Modelo predictivo basado en técnicas avanzadas de aprendizaje automático para la estimación precisa y personalizada del gasto calórico en actividades físicas

Andrés José Peña Gativa

Asesor

Felipe Alexander Pipicano Guzmán

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Resumen

Los métodos convencionales de estimación calórica, ecuaciones generalizadas y equivalentes metabólicos presentan limitaciones estructurales al ignorar el perfil fisiológico individual del usuario. Este proyecto desarrolla y evalúa un modelo predictivo basado en técnicas avanzadas de aprendizaje automático para la estimación precisa y personalizada del gasto calórico durante actividades físicas, desde el análisis exploratorio hasta el despliegue de un sistema interactivo.

Sobre el dataset público *gym_members_exercise_tracking* (Khorasani, 2023), se evaluaron cinco algoritmos de aprendizaje automático supervisado bajo tres estrategias arquitectónicas mediante validación cruzada K-Fold. El modelo seleccionado, XGBoost con el nivel de experiencia codificado como variable predictora, fue optimizado mediante búsqueda bayesiana de hiperparámetros con Optuna TPE, reduciendo el MAE de 19.79 a 9.22 kcal (−53.4 %, $R^2 = 0.9979$). La explicabilidad se instrumentó con SHAP TreeExplainer, particionada por segmento fisiológico. El sistema se despliega mediante una arquitectura FastAPI + Streamlit que integra predicción puntual, intervalos cuantílicos calibrados al 80 % y contribuciones SHAP por variable.

Los resultados demuestran la viabilidad técnica del enfoque, aunque la naturaleza semisintética del dataset limita la validez externa del modelo sobre datos biométricos reales.

***Palabras clave*:** Gasto calórico, XGBoost, optimización bayesiana, SHAP, aprendizaje automático supervisado, dashboard interactivo.

Abstract

Conventional methods for estimating calorie expenditure, generalized equations, and metabolic equivalents have structural limitations because they ignore the individual physiological profile of the user. This project develops and evaluates a predictive model based on advanced machine learning techniques for the accurate and personalized estimation of calorie expenditure during physical activity, from exploratory analysis to the deployment of an interactive system.

Using the public dataset gym_members_exercise_tracking (Khorasani, 2023), five supervised machine learning algorithms were evaluated under three architectural strategies using K-fold cross validation. The selected model, XGBoost with experience level coded as a predictor variable, was optimized using Bayesian hyperparameter search with Optuna TPE, reducing the MAE from 19.79 to 9.22 kcal (−53.4%, $R^2 = 0.9979$). Explainability was assessed using SHAP TreeExplainer, partitioned by physiological segment. The system is deployed using FastAPI + Streamlit architecture that integrates point prediction, quantile intervals calibrated to 80%, and SHAP contributions per variable.

The results demonstrate the technical feasibility of the approach, although the semi-synthetic nature of the dataset limits the external validity of the model on real biometric data.

Keywords**:** Calorie expenditure, XGBoost, Bayesian optimization, SHAP, supervised machine learning, interactive dashboard.

Tabla de Contenido

Introducción 9

Planteamiento del Problema 10

Justificación 12

Objetivos 13

Objetivo General 13

Objetivos Específicos 13

Marco de Referencia 14

Fundamentos Teóricos 14

Revisión de Literatura 15

Bases Conceptuales 16

Metodología 19

Enfoque de la Investigación 19

Diseño de la Investigación 19

Tipo y Alcance de la Investigación 19

Conjunto de Datos 20

Variables de Estudio 21

Herramientas y Entorno Tecnológico 22

Procedimiento General del Pipeline 23

Arquitectura de la Implementación 24

Análisis Exploratorio de Datos (EDA) 25

Segmentación y Entrenamiento de Modelos de Aprendizaje Automático. 43

Optimización de Hiperparámetros con Optuna y Explicabilidad SHAP 57

Dashboard Interactivo para la Estimación del Gasto Calórico 69

Limitaciones.....80

Conclusiones.....82

Recomendaciones 84

Referencias Bibliográficas.....86

Lista de Tablas

Tabla 1 *Resumen de las Variables de Estudio* 21

Tabla 2 *Stack Tecnológico y Versiones del Entorno de Desarrollo* 22

Tabla 3 *Resumen de Inspección Inicial del Dataset* 25

Tabla 4 *Clasificación de las Variables del Dataset* 25

Tabla 5 *Estadísticas Descriptivas de las Principales Variables Numéricas* 26

Tabla 6 *Distribución de Frecuencias de Variables Categóricas y Ordinales* 28

Tabla 7 *Transformaciones Aplicadas en la Etapa de Preprocesamiento* 40

Tabla 8 *Estadísticas Descriptivas de Calories_Burned por Nivel de Experiencia* 46

Tabla 9 *Pruebas Mann-Whitney U y d de Cohen por Pares de Niveles (Calories_Burned)* 47

Tabla 10 *Métricas de Validación Cruzada – Estrategia A: Modelo Global (K-Fold, k = 5)* 49

Tabla 11 *Métricas – Estrategia B y Variación Respecto a Estrategia A* 50

Tabla 12 *Métricas OOF Globales – Estrategia C: Modelos Especializados por Nivel* 51

Tabla 13 *Diagnóstico de Sesgo Sistemático por Nivel – Residuos XGBoost (Estrategia B)* 53

Tabla 14 *Espacio de Búsqueda de Hiperparámetros XGBoost Definido para Optuna* 57

Tabla 15 *Hiperparámetros XGBoost Óptimos Optuna e Importancia fANOVA* 59

Tabla 16 *Validación Cruzada K-Fold (k = 5) – Baseline vs. Optuna, Estrategia B* 60

Tabla 17 *Ranking de Importancia SHAP Global* 62

Tabla 18 *Importancia SHAP por Nivel de Experiencia* 64

Tabla 19 *Estadísticos de Residuos OOF Nivel de Experiencia – Modelo XGBoost Optimizado* 66

Tabla 20 *Stack Tecnológico del Sistema* 69

Tabla 21 *Payload de Entrada y Respuesta Completa del Endpoint POST /Predict* 76

Lista de Figuras

Figura 1 *Flujo Metodológico Secuencial del Proyecto* 24

Figura 2 *Histogramas con KDE: Variables Antropométricas y Cardíacas* 29

Figura 3 *Histogramas con KDE: Variables Cardíacas y de Sesión* 30

Figura 4 *Histogramas con KDE: Variables de Composición y Frecuencia* 30

Figura 5 *Distribución de Calories_Burned* 31

Figura 6 *Distribución de Frecuencias para Género y Tipo de Entrenamiento* 32

Figura 7 *Distribución de Frecuencias Nivel de Experiencia y Frecuencia de Entrenamiento Semanal* 33

Figura 8 *Distribución de Calories_Burned por Nivel de Experiencia (Boxplot)* 33

Figura 9 *Matriz de Correlación de Pearson* 35

Figura 10 *Ranking de Correlación de Pearson con Calories_Burned* 37

Figura 11 *Scatter Plots Predictores Principales vs. Calories_Burned por Género* 38

Figura 12 *Heatmap Calories_Burned Promedio × Workout_Type × Experience_Level* 39

Figura 13 *Distribución Calories_Burned por Nivel de Experiencia* 46

Figura 14 *MAE y R^2 por Estrategia y Modelo* 51

Figura 15 *MAE por Nivel de Experiencia en la Estrategia C (Modelos Especializados)* 52

Figura 16 *Diagnóstico de Residuos – XGBoost, Estrategia B* 54

Figura 17 *Comparativa MAE Pre y Post Optimización Optuna – XGBoost, Estrategia B* 61

Figura 18 *Importancia SHAP Global – Modelo XGBoost Optimizado, Estrategia B* 63

Figura 19 *Importancia SHAP por Feature y Nivel de Experiencia* 65

Figura 20 *Cobertura Empírica y Amplitud Media del Intervalo 80 % por Nivel de Experiencia* 70

Figura 21 *Popup Metodológico (@st.dialog)* 72

Figura 22 *Popup Metodológico (@st.dialog) Segunda Parte* 72

Figura 23 *Formulario de Inputs (7 Controles)* 73

Figura 24 *Panel de Predicción* 74

Figura 25 *Panel de Explicabilidad SHAP – Waterfall* 75

Figura 26 *Panel de Explicabilidad SHAP – Summary Bar* 76

Figura 27 *Waterfall SHAP – Ejemplo de Predicción* 78

Introducción

La convergencia entre ciencias del deporte, fisiología del ejercicio y analítica de datos ha abierto posibilidades concretas para personalizar el monitoreo del rendimiento físico. Estimar con precisión cuánta energía consume una persona durante una sesión de entrenamiento adaptada al perfil fisiológico del usuario, tiene implicaciones directas en la planificación nutricional, el control del peso corporal y la adherencia a programas de actividad física. Aunque el problema puede formularse de manera sencilla, su resolución requiere modelar múltiples variables fisiológicas interdependientes.

Los métodos tradicionales fallan en ese punto crítico: la personalización. Fórmulas y tablas de METs producen estimaciones válidas para una población promedio, pero imprecisas para un individuo concreto cuya frecuencia cardíaca, composición corporal o nivel de experiencia difieren del perfil de referencia. La tecnología wearable reduce esa brecha, pero a costa de algoritmos no auditables y validez externa no garantizada.

Este proyecto propone una respuesta basada en aprendizaje automático supervisado. Se desarrolló y evaluó un modelo predictivo que abarca desde el análisis exploratorio de datos hasta el despliegue de un sistema interactivo de estimación calórica. El modelo seleccionado, XGBoost con nivel de experiencia codificado como variable predictora (Estrategia B), fue optimizado mediante búsqueda bayesiana de hiperparámetros (Optuna TPE) y complementado con explicabilidad SHAP por nivel fisiológico. El sistema se expone al usuario final a través de una arquitectura FastAPI + Streamlit con intervalos de predicción cuantílicos calibrados.

Planteamiento del Problema

La estimación del gasto calórico durante la actividad física es una necesidad práctica en contextos de entrenamiento, nutrición y salud preventiva. Sin embargo, los métodos disponibles más extendidos como ecuaciones de Harris-Benedict o el uso de equivalentes metabólicos (METs) generalizados, fueron diseñados para estimaciones poblacionales, no individuales. Ignoran variables dinámicas como la intensidad real del esfuerzo, la frecuencia cardíaca o el nivel de condición física del usuario, produciendo estimaciones que pueden desviarse en rangos clínicamente significativos respecto al gasto energético real (Keytel et al., 2005).

Esta imprecisión tiene consecuencias concretas. En programas de entrenamiento, un error sistemático en la estimación calórica compromete el balance energético calculado, desvía los ajustes nutricionales y puede generar adherencia a rutinas inadecuadas para el perfil del usuario. Los dispositivos wearables comerciales intentan abordar este problema, pero sus algoritmos son propietarios, no validados públicamente y de baja adaptabilidad a perfiles fisiológicos heterogéneos (Miah et al., 2022).

El avance de la ciencia de datos y el aprendizaje automático ofrece una vía alternativa: construir modelos predictivos entrenados sobre datos fisiológicos reales, capaces de ajustarse al perfil individual y de operar con variables fácilmente obtenibles como peso, edad, frecuencia cardíaca, duración de la sesión y nivel de experiencia. No obstante, la mayoría de los desarrollos reportados en la literatura corresponden a prototipos de laboratorio sin arquitecturas funcionales de despliegue, mecanismos de explicabilidad ni cuantificación de la incertidumbre predictiva (Priscilla et al., 2024; Salanke & Sathyajeeth, 2024). Existe, por tanto, una brecha entre la capacidad técnica demostrada y la disponibilidad de sistemas usables, transparentes y metodológicamente rigurosos.

Este proyecto se sitúa en esa brecha. Se plantea el desarrollo y evaluación de un modelo predictivo basado en técnicas avanzadas de aprendizaje automático, desde el análisis exploratorio de datos hasta el despliegue de un dashboard interactivo que integre predicción puntual, intervalos de incertidumbre y explicabilidad por contribución de variables, sobre el conjunto de datos utilizado.

¿En qué medida un modelo predictivo basado en XGBoost con optimización bayesiana de hiperparámetros permite estimar el gasto calórico durante sesiones de actividad física a partir de variables fisiológicas y de entrenamiento accesibles, y qué tan determinante resulta el nivel de experiencia del usuario sobre la precisión de dicha estimación?

Justificación

Las soluciones actuales de estimación calórica presentan limitaciones relacionadas con interpretabilidad, accesibilidad y robustez metodológica. Desde el punto de vista científico, la literatura reciente respalda el uso de algoritmos de ensamble para tareas de predicción fisiológica. Salanke y Sathyajeeth (2024) y Santhiya et al. (2024) reportan resultados de alta precisión con XGBoost y SVR respectivamente; Priscilla et al. (2024) demuestran que arquitecturas híbridas mejoran la generalización del modelo. Sin embargo, estos desarrollos operan como pruebas de concepto: no incluyen mecanismos de incertidumbre, no exponen la contribución de cada variable a la predicción y carecen de interfaz de uso para el usuario final. Este proyecto aborda esas tres ausencias de forma simultánea.

Desde el punto de vista técnico, la integración de optimización bayesiana de hiperparámetros (Optuna TPE), explicabilidad SHAP por segmento poblacional e intervalos cuantílicos calibrados representa un nivel de sofisticación metodológica poco frecuente en proyectos de esta escala. Cada componente responde a una necesidad concreta: Optuna porque los hiperparámetros por defecto de XGBoost son subóptimos para datasets de tamaño moderado; SHAP porque la interpretabilidad del modelo es tan relevante como su precisión en aplicaciones de salud; los intervalos cuantílicos porque una predicción puntual sin cuantificación de incertidumbre traslada al usuario una falsa sensación de exactitud.

La plataforma desarrollada, funcional, documentada y con advertencia explícita sobre las limitaciones del dataset, constituye un prototipo replicable y extensible sobre datos clínicos reales.

Objetivos

Objetivo General

Desarrollar y evaluar un modelo predictivo basado en técnicas avanzadas de aprendizaje automático para la estimación precisa y personalizada del gasto calórico durante actividades físicas a partir de variables fisiológicas y de entrenamiento.

Objetivos Específicos

Realizar un análisis exploratorio de datos exhaustivo sobre el dataset gym_members_exercise_tracking (Khorasani, 2023), identificando patrones, distribuciones, relaciones estadísticas y variables relevantes que influyen en el gasto calórico.

Implementar modelos predictivos supervisados utilizando algoritmos de Machine Learning como regresión lineal múltiple, árboles de decisión, Random Forest y XGBoost, modelando de manera precisa la relación entre variables fisiológicas y el gasto calórico.

Evaluar el desempeño de los modelos desarrollados, empleando métricas cuantitativas como MAE, RMSE y coeficiente de determinación (R^2), seleccionando el modelo con mayor capacidad predictiva y menor error.

Optimizar el modelo seleccionado mediante técnicas de ajuste de hiperparámetros y validación cruzada, mejorando su generalización, robustez y precisión en datos no vistos.

Diseñar un dashboard interactivo e intuitivo, que permita la visualización dinámica de las predicciones del gasto calórico, facilitando la interpretación de resultados y la interacción del usuario con el sistema de manera clara y eficiente.

Marco de Referencia

Fundamentos Teóricos

Gasto Calórico y su Estimación

El gasto calórico total comprende el metabolismo basal, el efecto térmico de los alimentos y el gasto por actividad física, siendo este último el componente más variable entre individuos (Benton & Young, 2017). Los métodos clásicos de estimación, como la ecuación de Harris-Benedict o los equivalentes metabólicos (METs), fueron diseñados para poblaciones promedio y no capturan la variabilidad individual en composición corporal, condición cardiovascular o nivel de entrenamiento. Esta limitación produce errores clínicamente significativos al aplicarse en contextos de entrenamiento personalizado (Keytel et al., 2005).

