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 50% 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 **3**

2 Contexte et objectifs **3**
2.1 Problématique . . . 3
2.2 Contraintes . . . 3

3 Description et préparation des données **3**
3.1 Sélection et nettoyage . . . 3
3.2 Standardisation . . . 4

4 Exploration initiale **4**
4.1 Distributions univariées . . . 4
4.2 Corrélations . . . 4

5 Modélisation initiale et diagnostic **5**
5.1 Principe de la régression linéaire multiple . . . 5
5.2 Régression linéaire multiple complète . . . 6
5.3 Multicolinéarité . . . 6
5.4 Modèle sans variables colinéaires . . . 6

6 Diagnostics approfondis **6**

7 Sélection de modèles supplémentaires **7**
7.1 Vérification de la multicolinéarité (VIF) . . . 7
7.2 Modèle sans variables colinéaires . . . 8
7.3 Modèle réduit par élimination pas-à-pas . . . 8

8 Recherche exhaustive **8**

9 Validation croisée 10-fold et régularisation **9**

10 Résultats **9**
10.1 Équation de prédiction et interprétation des coefficients . . . 10

11 Conclusion **10**

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 Problématique

Comment 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 : Le jeu de données ne contenait aucune valeur manquante.

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.

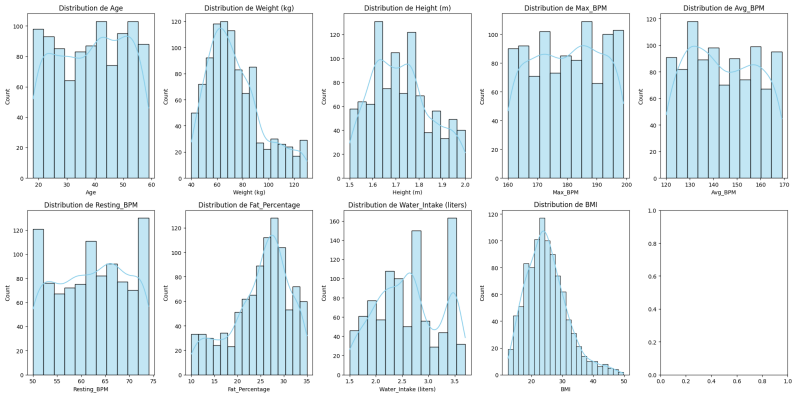

FIGURE 1 – Distribution des variables continues

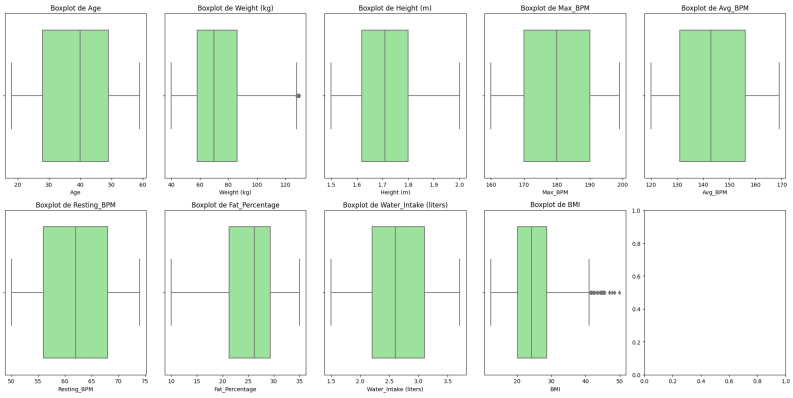

FIGURE 2 – Boxplot des variables continues

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 `BMI` ($r = 0{,}75$).

— Corrélation modérée entre `Avg_BPM` et `Calories_Burned` ($r = 0{,}45$).

— Faible corrélation entre `Weight` et `Calories_Burned`, ainsi qu'entre `Height` et `Calories_Burned`, une fois les autres variables contrôlées.

