Application au service de la santé publique

Samuel Pariente

Parcours Data Scientist – projet n°3

Sommaire

Aperçu du projet

Appel à projets

- Lancé par : Santé Publique France
- Objectif : trouver des idées innovantes d'applications en lien avec l'alimentation

Base de données

- Base de données collaborative Open Food Facts
- Extraction du 13/02/2022 à 17:36
- Composée de
  - 874,943 lignes : produits alimentaires
  - 191 colonnes : informations générales, tags spécifiques au produit, liste d'ingrédients, valeurs nutritionnelles pour 100g

Concept d'application

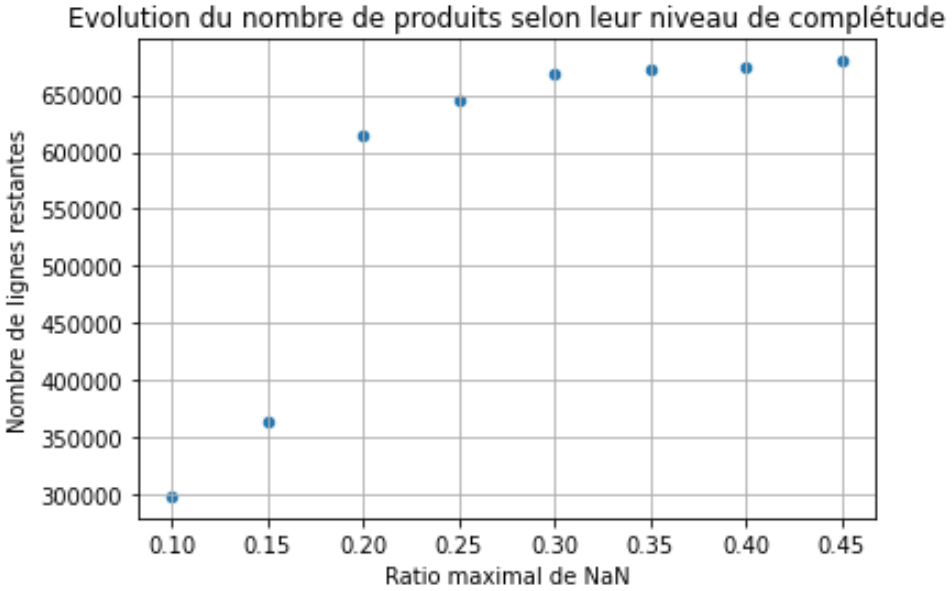
- Idée générale :
  - Aider l'utilisateur au moment de ses achats
  - Proposer des produits complémentaires à son panier pour une alimentation équilibrée
  - Permettre de choisir une catégorie de produit
  - Sélectionner uniquement les produits ayant la meilleure note Nutriscore
- Donnés nécessaires :
  - Eléments d'identification
  - Quantité de chaque produit
  - Valeurs nutritionnelles de chaque produit
  - Catégorie de chaque produit
  - Notes Nutriscore disponible

Filtrage de la base de données

Filtrage des données

- Extraction des données relatives à la France
- Complétion de la valeur énergétique en kcal
- Filtrage successif des données :
  1. Indicateurs : ≤ 70% valeurs manquantes
  2. Sélection d'indicateurs :
     - Code barre, nom du produit
     - Url de fiche openfoodfacts et image basse résolution, date de dernière mise à jour
     - Données numériques : quantité, valeurs nutritionnelles
  3. Produits : ≤ 15% valeurs manquantes

