Projet M1 Data Science en Santé

Pour M. Zitouni

2026

Analyse du Nutri-Score et des Nutriments des Produits Alimentaires en France

À partir des données de la Open Food Facts (OFF)

Tableau de bord

Git

Selma El Fagrouchi

Qais Kazimi

Tiziri Dridi

Mehdi Bououdina

Said Mahmah

Sommaire :

Analyse du Nutri-Score et des Nutriments des Produits Alimentaires en France.....1
I. Introduction.....3
II. Présentation des données.....4
a. Nettoyage, préparation et transformation des données.....4
b. Processus de catégorisation sémantique.....5
III. Méthodologie.....7
a. Page d'analyse du Nutriscore.....7
b. Page d'analyse des nutriments.....8
c. Page du modèle de prédiction du Nutri-Score.....9
IV. Résultats et discussion.....11
a. Page d'analyse du Nutriscore.....11
b. Page d'analyse des nutriments.....14
c. Page du modèle de prédiction du Nutri-Score.....20
V. Limites et difficultés techniques.....23
VI. Déploiement et reproductibilité.....24
VII. Conclusion générale.....25
Références.....26
Instructions.....27
Étapes d'utilisation :.....27
Étapes d'utilisation :.....27
Étapes d'utilisation :.....28

I. Introduction

L'alimentation constitue aujourd'hui l'un des déterminants majeurs de la santé publique. Une consommation excessive de sucre, de sel et de graisses saturées est reconnue comme un facteur de risque de développement de maladies chroniques telles que le diabète de type 2, l'obésité et les maladies cardiovasculaires, qui représentent un enjeu de santé publique considérable en France et dans le monde (Afshin et al., 2019 ; World Health Organization, 2020).

Pour aider les consommateurs à faire des choix alimentaires plus éclairés, le Nutri-Score a été instauré en France. Ce système classe les aliments sur une échelle allant de A à E, où A correspond à des produits alimentaires plutôt sain et E, l'inverse, plutôt mauvais pour la santé. Il prend en compte à la fois l'énergie et les nutriments défavorables, comme les sucres, les acides gras saturés et le sel, et les nutriments favorables, tels que les fibres, les protéines et la proportion de fruits et légumes (Julia & Hercberg, 2017).

Notre projet s'inscrit dans cette logique et vise à exploiter les données issues de la base Open Food Facts, qui rassemble des informations détaillées sur des millions de produits alimentaires vendus en France et à l'international. Cette base collaborative constitue une source précieuse pour analyser la composition nutritionnelle des produits, la distribution des nutriments et la pertinence du Nutri-Score dans le contexte des catégories alimentaires existantes.

La question centrale que nous avons choisie de traiter est la suivante : **comment l'analyse des nutriments permet-elle d'expliquer les différences de Nutri-Score entre les produits alimentaires ?**

Nous avons fait ce choix en considérant le lien direct entre la consommation excessive de sucre (et autres nutriments) et l'apparition du surpoids et du diabète, ainsi que l'importance grandissante de la prévention en santé publique pour limiter les complications à long terme.

Notre objectif principal est de concevoir un outil accessible à un public non spécialiste, capable de présenter des résultats fiables, compréhensibles et exploitables, en combinant analyse exploratoire, visualisation interactive et modélisation prédictive.

II. Présentation des données

Les données que nous avons exploitées proviennent de la base Open Food Facts, qui recense des produits alimentaires commercialisés dans le monde entier. Le volume initial était particulièrement important, comprenant plus de 1,2 million de produits en France et environ 250 variables, pour un poids total d'environ 12 gigaoctets. En raison de l'ampleur de ce jeu de données, le nettoyage a été réalisé sur Google Colab afin de faciliter la manipulation et le traitement des informations.

Par ailleurs, afin que chaque membre du groupe puisse travailler en local et pour simplifier le chargement du dashboard, nous avons généré plusieurs fichiers de données distincts. Chaque fichier était dédié à un objectif précis :

- l'analyse du Nutri-Score, contenant les colonnes du Nutri-Score et des catégories (Fichier1_Nutriscore_Categories.csv) ;
- l'analyse des nutriments, incluant les colonnes du Nutri-Score, des catégories et les taux des différents nutriments (Fichier4.csv);
- l'analyse des produits dont le Nutri-Score était inconnu, conservant la version non nettoyée du Nutri-Score (Fichier_3.csv)
- un fichier destiné à la modélisation prédictive, comprenant les colonnes du Nutri-Score et les taux des différents nutriments (Fichier4_.csv).

Dans un premier temps, nous avons cherché à nous familiariser avec les données à l'aide du logiciel R. Cela nous a permis de produire différents graphiques exploratoires et de constater que de nombreuses valeurs étaient manquantes, notamment pour le Nutri-Score et certains nutriments essentiels. L'utilisation de boxplots a ensuite révélé la présence de valeurs aberrantes, comme des taux de nutriments dépassant 100 grammes pour 100 grammes de produit, indiquant très probablement des erreurs de saisie. Ces constats ont montré qu'un travail approfondi de nettoyage et de structuration des données était indispensable avant toute analyse fiable.

a. Nettoyage, préparation et transformation des données

Pour rendre les données exploitables, nous avons procédé à plusieurs étapes de prétraitement. Nous avons commencé par convertir les variables numériques de manière sécurisée à l'aide de la fonction `pd.to_numeric`, en gérant les erreurs pour éviter les problèmes de conversion. Nous avons filtré les valeurs incohérentes, supprimé les lignes inexploitables et éliminé les valeurs de sucre inférieures à zéro ou supérieures à 100

grammes, ainsi que d'autres valeurs extrêmes qui pouvaient fausser les moyennes et les indicateurs.

Pour les analyses descriptives, nous n'avons pas appliqué d'imputation systématique afin de ne pas introduire de biais. En revanche, pour la modélisation prédictive, nous avons effectué une imputation médiane afin de traiter les valeurs manquantes et permettre au modèle de s'entraîner sur des données complètes. Cette étape a permis de constituer une base fiable, adaptée à la visualisation, aux calculs statistiques et à l'apprentissage automatique.

b. Processus de catégorisation sémantique

Pour les catégories déjà présentes dans la base de données, elles sont très hétérogènes et souvent trop détaillées (environ 200 catégories différentes), avec des libellés à la fois en français et en anglais. Pour rendre les analyses comparables, nous avons ainsi développé nos propres catégories, basées sur un dictionnaire de mots-clés que nous avons appliqué de manière itérative.

Lors de la première version, environ 60 % des produits étaient classés dans la catégorie « Autres », ce qui rendait les résultats peu exploitables. Pour améliorer cela, nous avons identifié les 20 catégories les plus fréquentes au sein de la catégorie « Autres », enrichi le dictionnaire de mots-clés avec ces termes, puis relancé l'algorithme. Nous avons répété ce processus huit fois, en analysant les résultats à chaque cycle et en ajustant les règles afin de réduire progressivement le nombre de produits non classés.

Ainsi, nous avons créé une fonction personnalisée, **clean_category**, permettant de normaliser les libellés, de convertir les textes en minuscules et d'attribuer une catégorie lorsqu'une correspondance était trouvée. L'utilisation de la méthode **apply()** de Pandas a permis d'appliquer cette transformation à l'ensemble du DataFrame. Grâce à cette démarche, la majorité des produits ont été répartis dans des familles cohérentes telles que boissons, viandes, produits laitiers, féculents et produits sucrés, améliorant la lisibilité et la pertinence des analyses. Un total de 18 catégories à été crée :

- **Animaux** : regroupe les produits destinés à l'alimentation ou aux soins des animaux domestiques, tels que les croquettes, pâtées, friandises et autres aliments pour chiens, chats et animaux de compagnie.
- **Boissons** : regroupe l'ensemble des boissons chaudes et froides, incluant les eaux, jus de fruits, sodas, thés, cafés, nectars, sirops, kombuchas et autres boissons prêtes à consommer.
- **Viande & poisson** : regroupe les produits à base de viande et de produits de la mer, comme le bœuf, le porc, le poulet, la dinde, les charcuteries, le saumon, le thon et autres poissons transformés ou frais.
- **Produits laitiers** : regroupe les aliments issus du lait et assimilés, tels que les fromages, yaourts, crèmes, beurres, laits ainsi que les œufs, souvent associés aux produits laitiers dans les bases de données.

- **Plats préparés** : regroupe les plats prêts à consommer ou nécessitant une simple réchauffe, comme les pizzas, lasagnes, soupes, sandwichs, burgers et plats surgelés.
- **Snacks sucrés** : regroupe les produits sucrés destinés au grignotage, notamment les biscuits, gâteaux, chocolats, bonbons, gaufres et autres confiseries.
- **Boulangerie** : regroupe les produits issus de la boulangerie et de la pâtisserie simple, tels que le pain, les toasts, les brioches, les crêpes et produits assimilés.
- **Fruits & légumes** : regroupe les fruits et légumes frais, transformés ou préparés, comme les salades, tomates, carottes, fruits découpés ou en conserve.
- **Féculents** : regroupe les aliments riches en glucides complexes servant de base aux repas, notamment les pâtes, le riz, les nouilles, les lentilles et autres légumineuses.
- **Sauces & tartinables** : regroupe les sauces d'accompagnement et produits à tartiner, tels que le ketchup, la mayonnaise, la moutarde, le houmous et diverses sauces culinaires.
- **Confitures & miel** : regroupe les produits sucrés d'origine fruitière ou apicole, comme les confitures, gelées et miels utilisés au petit-déjeuner ou en pâtisserie.
- **Fruits secs & graines** : regroupe les fruits à coque, oléagineux et graines, tels que les amandes, noix, noisettes, cajous et mélanges de graines.
- **Huiles & matières grasses** : regroupe les produits gras utilisés pour la cuisson ou l'assaisonnement, comme les huiles végétales, l'huile d'olive, la margarine et autres matières grasses.
- **Céréales & petit-déjeuner** : regroupe les produits consommés principalement au petit-déjeuner, tels que les céréales, mueslis, granolas et préparations similaires.
- **Produits végétaux** : regroupe les alternatives végétales et produits destinés aux régimes végétariens ou véganes, comme le tofu, le soja et les substituts de viande ou de produits laitiers.
- **Desserts & glaces** : regroupe les produits sucrés de fin de repas, incluant les crèmes desserts, puddings, sorbets, glaces et desserts préparés.
- **Épicerie** : regroupe les produits de base et ingrédients culinaires, tels que le sel, le sucre, la farine, les épices et autres produits secs utilisés en cuisine.
- **Autres :** regroupe les produits qui n'ont pas pu être classés automatiquement dans les catégories définies, soit en raison de libellés incomplets, ambigus ou trop spécifiques, soit à cause de l'absence de mots-clés correspondants dans le dictionnaire.

