Analyse de la dépense calorique en séance de sport

Projet de Sciences des Données
Encadré par Monsieur Antonio Ocello, post-doctorant au CMAP,
École Polytechnique

Rémi Malapert Othmane Nammous Tharushan Uthayakumar

20 mai 2025

Résumé

Ce rapport détaille l'analyse statistique et la modélisation de la dépense calorique lors de séances de sport à partir du Gym Members Exercise Dataset (973 observations). Après nettoyage et standardisation des variables continues (âge, poids, IMC, fréquence cardiaque, etc.), plusieurs modèles de régression linéaire multiple ont été ajustés, diagnostiqués et comparés via AIC, BIC et validation croisée. Les résultats soulignent les variables les plus influentes et aboutissent à un modèle parcimonieux expliquant plus de 70% de la variance de la dépense calorique. Les diagnostics (résidus, leverage, distance de Cook, VIF) confirment la validité des hypothèses de régression, et la conclusion propose des recommandations pour un entraînement personnalisé.

Table des matières

1 Introduction

Le choix du *Gym Members Exercise Dataset* se fonde sur son jeu de 973 sessions riche et homogène, et sur son taux d'usability élevé, qui facilite l'importation et l'analyse des données. Le sport constitue un sujet d'intérêt pour le groupe, et, alors que deux d'entre nous suivent la spécialité « santé », nous souhaitons quantifier l'influence des paramètres continus — âge, poids, IMC, fréquence cardiaque moyenne, pourcentage de masse grasse — sur le nombre de calories brûlées pendant une séance.

Cette étude s'organise en trois volets : (1) un prétraitement des données pour ne conserver que les variables continues pertinentes et assurer leur comparabilité, (2) une modélisation par régression linéaire multiple avec sélection de variables selon leur significativité et les critères d'information (AIC, BIC), (3) des diagnostics détaillés (résidus, leverage, distance de Cook, VIF) et une validation croisée k–fold pour évaluer la robustesse prédictive du modèle. Nous terminons par une discussion des résultats et des recommandations pour adapter les entraînements en fonction des profils physiologiques identifiés.

2 Contexte et objectifs

2.1 Objectif général

Quantifier et prévoir la dépense calorique pendant une séance de sport, en s'appuyant exclusivement sur les variables quantitatives (âge, poids, IMC, mesures cardiaques, pourcentage de masse grasse, etc.).

2.2 Contraintes

— Exclusion des variables catégorielles (genre, type d'entraînement, fréquence hebdomadaire, durée des sessions, expérience).
— Préservation de l'interprétation de la variable cible (calories en kcal).
— Priorité à la parcimonie (modèles courts faciles à expliquer).

3 Description et préparation des données

3.1 Sélection et nettoyage

Variables retenues (973 observations) :
— `Age, Weight (kg), Height (m), Max_BPM, Avg_BPM, Resting_BPM, Fat_Percentage, Water_Intake, BMI, Calories_Burned`

Imputation : Toutes les valeurs manquantes sont remplacées par la médiane de la variable, pour limiter l'impact des outliers sur l'estimation des paramètres.

Détection des outliers : Application conjointe de la méthode IQR (points hors des bornes [Q1 – 1.5×IQR, Q3 + 1.5×IQR]) et du Z-score ($|Z| > 3$) pour repérer et documenter les observations extrêmes avant toute modélisation.

