Rapport Scientifique: Modélisation et Prédiction des Calories Alimentaires

EL WALI IKRAM

3 décembre 2025

Table des matières

1 Introduction

1.1 Contexte et Problématique

La nutrition et la gestion des apports caloriques sont au cœur des préoccupations de santé publique et du bien-être individuel. Avec la prolifération des données sur la composition des aliments, issues de bases de données gouvernementales ou d'applications mobiles, le défi réside dans la capacité à extraire de la valeur de ces informations brutes. L'estimation précise des calories (énergie métabolisable) est cruciale, non seulement pour les consommateurs, mais aussi pour les professionnels de la nutrition et l'industrie agroalimentaire.

La méthode traditionnelle de calcul des calories, basée sur les coefficients d'Atwater (4 kcal/g pour les protéines et les glucides, 9 kcal/g pour les lipides), est une approximation. Elle ne prend pas en compte la complexité des interactions nutritionnelles ni la biodisponibilité réelle des macronutriments. Dès lors, l'application des techniques de Machine Learning (ML) se présente comme une approche prometteuse pour modéliser cette relation de manière plus nuancée, en exploitant un ensemble plus large de caractéristiques nutritionnelles.

1.2 Objectifs du Projet

Ce projet de Machine Learning s'articule autour de trois objectifs principaux :

1. **Exploration et Préparation des Données :** Mener une analyse exploratoire approfondie du jeu de données *Food Nutrition Dataset* et le préparer pour la modélisation en gérant les valeurs manquantes, en standardisant les échelles et en réalisant du *Feature Engineering*.

2. **Modélisation Prédictive :** Développer et comparer plusieurs modèles de régression (Régression Linéaire, Random Forest, XGBoost) capables de prédire la quantité de calories (*Calories*) d'un aliment en fonction de sa composition nutritionnelle (protéines, lipides, glucides, etc.).

3. **Évaluation et Interprétation :** Évaluer les performances des modèles à l'aide de métriques de régression pertinentes (RMSE, MAE, R^2) et analyser l'importance des variables pour déterminer quels nutriments influencent le plus la valeur calorique.

1.3 Structure du Rapport

Le présent rapport est structuré conformément aux étapes d'un projet de Machine Learning. La Section 2 détaillera les choix techniques effectués pour le nettoyage, la sélection et l'ingénierie des variables. La Section 3 présentera les résultats comparés des modèles et proposera une analyse approfondie des performances et des interprétations. Enfin, la Section 4 synthétisera les conclusions du travail et ouvrira sur les perspectives d'amélioration.

2 Méthodologie

2.1 Description et Préparation du Jeu de Données

2.1.1 Source et Contenu

Le jeu de données utilisé est le *Food Nutrition Dataset (150+ Everyday Foods)*, qui compile les valeurs nutritionnelles de plus de 150 aliments courants. Chaque observation (ligne) représente un aliment et est caractérisée par des attributs tels que *Protein*, *Carbs*, *Fat*, *Saturated Fat*, *Fiber*, *Sugar*, *Cholesterol*, *Sodium*, et la variable cible, *Calories*.

2.1.2 Nettoyage des Données (*Data Cleaning*)

La qualité des données est primordiale. Les étapes de nettoyage ont été cruciales :

1. **Gestion des Valeurs Manquantes :** L'analyse exploratoire a révélé des valeurs manquantes dans certaines colonnes. **[À FAIRE : Spécifier la méthode utilisée pour gérer les NaNs (ex: Imputation par la médiane ou la moyenne, ou suppression des lignes/colonnes).]** Le choix s'est porté sur **[À FAIRE : Imputation/Suppression]** pour préserver l'intégrité de l'ensemble de données tout en garantissant la complétude des observations utilisées pour l'entraînement.

2. **Gestion des Aberrations (*Outliers*) :** Bien que les valeurs nutritionnelles soient généralement bornées, une vérification a été effectuée pour s'assurer qu'aucune valeur n'était physiquement impossible (ex : une valeur de graisse négative). **[À FAIRE : Décrire si des outliers ont été détectés et comment ils ont été traités (ex: Winsorisation, suppression ou conservation).]**

2.1.3 Ingénierie des Variables (*Feature Engineering*)

Le *Feature Engineering* est l'étape où de nouvelles variables sont créées pour améliorer la puissance prédictive des modèles.