III. Méthodologie

Notre approche a été progressive. Nous avons commencé par explorer les principales variables comme le Nutri-Score, le sucre et les catégories principales des produits, avec des graphiques simples pour visualiser les données. Au départ, nous avions testé des visualisations plus complexes (violin plot…), mais elles étaient difficiles à comprendre pour un public non spécialiste. Nous avons donc choisi des indicateurs simples et des graphiques clairs pour montrer la répartition des nutriments et du Nutri-Score.

Après le nettoyage des données, nous nous sommes organisés pour travailler chacun sur un dashboard. Initialement, nous avions créé trois dashboards séparés, mais nous avons décidé de les fusionner en un seul dashboard multi-pages. Chaque page correspond à un thème précis : le Nutri-Score, les nutriments et la prédiction des données manquantes. Cette organisation en trois pages rend la navigation plus facile et évite de scroller.

a. Page d'analyse du Nutriscore

La première page a été conçue pour donner une vue globale sur le Nutri-Score des produits. Nous avons cherché à montrer des indicateurs représentatifs tout en permettant une exploration plus fine par catégorie. Pour cela, nous avons utilisé la librairie pandas, qui nous a permis de réaliser des agrégations et des calculs statistiques avec des fonctions telles que `groupby()`, `value_counts()`, `count()` et `sum()`. Les visualisations interactives ont été créées avec Plotly Express et Plotly Graph Objects, en utilisant des fonctions comme `px.histogram()`, `go.Figure()`, `add_trace()` et `update_layout()` pour personnaliser les titres, les axes et les couleurs. Nous avons également exploité Dash pour rendre l'interface interactive, en important notamment `dcc.Dropdown`, `dcc.Graph`, `html.Div` et en utilisant les callbacks via `@app.callback` pour mettre à jour dynamiquement les graphiques selon les sélections des catégories effectuées par l'utilisateur.

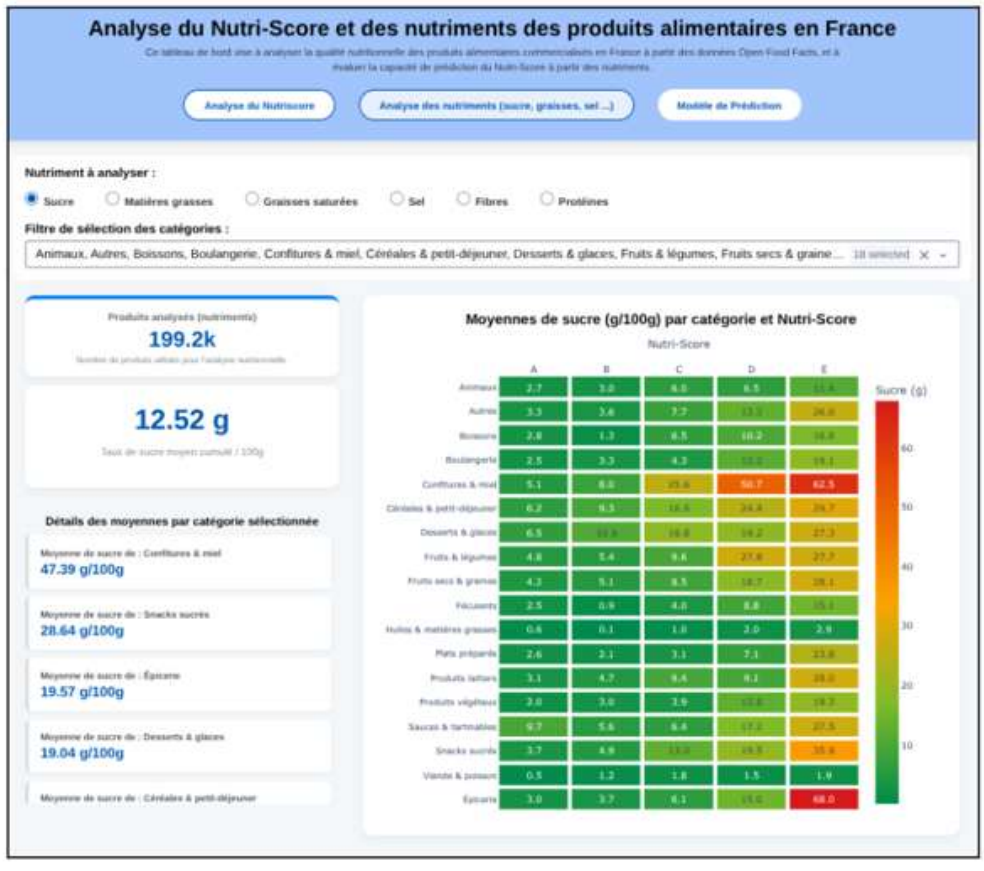
Dans cette page, nous avons mis des indicateurs clés tels que le nombre total de produits, la répartition des Nutri-Scores de A à E et les proportions de produits jugés sains ou moins sains. Nous avons intégré un menu déroulant permettant de filtrer les produits par catégorie, les valeurs étant générées automatiquement à partir des données via `df['category'].unique()`. Les graphiques se mettent à jour en temps réel grâce aux callbacks de Dash utilisant `Input()` et `Output()`. De plus, pour rendre la visualisation intuitive, nous avons appliqué un code couleur respectant les standards du Nutri-Score : vert pour A, jaune pour B, orange pour C, rouge clair pour D et rouge foncé pour E.



b. Page d'analyse des nutriments

La deuxième page a été pensée pour explorer plus en détail les nutriments principaux et leur impact sur le Nutri-Score. Nous avons cherché à visualiser la distribution des valeurs nutritionnelles par catégorie et à détecter rapidement des tendances ou des anomalies. Nous avons utilisé **pandas** pour l'agrégation et la préparation des données, avec des fonctions telles que `groupby()`, `pivot_table()`, `mean()` et `sort_values()` afin de calculer les moyennes par catégorie. Pour les visualisations interactives, nous avons de nouveau exploité **Plotly Express**, notamment les fonctions `px.box()` pour représenter la répartition d'un nutriment, `px.bar()` pour visualiser les moyennes par catégorie, et `px.imshow()` pour créer des heatmaps croisées entre catégories et Nutri-Scores. L'interactivité a été renforcée avec **Dash**, en utilisant `dcc.RadioItems` pour sélectionner le nutriment d'intérêt et `dcc.Graph` pour afficher le graphique correspondant. Les callbacks ont été programmés pour mettre à jour automatiquement les visualisations en fonction de la sélection de l'utilisateur.

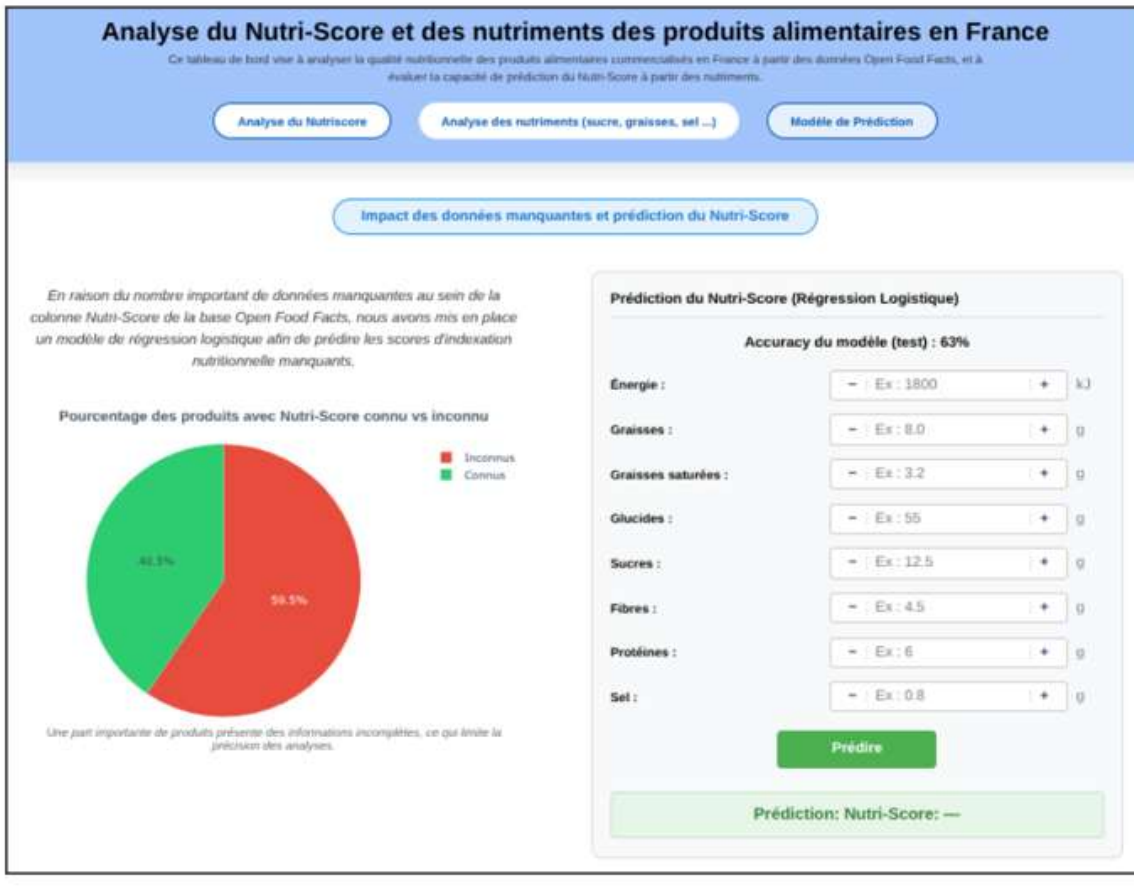
Cette page permet à l'utilisateur de choisir un nutriment spécifique tel que le sucre, les matières grasses, le sel, les fibres ou les protéines, et de visualiser immédiatement les statistiques correspondantes. La heatmap illustre la moyenne de ce nutriment pour chaque catégorie en fonction du Nutri-Score, permettant ainsi de repérer rapidement quels nutriments sont problématiques dans certaines catégories. Des filtres dynamiques ont également été mis en place afin de combiner catégories et nutriments dans une exploration visuelle unique, et des couleurs codées ont été appliquées pour rendre l'interprétation intuitive et immédiate.