Visualisation du nombre de produits retenus



Transformation initiale de la base de données

Conversion de types

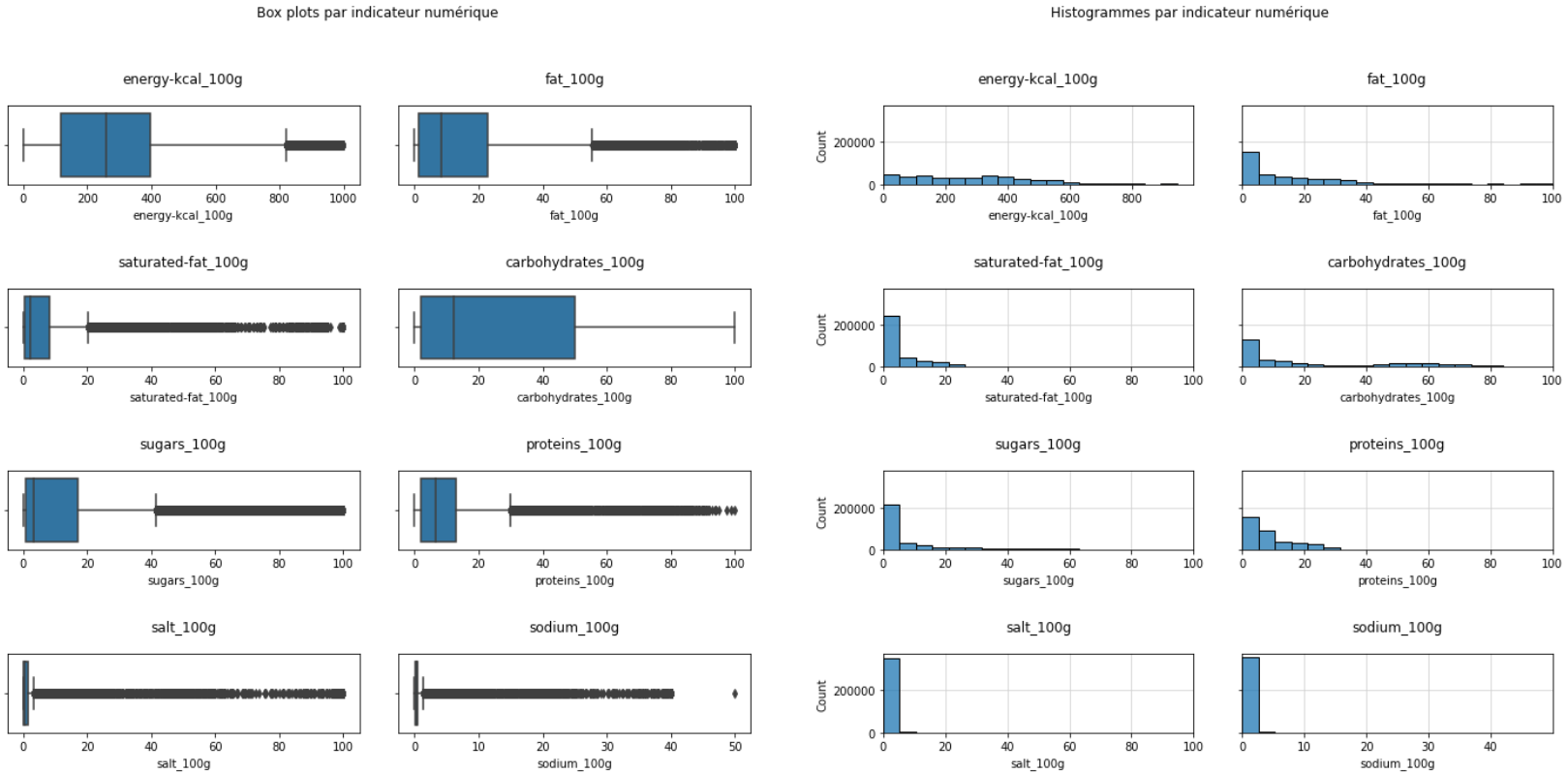
- Conversion des dates en format adapté
- Conversion de la donnée de quantité de produit en valeur numérique (g ou mL)

Traitement des valeurs aberrantes

- Valeurs nutritionnelles hors énergie :
  - Exclusion des produits dont la somme des valeurs nutritionnelles / 100g (hors énergie) est supérieure à 100g
  - Plafonnement de certaines valeurs nutritionnelles par celles du niveau supérieur
- Valeurs énergétiques :
  - Exclusion de valeurs énergétiques > 4,186.8
  - Conversion en kcal de valeurs en kJ
  - Suppression des valeurs nulles qui ne devraient pas l'être
- Exclusion des valeurs négatives