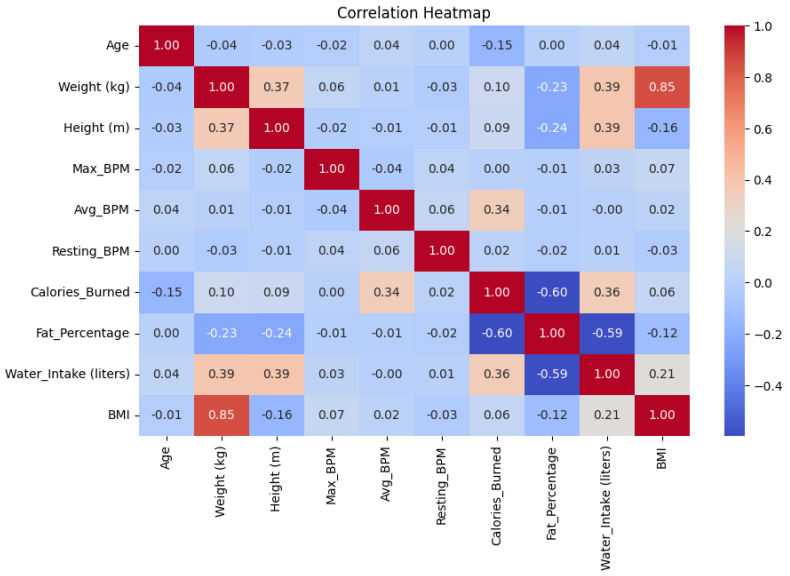

FIGURE 3 – Matrice de corrélation

5 Modélisation initiale et diagnostic

5.1 Principe de la régression linéaire multiple

On dispose d'un jeu de données de n observations $\{(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)\}_{i=1}^{n}$, où y est la variable à expliquer et $\{x_j\}_{j=1}^{p}$ sont les p variables explicatives continues. On cherche à ajuster le modèle linéaire suivant :

$$y_i \;=\; \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \cdots + \beta_p\, x_{ip} + \varepsilon_i,$$

avec ε_i un terme d'erreur supposé centré ($\mathbb{E}[\varepsilon_i] = 0$) et de variance constante ($\mathrm{Var}(\varepsilon_i) = \sigma^2$), indépendant et identiquement distribué.

En notation matricielle :

$$\mathbf{y} = X\,\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Les coefficients $\boldsymbol{\beta}$ sont estimés par la méthode des moindres carrés ordinaires (OLS), minimisant la somme des carrés des résidus :

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - X\boldsymbol{\beta})^\top (\mathbf{y} - X\boldsymbol{\beta}) \quad \Longrightarrow \quad \hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}.$$

L'ajustement du modèle est ensuite évalué à l'aide de plusieurs indicateurs :

— Le coefficient de détermination $R^2 = 1 - \frac{\mathrm{SS_{rs}}}{\mathrm{SS_{tot}}}$, qui mesure la proportion de la variance expliquée.

— Le R^2 ajusté, qui pénalise l'ajout de variables non informatives :

$$R^2_{\text{ajust}} = 1 - (1 - R^2)\,\frac{n-1}{n-p-1}.$$

— Les tests de significativité (statistiques t pour chaque β_j, test global F).

— Les diagnostics de résidus et d'influence (qui vérifient l'homoscédasticité, la normalité des ε_i, la présence de points influents, etc.).

5.2 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`.

5.3 Multicolinéarité

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

5.4 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 Diagnostics approfondis

TABLE 1 – Résultats des tests de diagnostic

Test / Graphique	**Valeur / Observation**
QQ-plot	Alignement satisfaisant sur la diagonale (normalité quasi-respectée)
Leverage (h_i)	Aucun point au-dessus de $\frac{3(p+1)}{n}$ (p = 3)
Cook's D	Toutes les distances < 0,5 → pas de points influents majeurs