c. Page du modèle de prédiction du Nutri-Score

La troisième page a été dédiée à la modélisation prédictive, avec pour objectif de compléter les valeurs manquantes de Nutri-Score et de démontrer à l'utilisateur comment une prédiction peut être effectuée en temps réel. Pour ce faire, nous avons utilisé la librairie **scikit-learn**, en mettant en œuvre un modèle de `LogisticRegression()` pour la classification multi-classes, et en séparant les données en ensembles d'entraînement et de test via `train_test_split()`. Nous avons appliqué `StandardScaler()` pour normaliser les valeurs nutritionnelles et `SimpleImputer()` pour gérer les données manquantes. Les fonctions `fit()`, `predict()` et `score()` ont permis d'entraîner le modèle et d'évaluer sa performance globale. Pour l'interface utilisateur, nous avons utilisé **Dash** avec des composants tels que `dcc.Input`, `html.Button` et `dcc.Graph`, ainsi que des callbacks permettant de calculer et d'afficher la prédiction en temps réel.

Sur cette page, toutes les colonnes nutritionnelles pertinentes ont été utilisées comme variables explicatives, incluant l'énergie, les matières grasses, les graisses saturées, les glucides, les sucres, les fibres, les protéines et le sel. Le modèle a été entraîné sur 80 % des données et testé sur 20 %, avec un paramètre `class_weight='balanced'` afin de corriger le déséquilibre des classes. L'utilisateur peut saisir les valeurs nutritionnelles d'un produit et obtenir instantanément la prédiction de son Nutri-Score. Nous avons également affiché des indicateurs de performance tels que la précision globale du modèle et la distribution des prédictions, tout en comparant les valeurs prédites aux valeurs réelles lorsque disponibles. Enfin, des filtres ont été intégrés pour analyser les prédictions par catégorie et identifier les cas où certaines catégories sont mieux ou moins bien prédites.



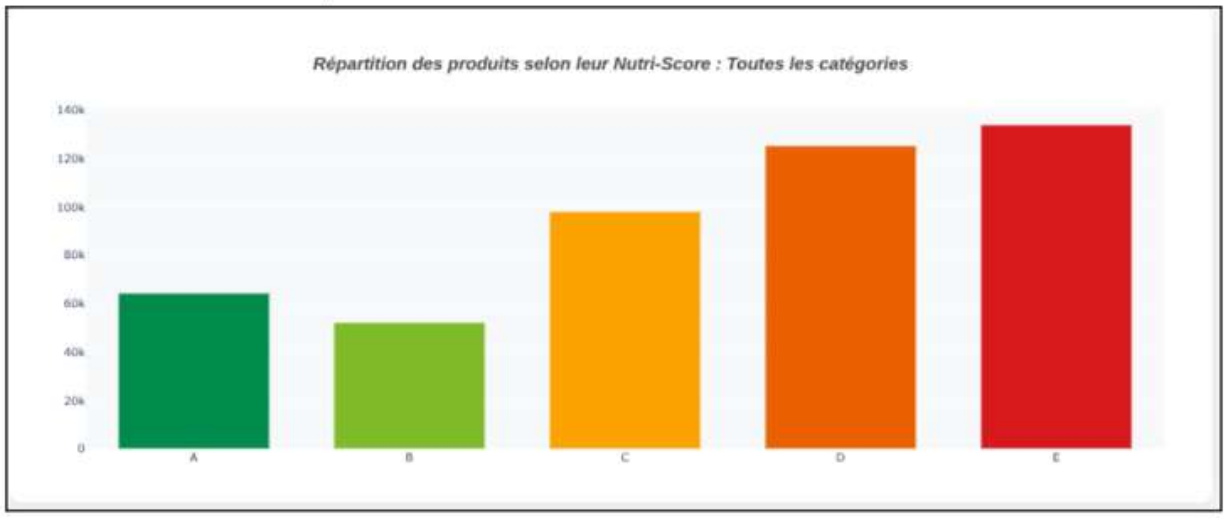
L'ensemble du dashboard repose sur la fusion de ces trois pages dans une interface unique, permettant une navigation intuitive et une exploration complète des données. Nous avons intégré une navigation simplifiée via des boutons, une interactivité dynamique grâce aux callbacks Dash et aux composants **Input** et **Output**, ainsi que des visualisations interactives avec Plotly permettant le zoom, la sélection et l'export des données. Les bibliothèques principales que nous avons utilisées incluent **pandas** et **numpy** pour la manipulation et le traitement des données, **dash** pour l'interface et l'interactivité, **plotly.express** et **plotly.graph_objects** pour les graphiques, **scikit-learn** pour la modélisation prédictive, et **joblib** pour sauvegarder et charger le modèle entraîné.

IV. Résultats et discussion

Après avoir présenté la méthodologie et la construction de notre dashboard, nous nous intéressons maintenant aux résultats obtenus à partir de notre outil. Les différentes visualisations nous ont permis d'obtenir de nombreuses informations sur le Nutri-Score et sur les produits alimentaires en France. Dans cette partie, nous cherchons donc à analyser ces données afin de mieux comprendre ce qu'elles révèlent, de répondre à notre problématique et d'identifier les principales tendances observées.

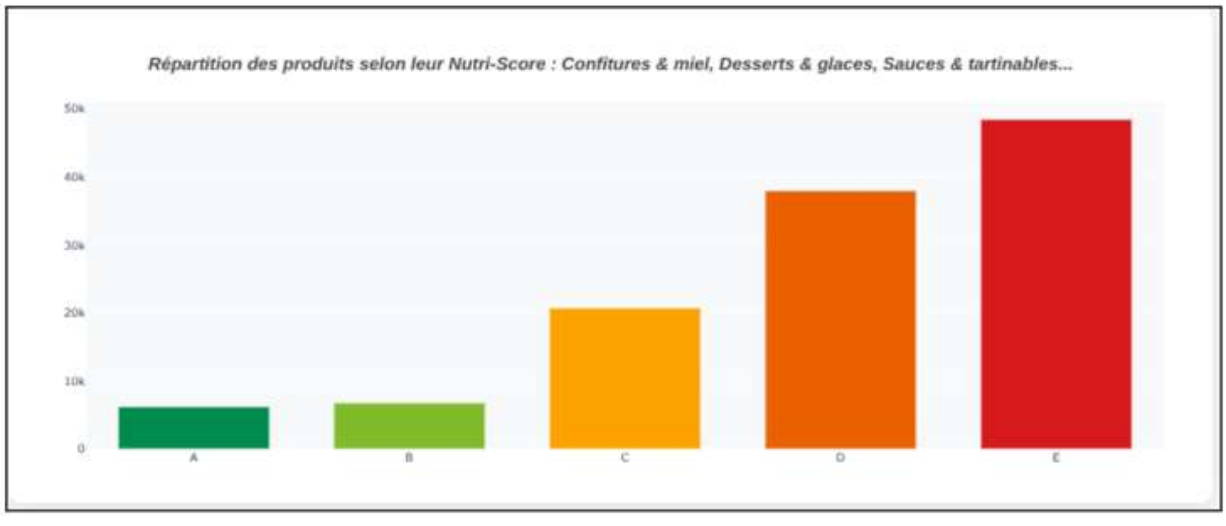
a. Page d'analyse du Nutriscore

Tout d'abord, nous avons analysé la situation globale des produits alimentaires référencés dans la base Open Food Facts en France. Sur les 473 300 produits disposant d'un Nutri-Score renseigné, seuls 24,58 % obtiennent le score A ou B, tandis que 54,7 % sont obtiennent D ou E. Cela met en évidence un déséquilibre important en faveur des produits les moins bien notés. En effet, plus on avance dans le score, plus le nombre de produits à l'air d'augmenter avec un pic particulièrement marqué pour le score E, dépassant les 130 000 produits. Nous constatons ainsi que le marché alimentaire est largement dominé par des produits de faible qualité nutritionnelle, ce qui souligne l'utilité du Nutri-Score comme outil d'aide à la décision pour les consommateurs.