Aprendizaje Automático Supervisado en el Rendimiento Humano

Los modelos de aprendizaje automático supervisado construyen funciones matemáticas que mapean variables de entrada (predictoras) a una variable de salida (objetivo) a partir de datos etiquetados. En el contexto del rendimiento físico, esta aproximación permite integrar variables fisiológicas heterogéneas como frecuencia cardíaca, peso, edad, duración del ejercicio en un único modelo adaptado al perfil del individuo, superando las limitaciones de las ecuaciones paramétricas clásicas. Algoritmos de ensamble como Random Forest y XGBoost han demostrado especial efectividad en datos tabulares con interacciones no lineales entre predictores (Chen & Guestrin, 2016).

Tecnología Wearable y su Rol en la Recolección de Datos

Los dispositivos portátiles (*wearables)* como bandas inteligentes, relojes deportivos, pulsómetros, permiten registrar en tiempo real variables como frecuencia cardíaca, aceleración y duración del ejercicio, generando flujos de datos adecuados para entrenar y validar modelos

predictivos de actividad física. Su accesibilidad y portabilidad los posicionan como la fuente de datos más viable para sistemas de estimación calórica en contextos fuera de laboratorio. Sin embargo, la calidad y consistencia de estos datos depende del protocolo de registro y del dispositivo utilizado (Lavanya & Sivaraman, 2024).

Revisión de Literatura

Aunque los trabajos recientes alcanzan desempeños predictivos elevados, persisten limitaciones metodológicas asociadas a explicabilidad, cuantificación de incertidumbre y aplicabilidad práctica, es decir, alcanzan alta precisión en sus contextos de evaluación, pero no implementan los tres componentes que un sistema de uso real requiere, cuantificación de incertidumbre, explicabilidad por variable y arquitectura de despliegue funcional.

Salanke y Sathyajeeth (2024) desarrollaron un modelo con XGBoost que integra frecuencia cardíaca, duración de la actividad, peso y edad, reportando alta precisión en un entorno controlado. Su trabajo constituye el referente más directo al presente proyecto en términos de algoritmo y variables predictoras. Priscilla et al. (2024) emplearon una arquitectura híbrida AutoEncoder + EfficientNet, demostrando que la combinación de técnicas de aprendizaje profundo mejora la capacidad de generalización, aunque a costa de interpretabilidad. Santhiya et al. (2024) confirmaron que algoritmos como SVR y árboles de decisión pueden alcanzar errores mínimos (MAE = 1.48 kcal) en entornos muy controlados.

En cuanto a la personalización por perfil de usuario, Clarinda et al. (2024) propusieron un sistema adaptativo de recomendación nutricional que ajusta dinámicamente sus salidas según el historial del usuario, reforzando la importancia de incorporar variables de nivel de experiencia o adaptación crónica al modelo. Miah et al. (2022) demostraron que datos de dispositivos móviles permiten predecir la aptitud física con precisión superior al 95 %, evidenciando el potencial de

fuentes de datos accesibles. Sobre validación y aplicabilidad, Yi (2024) subraya que los modelos deben ser interpretables, eficientes y adaptativos para tener impacto práctico real, condición que los prototipos anteriores no satisfacen de forma integral.

Ninguno de los trabajos revisados integra simultáneamente: optimización formal de hiperparámetros, explicabilidad por segmento poblacional, cuantificación probabilística del error e interfaz de usuario desplegada. Este proyecto responde exactamente a ese conjunto de ausencias.

Bases Conceptuales

Aprendizaje Automático Supervisado

Rama de la inteligencia artificial que permite a los sistemas aprender automáticamente de los datos y hacer predicciones sin ser explícitamente programados (Halilaj et al., 2018).

Gasto Calórico

Energía total consumida durante un periodo de tiempo, influenciada por factores como metabolismo basal, actividad física y efecto térmico de los alimentos. Es la variable objetivo del sistema predictivo desarrollado (Benton & Young, 2017).

Modelos Supervisados

Técnicas que utilizan datos etiquetados para construir modelos predictivos, incluyendo regresión lineal, árboles de decisión, Random Forest y XGBoost, entre otros.

Métricas de Evaluación (MAE, RMSE, R^2)

El Error Absoluto Medio (MAE) mide la desviación promedio entre predicción y valor real, en las mismas unidades del objetivo. El RMSE penaliza errores grandes por su estructura cuadrática. El coeficiente de determinación (R^2) indica la fracción de varianza del objetivo explicada por el modelo, con valor máximo de 1.0.

Optimización Bayesiana de Hiperparámetros

Estrategia de búsqueda que construye un modelo probabilístico (*surrogate model*) de la función objetivo (por ejemplo, MAE en validación cruzada) para proponer de forma secuencial los hiperparámetros con mayor probabilidad de mejorar el resultado. A diferencia de la búsqueda en grilla o aleatoria, aprovecha la historia de evaluaciones anteriores, resultando más eficiente en espacios de alta dimensionalidad. Optuna implementa esta estrategia mediante el muestreador TPE (*Tree-structured Parzen Estimator*) (Akiba et al., 2019).

SHAP (SHapley Additive exPlanations)

Marco de explicabilidad basado en la teoría de juegos cooperativos que asigna a cada variable predictora una contribución marginal (valor de Shapley, φ_i) a la predicción individual de un modelo. Garantiza consistencia, aditividad y exactitud local, propiedades que otros métodos de importancia de variables no satisfacen simultáneamente. La implementación *TreeExplainer* calcula los valores exactos de Shapley para modelos basados en árboles en tiempo polinomial (Lundberg & Lee, 2017).

Validación Cruzada K-Fold

Técnica de evaluación de modelos que particiona el conjunto de datos en k subconjuntos (*folds*) mutuamente excluyentes. En cada iteración, $k-1$ folds se usan para entrenamiento y 1 para evaluación, rotando hasta que todos los registros han servido como conjunto de prueba exactamente una vez. El resultado final es el promedio de las métricas sobre los k folds, produciendo un estimador de generalización con menor varianza que una partición simple entrenamiento/prueba (Hastie et al., 2009).

Wearables

Dispositivos electrónicos portátiles que registran variables fisiológicas en tiempo real como frecuencia cardíaca, aceleración, duración del ejercicio, facilitando la recolección de datos en contextos naturales de entrenamiento (Lavanya & Sivaraman, 2024).

XGBoost (Extreme Gradient Boosting)

Algoritmo de aprendizaje automático basado en conjuntos de árboles de decisión entrenados de forma secuencial mediante *gradient boosting*. Cada árbol corrige los errores residuales del anterior, optimizando una función de pérdida diferenciable. Su eficiencia computacional, capacidad para manejar valores faltantes y mecanismos de regularización integrados (L1/L2) lo posicionan como uno de los algoritmos de mayor desempeño en tareas de regresión y clasificación supervisada sobre datos tabulares (Chen & Guestrin, 2016).

Metodología

Enfoque de la Investigación

El presente proyecto se desarrolla bajo un enfoque cuantitativo y aplicado, orientado al diseño, implementación y evaluación de un modelo predictivo del gasto calórico individual a partir de variables fisiológicas y de entrenamiento.

Diseño de la Investigación

El diseño corresponde a un esquema no experimental de tipo correlacional-predictivo. Se denomina no experimental porque no se manipulan variables ni se interviene sobre los sujetos de estudio; en cambio, se trabaja sobre un conjunto de datos observacionales ya existente, el dataset *gym_members_exercise_tracking* (Khorasani, 2023), sin alterar las condiciones de registro. La investigación se limita al análisis de relaciones entre variables tal como ocurrieron naturalmente en el contexto de observación original.

El componente correlacional establece y cuantifica la asociación estadística entre las variables fisiológicas y demográficas (predictoras) y la variable objetivo Calories_Burned, mediante análisis de correlación de Pearson y Spearman, pruebas de significancia estadística y análisis de varianza. El componente predictivo trasciende la descripción de relaciones y construye un modelo matemático capaz de estimar el gasto calórico para nuevas observaciones, evaluado bajo métricas de error (MAE, RMSE) y bondad de ajuste (R^2).

Tipo y Alcance de la Investigación

Tipo: Aplicada y tecnológica.

Alcance: Descriptivo, correlacional y predictivo. El componente descriptivo caracteriza el comportamiento estadístico de las variables del dataset. El correlacional evalúa la fuerza y

dirección de las asociaciones entre predictores y gasto calórico. El predictivo desarrolla, optimiza y valida el modelo de estimación.

Conjunto de Datos

El conjunto de datos empleado como material base del proyecto es el dataset *gym_members_exercise_tracking* (Khorasani, 2023), disponible públicamente en la plataforma Kaggle. Reúne registros de miembros de un gimnasio, capturando variables fisiológicas, antropométricas y de comportamiento de entrenamiento.

El dataset contiene 973 registros y 15 variables en su versión preprocesada (14 originales + 1 variable derivada, BMI). Se verificó su completitud: 0 valores nulos y 0 registros duplicados, condición óptima que eliminó la necesidad de etapas de imputación. La distribución de género es prácticamente equitativa (52.5 % masculino / 47.5 % femenino), descartando sesgos de representatividad por sexo. Los cuatro tipos de entrenamiento registrados (Strength, Cardio, Yoga, HIIT) están distribuidos de forma homogénea (~25 % cada uno).

La variable objetivo es *Calories_Burned*, que representa el gasto calórico total en kilocalorías durante una sesión de actividad física (rango: 303 – 1783 kcal; $\mu = 905.4$ kcal; $\sigma = 272.6$ kcal). Cabe señalar un aspecto crítico de esta variable: el análisis de correlación reveló que *Session_Duration* la predice con $r = +0.908$, R^2 univariado = 0.824, un nivel de linealidad atípico para datos biológicos reales. Esta evidencia, coherente con lo reportado por Keytel et al. (2005), indica que *Calories_Burned* fue generada mediante una ecuación de tipo MET (Metabolic Equivalent of Task) en lugar de registros fisiológicos directos. Esta limitación condiciona la validez externa del modelo y se comunica explícitamente en el sistema desarrollado.

Las variables predictoras incluyen: duración de la sesión (*Session_Duration*, en horas), frecuencia cardíaca promedio (*Avg_BPM*), frecuencia cardíaca en reposo (*Resting_BPM*),

frecuencia cardíaca máxima (*Max_BPM*), peso corporal (kg), edad (años), género, porcentaje de grasa corporal, frecuencia semanal de entrenamiento, nivel de experiencia (ordinal: 1 – 3) y tipo de actividad física. Adicionalmente, se construyeron dos variables derivadas: el índice de masa corporal (IMC, calculado a partir de peso y talla) y la frecuencia cardíaca relativa (*FC_relativa* = Avg_BPM / Max_BPM), indicador fisiológico de la intensidad real de cada sesión.

El dataset opera como fuente de datos observacionales secundaria, no fue diseñado específicamente para este proyecto, pero sus características (diversidad de perfiles, balance de género y distribución homogénea de modalidades de entrenamiento) lo hacen adecuado para el objetivo de estimación correlacional-predictiva planteado. El análisis se desarrolló íntegramente sobre esta fuente, sin recolección de datos primarios ni intervención sobre los sujetos.

Variables de Estudio

Tabla 1

Resumen de las Variables de Estudio

Variable dependiente	Gasto calórico (Kcal) durante una sesión de actividad física.
Variables independientes	Nivel de experiencia (Principiante, Intermedio, Avanzado), edad, peso (Kg), género, altura (m), frecuencia cardíaca promedio (BPM), frecuencia cardíaca máxima (BPM), duración de la sesión (min), tipo de actividad física.
Variables Derivadas	Índice de masa corporal (IMC), intensidad estimada (derivada de la FC relativa).

Herramientas y Entorno Tecnológico

El pipeline se desarrolló íntegramente en Python 3.12 sobre el entorno de desarrollo Visual Studio Code, ejecutado en sistema operativo Windows 11. La Tabla 2 detalla las bibliotecas utilizadas con sus versiones exactas, garantizando la replicabilidad del entorno computacional.

Tabla 2

Stack Tecnológico y Versiones del Entorno de Desarrollo

Categoría	Biblioteca / Herramienta	Versión	Función principal
Lenguaje	Python	3.12	Lenguaje base del proyecto
Entorno	Visual Studio Code	–	IDE de desarrollo y ejecución
Manipulación de datos	pandas	2.3.3	Carga, transformación y análisis del dataset
Computación numérica	numpy	2.4.2	Operaciones matriciales y vectoriales
Modelado ML	scikit-learn	1.6.1	Pipeline, K-Fold, métricas, modelos base
Modelado ML	xgboost	3.0.0	Modelo principal y regresión cuantílica
Optimización	optuna	4.7.0	Búsqueda bayesiana de hiperparámetros (TPE)
Explicabilidad	shap	0.51.0	TreeExplainer, valores de Shapley
Análisis estadístico	scipy	1.15.2	Pruebas de hipótesis (Wilcoxon, Kruskal-Wallis)

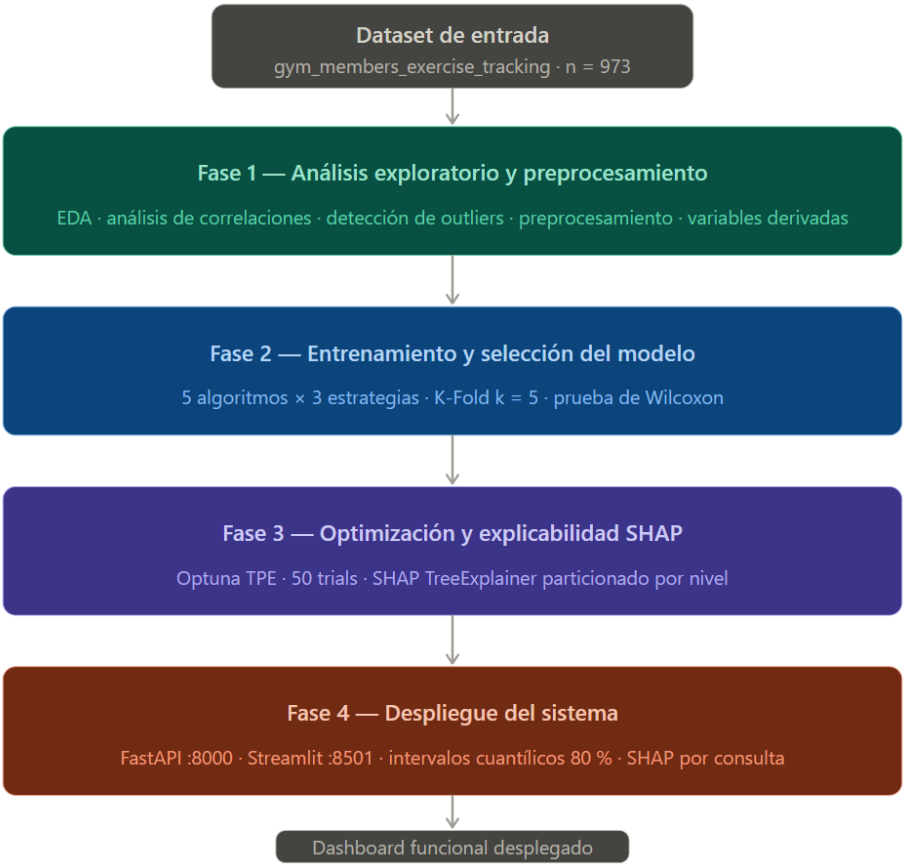
Análisis estadístico	statsmodels	0.14.4	VIF, diagnóstico de residuos
Visualización	matplotlib	3.10.1	Gráficos estáticos, waterfall SHAP
Visualización	seaborn	0.13.2	Mapas de calor, distribuciones, scatter plots
Backend API	fastapi	0.134.0	Endpoints de inferencia y validación Pydantic
Backend API	uvicorn	0.41.0	Servidor ASGI para FastAPI
Frontend	streamlit	1.54.0	Dashboard interactivo
Serialización	joblib	1.4.2	Persistencia de modelos (.pkl)
Validación	pydantic	2.12.5	Validación de entradas en el API

Procedimiento General del Pipeline

El desarrollo del proyecto siguió un pipeline secuencial de cuatro fases, cada una con entradas, salidas y criterios de avance definidos. La Figura 1 sintetiza el flujo metodológico general.

