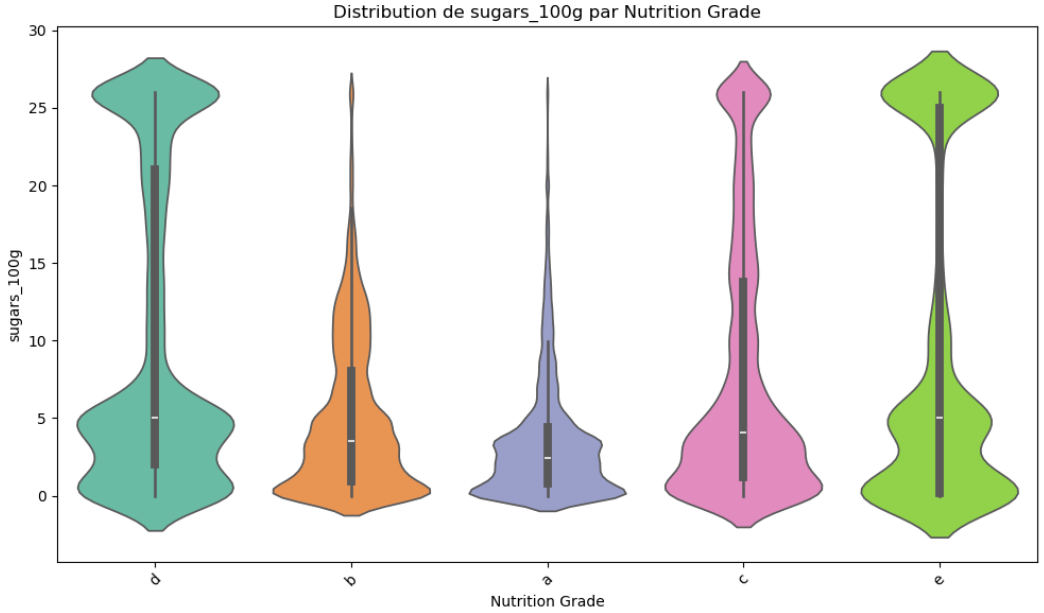
Analyse bivariée

1. Le graphique représente **le nuage de point** entre le nutriment gras saturés et nutrition score. Plus un rond est petit et bleu plus le nutrition score et faible et inversement. On voit se dessiner faiblement une relation : les aliments qui ont une quantité élevée en gras saturés ont un nutrition score élevés (rouge) et une faible relation linéaire.