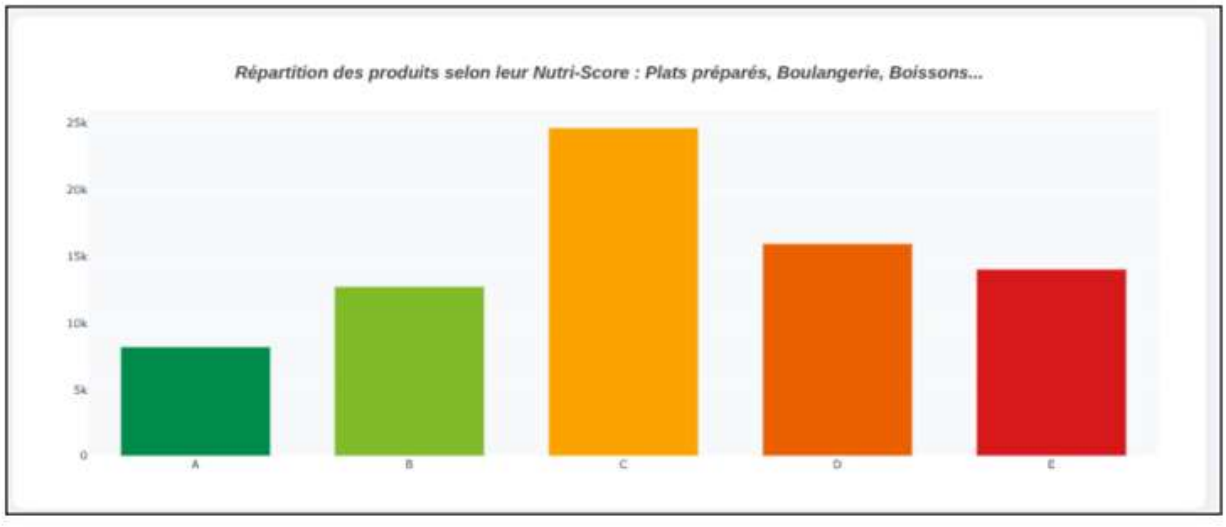


En analysant les répartitions par catégories, nous avons pu diviser nos résultats en trois grandes tendances :

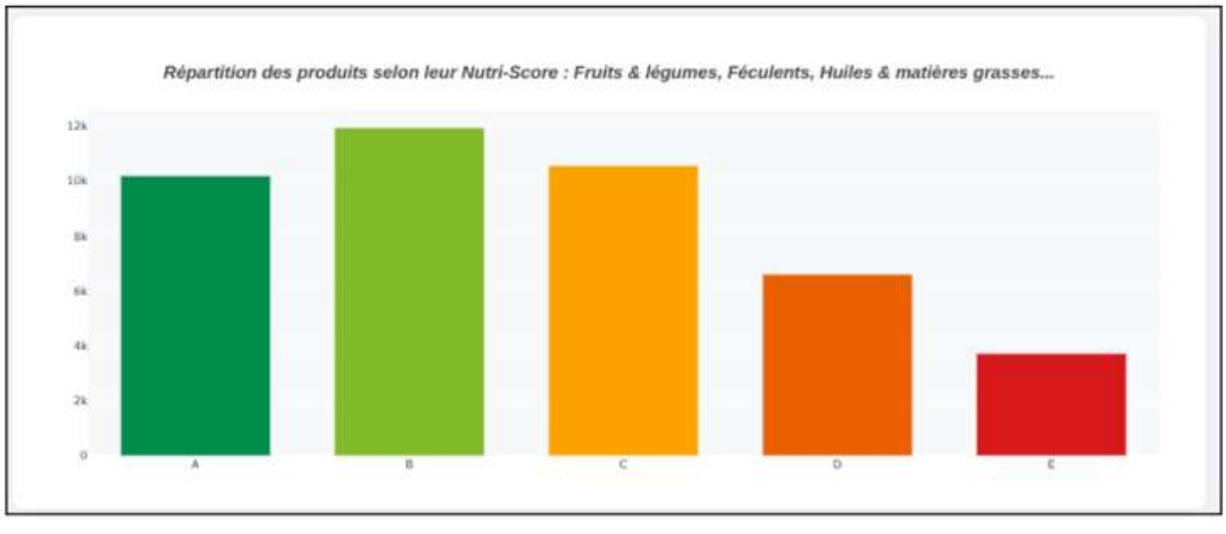
- Une première tendance concerne les catégories de produits **plutôt malsains**, avec une surreprésentation des scores D et E. C'est notamment le cas des snacks sucrés, des sauces et tartinables, des confitures et du miel, mais aussi d'une partie des produits laitiers et des desserts, ce qui représentent 120k produits. Dans ces catégories, les proportions de produits mal notés dépassent en moyenne les 70 %, et atteignent même plus de 90 % pour les snacks sucrés. Cela s'explique principalement par la forte présence de sucres ajoutés, de graisses saturées et de sel. Les confitures, par exemple, sont composées en grande partie de sucre, tandiS que les sauces industrielles cumulent souvent sucre et matières grasses. De même, les snacks sucrés combinent généralement plusieurs facteurs négatifs, sans réel apport nutritionnel compensateur. Dans ces familles de produits, le Nutri-Score joue pleinement son rôle en signalant clairement les aliments à consommer avec modération.



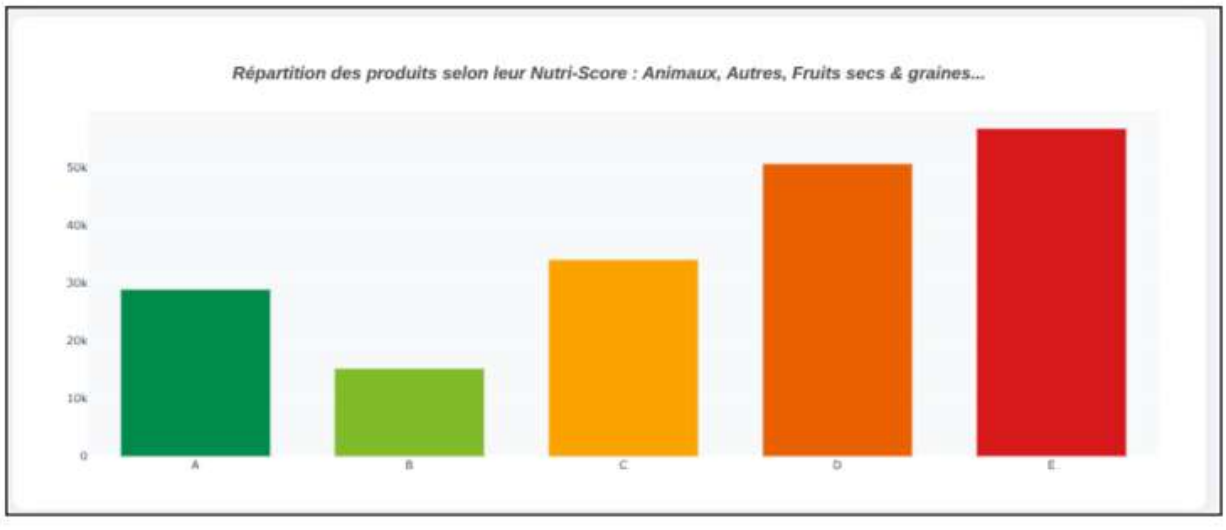
- Une deuxième tendance concerne les catégories plutôt **intermédiaires**, dans lesquelles le score C est dominant. Nous retrouvons dans ce groupe les plats préparés, la boulangerie, les boissons, les céréales du petit-déjeuner et l'autre partie des desserts avec un total de 75k produits. Les plats préparés, par exemple, peuvent contenir des légumes ou des protéines, mais restent riches en sel et en additifs. De même, certaines céréales apportent des fibres, mais sont fortement sucrées. Dans le cas des boissons, la coexistence de l'eau, des jus et des sodas explique sûrement ce score C. Ces résultats montrent que ces produits correspondent souvent à des choix de compromis, ni particulièrement sains, ni totalement défavorables.



- À l'opposé, certaines catégories se distinguent par une proportion élevée de produits **bien notés** (43k). Les fruits et légumes, les féculents, les produits végétaux et certaines huiles constituent les familles les plus favorables du point de vue nutritionnel. Dans ces catégories, les scores A et B sont majoritaires, atteignant parfois plus de 60 % pour les féculents. Ces bons résultats s'expliquent par une transformation limitée des produits, une richesse en fibres, ainsi qu'une faible teneur en sucres et en graisses saturées. Les huiles riches en acides gras insaturés, comme l'huile d'olive ou de colza, bénéficient également d'une meilleure valorisation depuis la mise à jour du Nutricore. Ces catégories représentent ainsi les piliers d'une alimentation plus équilibrée, et confirment l'intérêt du Nutri-Score pour orienter les choix quotidiens.



Certaines catégories présentent toutefois une forte hétérogénéité, ce qui rend leur interprétation plus complexe. C'est notamment le cas des fruits secs et graines, de l'épicerie, des produits destinés aux animaux et de la catégorie « Autres ». Les fruits secs illustrent bien cette variabilité, avec d'un côté des produits naturels bien notés, et de l'autre des versions salées ou grillées classées E. De même, l'épicerie regroupe à la fois des produits bruts comme les légumineuses, et des préparations industrielles très transformées. Les produits pour animaux, tels que les croquettes et pâtées, obéissent à des logiques nutritionnelles spécifiques, différentes de l'alimentation humaine, ce qui limite les comparaisons directes. Quant à la catégorie « Autres », elle rassemble des produits trop divers pour permettre une analyse fine. Ces tendances montrent ainsi les limites de notre catégorisation et soulignent l'importance du contexte dans l'interprétation des résultats.



b. Page d'analyse des nutriments

Pour mieux comprendre les impacts des différents nutriments sur le Nutri-Score, nous avons réalisé une analyse détaillée en nous basant sur les valeurs mesurées dans diverses catégories de produits. Afin de contextualiser ces résultats, nous avons utilisé les recommandations nutritionnelles de l'Organisation mondiale de la Santé (OMS - 2026), ce qui permet d'évaluer la qualité nutritionnelle relative des aliments par rapport aux apports journaliers recommandés.