Figura 1

Flujo Metodológico Secuencial del Proyecto



Arquitectura de la Implementación

El sistema adopta una arquitectura **cliente-servidor desacoplada**. FastAPI centraliza toda la lógica de inferencia, carga de modelos, validación de entradas vía Pydantic, cálculo de SHAP values e invocación de los tres modelos XGBoost, además, expone un único endpoint POST /predict. Streamlit opera exclusivamente como capa de presentación, consumiendo dicho endpoint mediante una llamada HTTP y renderizando los resultados en el dashboard. Este desacoplamiento garantiza que el modelo de predicción sea independiente del frontend, facilitando su reutilización o migración a otras interfaces sin modificar la lógica de inferencia.

Análisis Exploratorio de Datos (EDA)

El Análisis Exploratorio de Datos (EDA) constituye la fase inicial y fundamental del ciclo de ciencia de datos. Su propósito es comprender la estructura subyacente del conjunto de datos, caracterizar el comportamiento estadístico de cada variable, identificar relaciones entre predictores y la variable objetivo, detectar anomalías, y establecer las bases analíticas que orientarán el diseño y entrenamiento de los modelos predictivos.

Carga, Inspección Inicial y Calidad de Datos

La primera etapa consistió en cargar el dataset y realizar una inspección sistemática de su estructura, tipos de datos, completitud y presencia de registros duplicados. Los resultados obtenidos se resumen a continuación.

Tabla 3

Resumen de Inspección Inicial del Dataset

Atributo	Valor
Registros / Variables	973 · 15 (14 originales + BMI derivado)
Valores nulos	0 (0.0 %)
Registros duplicados	0
Balance de género	Male 52.52 % · Female 47.5 %
Tipos de entrenamiento	4 categorías con distribución homogénea (~25 % c/u)
Variable objetivo	Calories_Burned – continua, rango 303 – 1783 kcal

Tabla 4

Clasificación de las Variables del Dataset

Variable	Tipo	Rol
Age	Numérica continua	Predictor

Gender	Categórica nominal	Predictor
Weight (kg)	Numérica continua	Predictor
Height (m) / BMI	Numérica continua	Predictor
Max_BPM / Avg_BPM / Resting_BPM	Numérica discreta	Predictor
Session_Duration (hours)	Numérica continua	Predictor principal
Workout_Type	Categórica nominal	Predictor
Fat_Percentage	Numérica continua	Predictor
Water_Intake (liters)	Numérica continua	Predictor
Workout_Frequency (days/week)	Ordinal	Predictor
Experience_Level	Ordinal (1 – 3)	Predictor
Calories_Burned	Numérica continua	Objetivo

Nota. El dataset se encontró en condiciones óptimas de calidad sin valores faltantes ni duplicados, lo que eliminó la necesidad de etapas de imputación o limpieza agresiva, permitiendo proceder directamente al análisis estadístico.

Estadística Descriptiva

La Tabla 5 resume las estadísticas descriptivas de las variables numéricas más relevantes, incluyendo el coeficiente de variación (CV%) como medida de dispersión relativa y la asimetría (skewness) para evaluar la forma de cada distribución.

Tabla 5

Estadísticas Descriptivas de las Principales Variables Numéricas

Variable	Media	DE	CV%	Mediana	Rango	Skewness
Age (años)	38.68	12.18	31.5	38.0	18–59	-0.08 Simétrica
Weight (kg)	73.86	21.21	28.7	73.7	40–143.5	+0.77 Asim. dcha.

Session_Duration (h)	1.256	0.343	27.3	1.27	0.50–2.00	+0.03 Simétrica
Calories_Burned (kcal)	905.4	272.6	30.1	893.0	303–1 783	+0.28 Simétrica
Fat_Percentage (%)	24.98	6.26	25.1	25.9	10.0–38.0	-0.64 Asim. izq.
Avg_BPM	143.8	14.35	10.0	144	120–169	+0.09 Simétrica
Water_Intake (L)	2.627	0.600	22.9	2.6	1.5–3.7	+0.07 Simétrica
BMI	24.91	6.56	26.3	23.84	12.32–49.84	+0.76 Asim. dcha.
Experience_Level	1.810	0.740	40.9	2.0	1–3	+0.32 Simétrica
Workout_Freq. (d/sem)	3.322	0.913	27.5	3.0	2–5	+0.15 Simétrica

Nota. DE = desviación estándar.

Session_Duration: Presenta un rango de 0.5 a 2.0 horas con distribución prácticamente simétrica (skewness = +0.026) y un CV del 27.3%, indicando variabilidad moderada en los tiempos de entrenamiento de los miembros.

Calories_Burned: Muestra una distribución levemente asimétrica positiva (skewness = +0.278), con media de 905 kcal y mediana de 893 kcal, la cercanía entre ambas confirma una distribución aproximadamente normal. El rango total (303 – 1783 kcal) refleja la alta heterogeneidad en el gasto energético.

Weight (kg): Es la variable con mayor asimetría positiva (skewness = +0.772), sugiriendo la presencia de individuos con peso corporal elevado que desplazan la distribución hacia la derecha.

Fat_Percentage: Exhibe asimetría negativa (skewness = -0.635), lo que indica una concentración de individuos con porcentajes de grasa moderados – altos, con cola izquierda hacia valores bajos asociados a usuarios avanzados.

BMI: Tiene la mayor asimetría positiva del conjunto (skewness = +0.764), coherente con el comportamiento de Weight; la media de 24.91 se ubica en el límite superior del rango "normal" según la World Health Organization (2000).

Variables Categóricas y Ordinales

El dataset presenta un balance de género prácticamente equitativo (52.5% masculino / 47.5% femenino), lo que descarta sesgos de representatividad por sexo. Los cuatro tipos de entrenamiento están distribuidos de manera homogénea (~25% cada uno), garantizando representación suficiente de cada modalidad. La mayoría de los miembros pertenecen al nivel Intermedio (41.7%) o Principiante (38.6%), siendo los usuarios Avanzados el grupo minoritario (19.6%). La frecuencia de entrenamiento más común es de 3 días por semana (37.8%).

Tabla 6

Distribución de Frecuencias de Variables Categóricas y Ordinales

Variable	Categoría / Nivel	Frecuencia	Porcentaje (%)
Gender	Male	511	52.52
	Female	462	47.48
Workout_Type	Strength	258	26.52
	Cardio	255	26.21
	Yoga	239	24.56
	HIIT	221	22.71
Experience_Level	1 – Principiante	376	38.64

	2 – Intermedio	406	41.73
	3 – Avanzado	191	19.63
Workout_Frequency	2 días/semana	197	20.25
	3 días/semana	368	37.82
	4 días/semana	306	31.45
	5 días/semana	102	10.48

Análisis de Distribuciones y Detección de Outliers

A continuación, se presentan los histogramas con KDE (Kernel Density Estimation) para las 12 variables numéricas independientes. La superposición de media (línea discontinua) y mediana (línea punteada) permite identificar visualmente la simetría de cada distribución.

Figura 2

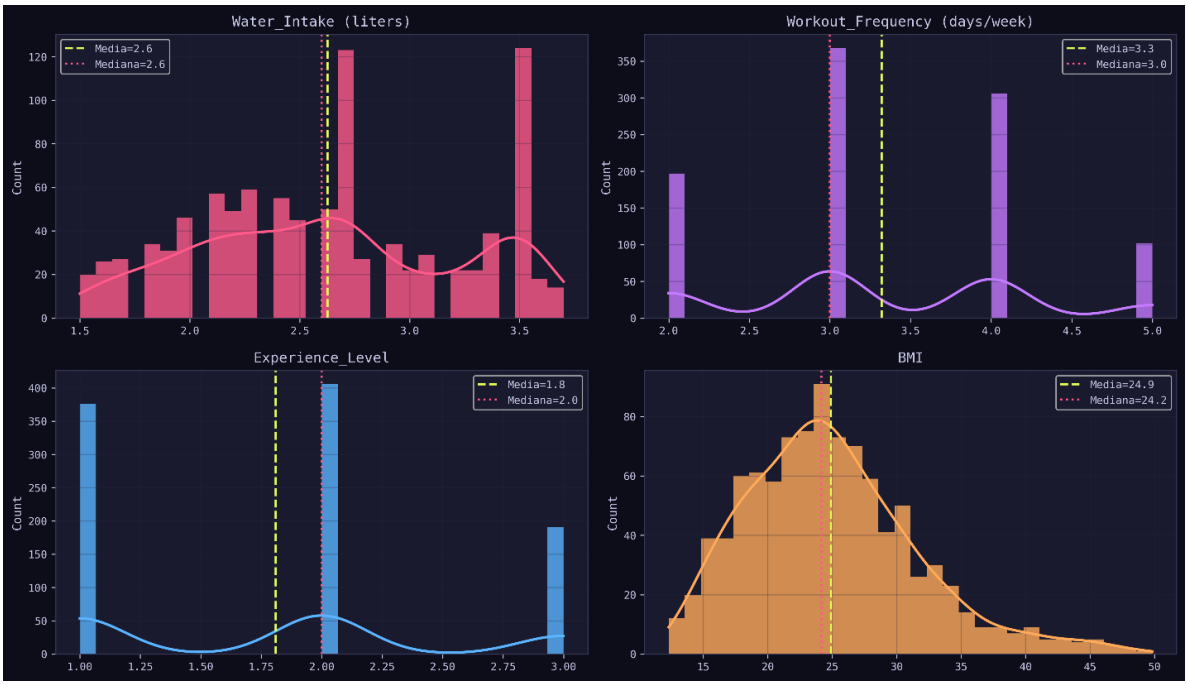
Histogramas con KDE Variables Antropométricas y Cardíacas

Figura 3

Histogramas con KDE Variables Cardíacas y de Sesión

Figura 4

Histogramas con KDE Variables de Composición y Frecuencia



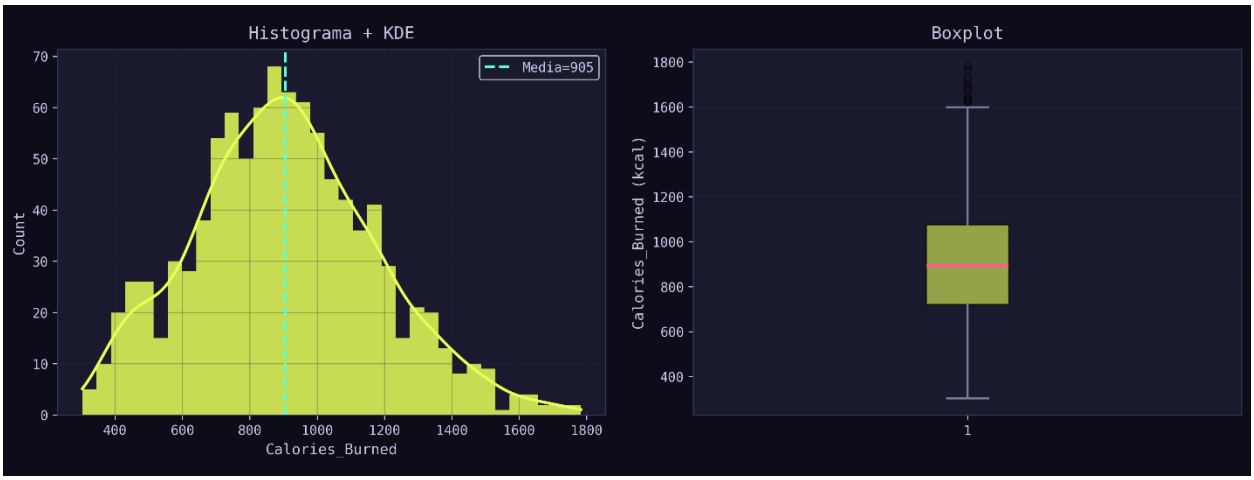
Las distribuciones revelan que la mayoría de las variables numéricas son aproximadamente simétricas, con media y mediana muy próximas entre sí. Las excepciones son Weight (kg) y BMI ambas con cola derecha pronunciada y Fat_Percentage, con cola izquierda leve. Session_Duration se concentra en el intervalo 1.0 – 1.5 horas, reflejando la duración típica de una sesión de entrenamiento en gimnasio. Las variables de frecuencia cardíaca (Max_BPM, Avg_BPM, Resting_BPM) exhiben distribuciones uniformes dentro de sus rangos fisiológicos esperados.

Análisis de la Variable Objetivo: Calories_Burned

Por ser la variable objetivo del modelo predictivo, se realizó un análisis de distribución más detallado sobre Calories_Burned, utilizando dos representaciones complementarias: histograma con KDE y boxplot.

Figura 5

Distribución de Calories_Burned



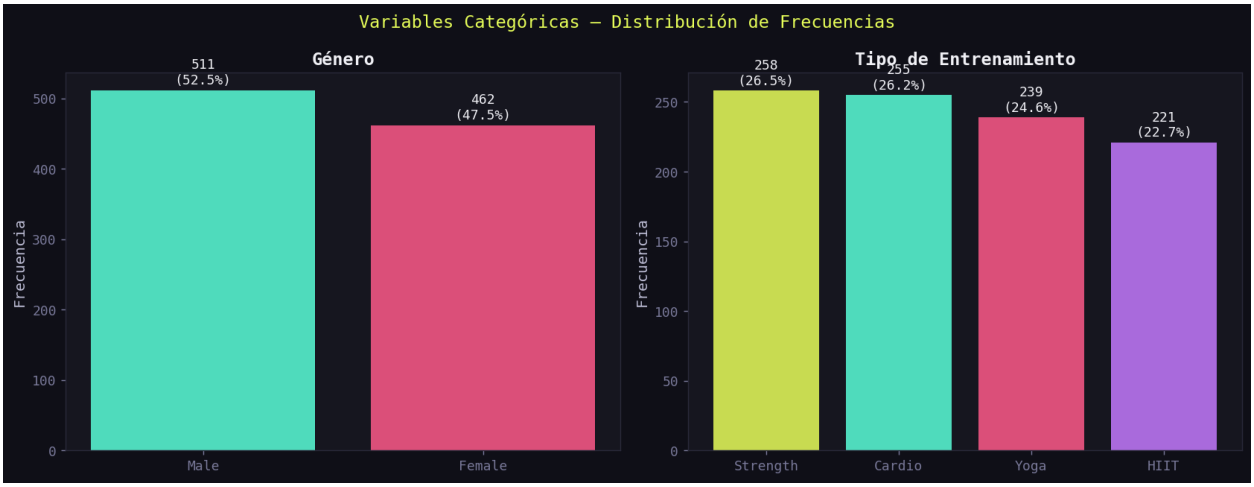
Nota. El boxplot identifica 10 valores atípicos por método IQR (1.03 %) en la cola superior, sesiones largas de usuarios avanzados y únicamente 9 en Weight (0.92 %). Ambos se conservan en el dataset por ser fisiológicamente plausibles y porque los algoritmos de árboles son inherentemente robustos frente a outliers.

Análisis de Variables Categóricas y su Impacto sobre Calories_Burned

La Figura 6 presenta la distribución de frecuencias para las dos variables categóricas nominales: Gender y Workout_Type.