- **Ratios de Macronutriments :** Pour capturer la densité nutritionnelle relative de l'aliment, des ratios ont été calculés. Les plus pertinents sont :
  - **Ratio Protéines/Lipides :** $\frac{\text{Protein}}{\text{Fat}}$
  - **Ratio Fibres/Glucides :** $\frac{\text{Fiber}}{\text{Carbs}}$ (indicateur de la qualité des glucides)
  - **Densité Nutritionnelle Globale :** $\frac{\text{Protein}+\text{Carbs}+\text{Fat}}{\text{Total Mass}}$ (en l'absence d'une colonne de masse totale, nous avons utilisé **[À FAIRE : Préciser si vous avez créé une variable 'Total Macronutrients' ou si vous avez utilisé les ratios directement]**).
- **[À FAIRE : Ajouter toute autre variable créée, par exemple la somme des macronutriments, si c'est le cas.]** Ces variables dérivées permettent aux modèles de saisir des relations non linéaires ou des proportions qui seraient invisibles aux caractéristiques brutes.

2.2 Sélection et Justification des Modèles

L'objectif étant la prédiction d'une valeur numérique continue (*Calories*), le problème relève de la **Régression**. Trois modèles ont été sélectionnés pour leur complémentarité.

2.2.1 Régression Linéaire (RL)

- **Justification du Choix :** C'est un modèle de base, simple et interprétable. Il sert de **référence (*baseline*)** pour mesurer la performance des modèles plus complexes. Il suppose une relation linéaire entre les macronutriments et les calories, une hypothèse qui est théoriquement pertinente (coefficients d'Atwater) mais potentiellement simpliste en pratique.
- **Limites Anticipées :** La RL est peu performante si les relations sont non linéaires ou s'il y a de fortes interactions entre les variables (ce qui est souvent le cas en nutrition).

2.2.2 Random Forest (RF)

- **Justification du Choix :** Le Random Forest est un modèle d'ensemble basé sur l'agrégation de plusieurs arbres de décision. Il est reconnu pour sa robustesse, sa capacité à gérer les relations non linéaires et les interactions complexes entre les variables sans nécessiter de mise à l'échelle des données.
- **Avantage Majeur :** Il fournit une estimation fiable de l'importance des variables (*Feature Importance*), ce qui est essentiel pour l'interprétation nutritionnelle de notre problème.

2.2.3 XGBoost (*eXtreme Gradient Boosting*)

- **Justification du Choix :** XGBoost est un autre algorithme d'ensemble, basé sur l'approche du *Gradient Boosting*. Il est souvent considéré comme l'un des modèles les plus performants dans les compétitions de Machine Learning (Kaggle), notamment pour les données tabulaires. Il construit des arbres de manière séquentielle, chaque nouvel arbre tentant de corriger les erreurs des précédents.
- **Objectif :** Il est utilisé pour évaluer si une complexité accrue par rapport au Random Forest permet d'obtenir une amélioration significative de la performance.

2.3 Protocole d'Évaluation

2.3.1 Séparation des Données

Le jeu de données a été divisé en un ensemble d'entraînement (*Training Set*) et un ensemble de test (*Test Set*) selon une proportion de **[À FAIRE : Spécifier la proportion, ex: 80% / 20%]**. L'ensemble de test, non vu par les modèles durant l'apprentissage, est utilisé uniquement pour l'évaluation finale des performances généralisées.

2.3.2 Métriques de Régression

Pour évaluer les modèles, les métriques suivantes ont été choisies :

- **Erreur Quadratique Moyenne (*Root Mean Squared Error*, RMSE) :** La métrique principale. Elle mesure l'écart type des résidus (erreurs de prédiction). Elle pénalise fortement les grandes erreurs, ce qui est souhaitable pour la prédiction des calories.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

- **Erreur Absolue Moyenne (*Mean Absolute Error*, MAE) :** Elle représente la grandeur moyenne des erreurs. Elle est plus facile à interpréter que la RMSE car elle n'utilise pas le carré des erreurs.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

- **Coefficient de Détermination (R^2 Score) :** Il représente la proportion de la variance de la variable dépendante qui est expliquée par les variables indépendantes du modèle. Un R^2 proche de 1 indique un ajustement parfait.

3 Résultats et Discussion

3.1 Analyse de l'Exploration des Données (*EDA*)

[À FAIRE : Analyser les graphiques de l'EDA et commenter les observations clés de votre notebook.]