Analyse univariée des valeurs nutritionnelles

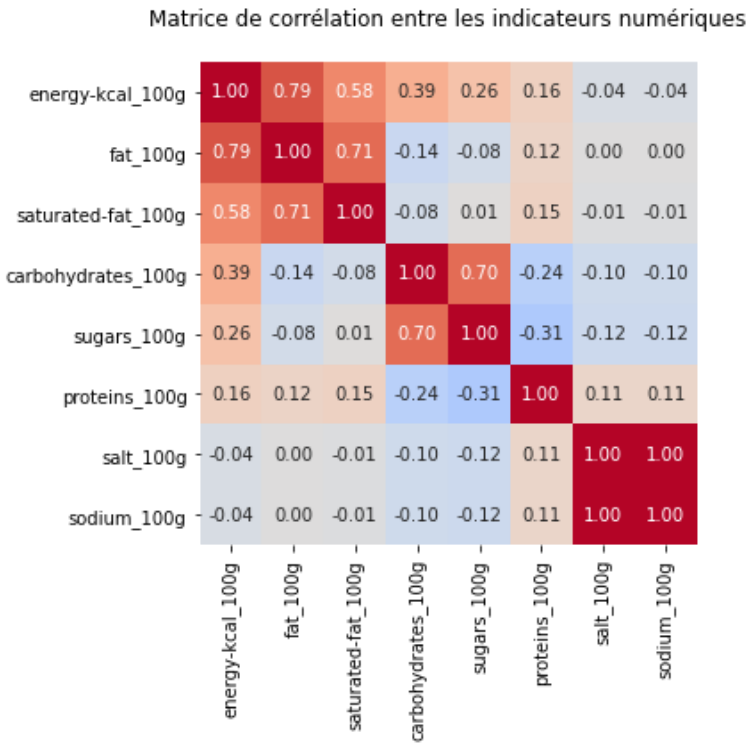
Aperçu de la distribution des différents indicateurs nutritionnels



- Distribution variant selon l'indicateur, tant en moyenne qu'en variance
- Plusieurs modes observables sur au moins certains indicateurs (énergie, graisses, glucides)

Corrélation entre indicateurs et ANOVA

Corrélation entre valeurs nutritionnelles



Nombreuses variables corrélées, p-values = 0 pour l'hypothèse nulle

Analyse de la variance (ANOVA)

Indicateur	Rapport de corrélation	p-value (ANOVA)	p-value (Alexander-Govern)
energy-kcal_100g	0.537911	0.0	0.0
fat_100g	0.568242	0.0	0.0
saturated-fat_100g	0.443399	0.0	0.0
carbohydrates_100g	0.594652	0.0	0.0
sugars_100g	0.533167	0.0	0.0
proteins_100g	0.564418	0.0	0.0
salt_100g	0.502124	0.0	0.0

Les variables ne suivent pas une distribution homogène entre toutes les catégorie

Traitements préliminaires

Valeurs aberrantes ou manquantes

- Corrélation directe entre sel et sodium :
  - Evidence par expertise métier : masse de sodium = 0,4 x masse de sel
  - Conclusion de la matrice de corrélation
- Exclusion des données aberrantes
- Imputation d'éventuelles données manquantes pour un des deux indicateurs mais saisie pour le second

Catégorisation manquante

- Valeurs « unknown » identifiant des catégories dont l'information n'est pas renseignée :
  - Erreur / oubli de saisie
  - Absence de catégorie adéquate pour le produit
- Oublis de saisie :
  - Traitement de quelques données dont l'information existe sous une autre forme à ce stade
  - Traitement systématique dans une étape ultérieure
- Absence de catégorie adéquate :
  - Création de 2 nouvelles catégories : sels et compléments protéinés
  - Attribution de cette catégorie à certains produits, sur la base d'une valeur nutritionnelle seuil, après analyse basée sur leur nom

Identification et traitement des outliers

Identification des outliers

- Outliers définis pour chaque indicateur nutritionnel, par catégorie de produits
- Approche de Tukey mise en œuvre :