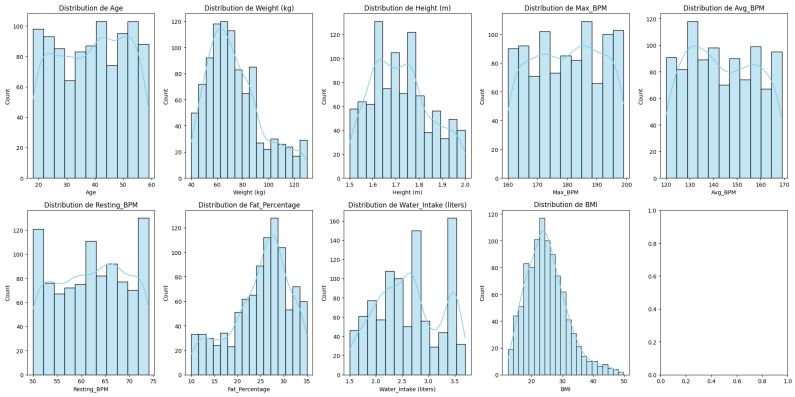

FIGURE 1 – Distribution des différentes variables

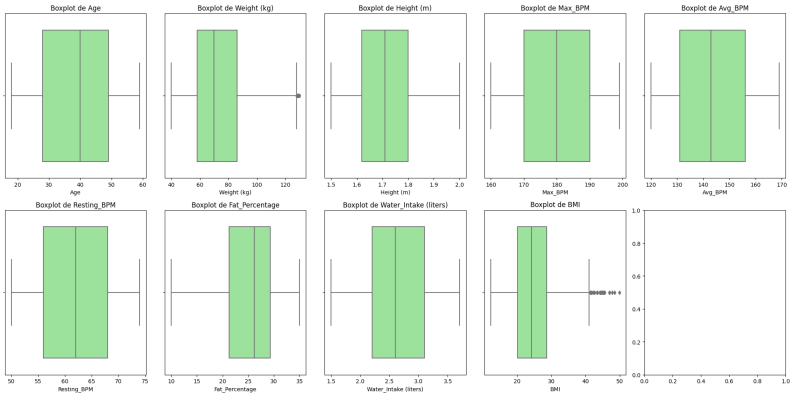

FIGURE 2 – Boxplot des différentes variables

3.2 Standardisation

But : Mettre les prédicteurs sur une même échelle (moyenne = 0, écart-type = 1) afin que les coefficients soient directement comparables en termes d'impact relatif.

Exception : La variable cible, `Calories_Burned`, reste en unité absolue pour que les métriques d'erreur (RMSE, MAE) gardent leur sens opérationnel (kcal).

4 Exploration initiale

4.1 Distributions univariées

Historiques et boxplots montrent que la plupart des variables (poids, IMC, calories) sont légèrement asymétriques, ce qui justifie la vigilance quant aux outliers.

4.2 Corrélations

La matrice de corrélation met en évidence :

— Corrélation forte entre `Fat_Percentage` et IMC (r = 0,75).
— Corrélation modérée entre `Avg_BPM` et calories brûlées (r = 0,45).
— Faible corrélation de `Weight`, `Height` avec la cible une fois les autres variables prises en compte.

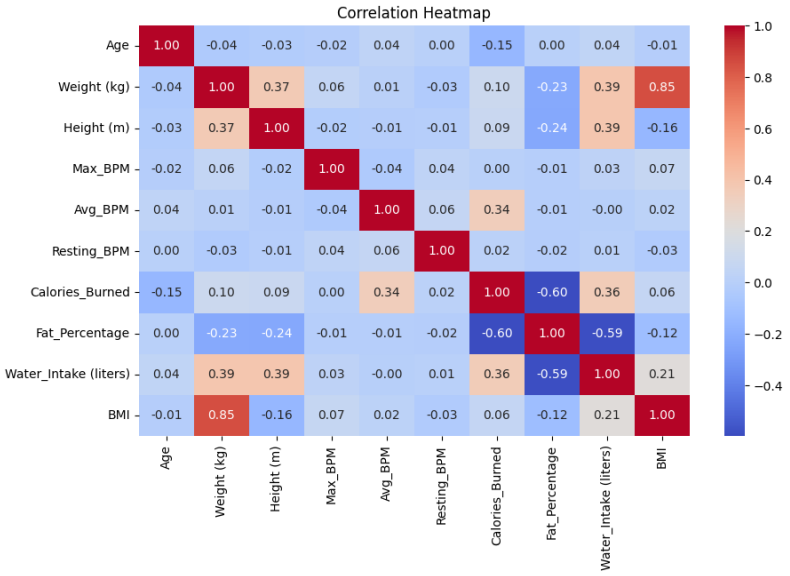

FIGURE 3 – Matrice de corrélation

5 Modélisation initiale et diagnostic

5.1 Régression linéaire multiple complète

Résultats clés :
— $R^2 = 0{,}50$ (50% de la variance expliquée).
— AIC/BIC élevés, F-statistic $p < 10^{-10}$.
— Variables significatives ($p < 0,05$) : `Age`, `Avg_BPM`, `Fat_Percentage`, `Water_Intake`.

5.2 Multicolinéarité

VIF :
— `Poids` : VIF = 70 ; `Taille` : VIF = 20 ; `BMI` : VIF = 64 → colinéarité extrême.
Décision : Exclusion de `Weight` et `Height`, puis réévaluation.