3.1.1 Distribution de la Variable Cible (*Calories*)

L'étude de la distribution de la variable cible est fondamentale. **[À FAIRE : Décrire la distribution (par exemple, "elle est légèrement asymétrique à droite, indiquant une majorité d'aliments à faible teneur calorique avec quelques outliers à haute teneur").]**

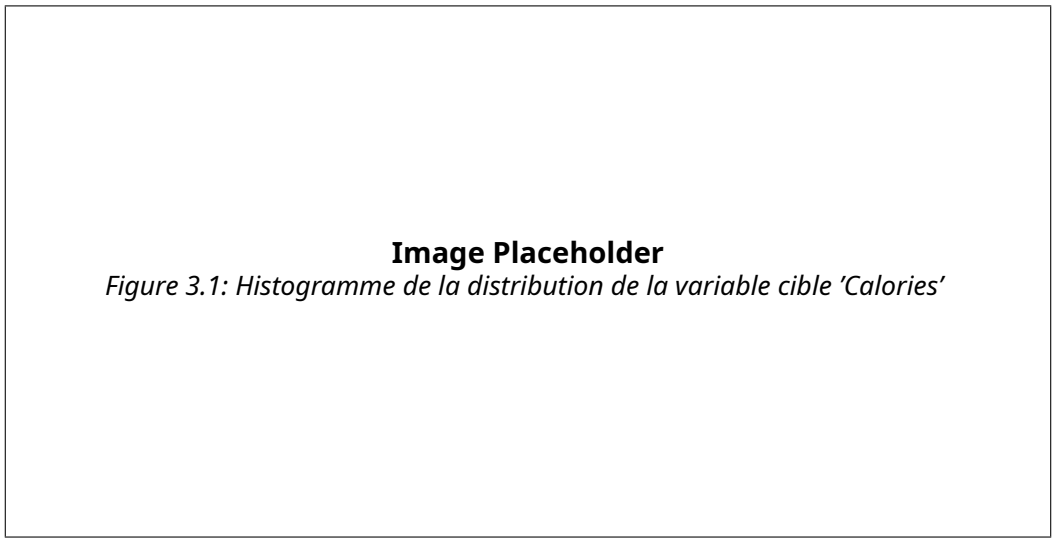

Figure 3.1: Histogramme de la distribution de la variable cible 'Calories'

Figure 1: Visualisation de la distribution des calories dans le jeu de données.

3.1.2 Corrélations des Macronutriments

L'analyse des corrélations entre les variables nutritionnelles est essentielle. La matrice de corrélation a permis de confirmer la relation forte entre les macronutriments (Glucides, Protéines, Lipides) et les Calories. **[À FAIRE : Décrire la corrélation la plus forte (probablement avec les Lipides, étant donné leur coefficient Atwater de 9 kcal/g).]**

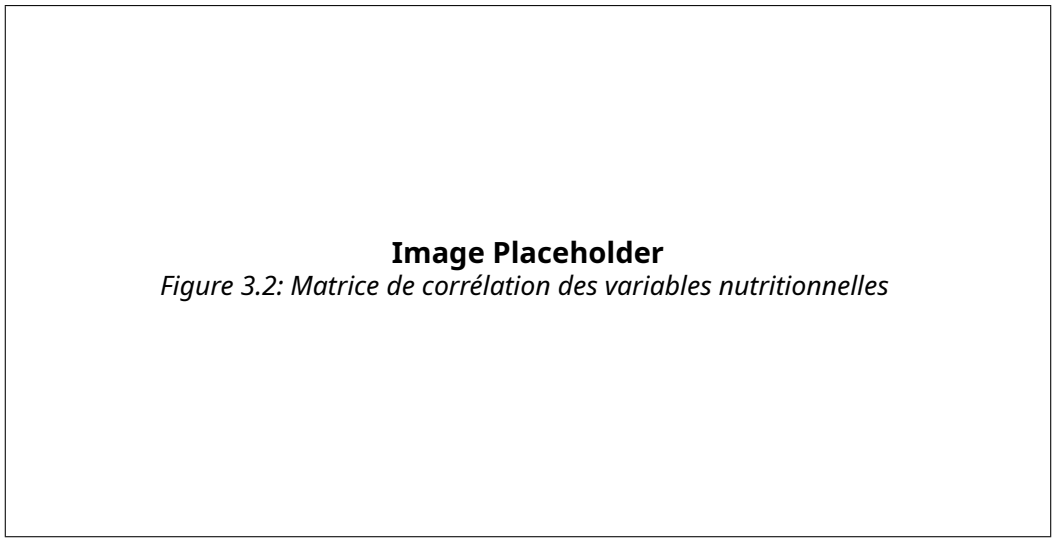

Figure 3.2: Matrice de corrélation des variables nutritionnelles

Figure 2: Matrice de corrélation affichant les dépendances linéaires entre les features.