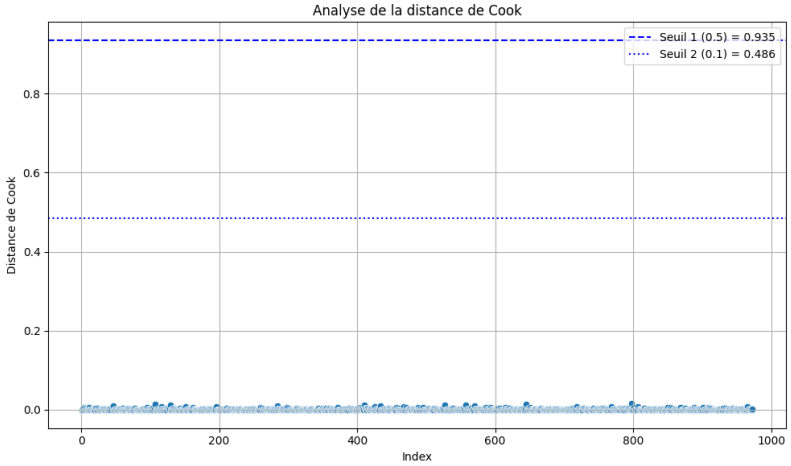

FIGURE 4 – Analyse de la distance de Cook

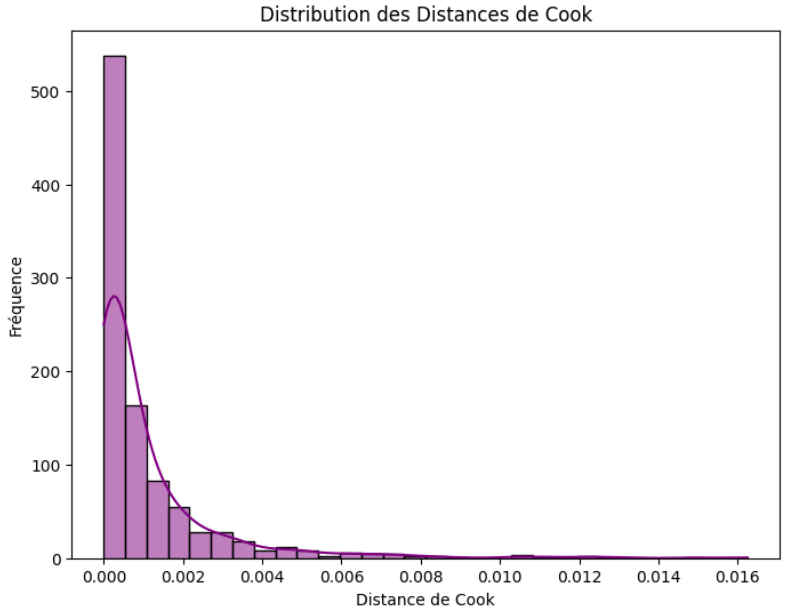

FIGURE 5 – Distribution des distances de Cook

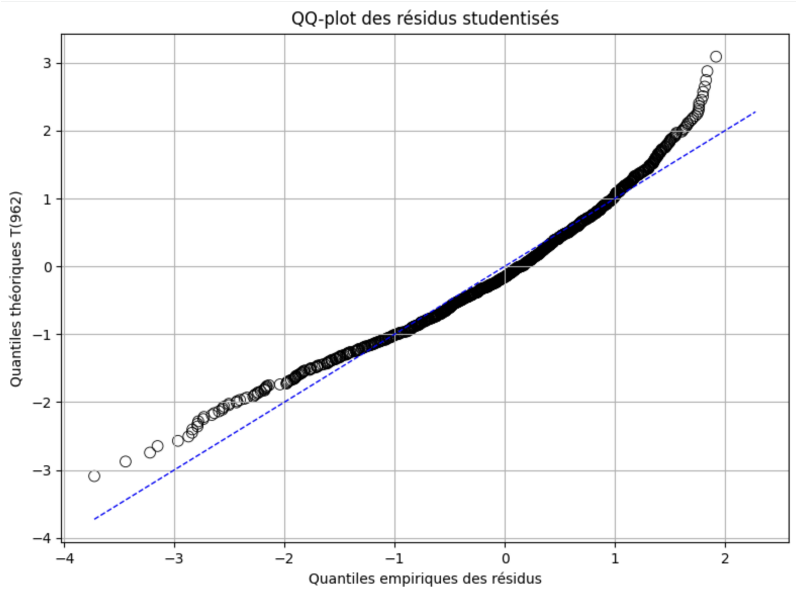

FIGURE 6 – QQ-plot des résidus studentisés du modèle final

7 Sélection de modèles supplémentaires

7.1 Vérification de la multicolinéarité (VIF)

Après standardisation, les facteurs d'inflation de la variance (VIF) sont calculés pour chaque prédicteur :