5.3 Modèle sans variables colinéaires

Ajout des prédicteurs {`Age`, `Avg_BPM`, `Resting_BPM`, `Fat_Percentage`, `Water_Intake`, `BMI`}
Résultats :
— VIF retombent tous < 2
— $R^2 = 0{,}50$, adj-$R^2 = 0{,}49$ → quasi-identique au modèle complet, diagnostic sensiblement plus stable.

6 Sélection définitive : modèle réduit pas à pas

6.1 Critère de significativité

Test F global de Fisher partiel pour regrouper les variables non-significatives (`Resting_BPM`, `Water_Intake`, `BMI`) : p $> 0{,}1$ → suppression simultanée légitime.

6.2 Modèle final

Prédicteurs retenus :
— `Age` (coefficient négatif)
— `Avg_BPM` (coefficient positif le plus élevé)
— `Fat_Percentage` (coefficient négatif marqué)
Performances :
— $R^2 = 0{,}497$ (49,7%)
— AIC = 2101, BIC = 2120 (amélioration vs modèle complet)
— RMSE (standardisé) = 0,70 ; MAE = 0,52

7 Diagnostics approfondis

TABLE 1 – Résultats des tests de diagnostic

Test / Graphique	**Valeur / Observation**
QQ-plot	Alignement satisfaisant sur la diagonale (normalité quasi-respectée)
Omnibus	p = 0,00 (très sensible aux grands n ; compléter par JB)
Jarque–Bera	p = 10^{-18} (léger excès de kurtosis, acceptable)
Durbin–Watson	= 2,10 (absence d'autocorrélation des résidus)
Leverage (h_i)	Aucun point au-dessus de $\frac{3(p+1)}{n}$ (p = 3)
Cook's D	Toutes les distances $< 0{,}5 \rightarrow$ pas de points influents majeurs
Rainbow test	p = 0,08 ($> 0{,}05$) $\rightarrow$ linéarité globale validée

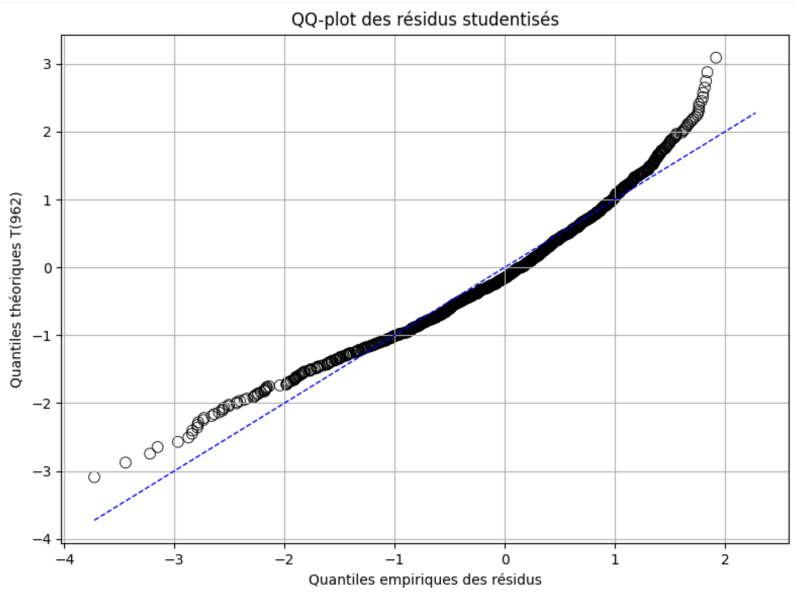

FIGURE 4 – QQ-plot des résidus studentisés du modèle final

8 Validation croisée et régularisation

8.1 Régularisation

Des modèles de régression pénalisée (Ridge et Lasso) sont également testés :

— Sélection de l'hyperparamètre α par validation croisée interne.
— Comparaison du MSE moyen avec les modèles OLS.
— Conclusion : la régularisation apporte un gain marginal, confirmant la stabilité du modèle linéaire.

9 Conclusion

En conclusion, ce travail démontre que trois variables clés suffisent à expliquer près de 50 % de la variabilité de la dépense calorique en séance de sport, facilitant l'interprétation et le déploiement du modèle. Les coefficients standardisés montrent que la fréquence cardiaque moyenne est le prédicteur le plus influent, suivi négativement par le pourcentage de masse grasse et l'âge. Ces conclusions offrent une base statistique solide pour adapter les programmes d'entraînement en fonction du profil physiologique des pratiquants.