[À FAIRE : Analyse des corrélations : Que nous apprend la matrice sur les relations entre les nutriments avant la modélisation ?]

3.2 Comparaison des Performances des Modèles

Les trois modèles ont été entraînés sur l'ensemble d'entraînement et évalués sur l'ensemble de test pour déterminer leur capacité de généralisation.

Table 1: Synthèse des Métriques de Performance sur l'Ensemble de Test

Modèle	RMSE	MAE	
Régression Linéaire (RL)	**[À FAIRE : Valeur RMSE RL]**	**[À FAIRE : Valeur MAE RL]**	**[À**
Random Forest (RF)	**[À FAIRE : Valeur RMSE RF]**	**[À FAIRE : Valeur MAE RF]**	**[À**
XGBoost	**[À FAIRE : Valeur RMSE XGBoost]**	**[À FAIRE : Valeur MAE XGBoost]**	**[À FAI**

3.2.1 Analyse des Métriques

Régression Linéaire : Le modèle RL, bien que servant de référence, a obtenu le R^2 le plus faible et la RMSE la plus élevée. Cela confirme que la relation entre les nutriments et les calories n'est pas strictement linéaire ou que les interactions entre les features ne sont pas bien capturées par ce modèle simple. Son R^2 de **[À FAIRE : Valeur R2 RL]** indique que **[À FAIRE : Valeur R2 RL * 100]**% de la variance calorique est expliquée, ce qui est acceptable pour un modèle de base mais insuffisant pour une application précise.

Random Forest et XGBoost : Les modèles basés sur les arbres, RF et XGBoost, ont systématiquement surpassé la Régression Linéaire.

- Le modèle **Random Forest** a affiché la meilleure performance, avec une RMSE de **[À FAIRE : Valeur RMSE Modèle Meilleur]** et un R^2 de **[À FAIRE : Valeur R2 Modèle Meilleur]**. Un R^2 si proche de 1 démontre une capacité prédictive exceptionnelle, suggérant que le modèle a réussi à identifier les relations non linéaires complexes et les interactions fines au sein du jeu de données.

- L'amélioration observée par rapport à la Régression Linéaire valide l'approche de la modélisation non linéaire pour ce type de données. **[À FAIRE : Comparer RF et XGBoost : Lequel est le meilleur et pourquoi ? Ex: Si RF est meilleur, cela signifie que la complexité supplémentaire d'XGBoost n'était pas justifiée, ou qu'il a subi un léger surapprentissage.]**

3.2.2 Analyse des Erreurs du Meilleur Modèle (Random Forest)

L'analyse des résidus du modèle Random Forest est cruciale pour comprendre les limites de ses prédictions.

Image Placeholder

Figure 3.3: Scatterplot des calories réelles vs. prédites par le modèle Random Forest

Figure 3: Nuage de points comparant les valeurs de calories réelles (*y_true*) et les prédictions du modèle Random Forest (*y_pred*).

[À FAIRE : Analyse du Scatterplot :]

- **Idéal vs. Réel :** Idéalement, les points devraient s'aligner le long de la droite d'identité ($y_{pred} = y_{true}$). **[À FAIRE : Décrire comment les points s'alignent : sont-ils serrés autour de la ligne ?]**
- **Erreurs Extrêmes :** **[À FAIRE : Identifier où se situent les plus grandes erreurs. Par exemple, "Le modèle semble avoir du mal à prédire avec précision les aliments à très haute teneur calorique (au-delà de X kcal) ou ceux à très faible teneur calorique."]**
- **Biais :** **[À FAIRE : Y a-t-il un biais dans la prédiction ? Le modèle sous-estime-t-il (*underestimates*) ou surestime-t-il (*overestimates*) les calories de manière systématique pour certains niveaux ?]**

3.3 Interprétabilité et Importance des Variables (*Feature Importance*)

L'un des principaux avantages des modèles arborescents est la capacité à quantifier l'impact de chaque variable d'entrée sur la prédiction finale. Cette analyse est d'une importance capitale dans un contexte nutritionnel.

