Cahier des charges - projet fil rouge – Prédiction des valeurs nutritionnels d'aliments pour les animaux monogastriques

Alan Turbot, Baptiste Granier, François Faramus, Pierre-Jean Gouze, Valentine Michelet

1. Besoin initial du projet

L'Association Française de Zootechnie dispose d'une base de données de références de valeurs nutritionnelles pour environ 200 produits. L'accès à ces données reste complexe : pour chaque produit, des dizaines d'équations successives sont nécessaires pour couvrir les valeurs nutritionnelles selon les différentes espèces. De plus, les bons prédicteurs ne sont pas toujours disponibles, et certains produits sont mal connus, rendant les calculs incertains ou impossibles. La mise en œuvre actuelle repose sur une base d'équations difficile à exploiter et à maintenir. Le besoin principal du projet est donc de simplifier l'accès aux données, d'améliorer la cohérence des calculs et de proposer une méthode robuste pour gérer les produits insuffisamment documentés.

2. Les données disponibles

Les données sont organisées sous forme d'un tableau où chaque ligne décrit un produit par deux informations structurantes : sa **classe**, qui correspond au type de plante (céréales, tourteaux, fabacées, etc.), et son **nom**, qui désigne l'espèce précise (blé tendre, avoine, etc.). Ces deux champs sont suivis d'un ensemble de **variables d'entrée** caractérisant la composition brute du produit : teneur en matière sèche (MS), protéines brutes (PB), cellulose brute (CB), matières grasses (MGR), matières minérales (MM), fractions fibreuses (NDF, ADF), lignine, amidon et sucres. À partir de ces composantes sont définies les **variables de sortie**, comprenant les différentes valeurs énergétiques (EB, ED, EM, EN pour le porc), les énergies métabolisables pour volailles (EMAn coq et poulet), ainsi que des indicateurs tels que l'UFL, les PDI et le bilan protéique ruminal. Ces sorties constituent les valeurs nutritionnelles à estimer ou prédire à partir de la composition initiale.

Classe	**Nom**	**MS** % brut	**PB** % brut	**CB** % brut	**MGR** % brut	**MM** % brut	**NDF** % brut	**ADF** % brut	**Lignine** % brut	**Amidon** % brut	**Sucres** % brut

Tableau 1 : Entrées des modèles à implémenter

EB (kcal) kcal/kg brut	**ED porc croissance** (kcal) kcal/kg brut	**EM porc croissance** (kcal) kcal/kg brut	**EN porc croissance** (kcal) kcal/kg brut	**EMAn coq** (kcal) kcal/kg brut	**EMAn poulet** (kcal) kcal/kg brut	**UFL 2018** par kg brut	**PDI 2018** g/kg brut	**BalProRu** 2018 g/kg brut

Tableau 2 : Sorties des modèles à implémenter

3. Besoins de prétraitements

Après une première discussion avec les experts et certaines analyses statistiques, nous notons de fortes corrélations entre certaines variables. De ce fait, il existe de la redondance au sein du jeu de données et nous avons par conséquent une volonté de réduire au maximum le nombre de variables d'entrée tout en garantissant les mêmes performances. L'objectif est de faciliter l'interaction Homme/Machine, notamment en vue d'une mise en production des modèles créés.

De plus, il est important de remarquer que l'ensemble du jeu de données est très propre (aucune valeur manquante dans le dataset, absence de valeurs aberrantes). Cela est normal car les résultats sont issus de tables Excel qui contiennent des équations déjà définies et robustes.

Les analyses exploratoires (ACP) montrent que les produits se répartissent en groupes distincts et homogènes, reflétant des signatures chimiques propres à chaque produit. Toutefois, le jeu de données actuel, limité à six produits, ne couvre probablement qu'une faible partie des compositions chimiques potentiellement admissibles. Cette situation pourrait limiter la capacité du modèle à prédire correctement des produits nouveaux ou des formulations différentes de celles observées.

Afin d'améliorer la généralisation du modèle, il est envisagé de mettre en place **une augmentation de données**. L'objectif serait de générer, pour chaque produit, des compositions chimiques supplémentaires considérées comme plausibles, puis de leur appliquer les équations nutritionnelles associées. Cette démarche permettrait d'élargir la variabilité accessible pendant l'entraînement tout en conservant la cohérence propre à chaque produit.