Nutriment	Recommandation journalière OMS	pour 100 g (régime à 2000 kcal/jour)
Sucre	< 50 g/jour (idéal < 25 g)	< 5 g / 100 g (idéal < 2,5 g)
Matières grasses	20–35 % des apports énergétiques (~44–78 g/jour pour 2000 kcal)	22–39 g / 100 g
Gras saturés	< 10 % des apports énergétiques (~22 g/jour)	< 11 g / 100 g

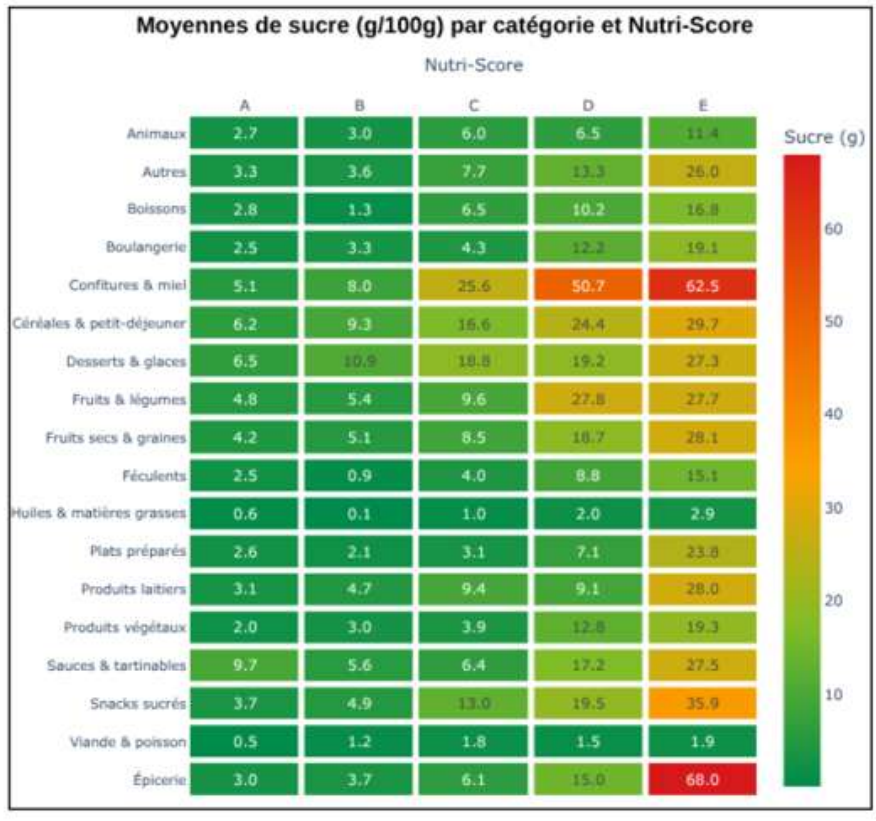
Sel	< 5 g/jour (soit < 2 g de sodium)	< 0,25 g / 100 g
Fibres	≥ 25 g/jour	≥ 2,5 g / 100 g
Protéines	10–15 % des apports énergétiques (~50–75 g/jour)	5–7,5 g / 100 g

- Sucre

La consommation de sucre constitue un facteur déterminant du Nutri-Score, avec des écarts très marqués selon les familles de produits. En nous basant sur les recommandations de l'OMS, qui suggèrent de limiter l'apport en sucres libres à moins de 25 g par jour pour un adulte, nous observons que de nombreuses catégories dépassent largement ce seuil.

La catégorie « Confitures & miel » affiche une moyenne record de 47,39 g pour 100 g, allant de 5,1 g pour les produits A à 62,5 g pour les produits E, montrant que le sucre constitue presque l'unique critère de classement. Les snacks sucrés et les desserts/glaces suivent une tendance similaire : les snacks présentent en moyenne 28,64 g, atteignant 36 g pour les produits E, tandis que les desserts et glaces passent de 6,5 g pour les A à plus de 27 g pour les E. Les céréales du petit-déjeuner illustrent également cette dynamique, avec 6,4 g pour les produits A contre plus de 27 g pour les produits E.

Même dans des catégories comme les sauces, le sucre joue un rôle central, avec une moyenne de 15,73 g et des pics à plus de 27 g pour les produits E. Globalement, dès que le seuil de 15 à 20 g de sucre est dépassé, l'accès aux scores A et B devient très difficile, confirmant que le respect des recommandations OMS est un facteur clé pour obtenir un Nutri-Score favorable.



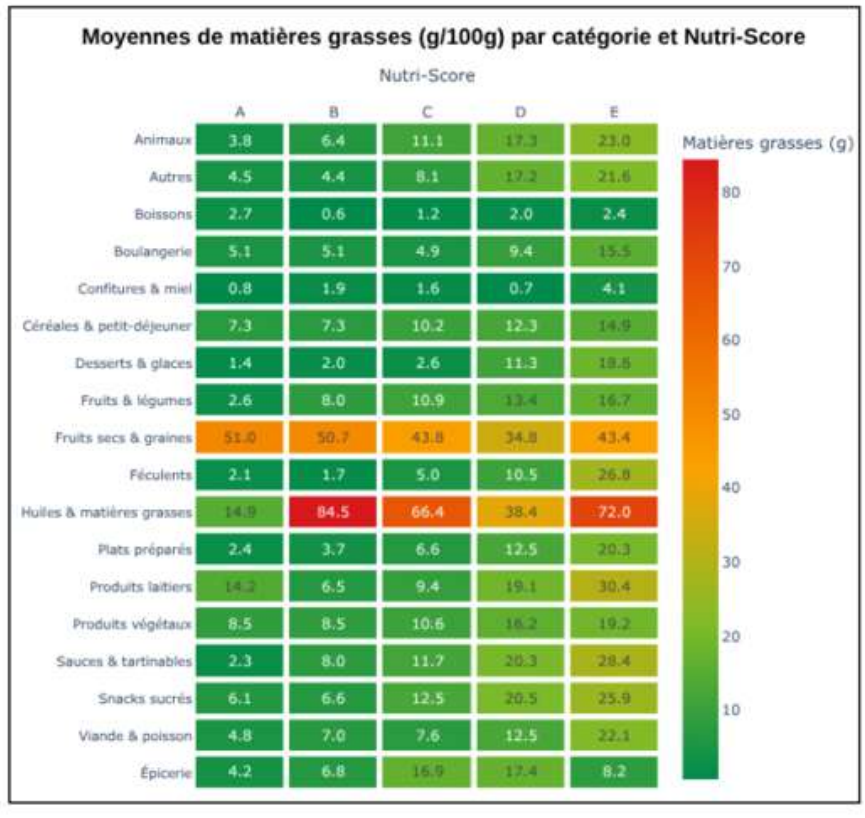
- Matières grasses

Les taux de matières grasses jouent également un rôle dans la détermination du Nutri-Score, mais leur impact dépend fortement de leur nature et de leur association avec les autres nutriments. Selon les recommandations de l'OMS, les lipides doivent représenter une part modérée des apports énergétiques quotidiens, ce qui implique une consommation raisonnée et privilégiant les graisses de bonne qualité.

Les huiles et matières grasses constituent logiquement la catégorie la plus riche en matières grasses, avec une moyenne de 67,32 g pour 100 g. Les produits classés B, majoritairement composés d'huiles végétales, atteignent même 84,5 g, tandis que certains produits classés E descendent à 72 g. Cette configuration montre que le Nutri-Score ne pénalise pas automatiquement une forte teneur en lipides, mais prend en compte d'autres nutriments également. Les fruits secs et graines présentent également une teneur élevée, avec une moyenne de 45,06 g, relativement stable selon les scores, ce qui confirme que la quantité seule n'est pas déterminante.

Dans les catégories plus transformées, comme les snacks sucrés et les produits laitiers, les matières grasses atteignent respectivement 22,68 g et 17,55 g en moyenne. Dans ces cas, leur association avec le sucre ou les graisses saturées renforce leur effet négatif sur le score. Les produits destinés aux animaux affichent quant à eux une moyenne de 16,37 g, pouvant atteindre 23 g pour les scores E, traduisant une forte densité énergétique.

Globalement, les matières grasses ne sont pas systématiquement pénalisantes. Leur impact dépend avant tout de leur composition et de leur interaction avec les autres nutriments, ce qui confirme l'approche qualitative adoptée par le Nutri-Score.



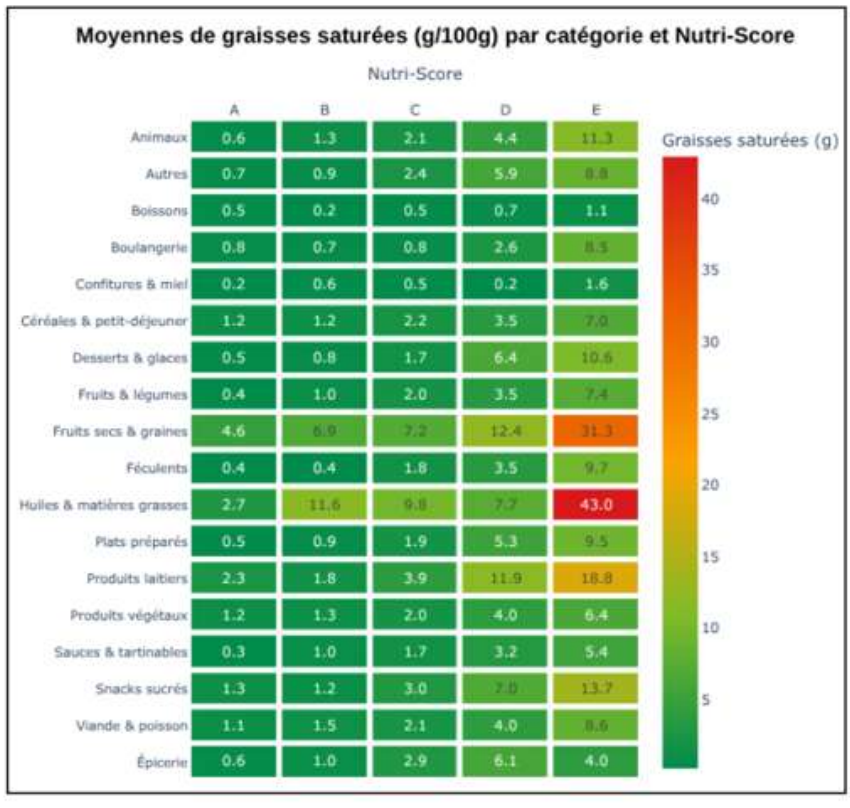
- Graisses saturées

Contrairement aux matières grasses, les graisses saturées constituent l'un des nutriments les plus pénalisants dans le calcul du Nutri-Score. L'OMS recommande d'en limiter fortement la consommation en raison de leur lien avec les maladies cardiovasculaires, ce qui explique leur poids important dans l'évaluation nutritionnelle.