Figura 6

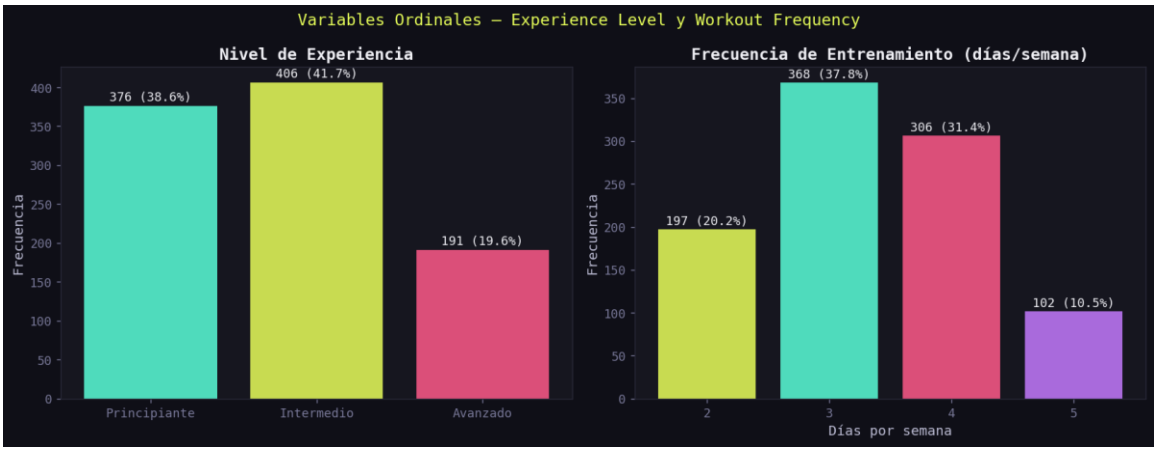
Distribución de Frecuencias para Género y Tipo de Entrenamiento



Nota. El balance de género es prácticamente equitativo: hombres representan el 52.52% (n = 511) y mujeres el 47.48% (n = 462). Esta distribución descarta sesgos de género en el entrenamiento del modelo. Los cuatro tipos de entrenamiento presentan frecuencias muy similares (Strength 26.52%, Cardio 26.21%, Yoga 24.56%, HIIT 22.71%) garantizando que ninguna modalidad esté minorizada en el dataset.

Figura 7

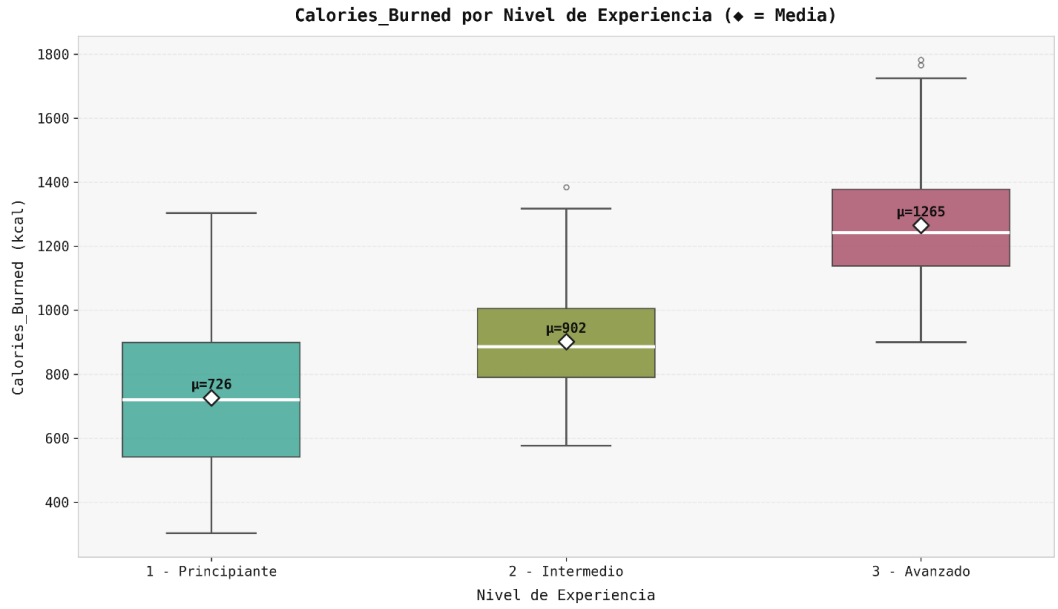
Distribución de Frecuencias Nivel de Experiencia y Frecuencia de Entrenamiento Semanal



Nota. El nivel Intermedio concentra el mayor número de miembros (41.73%, n=406), seguido por Principiante (38.64%, n=376) y Avanzado (19.63%, n=191). La menor representación del nivel Avanzado es esperable dado que este perfil corresponde a usuarios con alta adherencia y progresión sostenida. En cuanto a la frecuencia de entrenamiento, el 69.27% de los miembros entrena entre 3 y 4 días por semana, lo que refleja un perfil de entrenador moderado-frecuente.

Figura 8

Distribución de Calories_Burned por Nivel de Experiencia (Boxplot)



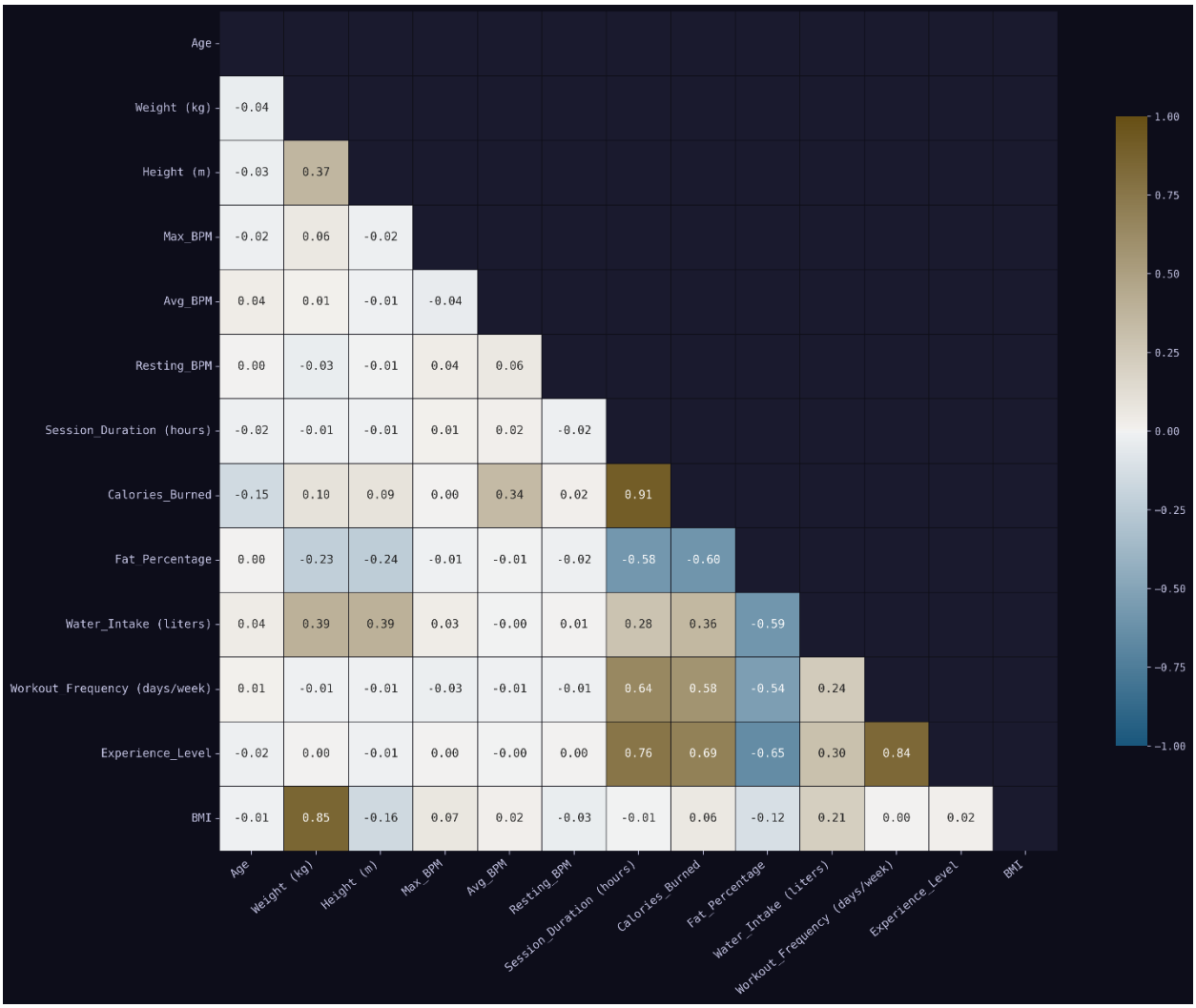
Nota. La prueba de Kruskal-Wallis ($H = 459.49$, $p < 10^{-100}$) confirma diferencias estadísticamente significativas entre niveles, con una separación de aproximadamente 539 kcal entre Principiante ($\mu = 726$ kcal) y Avanzado ($\mu = 1265$ kcal), coherente con la mayor capacidad aeróbica y eficiencia metabólica de los usuarios con mayor trayectoria de entrenamiento. Se empleó prueba no paramétrica ante la ausencia de verificación de normalidad intragrupo.

Matriz de Correlación e Identificación de Predictores

Se calculó la matriz de correlación de Pearson sobre las 13 variables numéricas. Adicionalmente se validaron los resultados con correlación de Spearman, encontrando diferencias mínimas ($\Delta = 0.091$ en Fat_Percentage), lo que confirma la robustez de los coeficientes frente a la no-normalidad de algunas variables.

Figura 9

Matriz de Correlación de Pearson



La inspección de la matriz revela patrones de correlación significativos y esperados desde el punto de vista fisiológico. Entre las correlaciones de interés práctico para el modelado se destacan:

Session_Duration – Calories_Burned: $r = +0.908$ (muy fuerte positiva). Relación lineal dominante del dataset.

Experience_Level – Water_Intake: $r = +0.712$. Los usuarios avanzados mantienen mayor hidratación, coherente con sesiones más intensas.

Experience_Level – Workout_Frequency: $r = +0.590$. La frecuencia de entrenamiento aumenta progresivamente con el nivel de experiencia.

Fat_Percentage – BMI: $r = +0.610$. Relación esperable entre índice de masa corporal y composición corporal.

Fat_Percentage – Experience_Level: $r = -0.524$. Los usuarios avanzados presentan menor porcentaje de grasa, reflejo de mayor adaptación metabólica.

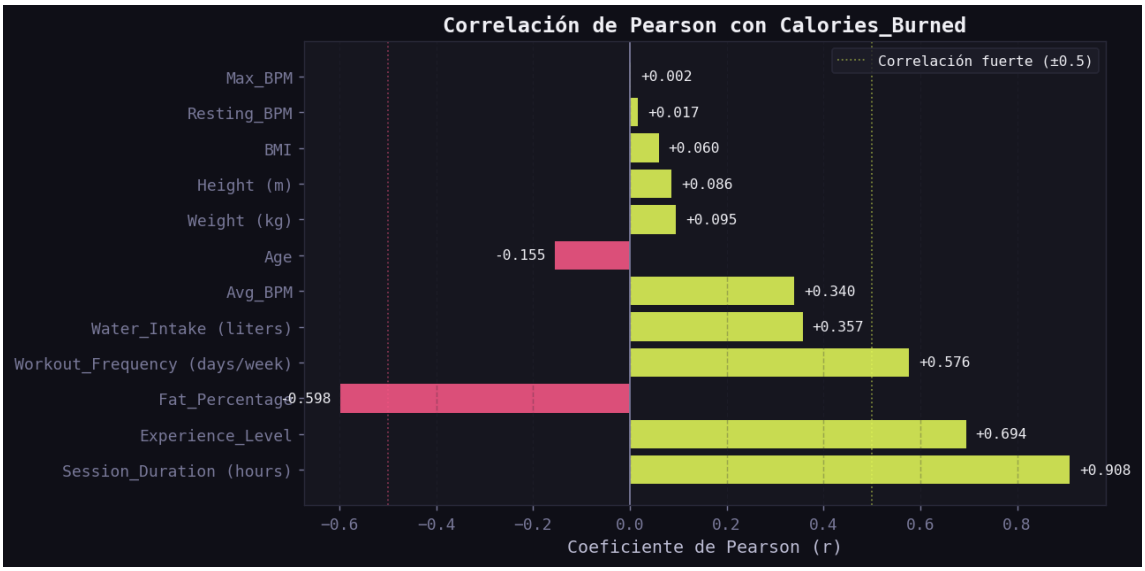
Max_BPM – Avg_BPM: Presentan correlación débil con Calories_Burned ($r < 0.01$ y $r = +0.34$ respectivamente), lo que sugiere que la frecuencia cardíaca por sí sola no captura completamente el esfuerzo energético.

Ranking de Predictores para Calories_Burned

La Figura 10 presenta el ranking de correlación de Pearson de todas las variables independientes con la variable objetivo Calories_Burned, ordenadas por valor absoluto. Se incluye la línea de referencia en r = ±0.5 para distinguir correlaciones fuertes de moderadas.

Figura 10

Ranking de Correlación de Pearson con Calories_Burned



Nota. Barras amarillas hacen referencia a correlación positiva; barras rosadas a la correlación negativa.

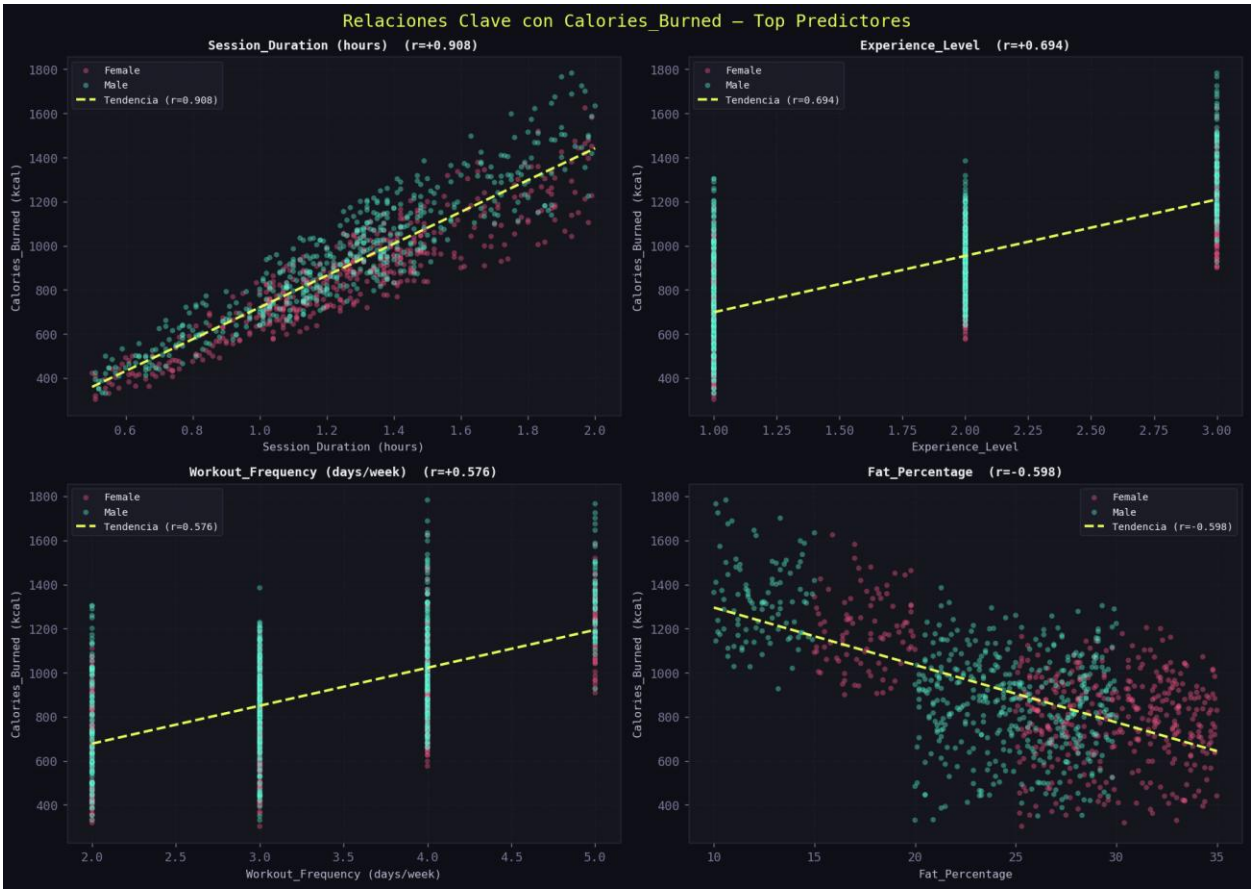
Session_Duration (hours) es el predictor primario y dominante, con $r = +0.908$, una correlación lineal excepcionalmente fuerte que explica en términos individuales el 82.4% de la varianza de Calories_Burned ($r^2 = 0.824$). Los predictores secundarios de alta relevancia son Experience_Level ($r = +0.694$), Fat_Percentage ($r = -0.598$) y Workout_Frequency ($r = +0.576$).