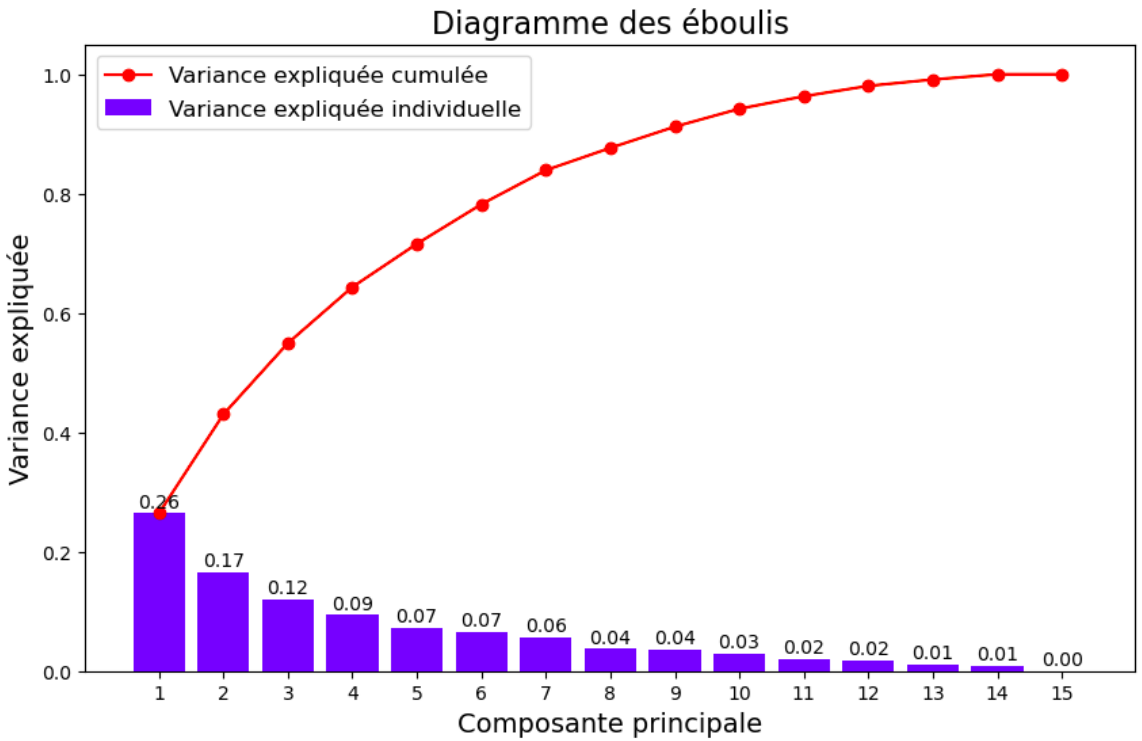


2. Le violin plot représente la distribution du nutriment sucres en fonction du nutrition grade. Les aliments classés en **D** et **E** ont des niveaux de sucres plus élevés, avec des distributions larges et une médiane (ligne blanche) plus élevée.

Les aliments classes en A et B ont une teneur en sucre plus faible. Les aliments classes en A ont une quantité en sucre ne dépassant pas 15g.