Dans la catégorie des huiles, nous observons un contraste très marqué entre les produits A, qui contiennent en moyenne 2,7 g de graisses saturées, et les produits E, qui atteignent jusqu'à 43 g. Cette différence illustre clairement la prise en compte de la qualité des lipides par l'algorithme en charge du Nutriscore. Les fruits secs et graines présentent également des écarts, notamment lorsque certains produits riches en noix de coco ou en mélanges transformés apparaissent dans les scores les plus faibles.

Les snacks sucrés et les desserts/glaces affichent en moyenne 10,61 g et plus de 10 g de graisses saturées pour les produits E, contribuant fortement à leur mauvais classement. Dans les produits laitiers, la moyenne atteint 9,45 g et monte jusqu'à 15,2 g pour les scores E, malgré une teneur stable en protéines.

Ainsi, dans presque toutes les catégories, les graisses saturées apparaissent comme un facteur de déclassement majeur. Le dépassement du seuil de 11g constitue un signal fort, prioritaire dans l'évaluation du Nutri-Score.



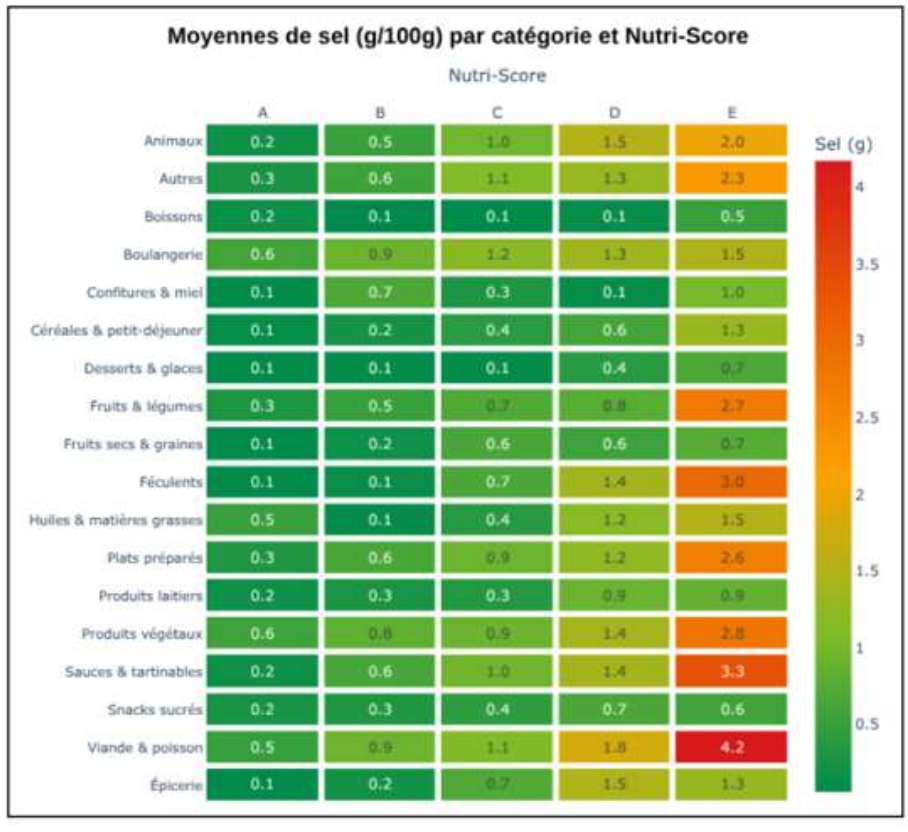
- Sel

Le sel quant à lui, est souvent sous-estimé par les consommateurs, mais fortement pris en compte par le Nutri-Score. L'OMS recommande de limiter l'apport quotidien à moins de 5 g, ce qui souligne l'importance d'une consommation modérée.

Les sauces et tartinables présentent les teneurs les plus élevées, avec une moyenne de 1,62 g pour 100 g et des pics à 3,3 g pour les produits classés E. La viande et le poisson, en particulier les charcuteries, affichent également des niveaux importants, avec une moyenne de 1,53 g et jusqu'à 4,2 g pour les scores E. Dans ces catégories, le déclassement est principalement lié à l'excès de sel.

La boulangerie constitue un autre point critique, avec une moyenne de 1,14 g, supérieure à celle des plats préparés (0,92 g). Nos données montrent que le sel est souvent le premier facteur de dégradation du score pour le pain et les produits assimilés. Les produits pour animaux présentent un écart particulièrement marqué, passant de 0,2 g en score A à 2 g en score E.

Dans l'ensemble, le sel apparaît comme un facteur déterminant dans les produits transformés. Il joue un rôle central dans le déclassement du Nutri-Score, souvent de manière peu visible pour le consommateur.



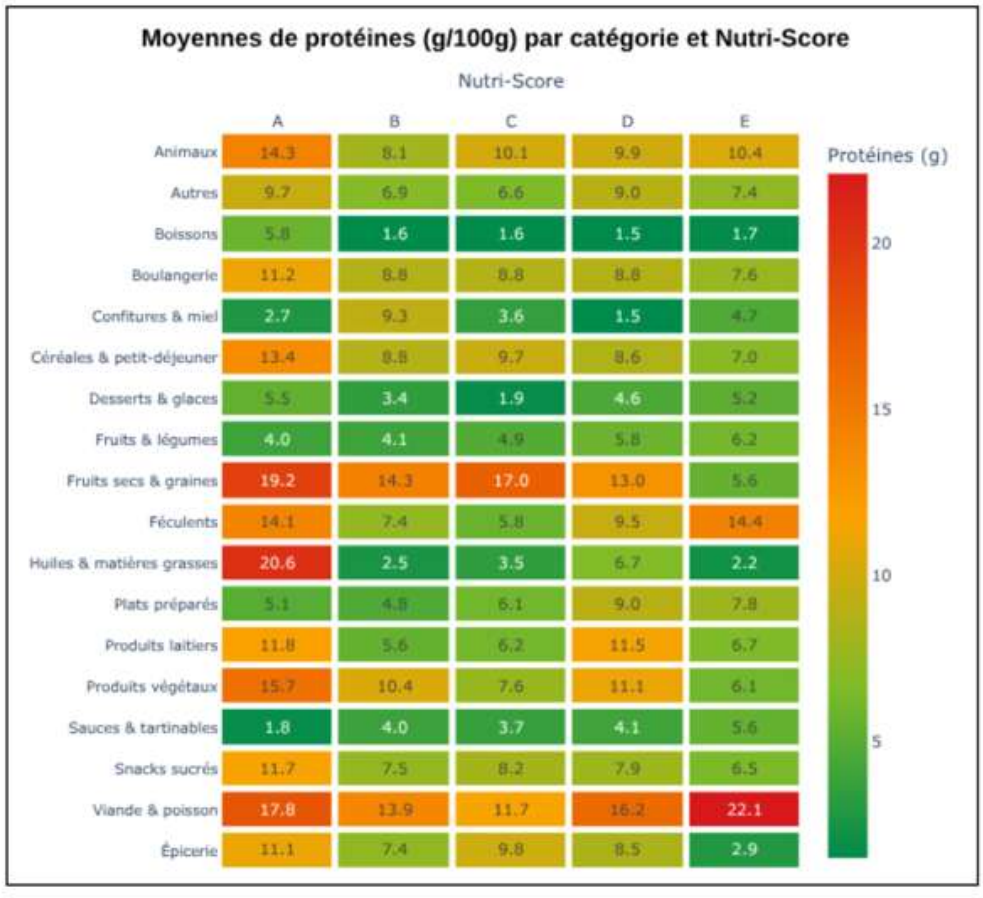
- Protéines

A l'inverse, les protéines constituent l'un des nutriments les plus valorisés par le Nutri-Score, en raison de leur rôle essentiel dans l'organisme. Selon l'OMS, un apport suffisant est nécessaire au bon fonctionnement du corps, mais il ne doit pas masquer un déséquilibre nutritionnel global.

La catégorie « Viande & poisson » domine largement avec une moyenne de 15,85 g pour 100 g, atteignant plus de 22 g pour les produits E. Les féculents et les produits végétaux affichent également des apports intéressants, avec respectivement 10,17 g et 11,21 g, contribuant à leurs bons scores. Les fruits secs et graines se distinguent aussi avec une moyenne de 14,71 g.

En revanche, dans les snacks sucrés et les desserts, le taux de protéines reste relativement stable, autour de 7 g, quel que soit le score. Les produits laitiers présentent une moyenne de 8,32 g, également constante. Dans ces catégories, les protéines ne suffisent pas à compenser des teneurs élevées en sucre, en sel ou en graisses saturées.

Nous constatons ainsi que les protéines constituent un atout important, mais qu'elles ne peuvent pas, à elles seules, sauver un produit nutritionnellement déséquilibré.