Análisis Bivariado: Relaciones Clave con Calories_Burned

La Figura 11 muestra los diagramas de dispersión de los cuatro predictores principales frente a Calories_Burned, diferenciados por género. Se incluye la línea de regresión lineal simple para cuantificar la dirección y fuerza de cada relación.

Figura 11

Scatter Plots Predictores Principales vs. Calories_Burned por Género

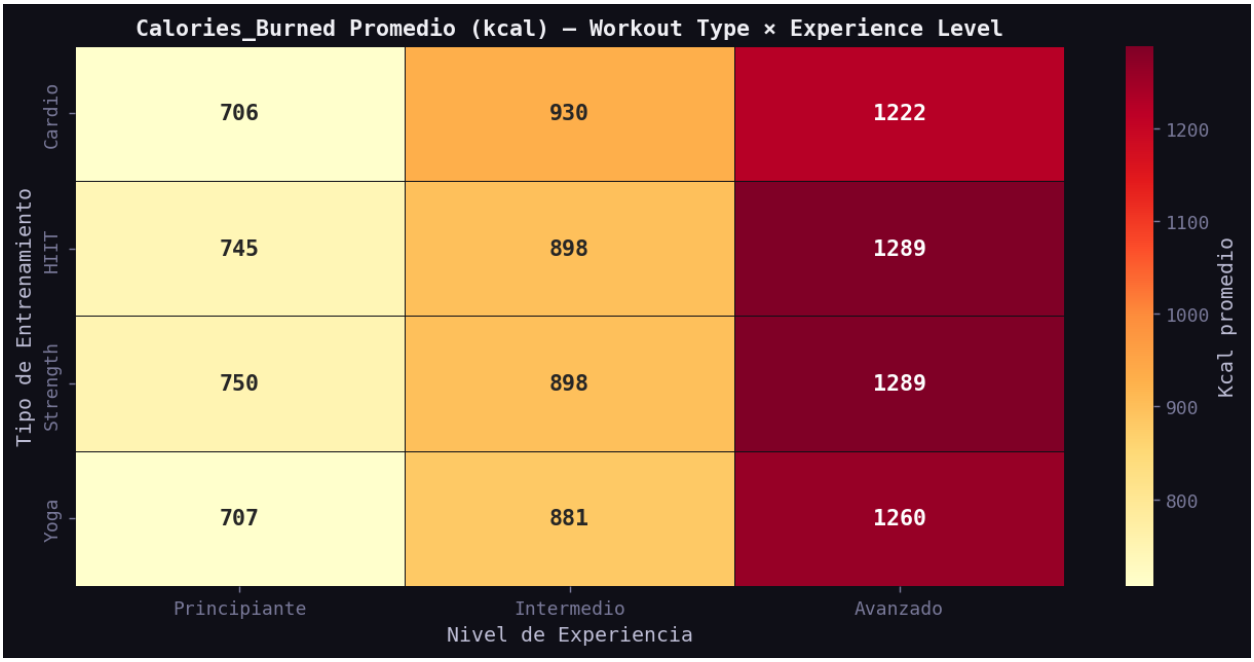


Session_Duration exhibe la relación más limpia y lineal, consistente entre géneros, con baja dispersión alrededor de la tendencia. Experience_Level muestra dispersión esperada por ser variable ordinal discreta, pero la tendencia ascendente es robusta. Fat_Percentage presenta relación negativa clara: a mayor porcentaje de grasa, menor masa muscular activa y menor gasto calórico por sesión. Workout_Frequency actúa como proxy de adaptación metabólica crónica, usuarios que entrenan con mayor frecuencia desarrollan mayor capacidad cardiovascular y queman más calorías por sesión.

Por otra parte, La Figura 12 presenta el efecto combinado de Workout_Type y Experience_Level sobre el gasto calórico promedio, revelando la interacción entre ambas variables.

Figura 12

Heatmap de Calories_Burned × Workout_Type × Experience_Level



El heatmap confirma que el efecto de Experience_Level es transversal a todos los tipos de entrenamiento: los usuarios Avanzados superan en más de 500 $kcal$ promedio a los Principiantes, independientemente de la modalidad. El incremento progresivo de Principiante a Intermedio a Avanzado es consistente y uniforme. La práctica de HIIT y Strength produce marginalmente mayor gasto calórico que Yoga en el nivel Avanzado, aunque las diferencias son pequeñas.

Variable Derivada: FC_relativa e Intensidad de Sesión

La variable FC_relativa ($Avg\_BPM \ / \ Max\_BPM$) se construyó como indicador fisiológico de la intensidad real de cada sesión. Esta métrica es ampliamente utilizada en ciencias

del deporte para clasificar zonas de entrenamiento, ya que normaliza la frecuencia cardíaca respecto al máximo individual de cada usuario (Garber et al., 2011), eliminando el sesgo que introduce la variabilidad fisiológica entre personas.

La distribución de la intensidad derivada muestra que ningún usuario entrenó en la zona baja (<60% FCmáx): el 34.6% entrenó en zona moderada (60 – 75%), el 34.7% en zona alta (75 – 85%) y el 28.9% en zona muy alta (>85%). Esto confirma que el dataset representa una población de miembros activos con niveles de esfuerzo moderado-alto, consistente con el contexto de un gimnasio.

Preprocesamiento del Dataset

A partir de los hallazgos del EDA, se ejecutó un proceso de preprocesamiento orientado a preparar el dataset para el entrenamiento de los modelos predictivos de Machine Learning. Las transformaciones aplicadas se detallan a continuación:

Tabla 7

Transformaciones Aplicadas en la Etapa de Preprocesamiento

Transformación	Variables afectadas	Justificación
Estandarización Z-Score (StandardScaler)	Las 12 variables numéricas independientes	Necesario para algoritmos sensibles a escala (regresión regularizada). Media ≈ 0, DE. ≈ 1 verificado post-transformación.
Categorización IMC (OMS)	BMI → BMI_Category (4 niveles)	Captura la no-linealidad del riesgo metabólico: Bajo peso / Normal / Sobrepeso / Obesidad.
Feature derivada: FC_relativa	FC_relativa = Avg_BPM / Max_BPM	Mide la intensidad relativa de la sesión como proporción de la FCmáx. Predictor fisiológico más informativo que los BPM absolutos.

Categorización de intensidad	FC_relativa → Intensidad_Sesion (3 niveles)	Baja (<60%), Moderada (60-75%), Alta (75-85%), Muy alta (>85%). Basado en zonas de entrenamiento reconocidas en fisiología del ejercicio.
Codificación LabelEncoder	Gender (Female=0, Male=1) Workout_Type (Cardio=0, HIIT=1, Strength=2, Yoga=3)	Conversión a formato numérico requerido por los algoritmos de ML.
Cuartiles de Calories_Burned	Calories_Burned → Calories_Quartile (Q1–Q4)	Variable auxiliar para análisis descriptivo y segmentación de usuarios.

Implicaciones para el Modelado Predictivo

Features prioritarias para ML: Session_Duration, Experience_Level, Fat_Percentage, Workout_Frequency, Avg_BPM, FC_relativa. Estas seis variables concentran el mayor poder predictivo individual.

Interacciones para explorar: Session_Duration × Experience_Level y Session_Duration × Workout_Type, dado el comportamiento diferenciado por nivel de experiencia observado en el heatmap.

Algoritmos recomendados: La fuerte correlación lineal de Session_Duration sugiere que modelos lineales pueden lograr buen desempeño base. Sin embargo, las no linealidades en Experience_Level y Fat_Percentage favorecen la aplicación de Random Forest y XGBoost para capturar interacciones complejas. Se espera un $R^2 > 0.90$ con XGBoost.

Normalización: La estandarización Z-Score aplicada es recomendable para los modelos sensibles a escala. Los algoritmos de árboles (Random Forest, XGBoost) no la requieren, pero el dataset preprocesado la incluye para flexibilidad.

Dataset Final para Modelado

El dataset preprocesado resultante contiene 973 registros con 15 features y 1 variable objetivo, sin valores faltantes. El archivo exportado (gym_members_preprocessed.csv) constituye la entrada directa para la fase de entrenamiento de modelos.

Segmentación y Entrenamiento de Modelos de Aprendizaje Automático.

Variables Predictoras del Vector de Entrenamiento

El modelo emplea seis variables fisiológicas y demográficas como predictores directos: *Session_Duration*, *Avg_BPM*, *Resting_BPM*, *Weight*, *Age* y *Gender_enc*. La justificación científica proviene de Keytel et al. (2005), quienes demostraron que el gasto calórico durante ejercicio submáximo puede estimarse con precisión a partir de la frecuencia cardíaca, el peso, la edad y el género, exactamente las variables del vector de predicción de este proyecto. Su ecuación, validada sobre calorimetría directa en sujetos de distintos niveles de aptitud, establece que la frecuencia cardíaca promedio y la duración de la sesión son los determinantes primarios del gasto energético, mientras que el peso, la edad y el género actúan como moduladores del metabolismo basal.

Por otra parte, *Session_Duration* emerge como el predictor dominante del dataset (r = +0.908), coherente con la ecuación tipo MET subyacente que multiplica la tasa metabólica por el tiempo de actividad. *Avg_BPM* captura la intensidad cardíaca de la sesión como proxy del esfuerzo aeróbico; *Resting_BPM* incorpora información sobre la condición cardiovascular basal del individuo (Booth et al., 2012). *Weight* y *Age* determinan el componente basal del metabolismo según las ecuaciones de referencia de la literatura fisiológica (Benton & Young, 2017). *Gender_enc* integra la diferencia metabólica documentada entre sexos, reflejada en los coeficientes diferenciales de las ecuaciones de gasto energético.

Variable de Segmentación

Experience_Level no opera como predictor continuo, sino como variable de segmentación codificada mediante dos dummies (*Level_2*, *Level_3*; referencia: Principiante). Esta decisión se sustenta en la heterogeneidad fisiológica documentada entre niveles de aptitud

física: los usuarios avanzados exhiben mayor VO_2máx, eficiencia metabólica y capacidad de trabajo aeróbico que los principiantes, lo que se traduce en pendientes de gasto calórico sustancialmente distintas respecto a la duración de la sesión: 427, 589 y 669 kcal/h para Principiante, Intermedio y Avanzado respectivamente. Esta diferencia de pendientes (56.8 % entre extremos) es evidencia empírica directa de que la relación *features → target* no es homogénea entre grupos, condición que el análisis estadístico confirma con alta significancia (Kruskal-Wallis $H = 459.49$, $p < 10^{-100}$; d de Cohen hasta 2.59 entre niveles extremos). La codificación como dummies, en lugar de incluir el valor ordinal continuo preserva la interpretabilidad del intercepto diferencial por grupo sin imponer una distancia métrica artificial entre niveles.

Variables Excluidas del Vector de Predicción

Varias variables disponibles en el dataset fueron excluidas del modelo final por razones técnicas explícitas. *Fat_Percentage* y *BMI* presentan alta colinealidad entre sí ($r = +0.610$) y con *Weight*; su inclusión incrementaría el VIF sin aportar información predictiva independiente. *Water_Intake* exhibe causalidad invertida respecto al gasto calórico: no determina la energía consumida, sino que es una consecuencia observable de la intensidad de la sesión, incorporarla como predictor introduciría un sesgo causal en el modelo. *Workout_Type* y *Workout_Frequency* mostraron correlaciones moderadas con el target ($r < 0.58$) y alta colinealidad con *Experience_Level* ($r = +0.59$ para frecuencia). El diagnóstico VIF sobre el vector final confirmó valores inferiores a 2.0 en las seis variables seleccionadas, validando la ausencia de colinealidad problemática.

Se entrenaron cinco algoritmos: LinearRegression, Ridge ($\alpha = 1.0$), DecisionTree (max_depth = 8), RandomForest (n_estimators = 200, max_depth = 12, min_samples_leaf = 2) y

XGBoost (n_estimators = 150, learning_rate = 0.05, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8). Todos los modelos se evaluaron bajo tres estrategias arquitectónicas mediante K-Fold con k = 5 pliegues, mezcla aleatoria y semilla fija (random_state = 42):

Estrategia A – Modelo global: un único modelo sobre los 973 registros con las 6 features, sin información de nivel.

Estrategia B – Experience_Level como feature: se añaden dos variables dummy (drop_first = True, referencia: Principiante) al espacio de predicción. Mide el aporte ortogonal del nivel de experiencia respecto a las 6 features originales.

Estrategia C – Modelos especializados: un modelo independiente por nivel, evaluado con predicciones out-of-fold (OOF) concatenadas para evitar la subestimación del error propia del MAE ponderado intragrupo.

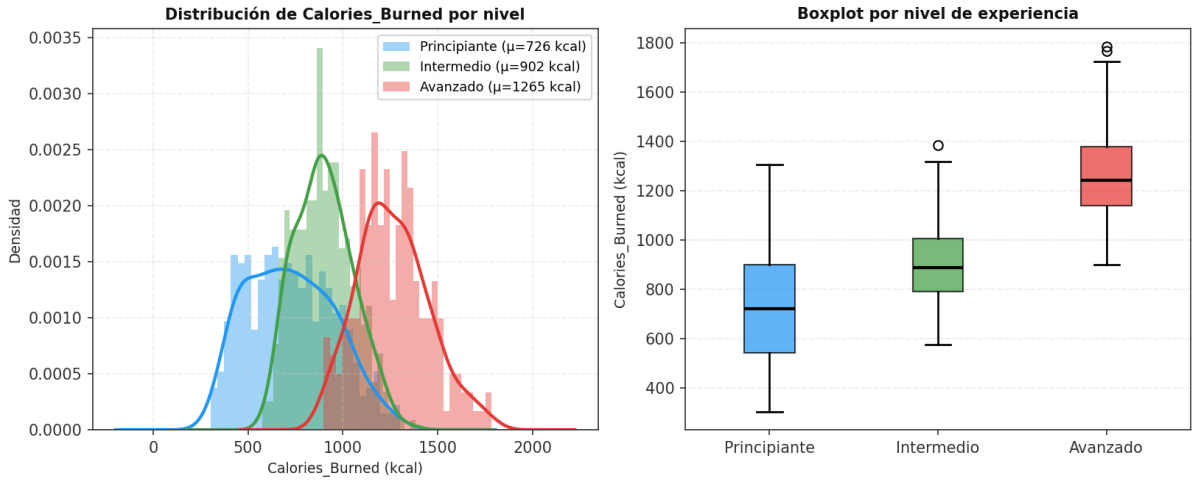
Las métricas reportadas son el Error Absoluto Medio (MAE ± DE entre pliegues), el R^2 y la ganancia relativa de cada estrategia sobre el baseline global (Δ%). La estrategia A constituye el techo de referencia. Las hipótesis evaluadas son: H_0 la segmentación no aporta información predictiva (Δ < 2 % en B y C); H_1 Experience_Level como feature mejora el modelo global (Δ_B > 2 %, B ≈ C); H_2 la relación features→Calories es heterogénea entre niveles (Δ_C > Δ_B + 1 %).