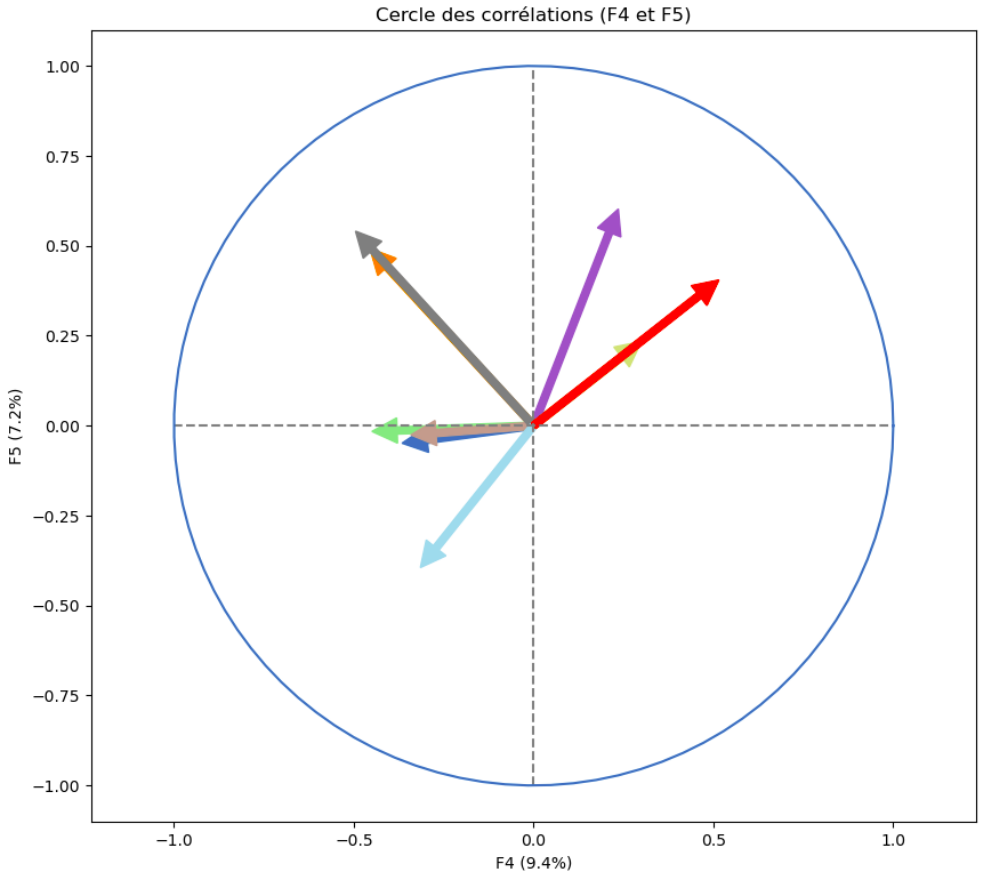


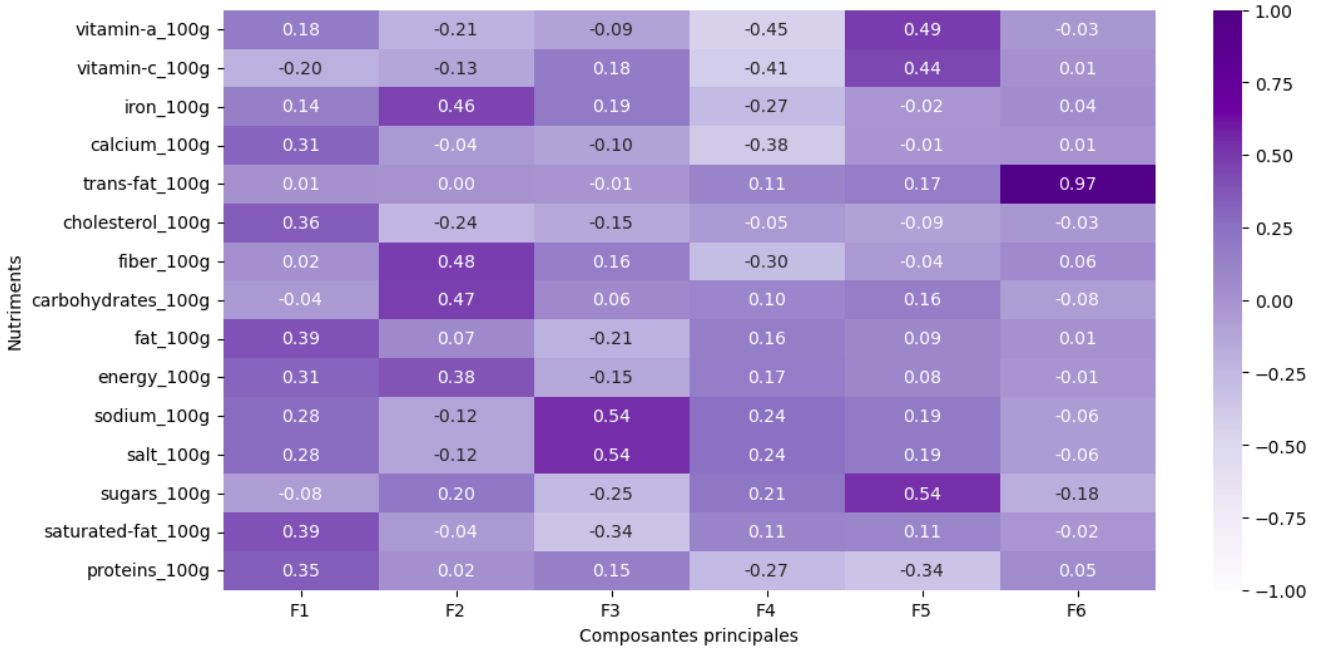
Analyse multivariée



- Nous utilisons la méthode l'Analyse des Composantes Principales (ACP) qui va nous permettre de réduire la dimensionnalité (ici 15 features) tout en gardant un maximum d'informations.
- En effet l'ACP va créer de nouvelles variables , appelées composantes principales, qui vont nous permettre de mieux comprendre comment les nutriments influencent le nutrition score, tout en réduisant le nombre de variables à analyser.
- Grâce au diagramme des éboulis nous pouvons choisir le nombre de composantes principales nécessaires pour ne pas voir une perte d'informations importantes.
- Le diagramme montre que 6 composantes principales sont nécessaires pour expliquer 78 % de la variance totale des données.

Analyse multivariée: ACP





- Le tableau des composantes principales (représenté avec la matrice de corrélation) permet de visualiser la contribution de chaque nutriment aux nouvelles composantes principales.
- Par exemple, pour la composante principale F1:

  *F1 = 0.18 * vitamin_a_100g + (-0.20) * vitamin_c_100g + 0.14 * iron_100g + 0.31 * calcium_100g + 0.01 * trans-fat_100g +etc.*

- Le cercle de corrélation met en avant les corrélations selon les projections. Ici, nous voyons que le sel et le sucres sont corrélés avec le nutrition score. En revanche on peut voir se dessiner une corrélation négative avec les protéines.

Prédiction de nutrition score

Utilisation de la régression linéaire	**Utilisation du modèle Random Forest**	**Utilisation du modèle Gradient Boosting**
MAE: 3.0908300806441615, MSE: 17.264043629874557, RMSE: 4.155002241861556, R2: 0.7710030658756062 Avec un R^2 de 0.771, le modèle explique une bonne partie de la variance du nutrition score. Le modèle fait en revanche une erreur moyenne de 3 unités dans sa prédiction. Par exemple, lorsque le modèle donne un nutrition score de 20 , le vrai nutrition score est plus ou moins 3 points.	Random Forest - MAE: 1.1865109379263028 Random Forest - MSE: 4.610254488012868 Random Forest - RMSE: 2.1471503179826206 Random Forest - R2: 0.9388478061152905 Avec un R2 de 0,938 le modèle explique la majorité de la variance totale de nutrition score. Le modèle fait une erreur moyenne de 1,18 unités.	GB- MAE: 2.5092885595916634 GB- MSE: 12.207627826611008 GB- RMSE: 3.4939415888951273 GB- R2: 0.838073315547693 Avec un R2 de 0,838 , le modèle explique aussi une grande majorité de la variance du nutrition score mais plus faible que le modèle RF. L'erreur moyenne est aussi plus faible que le modèle de régression linéaire.