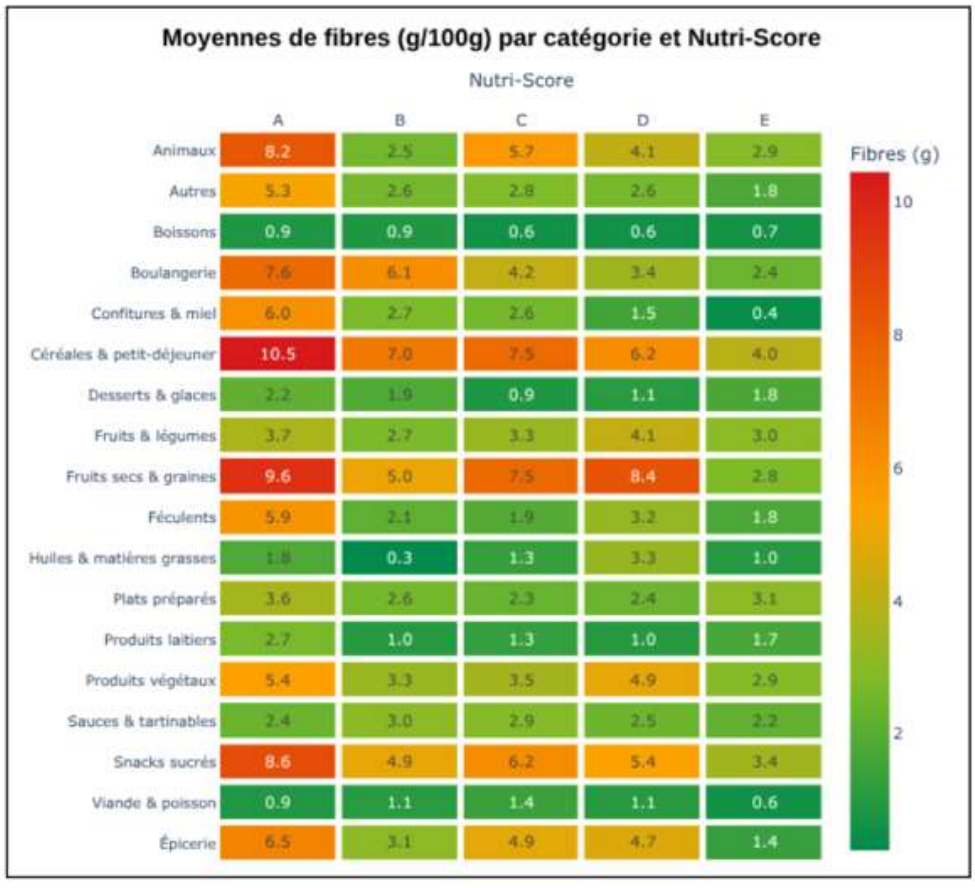
- Fibres

Les fibres jouent un rôle central dans la valorisation des produits peu transformés et équilibrés. L'OMS recommande un apport quotidien suffisant en fibres pour favoriser la digestion et réduire les risques de maladies chroniques.

Les céréales du petit-déjeuner affichent la moyenne la plus élevée, avec 7,40 g, suivies des fruits secs et graines avec 7,22 g, et des produits végétaux avec 4,26 g. Les féculents présentent une moyenne de 3,56 g, pouvant atteindre près de 6 g pour les produits classés A. Ces catégories illustrent l'intérêt des produits complets et peu raffinés.

Dans les céréales, l'écart entre les scores est particulièrement marqué : un produit A contient en moyenne 9,8 g de fibres, contre seulement 3,2 g pour un produit E. Dans les snacks sucrés, malgré une moyenne de 4,28 g, les produits restent majoritairement mal classés en raison d'un excès de sucre et de graisses. Dans les féculents, le passage du score A au score B correspond à une forte diminution du taux de fibres, liée au raffinage.

De manière générale, les fibres constituent un levier majeur d'amélioration du Nutri-Score, mais uniquement lorsqu'elles s'inscrivent dans un produit globalement équilibré.



Ainsi, l'analyse des six nutriments montre que le Nutri-Score repose sur un certain équilibre entre des apports nutritifs considérés favorables et défavorables. Le sucre, les graisses saturées et le sel apparaissent comme les principaux facteurs de déclassement, notamment dans les produits transformés comme les snacks, les desserts, les sauces et les plats préparés. Lorsque ces nutriments sont présents en grande quantité, l'accès aux meilleurs scores devient très difficile, même en présence de fibres ou de protéines.

Les matières grasses ont un impact plus nuancé, car leur effet dépend surtout de leur nature. Les huiles végétales et les fruits secs peuvent obtenir de bons scores malgré une forte teneur en lipides, tandis que les produits riches en graisses saturées sont davantage pénalisés. Les protéines et les fibres jouent un rôle positif, en particulier dans les produits végétaux, les céréales complètes et les féculents, mais leur capacité à compenser un déséquilibre reste limitée.

Globalement, le Nutri-Score valorise les produits simples, peu transformés et bien équilibrés. Il limite l'effet des nutriments positifs lorsque plusieurs nutriments défavorables sont présents simultanément, ce qui empêche qu'un produit déséquilibré soit artificiellement bien noté.

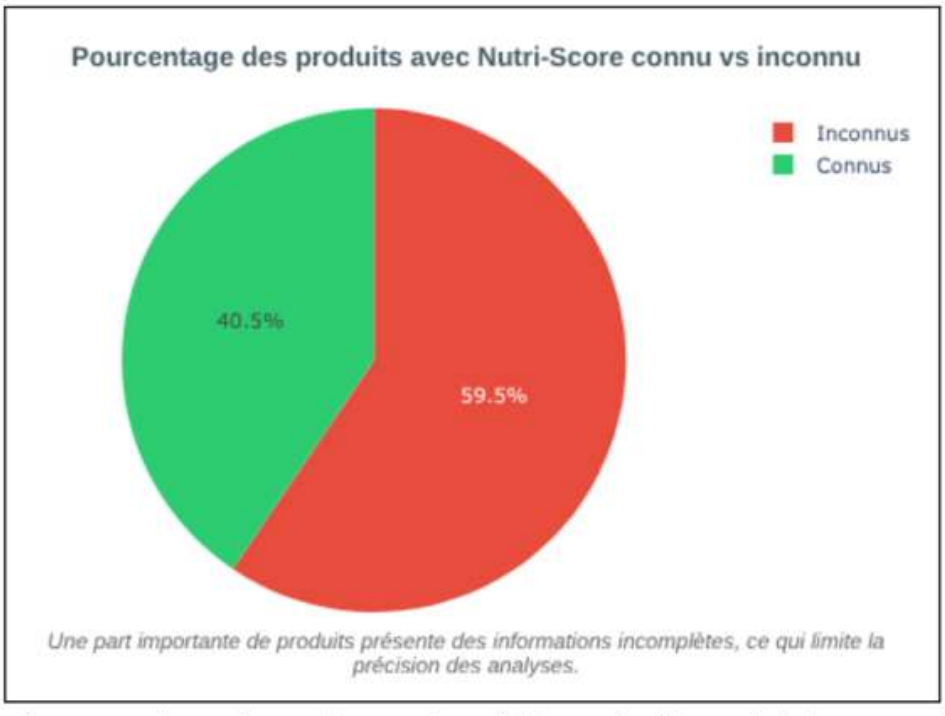
c. Page du modèle de prédiction du Nutri-Score

Au cours de notre étude, nous avons constaté que plus de la moitié des produits présents dans la base de données ne disposaient pas d'un Nutri-Score renseigné (environ 60%). Cette absence d'information limitait fortement nos possibilités d'analyse et risquait de biaiser certains résultats. Afin de pallier ce manque, nous avons donc choisi de développer notre propre modèle de prédiction du Nutri-Score, basé sur les caractéristiques nutritionnelles des produits.

Après entraînement et validation, notre modèle atteint un taux de fiabilité d'environ 63 %. Nous l'avons testé sur plusieurs catégories de produits, notamment les snacks, les produits laitiers, les céréales, les plats préparés et les produits végétaux. Les résultats montrent que le modèle fonctionne globalement de manière satisfaisante. Lorsqu'il se trompe, l'erreur reste généralement limitée à un seul niveau de score, par exemple en attribuant un B à la place d'un A, ou un C au lieu d'un D. Cela indique que le modèle parvient à capter correctement les grandes tendances nutritionnelles.

Cependant, une limite importante apparaît dans le cas des huiles et matières grasses. Le Nutri-Score ayant récemment évolué pour mieux prendre en compte la qualité des différentes huiles, notre modèle, a tendance à classer ces produits dans des catégories très défavorables, souvent en E, alors qu'ils peuvent désormais obtenir de meilleurs scores. Cette difficulté montre que notre modèle ne prend pas suffisamment en compte les spécificités de certaines catégories.

Ces résultats mettent en évidence les limites de notre approche. Le recours à un modèle unique pour l'ensemble des produits ne permet pas toujours de refléter correctement la diversité des profils nutritionnels. Pour améliorer la précision, une solution aurait été de développer des modèles de prédiction spécifiques par catégorie de produits, capables d'intégrer leurs particularités. Malgré ces limites, notre modèle reste un outil pertinent pour compléter les données manquantes et renforcer l'analyse globale du projet, tout en ouvrant des perspectives d'amélioration pour de futurs travaux.



Une part importante de produits présente des informations incomplètes, ce qui limite la précision des analyses.

De manière transversale, notre analyse met en évidence le rôle central du sucre et du sel dans la dégradation du Nutri-Score. Les catégories les plus pénalisées sont presque systématiquement celles où ces éléments sont fortement présents. À l'inverse, les produits peu transformés, riches en fibres et en nutriments naturels, obtiennent les meilleures notes. Le Nutri-Score apparaît ainsi globalement cohérent avec les connaissances nutritionnelles actuelles. Toutefois, certains mécanismes de compensation, observés notamment dans les desserts ou les boissons, peuvent parfois atténuer l'impact négatif du sucre grâce à la présence de protéines ou de minéraux.

Par ailleurs, notre dashboard montre que l'intérêt principal du Nutri-Score réside dans la comparaison à l'intérieur d'une même catégorie. Il est plus pertinent de comparer deux céréales, deux plats préparés ou deux yaourts entre eux, plutôt que de comparer des produits appartenant à des familles différentes. De plus, le calcul du score pour 100 grammes rappelle l'importance des portions, en particulier pour les catégories très pénalisées comme les sauces ou les snacks.