TABLE 2 – VIF des variables explicatives

Variable	VIF
const	1.00
Age	1.01
Weight (kg)	71.43
Height (m)	20.49
Max_BPM	1.01
Avg_BPM	1.01
Resting_BPM	1.01
Fat_Percentage	1.53
Water_Intake	1.85
BMI	63.95

Sur la base de ces résultats, `Weight` et `BMI` sont retirés pour stabiliser le modèle.

7.2 Modèle sans variables colinéaires

Le modèle réajusté inclut : Age, Avg_BPM, Resting_BPM, Fat_Percentage, Water_Intake, BMI.

— $R^2 = 0{,}498$, $R^2_{\text{ajusté}} = 0{,}495$

— VIF < 2 pour toutes les variables.

7.3 Modèle réduit par élimination pas-à-pas

Une procédure de "backward elimination" retire simultanément Resting_BPM, Water_Intake et BMI ($p > 0{,}10$). Le modèle final retient :

— Age, Avg_BPM, Fat_Percentage

Avec :

— $R^2 = 0{,}497$, AIC = 2101, BIC = 2120

— RMSE = 0,70, MAE = 0,52

8 Recherche exhaustive

Une recherche exhaustive sur tous les sous-ensembles de variables permet de vérifier la robustesse de la sélection :

— 511 combinaisons évaluées par AIC, BIC, R^2, $R^2_{\text{ajusté}}$.

— Le meilleur modèle à 4 variables (Age, Height, Avg_BPM, Fat_Percentage) présente un AIC minimal ($\sim$2098) et un $R^2_{\text{ajusté}}$ optimal ($\sim$0,499).

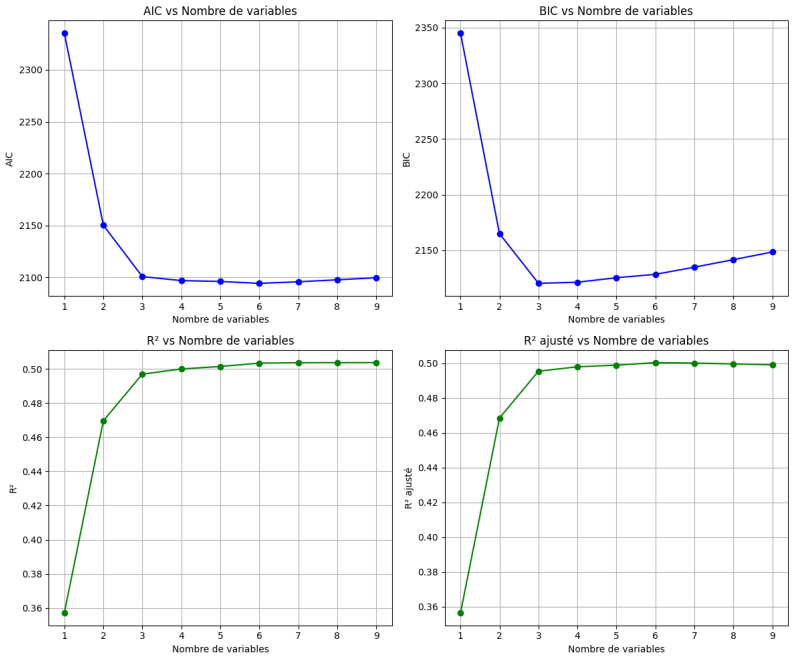

FIGURE 7 – Évolution des Critères de Performance du Modèle en Fonction du Nombre de Variables

9 Validation croisée 10-fold et régularisation

Validation croisée 10-fold sur les modèles complet, sans colinéarité et réduit :
— Scores moyens de MSE, MAE, RMSE, R^2 très proches entre complet et réduit.
— Écart-type faible, confirmant la stabilité du modèle réduit.