Le modèle le plus performant pour prédire le Nutrition Score est le Random Forest, il explique une bonne partie de la variance totale, avec une erreur moyenne minimale de 1,18 unités.

Prédiction de nutrition grade

Utilisation de la régression logistique	**Utilisation de RandomForestClassifier**	**Utilisation du modèle XGBoost**
Accuracy Rég log: 0.6097671491504091 61% des observations ont été correctement classifiées dans le bon grade.	Accuracy Random Forest : 0.8664065449968533 86,6% des observations sont bien classées dans le bon grade.	Accuracy XGBoost : 0.8175456261799874 81,7% des observations sont bien classées.

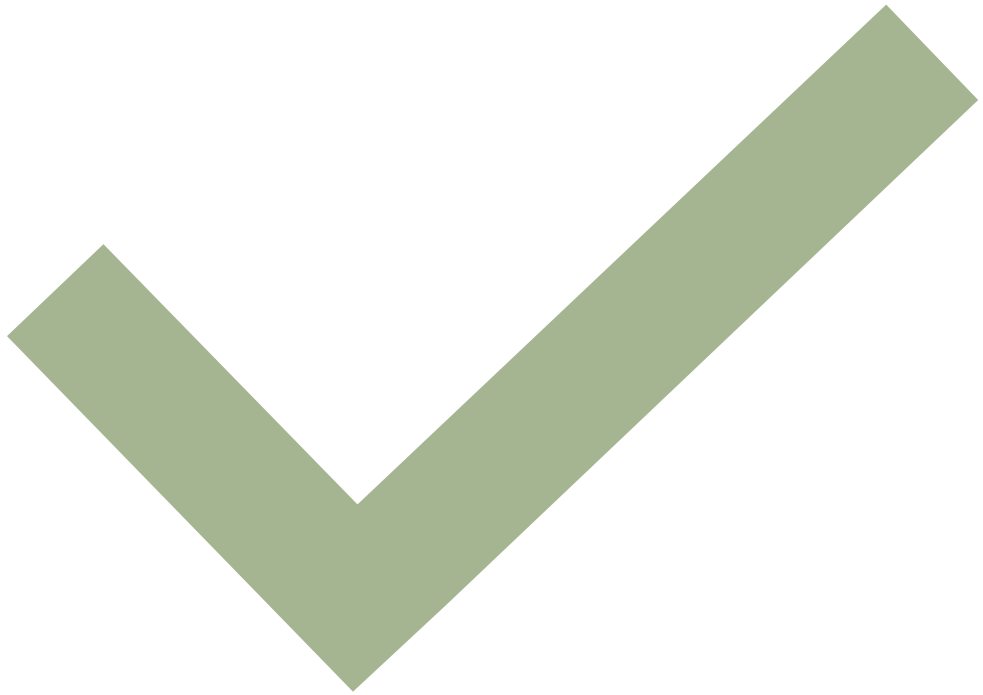
Pour prédire le Nutrition Grade, le modèle RandomForestClassifier s'avère également le plus robuste. Il atteint une précision de 86 %, ce qui signifie qu'il classe correctement les aliments dans la catégorie de Nutrition Grade dans 86 % des cas.

RGPD

1. Collecte des données nécessaires: les données collectées sont utiles pour un but bien précis. Ici pour la classification des aliments.
2. Transparence: Information claire sur l'utilisation des données des personnes.
3. Le droit des personnes: Il faut communiquer clairement sur le droit des personnes (rectification, suppression des données etc.).
4. Durée de conservation: Il faut préciser clairement la durée de conservation légale des données recueillies.
5. Sécurité des données: Mesures prises pour préserver la sécurité des données.

Source: CNIL

Conclusion



- Après analyse, l'application qui permet de prédire le nutrition score et le nutrition grade est **faisable** et **pertinente**.
- 1ère observation: Elle est **techniquement faisable** puisque nous avons observé des relations entre les nutriments sélectionnés et les cibles.
- 2ème observation: En effectuant une ACP les **données restent pertinentes** avec un maximum d'information conservée.
- 3ème observation: Les modèles retenus sont **performants** et minimisent la marge d'erreur.

Merci

Antonine Laroze-Cervetti