Enfin, certaines limites doivent être prises en compte dans l'interprétation des résultats. Le Nutri-Score ne renseigne pas sur le degré de transformation, la présence d'additifs ou l'origine des matières premières. Un produit bien noté peut donc rester discutable sur d'autres aspects. Néanmoins, notre analyse montre que, dans une grande majorité de cas, le Nutri-Score constitue un indicateur fiable de la qualité nutritionnelle globale.

En conclusion, notre dashboard met en évidence une segmentation claire du marché alimentaire. Certaines catégories apparaissent structurellement à risque, comme les snacks, les sauces et les produits très sucrés, tandis que d'autres offrent de réelles opportunités pour une alimentation plus équilibrée, notamment les fruits et légumes, les féculents et les produits végétaux. En regroupant les catégories selon leurs grandes tendances, notre travail permet de dépasser une simple lecture descriptive des données et d'apporter une compréhension globale des enjeux nutritionnels actuels.

V. Limites et difficultés techniques

Lors du développement du dashboard, nous avons rencontré plusieurs difficultés liées à la qualité et à l'organisation des données. Une grande partie des informations issues d'Open Food Facts était incomplète ou incohérente, notamment concernant les nutriments et le Nutri-Score. De nombreuses valeurs étaient manquantes ou mal renseignées, ce qui nous a obligés à effectuer un important travail de nettoyage, de vérification et de mise en forme avant de pouvoir les exploiter correctement.

Nous avons également rencontré des difficultés dans la catégorisation des produits. Les catégories n'étant pas toujours renseignées de manière claire et homogène, c'est pour cette raison que nous avons mis en place un système basé sur des mots-clés pour classer automatiquement les produits. Cette méthode fonctionne globalement bien, mais elle reste approximative et dépend fortement du nom des produits. Certains ont ainsi pu être mal classés ou regroupés dans des catégories trop larges, ce qui limite parfois la précision de nos analyses.

Le développement du modèle de prédiction a aussi représenté un défi important. La diversité des profils nutritionnels, la présence de données déséquilibrées et les valeurs manquantes ont limité ses performances. Malgré nos efforts pour améliorer le traitement des données, la fiabilité du modèle reste modérée. De plus, nous avons constaté qu'il prenait mal en compte certaines catégories, notamment les huiles, dont la classification a évolué, pour ne plus seulement prendre en compte ses nutriments mais également sa nature.

Enfin, la conception de l'interface interactive a demandé une organisation rigoureuse du code. Nous avons dû gérer de nombreux filtres, graphiques et mises à jour en temps réel tout en assurant la stabilité de l'application. Certaines fonctionnalités ont été simplifiées afin de garantir de bonnes performances, notamment lors du traitement de grandes quantités de données.

Dans l'ensemble, ces difficultés illustrent les contraintes liées à l'exploitation de données ouvertes, hétérogènes et partiellement incomplètes dans un projet de data analyse. Elles mettent également en évidence plusieurs pistes d'amélioration. À l'avenir, le projet pourrait être renforcé par l'utilisation de modèles de prédiction plus complexes et plus performants, tels que les Random Forest ou les méthodes de Gradient Boosting, afin d'améliorer la précision des résultats. L'intégration de techniques de traitement automatique du langage naturel (NLP) permettrait également d'affiner la catégorisation des produits à partir de leurs descriptions. De plus, l'ajout d'analyses temporelles offrirait la possibilité d'étudier l'évolution des produits, de leurs compositions et de leurs scores nutritionnels au fil du temps.

Le déploiement du dashboard à plus grande échelle sur le web permettrait également de toucher un public plus large. Enfin, le développement d'un système de recommandation nutritionnelle personnalisé, basé sur les profils des utilisateurs et leurs habitudes de consommation, constituerait une extension pertinente du projet.

Malgré ces limites, nous avons réussi à concevoir un dashboard fonctionnel, interactif et cohérent, capable de soutenir efficacement les objectifs initiaux du projet. Ce travail constitue ainsi une base solide pour de futures améliorations et pour le développement d'outils d'aide à la décision en matière de nutrition.

VI. Déploiement et reproductibilité

Nous avons publié l'application en ligne afin de garantir son accessibilité. Le code source est disponible sur GitHub, avec des instructions détaillées pour reproduire l'ensemble des analyses et visualisations. L'application interactive a été déployée sur la plateforme Render, permettant une consultation publique sans installation locale.

Les liens sont les suivants :

- GitHub : https://github.com/qaiskazimi1-tech/Dashboard-d-analyse-des-indicateurs-nutritionnels-en-France
- Application Render : https://dashboard-d-analyse-des-indicateurs.onrender.com

Le déploiement a nécessité la configuration d'un environnement Python approprié, l'installation des dépendances via `requirements.txt`, l'exposition du serveur Dash via l'objet `server = app.server`, et l'adaptation de la commande de démarrage pour un environnement de production. Cette mise en ligne démontre notre capacité à transformer une analyse locale en application web fonctionnelle et accessible à tous.

VII. Conclusion générale

Pour finir, nous pouvons rappeler notre problématique, qui était la suivante : **comment l'analyse des nutriments permet-elle d'expliquer les différences de Nutri-Score entre les produits alimentaires ?** À travers ce projet, nous avons cherché à comprendre dans quelle mesure la composition nutritionnelle des produits influence leur classement, en nous appuyant sur l'exploitation de données issues d'Open Food Facts, sur des visualisations interactives et sur un modèle de prédiction.

Nos résultats montrent que les écarts de Nutri-Score s'expliquent en grande partie par la présence de nutriments défavorables, en particulier le sucre, le sel et les graisses saturées. Les produits peu transformés, riches en fibres et en protéines, obtiennent généralement de meilleurs scores, tandis que les produits très transformés sont plus fortement pénalisés. Le Nutri-Score apparaît ainsi globalement cohérent avec les connaissances actuelles en nutrition et constitue un outil pertinent pour comparer les produits au sein d'une même catégorie.

Le développement du dashboard et du modèle de prédiction nous a également permis de mettre en évidence les limites liées à la qualité des données, à la diversité des catégories et à l'évolution des méthodes de calcul du Nutri-Score. Malgré ces contraintes, notre travail a permis de proposer un outil accessible, informatif et fiable, capable d'aider à la compréhension des enjeux nutritionnels actuels.

En conclusion, ce projet montre que l'analyse des données alimentaires constitue un levier important pour améliorer l'information des consommateurs et encourager des choix plus équilibrés. Il met en évidence l'intérêt de combiner data analyse, visualisation et modélisation pour aborder des problématiques de santé publique. Les perspectives d'amélioration identifiées ouvrent la voie à de futurs développements, visant à renforcer encore la précision, la pertinence et l'impact de ce type d'outils dans le domaine de la nutrition.

Références

- Afshin, A., Sur, P. J., Fay, K. A., et al. (2019). Health effects of dietary risks in 195 countries, 1990–2017: A systematic analysis. *The Lancet, 393*(10184), 1958–1972.
- Julia, C., & Hercberg, S. (2017). Research and lobbying conflicting on the issue of a front-of-pack nutrition labelling in France. *Archives of Public Health, 75*(1), 51.
- Monteiro, C. A., Cannon, G., Levy, R. B., et al. (2019). Ultra-processed foods: What they are and how to identify them. *Public Health Nutrition, 22*(5), 936–941.
- Santé Publique France. (2021). *L'obésité en France : état des lieux et tendances récentes.*
- World Health Organization (WHO). (2020). *Healthy diet – Fact sheet.*
- *https://www.who.int/news-room/fact-sheets/detail/healthy-diet (recommendation OMS 2026)*

Instructions

Le dashboard est composé de trois pages principales, accessibles grâce à des boutons situés en haut de l'interface. Chaque page correspond à un type d'analyse spécifique.

1. Accès au dashboard

Le tableau de bord est accessible en ligne via la plateforme Render. Lors du premier chargement, un temps d'attente pouvant aller jusqu'à plusieurs minutes (environ 5 à 7 minutes) peut être nécessaire, en raison du redémarrage automatique du serveur.

Une fois le chargement terminé, l'utilisateur accède à la page d'accueil du dashboard, qui présente le titre du projet, son objectif, ainsi que les boutons de navigation.

2. Navigation entre les pages

En haut de chaque page, trois boutons permettent d'accéder aux différentes sections du dashboard :

- **Analyse du Nutri-Score**
- **Analyse des nutriments**
- **Modèle de prédiction**

3. Page 1 : Analyse du Nutri-Score

Cette première page permet d'analyser la répartition des Nutri-Scores selon les catégories de produits.

Étapes d'utilisation :

1. Sélectionner une ou plusieurs catégories à l'aide du menu déroulant.
2. Observer les indicateurs (KPI) qui s'actualisent en fonction de la sélection.
3. Analyser l'histogramme représentant la distribution des scores de A à E.

4. Page 2 : Analyse des nutriments

Cette page est dédiée à l'étude des principaux nutriments (sucre, graisses, sel, fibres, protéines).

Étapes d'utilisation :

1. Sélectionner un nutriment à analyser à l'aide des boutons radio.
2. Choisir une ou plusieurs catégories via le menu déroulant.
3. Consulter les indicateurs affichés à gauche.
4. Faire défiler la liste des moyennes par catégorie à l'aide de la barre de défilement si nécessaire.
5. Analyser la carte thermique (heatmap) à droite.

5. Page 3 : Modèle de prédiction du Nutri-Score

Cette page permet d'explorer les données manquantes et d'utiliser le modèle de prédiction.

Étapes d'utilisation :

1. Observer le graphique circulaire indiquant la proportion de Nutri-Score connus et inconnus.
2. Renseigner toutes les valeurs nutritionnelles dans les champs de saisie.
3. Cliquer sur le bouton « Prédire ».
4. Lire le Nutri-Score estimé par le modèle.