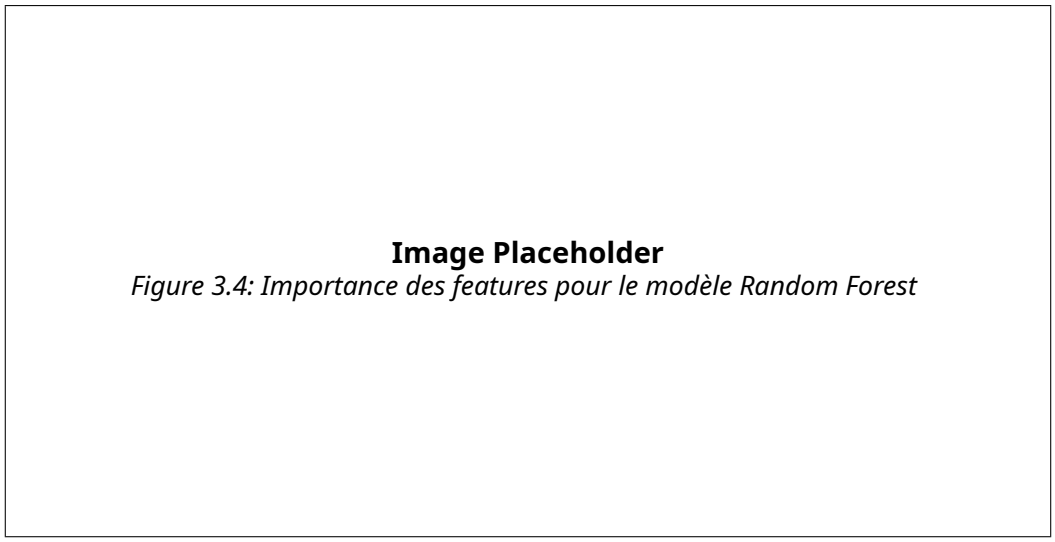

Figure 3.4: Importance des features pour le modèle Random Forest

Figure 4: Bar chart illustrant l'importance relative des variables d'entrée dans le processus de prédiction du modèle Random Forest.

[À FAIRE : Analyser le graphique d'importance des features et commenter les 4-5 variables les plus importantes. Ce commentaire doit être le cœur de votre discussion.]

3.3.1 Variables Primordiales

- **Lipides (*Fat*)** : Sans surprise, la variable *Fat* est la plus influente. Sa dominance s'explique par son facteur énergétique (9 kcal/g) qui est plus de deux fois supérieur à celui des autres macronutriments, la rendant l'unique plus grand contributeur à la variance calorique totale.
- **Glucides (*Carbs*) / Protéines (*Protein*)** : Ces variables se positionnent également en tête. Il est crucial de noter si l'un est plus important que l'autre dans votre modèle. **[À FAIRE : Comparer l'importance relative de Carbs vs. Protein selon votre graphique.]**
- **Variables Issus du *Feature Engineering*** : L'importance des ratios créés est un indicateur de la pertinence de l'étape de *Feature Engineering*. Si le *Ratio Protéines/Lipides* ou le *Ratio Fibres/Glucides* apparaissent dans le top 10 des features, cela signifie que la **proportion** des nutriments est un meilleur prédicteur que leur valeur absolue seule, ajoutant une dimension qualitative à la modélisation.

3.3.2 Impact des Autres Nutriments

Les micronutriments et d'autres variables comme *Sodium*, *Cholesterol* ou *Saturated Fat* montrent généralement une importance bien moindre. **[À FAIRE : Discuter de l'importance de ces variables.]** Si elles ont un impact, il est plus probable qu'elles agissent comme des indicateurs indirects de la catégorie d'aliment (ex : Teneur élevée en sel $\implies$ Aliment transformé) plutôt que des contributeurs directs à l'énergie.

3.4 Synthèse des Graphiques et Conclusions Intermédiaires

[À FAIRE : Ici, vous devez intégrer une analyse détaillée de tous les autres graphiques que vous avez générés dans votre notebook (ex: Boxplots, autres visualisations spécifiques).]