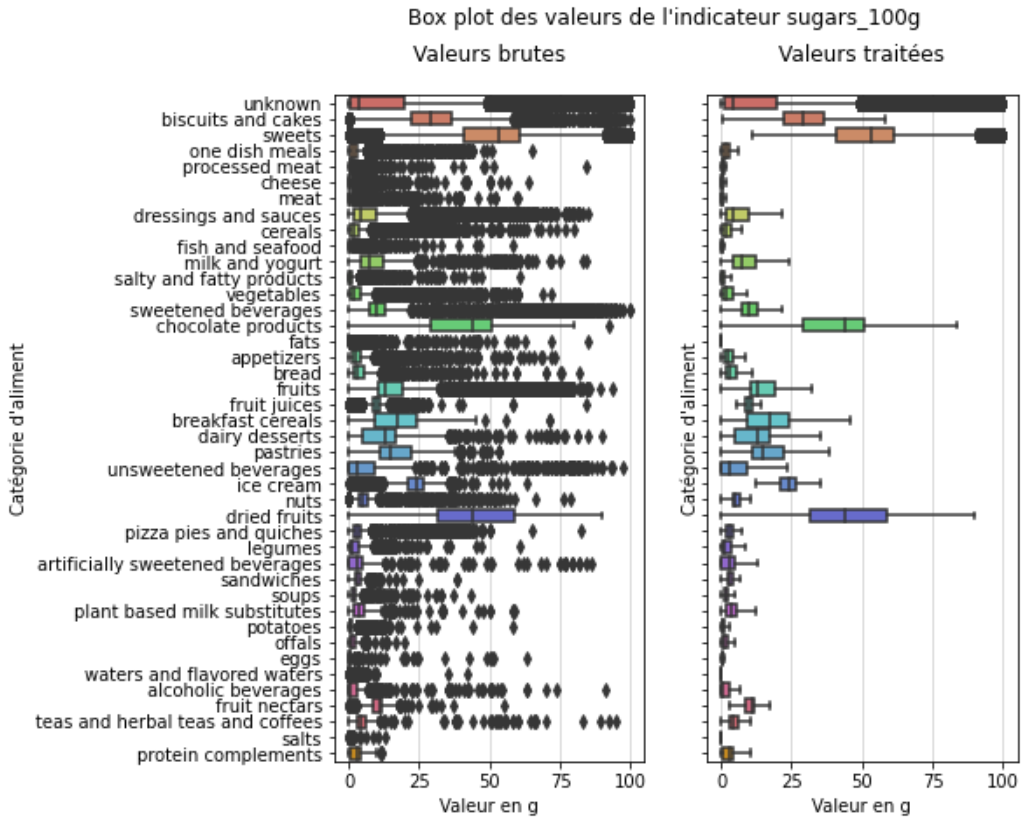
$$f(x) = \begin{cases} 0 \; si \; x \; \in [q_{0,25} - 1{,}5.IQR, q_{0,75} + 1{,}5.IQR] \\ 1 \; sinon \end{cases}$$

- Bornes modifiées pour catégories contenant des produits pouvant être purs (sel, sucres, protéines, graisses)

Traitement des outliers

- Seuil / plafond appliqué aux données identifiées comme outliers

Visualisation du traitement effectué



Traitement des valeurs manquantes

Données matériellement corrélées

- Approche par catégorie
- Algorithme de régression linéaire pour chaque indicateur en fonction de tous les autres
- Remplacement des données manquantes par les valeurs prédites par l'algorithme

Données faiblement corrélées

- Approche par catégorie
- Remplacement des données manquantes par la moyenne de l'indicateur pour la catégorie

Catégorie « unknown »

- Supposition que la catégorie peut être induite à partir des valeurs nutritionnelles du produit
- Algorithme de KNN entraîné en utilisant un échantillon
- Remplacement des valeurs « unknown » par les valeurs prédites

Quantités

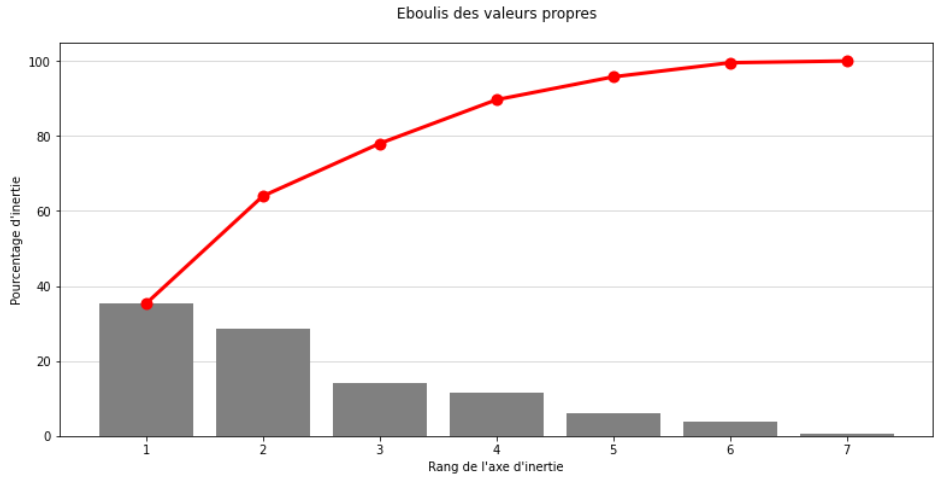
- Approche par catégorie
- Remplacement des données manquantes par le mode de l'indicateur pour la catégorie