Pour conclure cette partie, il n'y a pas de pré traitements lourds à mettre en place pour les enjeux actuels. Pour information, sans prétraitements, nous avons des valeurs de R^2 et de MAE excellentes avec comme modèle sélectionné XGBoost Regressor.

4. Objectifs du projet

Les objectifs du projet sont multiples :

- Simplification de l'interaction Homme/Machine : au lieu d'utiliser des tables Excel et des bases de données Access, l'idée est de déployer simplement des modèles d'apprentissage automatique pour prédire les valeurs nutritionnelles d'aliments,
- Réduire le nombre de variables utilisées tout en garantissant de bonnes performances en prédiction,

Ces objectifs proviennent des études préliminaires et sont voués à évoluer au fur et à mesure du projet.

5. Propositions d'approfondissement du projet

Au vu de la durée du projet, il est nécessaire de prévoir des pistes d'approfondissements et d'explorations. En effet, nous avons actuellement d'excellentes performances sur un modèle de régression qui prévoit correctement les valeurs nutritionnelles d'aliments connus, donc des choses « prévisibles ». ¨Pour réussir à avoir une qualité de prédiction convenable sur des

aliments inconnus, il est indispensable que les experts fournissent un plus large éventail de données, fortement diversifié. Avec cela, l'amélioration de notre modèle devrait être relativement rapide.

C'est pourquoi, nous avions pensé à d'autres pistes à explorer par la suite :

- Implémentation d'apprentissage profond dans le modèle (voir avec Vincent Guigue) :
  - Système de recommandation
  - LLM pour intégrer des variables textuelles au sein de modèles et simplifier l'interface Homme/Machine (que les utilisateurs aient juste à rentrer le nom des aliments par exemple)

6. Ordonnancement et planning prévisionnel

Ci-dessous différents axes à explorer pour l'avancement du projet :

Axe 1 - Analyse et compréhension des données + 1er test de modèle

Exploration approfondie du jeu de données fourni : corrélations globales et par classes / sous-classes, ACP, visualisation des structures et premières hypothèses. Puis premier test d'un modèle XGBoost Regressor qui servira de point de référence pour la suite du projet.

Ces premières investigations ont permis d'identifier la redondance entre variables, la cohérence des classes et la facilité d'apprentissage du modèle.

Axe 2 - Test de la robustesse du modèle

Évaluer la capacité à généraliser au-delà des aliments déjà présents dans la base. On souhaite vérifier que le modèle capture réellement la structure nutritionnelle, et non par simple répétition/ proximité.

- Tests de validation par sous-classes (leave-one-out-group-out).
- Splits structurés par classes, familles chimiques …
- Analyse d'erreurs selon la nature des aliments.

Axe 3 - Réduction de la dimension par sélection de variables

Identifier le plus petit set de variables d'entrée permettant de maintenir les performances du modèle, afin de proposer une version du modèle utilisable en production ou par des non-spécialistes avec un nombre limités de variables.

- Reprendre l'analyse des corrélations (matrices de corrélations, ACP)
- Tests de modèles entraînés sur les combinaisons réduites de variables

Axe 4 - Extension et enrichissement du jeu de données

Pour permettre au modèle de prédire des aliments atypiques ou peu documentés et ainsi améliorer la capacité de généralisation du modèle sur des aliments hors catalogue.

- Obtention de nouvelles classes/sous-classes auprès de M. Tran
- Intégration de produits plus variés (céréales secondaires, produits rares, mélanges …)
- Réflexion sur des stratégies d'augmentation de données.

Axe 5 - Approfondissements possibles

Plusieurs autres pistes d'explorations après celles ci-dessus sont envisagées :

- Système de recommandation nutritionnelle.
- Intégration de descriptions textuelles d'aliments via des modèles multimodaux (CLIP, embedding textuels).
- Développement d'une interface simple permettant de saisir les caractéristiques d'un aliment et d'obtenir automatiquement ses valeurs nutritionnelles estimées.

7. Organisation prévue au sein du groupe d'étudiants

Le projet se décompose en modules indépendants, il est donc possible de les scinder en développements autonomes.

Présentement, avec les deux fonctionnalités considérées pour la suite, un sous-groupe sera chargé de développer le modèle de réseau de neurones pour le système de recommandation et un autre sous-groupe développera le LLM permettant la simplification de l'interface Homme-Machine.