Segmentación Fisiológica por Nivel de Experiencia

La distribución de Calories_Burned difiere sustancialmente entre los tres niveles de experiencia, como ilustra la Figura 13. Los 973 registros se distribuyen en 376 Principiantes (38.6 %), 406 Intermedios (41.7 %) y 191 Avanzados (19.6 %), con un ratio de balance máximo/mínimo de 2.13:1, suficiente para garantizar K-Fold estable en todos los grupos.

Figura 13

Distribución de Calories_Burned por Nivel de Experiencia



Nota. Histograma con estimación de densidad kernel (izquierda) y boxplot (derecha).

Tabla 8

Estadísticas Descriptivas de Calories_Burned por Nivel de Experiencia

Nivel (n)	Media (kcal)	DE (kcal)	Mínimo	Máximo	Session_Duration
Principiante (376)	726.4	227.3	303	1304	1.01
Intermedio (406)	901.9	152.6	576	1385	1.25
Avanzado (191)	1265.3	186.8	900	1783	1.76

Nota. DE = desviación estándar. Los rangos de Calories_Burned se solapan parcialmente entre Principiante e Intermedio ([576 – 1304] kcal, 728 kcal de rango compartido) y entre Intermedio y

Avanzado ([900 – 1385] kcal, 485 kcal). Session_Duration (hours) y Experience_Level exhiben rangos casi disjuntos entre niveles.

La prueba de Kruskal-Wallis aplicada sobre Calories_Burned arroja $H = 459.49$ ($p < 10^{-100}$), rechazando la hipótesis nula de distribuciones idénticas con certeza prácticamente absoluta. Las comparaciones por pares mediante Mann-Whitney U (ver Tabla 9) confirman efectos de gran magnitud en todos los pares.

Tabla 9

Pruebas Mann-Whitney U y d de Cohen por Pares de Niveles (Calories_Burned)

Par comparado	Estadístico U	Valor p	D de Cohen	Magnitud
Principiante vs. Intermedio	42.614	< 0.001	+0.907	Grande
Intermedio vs. Avanzado	20.085	< 0.001	+2.130	Grande
Principiante vs. Avanzado	17.093	< 0.001	+2.590	Grande

Nota. La d de Cohen clasifica el efecto como grande cuando d > 0.80. La separación Principiante – Avanzado (d = 2.59) indica una distancia de 2.6 desviaciones estándar entre poblaciones, respaldando la heterogeneidad fisiológica entre niveles.

Un indicador clave de heterogeneidad es la pendiente de la regresión lineal simple de Session_Duration sobre Calories_Burned por nivel: Principiante = 427 kcal/h, Intermedio = 589 kcal/h y Avanzado = 669 kcal/h. La diferencia de pendientes (56.8 % entre extremos) es la

evidencia empírica directa de que la relación features → target no es homogénea entre grupos, lo que justifica los modelos especializados para algoritmos lineales.

Diagnóstico de Sesgos Estructurales

Previo al entrenamiento se cuantificaron dos sesgos con capacidad de distorsionar la interpretación de resultados.

El Factor de Inflación de Varianza (VIF), calculado sobre features estandarizadas, es inferior a 2.0 en las seis variables de predicción (Session_Duration: 1.000; Avg_BPM: 1.010; Resting_BPM: 1.010; Age: 1.010; Weight: 1.520; Gender_enc: 1.510), confirmando ausencia de colinealidad problemática. La equivalencia empírica entre LinearRegression y Ridge (ΔMAE = 0.02 kcal) corrobora este resultado.

Dominancia de Session_Duration

La correlación de Pearson entre Session_Duration (hours) y Calories_Burned es r = +0.908 (R^2 univariado = 82.5 %). Esta variable explica por sí sola el 82.5 % de la varianza del target, dejando a las otras cinco features en competencia por el 17.5 % restante. Consecuencias operativas: (a) los algoritmos de árbol posicionarán Session_Duration en el primer split en la mayoría de los árboles entrenados; (b) el análisis SHAP del Capítulo 3 reflejará esta dominancia. Este sesgo es una propiedad intrínseca del dataset que sugiere generación semi-sintética (Keytel et al., 2005), no un artefacto del pipeline.

Sesgo de MAE Intra – Nivel

Reportar el error promedio ponderado por tamaño de grupo en la Estrategia C subestimaría el MAE global porque aprovecha la menor varianza del target dentro de cada nivel. Por ello, todas las métricas de la Estrategia C se calculan sobre predicciones OOF concatenadas de los 973 registros, produciendo un estimador comparable con las Estrategias A y B.

Resultados por Estrategia de Entrenamiento

Estrategia A: Modelo Global (Baseline)

Tabla 10

Métricas de Validación Cruzada – Estrategia A: Modelo Global (K-Fold, k = 5)

Modelo	MAE (kcal)	±DE	RMSE (kcal)	R^2
LinearRegression	30.16	0.69	39.94	0.9784
Ridge (α = 1.0)	30.18	0.89	40.05	0.9783
DecisionTree	44.87	3.51	58.24	0.9541
RandomForest	28.65	1.35	36.10	0.9808
XGBoost	31.20	1.91	39.40	0.9772

Nota. DE = desviación estándar del MAE entre pliegues. RandomForest es el mejor modelo global (MAE = 28.65 kcal, R^2 = 0.9808). La equivalencia LinearRegression–Ridge (ΔMAE = 0.02 kcal) confirma el diagnóstico VIF. DecisionTree presenta la mayor varianza entre pliegues (DE = 3.51 kcal), indicando sensibilidad al solapamiento de rangos del target entre niveles.

Estrategia B: Experience_Level como Feature Predictora

Al incorporar dos dummies de Experience_Level (Intermedio y Avanzado, referencia: Principiante), el espacio de predicción pasa de 6 a 8 features. El impacto es radicalmente distinto según el algoritmo (ver Tabla 11).

Tabla 11

Métricas – Estrategia B y Variación Respecto a Estrategia A

Modelo	MAE B (kcal)	R^2	Δ MAE (B−A)	Δ% relativo	Wilcoxon p
LinearRegression	30.20	0.9784	+0.04 kcal	−0.1%	0.4375
Ridge (α = 1.0)	30.34	0.9781	+0.16 kcal	−0.5%	0.8125
DecisionTree	44.12	0.9554	−0.76 kcal	+1.7%	0.1250
RandomForest	28.96	0.9806	+0.31 kcal	−1.1%	—
XGBoost	21.44	—	−9.76 kcal	+31.3%	—

Nota. Δ% positivo = reducción de MAE respecto a Estrategia A (mejora). La prueba de Wilcoxon pareada (k = 5 pares) arroja p > 0.05 para LinearRegression, Ridge y DecisionTree (p = 0.4375, 0.8125 y 0.1250), indicando diferencias no significativas. Las celdas con — para RandomForest y XGBoost indican que el test no fue computable: en RandomForest, las diferencias fold-a-fold son cercanas a cero, generando n efectivo insuficiente; en XGBoost, con k = 5 pares de signo uniforme, la p-value mínima alcanzable es p = 0.0625 (α = 0.05). La ganancia de XGBoost (+31.3 %, Δ = −9.76 kcal) confirma la hipótesis H_1: las dummies de nivel actúan como particiones de alto poder discriminativo incorporadas en los primeros splits del ensamble.

Estrategia C: Modelos Especializados por Nivel

Los resultados de la Tabla 12 corresponden al MAE calculado sobre predicciones OOF globales concatenadas. El número de pliegues se ajusta dinámicamente: k = 5 para Principiante (n = 376) e Intermedio (n = 406), k = 5 para Avanzado (n = 191).

Tabla 12

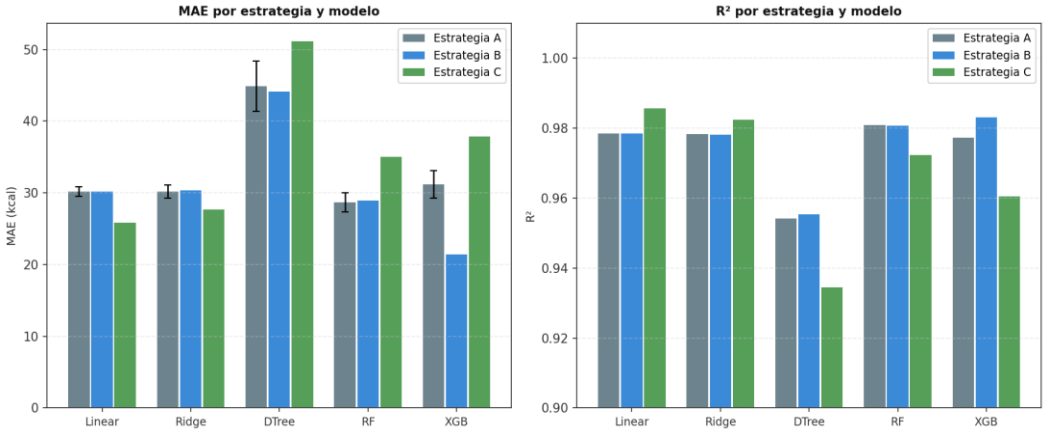
Métricas OOF globales – Estrategia C: modelos especializados por nivel

Modelo	MAE OOF (kcal)	RMSE (kcal)	R^2 OOF	Δ MAE (C−A)	Δ% relativo
LinearRegression	25.85	32.75	0.9856	−4.31 kcal	+14.3%
Ridge (α = 1.0)	27.64	36.28	0.9823	−2.54 kcal	+8.4%
DecisionTree	51.12	69.80	0.9344	+6.25 kcal	−13.9%
RandomForest	34.87	45.06	0.9727	+6.45kcal	−22.7%
XGBoost	30.90	41.01	0.9774	+7.86 kcal	−34.1%

Nota. Δ% positivo = mejora de C sobre A. Los modelos lineales confirman la hipótesis H_2: la especialización por nivel captura las pendientes diferenciales de *Session_Duration* (427 / 589 / 669 kcal/h), con mejora del +14.3 % para LinearRegression. RandomForest y XGBoost se degradan en Estrategia C (−22.7 % y −34.1 %, respectivamente) porque la reducción de datos por nivel, especialmente en Avanzado (n = 191, ~38 registros por pliegue) genera sobreajuste en modelos de alta complejidad.

Figura 14

MAE y R² por Estrategia y Modelo



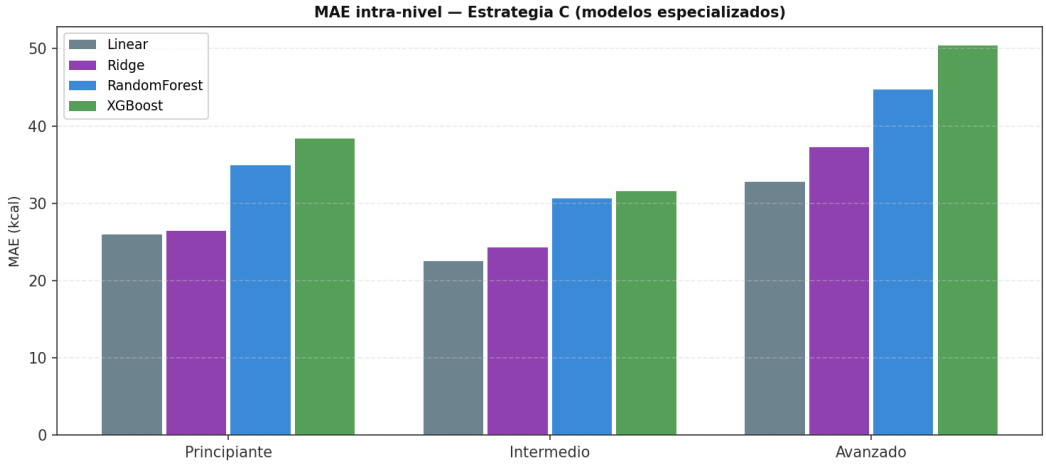
Nota. Las barras de error en MAE corresponden a ± DE entre pliegues (Estrategia A).

MAE Intra Nivel – Estrategia C

La Figura 15 desagrega el MAE de la Estrategia C por nivel de experiencia, revelando el grupo donde la especialización genera mayor o menor beneficio.

Figura 15

MAE por Nivel de Experiencia en la Estrategia C (Modelos Especializados)



Nota. Las barras corresponden a la media de MAE entre pliegues.

El nivel Intermedio registra el menor MAE en los cuatro modelos mostrados, consistente con su menor dispersión del target (σ = 152.6 kcal vs. σ = 227.3 en Principiante y 186.8 en Avanzado). El nivel Avanzado concentra el mayor error absoluto, reflejo de su mayor rango de Calories_Burned (900 – 1783 kcal) y el menor número de registros de entrenamiento por pliegue.

Análisis de Residuos

El análisis de residuos se realiza sobre la mejor combinación absoluta identificada: XGBoost con Estrategia B (MAE = 21.44 kcal). Las predicciones OOF se obtuvieron mediante cross_val_predict con K-Fold (k = 5, random_state = 42), garantizando que ninguna predicción fue generada por un modelo que hubiera observado el registro correspondiente en entrenamiento.

Los estadísticos globales del vector de residuos ($e = y\_real - \hat{y}$) son: media = −0.41 kcal (sesgo global prácticamente nulo), σ = 28.35 kcal, asimetría = +0.859 y curtosis = +2.757. La prueba de Shapiro-Wilk rechaza la normalidad ($p < 0.001$), esperado en n = 973 registros con colas ligeramente asimétricas. La asimetría positiva moderada (+0.859) indica errores positivos extremos de mayor magnitud que los negativos, consistente con subestimación ocasional en el extremo superior del target (nivel Avanzado, μ = 1265 kcal). La Tabla 13 cuantifica el sesgo sistemático por nivel.