Analyse en composantes principales (ACP)

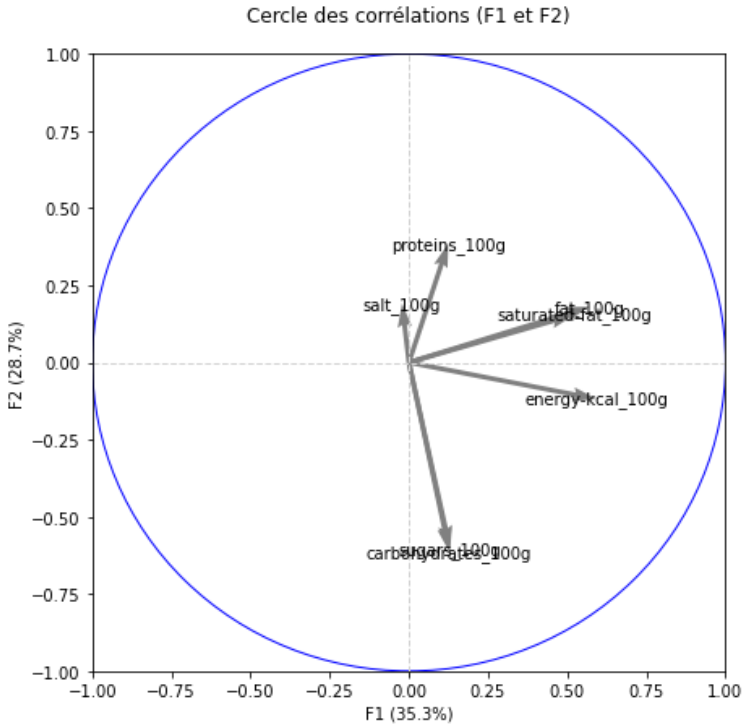
Approche

- 7 variables numériques (valeurs nutritionnelles)
- Recherche de dimensions composites maximisant la variance de la projection des données observées

Variance expliquée par chaque axe d'inertie



Composition des 2 premiers axes d'inertie



- F1 : apport énergétique
- F2 : forte composition en glucides

Exploitation pour l'application

Bases de données importées

- Importation de la base filtrée et traitée
- Importation d'une base complémentaire : valeurs nutritionnelles équilibrées par kcal

Calcul des valeurs nutritionnelles du panier

- Lecture des codes barres et ajout à la liste
- Extraction des valeurs pour 100g et multiplication par la quantité

Calcul des valeurs nutritionnelles équilibrées

- Multiplication du vecteur des valeurs équilibrées par la valeur énergétique du panier

Identification des produits adaptés

1. Sur la base de filtres utilisateur (catégorie, sous-catégorie), extraction des produits de la base principale
2. Filtre sur les produits ayant la meilleure note Nutriscore
3. Calcul des valeurs nutritionnelles totales pour chaque produit
4. Transposition de toutes les données dans un espace normé par les valeurs équilibrées
5. Calcul de la distance entre un vecteur complémentaire idéal et chaque produit

Exemple d'utilisation

Données de test

- 17 images de produits alimentaires contenant un code barres sourcées sur Google images

Aperçu de la composition du panier

Valeur nutritionnelle totale du panier
(16 produits retenus)

Composition du panier (17 produits)

Catégorie de produits recherchée

- A la main de l'utilisateur :
  - Catégorie : fruits et légumes
  - Sous-catégorie : fruits

Proposition de produits

- 3 produits proposés, les plus adaptés au panier actuel
- Choix d'un produit pour afficher des informations :
  - Nom et lien vers la fiche sur site openfoodfacts
  - Nutriscore
  - Quantité de produit recommandée
  - Valeurs nutritionnelles / 100g du produit