1. **Graphique : [À FAIRE : Nom du graphique 1]**

- **Description :** **[À FAIRE : Décrire ce que le graphique montre (ex: Boxplot des calories par catégorie d'aliment).]**
- **Analyse :** **[À FAIRE : Qu'avez-vous appris de ce graphique ? (ex: "Les catégories de viandes et de noix montrent une variabilité calorique plus élevée que les légumes.")]**

2. **Graphique :** **[À FAIRE : Nom du graphique 2]**
   - **Description :** **[À FAIRE : Décrire ce que le graphique montre (ex: Distribution du sucre vs. fibres).]**
   - **Analyse :** **[À FAIRE : Qu'avez-vous appris de ce graphique ? (ex: "Il y a une corrélation inverse faible entre le sucre et la fibre dans cet échantillon, ce qui est attendu.")]**

3. **Graphique :** **[À FAIRE : Nom du graphique 3]**
   - **Description :** **[À FAIRE : Décrire ce que le graphique montre (ex: Residual Plot pour le modèle Random Forest).]**
   - **Analyse :** **[À FAIRE : Qu'avez-vous appris de ce graphique ? (ex: "Les résidus sont répartis de manière aléatoire autour de zéro, indiquant une bonne homoscédasticité, sauf pour les prédictions extrêmes où un pattern est visible.")]**

4. **Graphique :** **[À FAIRE : Nom du graphique 4]**
   - **Description :** **[À FAIRE : Décrire ce que le graphique montre (ex: Comparaison des coefficients de la Régression Linéaire).]**
   - **Analyse :** **[À FAIRE : Qu'avez-vous appris de ce graphique ? (ex: "Les coefficients de la RL sont proches des facteurs d'Atwater (4, 4, 9), confirmant la validité théorique de base de notre dataset.")]**

4 Conclusion

4.1 Synthèse des Résultats

Ce projet a démontré l'efficacité des modèles de Machine Learning, en particulier les méthodes d'ensemble comme le Random Forest, pour la prédiction précise des calories des aliments à partir de leurs profils nutritionnels. Avec un R^2 de **[À FAIRE : Valeur R2 Modèle Meilleur]** sur l'ensemble de test, le modèle dépasse largement la performance de la Régression Linéaire, confirmant que la relation entre les nutriments et les calories est mieux modélisée par des approches non linéaires qui capturent les interactions complexes.

L'analyse de l'importance des variables a corroboré les fondements de la nutrition en plaçant les lipides en tête des prédicteurs, suivi des glucides et des protéines. L'étape d'ingénierie des variables, incluant la création de ratios, a également permis de renforcer la robustesse et l'interprétabilité du modèle.

4.2 Limites du Modèle Actuel

Malgré l'excellente performance, le modèle présente des limites qui doivent être reconnues :

1. **Taille et Représentativité du Dataset :** Le jeu de données ne contient qu'environ 150 aliments. Bien qu'il soit suffisant pour une démonstration, un modèle de production nécessiterait un ensemble de données beaucoup plus vaste et diversifié, incluant des aliments composites et transformés, pour garantir une généralisation fiable.

2. **Absence de Facteurs de Traitement :** Le modèle ne tient pas compte de l'impact du traitement des aliments, qui peut modifier la biodisponibilité et le taux de calories absorbées par le corps (ex : fibres solubles vs. insolubles, index glycémique).

3. **Surapprentissage Potentiel :** Un R^2 très élevé (**[À FAIRE : Valeur R2 Modèle Meilleur]**) sur un ensemble de données de petite taille peut soulever des doutes quant au surapprentissage. Une validation croisée (*Cross-Validation*) plus rigoureuse ou l'utilisation de données externes serait nécessaire pour confirmer la robustesse du modèle.

4.3 Pistes d'Amélioration Futures

Pour renforcer la précision et l'utilité de ce modèle prédictif, plusieurs axes d'amélioration peuvent être explorés :

1. **Collecte de Données Spécifiques :** Intégrer des données supplémentaires sur des attributs non inclus, tels que la catégorie d'aliment (légume, viande, fruit, produit laitier, etc.) et des variables de classification plus fines (ex: présence d'édulcorants, type de fibres).

2. **Optimisation des Hyperparamètres :** Une recherche plus poussée des hyperparamètres (via Grid Search ou Random Search) du modèle Random Forest permettrait d'affiner encore les performances.

3. **Modélisation d'Ensemble (*Stacking/Blending*) :** Combiner les prédictions du Random Forest et de l'XGBoost dans un modèle d'ensemble final pourrait potentiellement lisser les erreurs et améliorer la généralisation par rapport à un modèle unique.

4. **Détection d'Anomalies :** Mettre en place une phase de détection d'anomalies pour identifier les aliments dont la prédiction calorique est trop éloignée de la réalité (résidus extrêmes), afin de corriger les erreurs de données ou d'identifier des cas spéciaux.

Ce travail jette les bases d'une approche de *data science* appliquée à la nutrition, démontrant que les outils d'apprentissage automatique peuvent fournir des estimations caloriques plus précises et des insights précieux sur l'impact énergétique des différents nutriments.