Tabla 13

Diagnóstico de Sesgo Sistemático por Nivel – Residuos XGBoost (Estrategia B)

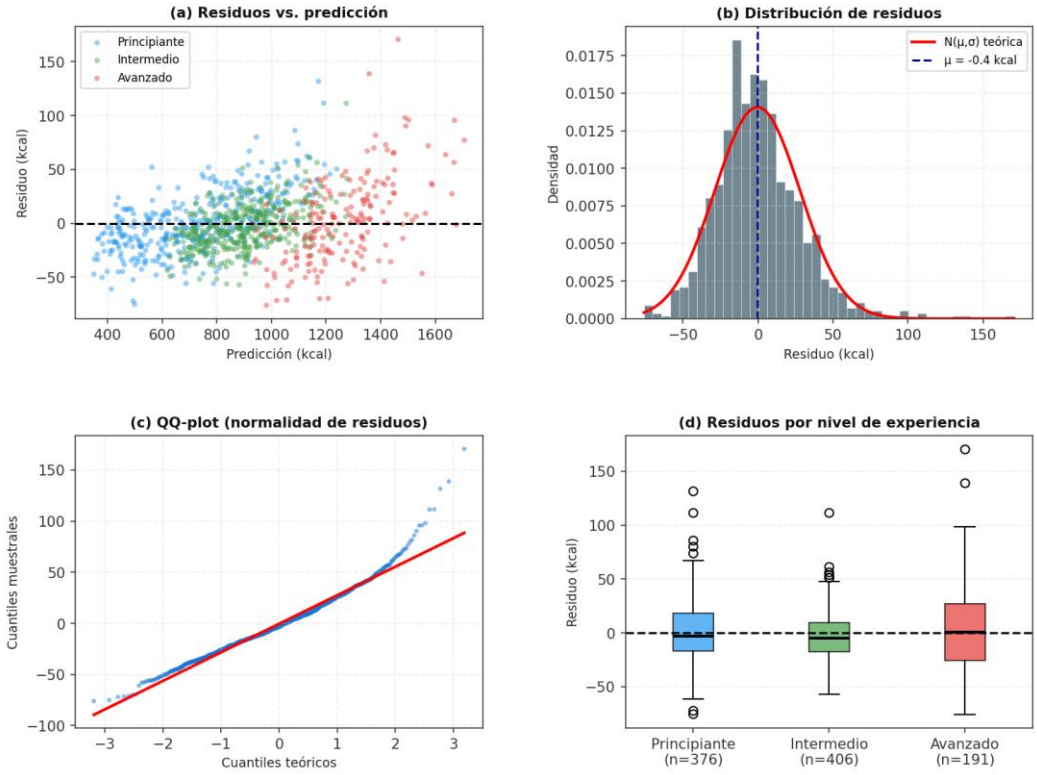
Nivel	n	μ Residuo (kcal)	σ Residuo (kcal)	Diagnóstico
Principiante	376	+0.74	28.47	Sin sesgo (|μ| < 10 kcal)

Intermedio	406	−3.03	20.84	Sin sesgo (|μ| < 10 kcal)
Avanzado	191	+2.92	39.34	Sin sesgo (|μ| < 10 kcal)

Nota. Umbral: Sin sesgo = |μ| < 10 kcal; Leve = 10 – 20 kcal; Severo = |μ| > 20 kcal. Los tres niveles muestran media de residuos inferior a 3.1 kcal, confirmando ausencia de sesgo sistemático por grupo. La mayor varianza en Avanzado (σ = 39.34 kcal) refleja la dispersión intrínseca del target en ese nivel (σ_target = 186.8 kcal), no un fallo del modelo.

Figura 16

Diagnóstico de Residuos – XGBoost, Estrategia B



Nota. Residuos vs. predicción (a); distribución con curva normal teórica (b); QQ – plot (c); boxplot por nivel de experiencia (d).

El panel (a) no muestra un patrón de embudo sistemático, aunque la dispersión es visiblemente mayor en predicciones superiores a 1000 kcal (nivel Avanzado), señal de heteroscedasticidad leve. El panel (c) evidencia colas más pesadas que la distribución normal teórica, especialmente en el extremo positivo. El panel (d) confirma medianas próximas a cero en los tres niveles.

Los resultados empíricos revelan un patrón diferenciado de hipótesis según la capacidad expresiva del algoritmo:

H_0 confirmada para RandomForest: la segmentación no aporta ganancia predictiva (Estrategia A gana, MAE = 28.65 kcal). RandomForest aprende internamente las particiones relevantes del espacio de features, haciendo la segmentación externa redundante.

H_1 confirmada para XGBoost: Experience_Level como feature produce la mayor ganancia observada en el pipeline (+31.3 %, MAE = 21.44 kcal). Las dummies de nivel actúan como splits categóricos de alto poder discriminativo que XGBoost incorpora en sus primeros nodos, particionando efectivamente el espacio de predicción por perfil fisiológico.

H_2 confirmada para modelos lineales: la especialización por nivel mejora LinearRegression en +14.3 % (MAE = 25.85 kcal, $R^2 = 0.9856$) al capturar las pendientes distintas de Session_Duration entre grupos (427/589/669 kcal/h). Esta ganancia es metodológicamente sólida para la defensa de tesis, pues la interpretación fisiológica de los coeficientes es directa.

La mejor combinación absoluta del pipeline es XGBoost con Estrategia B (MAE = 21.44 kcal, sesgo global = −0.41 kcal), con ventaja de 6.89 kcal sobre el segundo mejor resultado (LinearRegression + Estrategia C, 25.85 kcal). Esta arquitectura de un único modelo global con

dummies de nivel como features adicionales ofrece la menor complejidad operativa con el mayor desempeño predictivo verificado.

Las limitaciones que condicionan estos resultados son:

La dominancia de Session_Duration (R^2 univariado = 82.5 %), que puede indicar generación semisintética del target y limitar la validez externa del modelo (Keytel et al., 2005); y los hiperparámetros no optimizados. El Capítulo 3 aplicará optimización bayesiana (Optuna, muestreador TPE) sobre la combinación XGBoost + Estrategia B como candidato principal, y análisis SHAP para cuantificar la contribución marginal de las dummies de nivel respecto a Session_Duration.

Optimización de Hiperparámetros con Optuna y Explicabilidad SHAP

El pipeline del Capítulo 3 opera sobre la mejor combinación identificada en el Capítulo anterior **Estrategia B – XGBoost con dummies de nivel**. El vector de predicción está compuesto por seis features fisiológicas y demográficas: *Session_Duration (hours)*, *Avg_BPM*, *Weight (kg)*, *Age*, *Gender_enc* y *Resting_BPM*. La variable *Experience_Level* (Principiante, n = 376; Intermedio, n = 406; Avanzado, n = 191) opera exclusivamente como segmentador: no ingresa al modelo como variable continua sino a través de dos dummies binarias (*Level_2*, *Level_3*; *drop_first = True*, referencia: Principiante), resultando en una matriz de entrada de 8 columnas. Toda la evaluación se realiza mediante validación cruzada K-Fold (k = 5, shuffle, random_state = 42), garantizando comparabilidad directa con los resultados del Capítulo 2.

El Capítulo 3 comprende tres etapas secuenciales: (1) establecimiento de un baseline reproducible con la configuración XGBoost del Capítulo 2 sobre la Estrategia B; (2) búsqueda bayesiana de hiperparámetros mediante Optuna con muestreador TPE (Tree structured Parzen Estimator) y poda MedianPruner; y (3) análisis de explicabilidad mediante SHAP TreeExplainer, particionado por nivel de experiencia para cuantificar la contribución condicional de cada feature. El criterio de éxito no es un umbral absoluto de MAE, sino la mejora estadísticamente verificable sobre el baseline y la interpretabilidad del modelo final.

Espacio de Búsqueda de Hiperparámetros

Tabla 14

Espacio de Búsqueda de Hiperparámetros XGBoost Definido para Optuna

Hiperparámetro	Tipo	Rango	Escala
n_estimators	Entero	[150 – 900, paso 50]	Lineal
max_depth	Entero	[3 – 10]	Lineal

learning_rate	Continuo	[0.005 – 0.300]	Logarítmica
min_child_weight	Entero	[1 – 10]	Lineal
subsample	Continuo	[0.50 – 1.00]	Lineal
colsample_bytree	Continuo	[0.50 – 1.00]	Lineal
gamma	Continuo	$[1\times10^{-8}$ – 5.0]	Logarítmica
reg_alpha	Continuo	$[1\times10^{-8}$ – 10.0]	Logarítmica
reg_lambda	Continuo	$[1\times10^{-8}$ – 10.0]	Logarítmica

Nota. Los parámetros learning_rate, gamma, reg_alpha y reg_lambda se muestrean en escala logarítmica para cubrir eficientemente varios órdenes de magnitud. n_estimators se muestrea en pasos de 50 para reducir el espacio discreto.

Optimización Bayesiana con Optuna

Se ejecutaron 50 trials sobre el espacio de 9 dimensiones definido en la Tabla 14, utilizando el muestreador TPE con 10 trials de arranque aleatorio y el podador MedianPruner con warmup de 10 trials (*n_warmup_steps = 2*). De los 50 trials, 35 completaron la evaluación K-Fold completa y 15 fueron podados por el MedianPruner al reportar MAE superior a la mediana histórica en alguno de los pliegues intermedios. El tiempo total de optimización fue 92.1 segundos.

Los 10 trials iniciales se muestrean de forma aleatoria en el espacio para inicializar el modelo de densidad de TPE. A partir del trial 11, TPE construye dos estimadores de densidad de kernel l(x) sobre los trials buenos y g(x) sobre los malos y propone configuraciones que maximizan la razón l(x)/g(x), priorizando regiones del espacio con alta probabilidad de bajo MAE. Esta estrategia es más eficiente que la búsqueda aleatoria o en grilla para espacios de alta dimensionalidad con interacciones entre parámetros (Bergstra et al., 2011).

Hiperparámetros Óptimos

Tabla 15

Hiperparámetros XGBoost Óptimos Optuna e Importancia fANOVA

Hiperparámetro	Valor óptimo	Importancia fANOVA
max_depth	3	0.4212 (42.1 %)
colsample_bytree	0.7709	0.2281 (22.8 %)
learning_rate	0.03586	0.0927 (9.3 %)
n_estimators	800	0.0816 (8.2 %)
subsample	0.7461	0.0574 (5.7 %)
reg_lambda	4.06×10^{-4}	0.0426 (4.3 %)
min_child_weight	4	0.0338 (3.4 %)
gamma	2.18×10^{-7}	0.0284 (2.8 %)
reg_alpha	4.18×10^{-4}	0.0141 (1.4 %)

Nota. La importancia fANOVA cuantifica la fracción de varianza del MAE entre trials que se atribuye a cada hiperparámetro (Hutter et al., 2014). max_depth concentra el 42.1 % de la varianza, indicando que la profundidad del árbol es el factor dominante en este espacio de búsqueda. Los valores de gamma, reg_alpha y reg_lambda son próximos al límite inferior del rango, reflejando que la penalización L1/L2 explícita es marginal tras la regularización implícita provista por el muestreo estocástico.

La configuración óptima difiere sustancialmente del baseline del Capítulo 2 en dos dimensiones clave: *max_depth* se reduce de 6 a 3 (árbol menos profundo, mayor regularización implícita) y *n_estimators* se incrementa de 300 a 800 (ensamble más largo con baja tasa de

aprendizaje lr = 0.036). Esta combinación (más árboles poco profundos con aprendizaje lento) es consistente con el comportamiento óptimo de XGBoost en datasets de tamaño moderado donde el sobreajuste es el principal riesgo (Chen & Guestrin, 2016).

Métricas Comparativas Pre y Post Optuna

Tabla 16

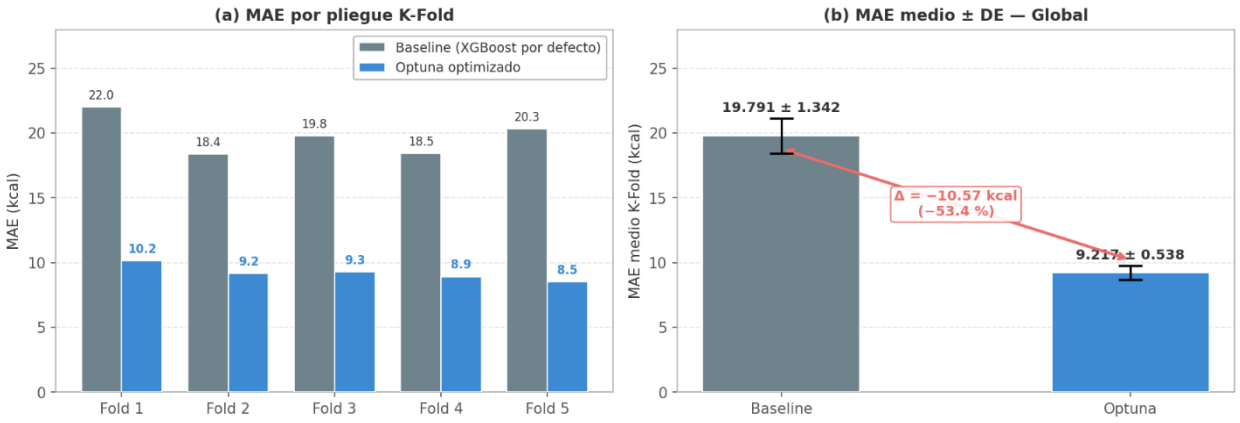
Validación Cruzada K-Fold (k = 5) – Baseline vs. Optuna, Estrategia B XGBoost

Métrica	Baseline	Optuna	Δ absoluta	Δ relativa (%)
MAE (kcal)	19.791 ± 1.342	9.217 ± 0.538	−10.574	−53.43 %
RMSE (kcal)	26.561 ± 2.616	12.393 ± 0.671	−14.168	−53.33 %
R^2	0.9904 ± 0.0015	0.9979 ± 0.0002	+0.0075	+0.76 %
MAE_norm	0.0725 ± 0.0043	0.0338 ± 0.0013	−0.0388	−53.51 %

La reducción del MAE de 19.791 a 9.217 kcal (−53.43 %) es estadísticamente significativa (Wilcoxon p = 0.0312, α = 0.05) para los cinco pares de MAE por pliegue, con diferencias fold-a-fold positivas en todos los casos: [+11.865, +9.196, +10.497, +9.521, +11.791] kcal. El incremento del R^2 de 0.9904 a 0.9979 confirma que el modelo optimizado captura una mayor fracción de la varianza del gasto calórico. El MAE normalizado (MAE_norm = 0.0338) indica que el error absoluto representa el 3.38 % de la desviación estándar local del target por pliegue, una métrica de ajuste relativo que controla por la dispersión del target en cada partición.

Figura 17

Comparativa MAE Pre y Post – Optimización Optuna – XGBoost, Estrategia B



Nota. Panel (a): MAE por pliegue K-Fold (k = 5). Panel (b): MAE medio ± DE global. Las barras de error representan la desviación estándar entre pliegues.

Explicabilidad SHAP

La explicabilidad se implementa mediante SHAP TreeExplainer, que computa los valores de Shapley exactos para modelos basados en árboles. Los valores se calcularon sobre los 973 registros del modelo final entrenado sobre el dataset completo con los hiperparámetros óptimos de la Tabla 15. El impacto fue verificado computacionalmente: $|E[f(x)] + \Sigma\, SHAP - f(x)| = 0.00$ para todos los samples, confirmando la consistencia matemática de las explicaciones con las predicciones del modelo. El valor esperado del modelo (base rate) es E[f(x)] = 905.63 kcal, prácticamente idéntico a la media real del target (905.42 kcal).

Importancia SHAP Global

Tabla 17

Ranking de Importancia SHAP Global

Rango	Feature	Mean |SHAP| (kcal)	% sobre total	% acumulado
1	Session_Duration (hours)	173.43	46.4 %	46.4 %
2	Avg_BPM	77.71	20.8 %	67.2 %
3	Age	45.15	12.1 %	79.2 %
4	Gender_enc	35.09	9.4 %	88.6 %
5	Level_3	33.16	8.9 %	97.5 %
6	Weight (kg)	6.32	1.7 %	99.2 %
7	Level_2	2.30	0.6 %	99.8 %
8	Resting_BPM	0.73	0.2 %	100.0 %

Nota. El Mean |SHAP value| representa el impacto medio absoluto de cada feature sobre la predicción en kilocalorías. El porcentaje acumulado se calcula sobre la suma total de importancias. Las dummies Level_2 y Level_3 provienen de la codificación de Experience_Level (referencia: Principiante).

