# Amélioration de la base de données Open Food Facts



*Ce sujet de projet est librement inspiré du projet 3 “Open Food Facts” de OpenClassrooms (parcours AI Engineer)*

## 

## Contexte

L'agence Santé publique France souhaite améliorer la qualité de sa base de données Open Food Facts, une ressource cruciale pour informer le public sur la composition nutritionnelle des produits alimentaires. Cependant, de nombreuses valeurs manquantes limitent son utilité.

## Livrables

* Un notebook Jupyter ou Colab documenté contenant l'ensemble de votre analyse, de votre code et de vos résultats.
* Une présentation claire et concise de vos conclusions et de votre modèle.
* Un document expliquant la conformité de votre projet au RGPD.

## 

## Objectif

Votre mission est de développer un système de suggestion ou d'auto-complétion pour aider les utilisateurs à renseigner plus facilement les informations manquantes. Vous devrez explorer, nettoyer et analyser les données existantes afin de construire un modèle capable de prédire intelligemment ces valeurs manquantes.

## Données

Vous travaillerez avec le jeu de données Open Food Facts, qui contient une multitude de variables sur les produits alimentaires. Le jeu de données est volumineux, il faudra donc optimiser son traitement.

## 

## Questions

### **Exploration et nettoyage des données**

1. Chargez le jeu de données en utilisant des techniques adaptées pour gérer sa taille (chargement partiel, optimisation de la mémoire, etc.).
2. Familiarisez-vous avec sa structure, ses variables et leur signification en utilisant des fonctions comme **head()**, **info()**, **describe()**.
3. Identifiez les variables pertinentes pour votre analyse et votre modèle, en vous concentrant sur les variables nutritionnelles présentant des valeurs manquantes.
4. Nettoyez les données :
   1. Analysez et gérez les valeurs manquantes de manière appropriée (suppression, imputation, etc.). Justifiez vos choix en fonction du type de variable et du contexte.
   2. Détectez et traitez les valeurs aberrantes et incohérentes à l'aide de visualisations et de méthodes statistiques.
   3. Créez des fonctions pour automatiser ces étapes de nettoyage afin de faciliter la reproductibilité de votre travail.

### Analyse exploratoire

1. Réalisez une analyse univariée de chaque variable pertinente :
   1. Calculez les statistiques descriptives (moyenne, médiane, écart-type, etc.).
   2. Visualisez les distributions à l'aide de graphiques adaptés (histogrammes, boxplots, etc.).
2. Effectuez une analyse bivariée pour identifier les relations entre les variables :
   1. Utilisez des graphiques tels que des nuages de points, des heatmaps de corrélation, etc.
   2. Calculez des mesures d'association appropriées (corrélation, tests statistiques, etc.).
3. Interprétez les résultats de vos analyses et tirez des conclusions sur les caractéristiques des données et les relations entre les variables.

### Analyse multivariée

1. Explorez les relations plus complexes entre plusieurs variables à l'aide de techniques telles que l'Analyse en Composantes Principales (ACP).
2. Si nécessaire, créez de nouvelles variables pertinentes pour votre modèle.
3. Sélectionnez les variables les plus informatives pour prédire les valeurs manquantes, en justifiant vos choix.

### Modélisation prédictive

1. Choisissez un ou plusieurs algorithmes de machine learning adaptés à la prédiction des valeurs manquantes (régression, classification, etc.).
2. Entraînez et évaluez votre modèle sur un échantillon représentatif des données, en utilisant des techniques de validation appropriées (validation croisée, etc.).
3. Expliquez clairement votre choix d'algorithme, les étapes de prétraitement des données nécessaires, les hyperparamètres choisis et les résultats obtenus en termes de performance.
4. Comparez les performances de différents modèles si vous en avez testé plusieurs.

### Présentation des résultats

1. Rédigez un rapport d'exploration clair et concis, présentant vos principales conclusions, vos visualisations et les résultats de votre modèle.
2. Préparez une présentation visuelle et pédagogique pour expliquer vos analyses et vos recommandations à un public non-expert (Santé publique France).
3. Expliquez comment votre projet respecte les principes du RGPD (confidentialité, intégrité, disponibilité, etc.), même en l'absence de données personnelles directement identifiantes.