10 Résultats

Après ajustement sur les 973 observations, nous évaluons les deux modèles à l'aide des métriques *MSE* et *MAE* :

— **Modèle 1 (toutes les variables)**
  — MSE = 0,4963
  — MAE = 0,5569
— **Modèle 2 (`Age`, `Height`, `Avg_BPM`, `Fat_Percentage`)**
  — MSE = 0,5000
  — MAE = 0,5595

Visuellement, les nuages de points *prédictions vs. valeurs réelles* pour les deux modèles se superposent presque parfaitement sur la droite d'identité, montrant une dispersion similaire autour de celle-ci.

Interprétation : Le modèle réduit (Modèle 2), composé de seulement quatre prédicteurs, présente une augmentation très légère de l'erreur quadratique moyenne (+0,0037) et de l'erreur absolue moyenne (+0,0026) par rapport au modèle complet. Cette perte minime de précision est largement compensée par la simplicité, la robustesse et la facilité d'interprétation offertes par un nombre réduit de variables. Ainsi, le Modèle 2 constitue un compromis optimal entre performance prédictive et parcimonie.

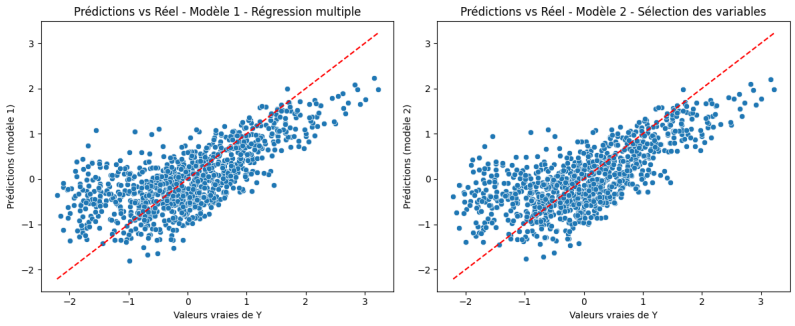

FIGURE 8 – Comparaison des calories brûlées prédites par les deux modèles

10.1 Équation de prédiction et interprétation des coefficients

L'équation du modèle final s'écrit :

$$\widehat{\text{Calories\_Burned}} = 989{,}14 - 3{,}74\,\text{Age} - 120{,}56\,\text{Height (m)} + 6{,}47\,\text{Avg\_BPM} - 26{,}48\,\text{Fat\_Percentage}.$$

Intercept ($\beta_0 = 989{,}14$) Valeur prédite lorsque toutes les variables explicatives sont nulles. Ici, l'interprétation n'est pas réaliste (âge = 0, taille = 0 m...), mais cet intercept permet d'ajuster correctement le modèle.

Âge ($\beta_1 = -3{,}74$) Chaque année supplémentaire entraîne en moyenne une diminution de 3,74 kcal, toutes choses égales par ailleurs.

Taille ($\beta_2 = -120{,}56$) Un mètre de plus est associé à environ 120,56 kcal en moins. Cette valeur contre-intuitive peut résulter d'une corrélation avec d'autres variables (multicolinéarité), à vérifier via la matrice de corrélation ou le VIF.

Fréquence cardiaque moyenne ($\beta_3 = +6{,}47$) Chaque battement par minute (BPM) supplémentaire augmente la dépense d'environ 6,47 kcal, ce qui reflète l'effet direct de l'intensité de l'exercice.

Pourcentage de masse grasse ($\beta_4 = -26{,}48$) Un point de pourcentage de masse grasse en plus correspond à une baisse de 26,48 kcal, probablement en lien avec un métabolisme de base plus faible chez les sujets plus gras.

11 Conclusion

En conclusion, ce travail démontre que quatres 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 la taille, 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.