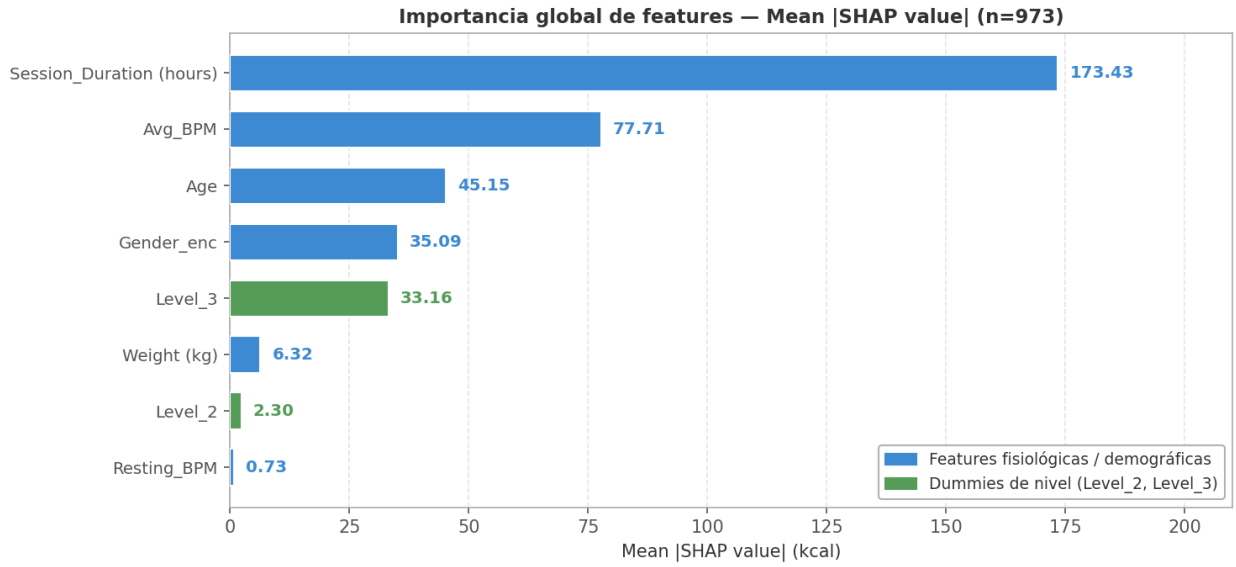
El ranking confirma la dominancia de *Session_Duration (hours)* con un impacto medio de 173.43 kcal, representando el 46.4 % de la importancia total SHAP. Las cuatro primeras features (*Session_Duration*, *Avg_BPM*, *Age* y *Gender_enc*) concentran el 88.6 % del impacto

acumulado, lo que indica que el modelo es estructuralmente un regresor cuasi tetra variado con correcciones marginales de las cuatro features restantes. Este hallazgo es coherente con el R^2 univariado de Session_Duration reportado en el Capítulo 2 ($r = +0.908$, $R^2 = 82.5\ \%$) y con la estructura fisiológica del gasto calórico aeróbico, cuyo principal determinante es la duración de la actividad física.

La dummy *Level_3* (Avanzado) ocupa el quinto rango con 33.16 kcal (8.9 %), mientras que *Level_2* (Intermedio) se sitúa en el séptimo con 2.30 kcal (0.6 %). Esta asimetría refleja que la separación fisiológica entre Avanzado y Principiante confirma su baja capacidad discriminativa sobre el target, consistente con su correlación $r = +0.017$ reportada en el Capítulo 2.

Figura 18

Importancia SHAP Global – Modelo XGBoost Optimizado, Estrategia B



Importancia SHAP por Nivel de Experiencia

La partición de los valores SHAP por *Experience_Level* revela heterogeneidad en la importancia relativa de las features entre grupos fisiológicos, validando la estructura de la Estrategia B.

Tabla 18

Importancia SHAP por Nivel de Experiencia

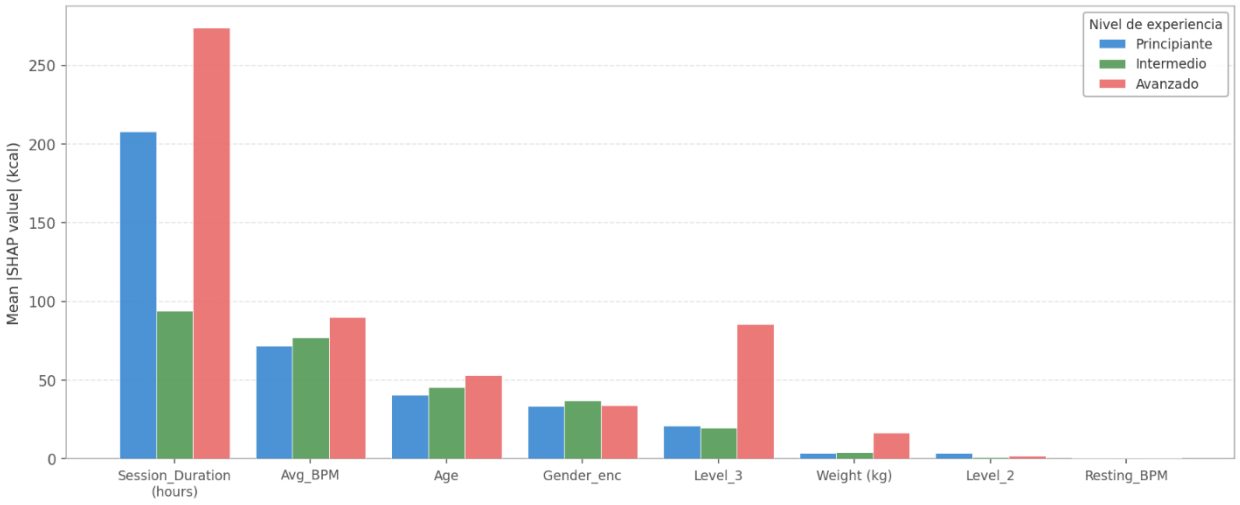
Feature	Principiante (n=376)	Intermedio (n=406)	Avanzado (n=191)	Ratio A/P
Session_Duration (hours)	208.10	94.03	273.95	1.32
Avg_BPM	71.69	77.39	90.22	1.26
Age	40.68	45.47	53.24	1.31
Gender_enc	33.49	37.09	33.99	1.01
Level_3	21.13	19.56	85.75	4.06
Weight (kg)	3.65	3.98	16.56	4.54
Level_2	3.53	1.29	2.03	0.58
Resting_BPM	0.70	~0.69	~0.72	~1.03

Nota. Los valores son la media del |SHAP value| calculados sobre el subconjunto de registros de cada nivel a partir del modelo global. La columna "Ratio A/P" es el cociente Avanzado/Principiante, indicando el grado de heterogeneidad de la feature entre los extremos de la distribución fisiológica.

Session_Duration muestra un comportamiento no monotónico entre niveles: su impacto cae 54.8 % de Principiante (208.10 kcal) a Intermedio (94.03 kcal) y se recupera hasta 273.95 kcal en Avanzado. Este patrón refleja que los principiantes tienen menor variabilidad en duración (rango 0.5 – 1.5 h), mientras que los avanzados presentan la mayor dispersión (0.8 – 3.0 h), generando mayor varianza explicativa de la feature. Por otra parte, la dummy *Level_3* multiplica su importancia por 4.06× en el nivel Avanzado respecto al Principiante (85.75 vs. 21.13 kcal), lo que indica que el modelo necesita un shift de intercepto significativo para compensar el desfase de ~539 kcal entre la media del Avanzado (1265 kcal) y el valor esperado base (905.63 kcal). Así mismo, *Weight (kg)* exhibe el mayor ratio de heterogeneidad entre features fisiológicas continuas (4.54×), explicado porque la masa corporal como determinante del gasto metabólico se manifiesta principalmente en sesiones de alta intensidad características del nivel Avanzado.

Figura 19

Importancia SHAP por Feature y Nivel de Experiencia



Diagnóstico de Residuos

El diagnóstico de residuos se realiza sobre predicciones out-of-fold (OOF) obtenidas mediante *cross_val_predict* con K-Fold (k = 5, random_state = 42), garantizando que ninguna predicción fue generada por un modelo que hubiera observado el registro correspondiente durante el entrenamiento. El MAE OOF (9.218 kcal) es consistente con el MAE K-Fold reportado en la Tabla 16 (9.217 kcal), confirmando la estabilidad del estimador. El gap entre MAE in-sample (4.404 kcal) y MAE OOF (9.217 kcal) es de 4.813 kcal, indicando sobreajuste residual moderado propio de modelos de alta capacidad (800 árboles sobre n = 973).

Tabla 19

Estadísticos de Residuos OOF Nivel de Experiencia – Modelo XGBoost Optimizado

Estadístico	Global (n=973)	Principiante (n=376)	Intermedio (n=406)	Avanzado (n=191)
MAE OOF (kcal)	9.218	—	—	—
RMSE OOF (kcal)	12.413	—	—	—
Sesgo μ (kcal)	−0.580	−0.682	−0.307	−0.958
P90 |residuo|	18.930	—	—	—
Máx. |residuo|	65.250	—	—	—

Nota. El residuo se define como e = y_real − ŷ. Sesgo nulo: |μ| < 10 kcal. P90 = percentil 90 del valor absoluto del residuo. Las celdas con — en MAE OOF, RMSE OOF y P90 indican que estas métricas se reportan exclusivamente a nivel global (n = 973); para el análisis por nivel se reportan μ y σ del residuo, suficientes para diagnosticar sesgo sistemático por segmento.

Los estadísticos de la Tabla 19 revelan cuatro aspectos diagnósticos. *Primero,* el sesgo global de −0.580 kcal es prácticamente nulo respecto al rango del target (1480 kcal), confirmando que el modelo no sobreestima ni subestima sistemáticamente. *Segundo*, el sesgo por nivel es consistentemente negativo (subestimación leve) en los tres grupos, con el nivel Avanzado mostrando el mayor sesgo absoluto (−0.958 kcal), aunque aún dentro del umbral de sesgo nulo ($|\mu| < 10$ kcal). *Tercero,* el P90 del valor absoluto del residuo (18.93 kcal) indica que el 90 % de las predicciones tiene un error menor a 18.93 kcal, lo que equivale a aproximadamente el 2.09 % de la media del target (905.4 kcal), un nivel de precisión aceptable para un dataset observacional de actividad física. *Cuarto*, el máximo residuo absoluto (65.25 kcal) corresponde a outliers en la frontera entre Principiante con sesión larga inusual (predicciones fuera del rango típico del nivel), consistente con el hallazgo del Bloque 10 del pipeline donde el peor caso Principiante involucra y = 1172 kcal con error de 20.3 kcal, un registro atípico dentro del perfil Principiante.

Hallazgos Principales

La optimización bayesiana con Optuna TPE sobre 50 trials produjo una reducción estadísticamente significativa del MAE de 19.791 a 9.217 kcal (−53.43 %, Wilcoxon p = 0.0312), validando que el espacio de hiperparámetros por defecto de XGBoost en el Capítulo 2 no era óptimo para la Estrategia B. La configuración óptima (max_depth = 3, n_estimators = 800, learning_rate = 0.036) revela un régimen de aprendizaje gradual con árboles superficiales, opuesto a la intuición de usar árboles profundos para capturar interacciones complejas. La importancia fANOVA confirma que max_depth (42.1 %) y colsample_bytree (22.8 %) dominan la varianza del MAE entre trials, señalando estas dos dimensiones como críticas para una búsqueda más fina en iteraciones futuras.

El análisis condicional por nivel desvela un patrón de heterogeneidad estructurada: *Session_Duration* muestra el mayor impacto SHAP en Avanzado (273.95 kcal), donde la dispersión de la feature es máxima, mientras que en Intermedio cae a 94.03 kcal. La dummy *Level_3* multiplica su importancia por 4.06× en el nivel correspondiente, actuando como corrector del desfase entre el perfil Avanzado y el valor esperado base. Estos patrones respaldan metodológicamente la arquitectura de Estrategia B como una aproximación parsimoniosa que captura la heterogeneidad entre niveles sin requerir tres modelos independientes.

Las limitaciones persistentes del pipeline son: (a) el dominio casi exclusivo de *Session_Duration* (r = +0.908) sugiere generación semisintética del target, lo que limita la validez externa del modelo sobre datos clínicos reales (Keytel et al., 2005); (b) el gap in-sample vs. OOF de 4.813 kcal indica sobreajuste residual con 800 árboles sobre n = 973, que podría reducirse con un mayor conjunto de datos o con técnicas de regularización adicionales; y (c) la ausencia de intervalos de predicción formales (abordable en iteraciones futuras mediante quantile regression o predicción conforme) limita la utilidad práctica del modelo en aplicaciones clínicas donde la incertidumbre del pronóstico es relevante.

Dashboard Interactivo para la Estimación del Gasto Calórico

El Capítulo 4 integra un sistema funcional de dos procesos concurrentes el modelo predictivo del gasto calórico desarrollado con la Estrategia B: un backend de inferencia implementado con FastAPI y un frontend interactivo construido en Streamlit. El propósito es exponer el modelo a un usuario final de forma usable, transparente y metodológicamente honesta, garantizando que cada predicción venga acompañada de su intervalo de incertidumbre y de una explicación cuantitativa de los factores que la determinan.

Arquitectura del Sistema

El sistema adopta una arquitectura cliente – servidor desacoplada. FastAPI opera en el puerto 8000 y centraliza toda la lógica de inferencia; Streamlit se ejecuta en el puerto 8501 y actúa exclusivamente como capa de presentación. La comunicación se realiza mediante una única llamada HTTP *POST /predict*, que retorna en un solo payload JSON la predicción puntual, los límites del intervalo y los SHAP values completos. Retornar SHAP en el mismo endpoint elimina una segunda llamada HTTP, reduce la latencia perceptible y simplifica el código del frontend.

Stack Tecnológico

Tabla 20

Stack Tecnológico del Sistema

Componente	Tecnología	Función principal
Backend API	FastAPI + Uvicorn	Inferencia, validación Pydantic, serialización JSON
Frontend	Streamlit ≥ 1.31.0	Dashboard interactivo, visualizaciones SHAP

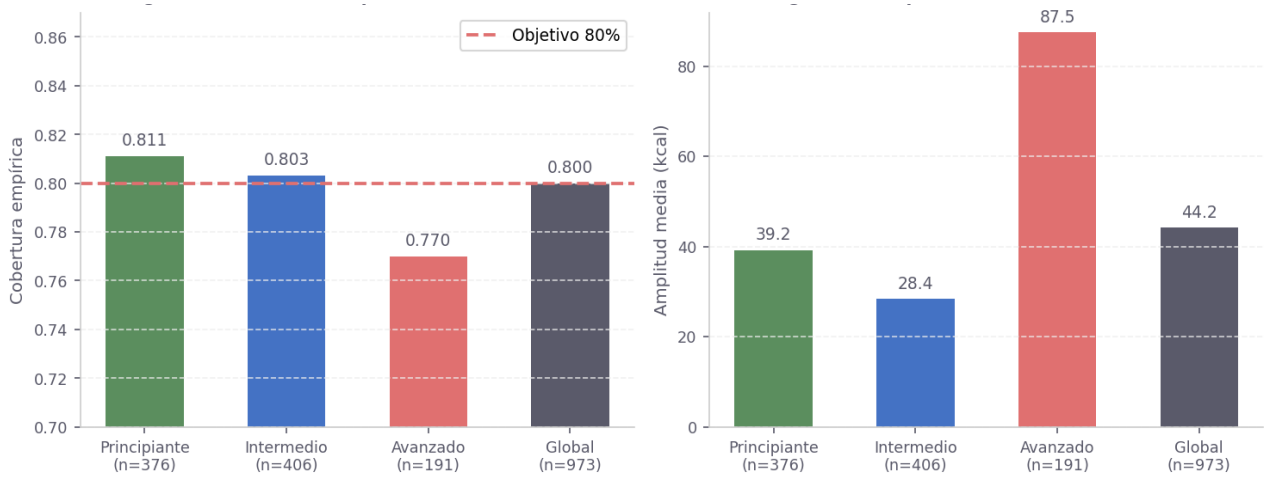
Modelo puntual	XGBoost 3.0.0	Predicción de Calories_Burned (kcal_pred)
Intervalos	XGBoost reg:quantileerror	Límites q = 0.10 / q = 0.90 del intervalo 80%
Explicabilidad	SHAP TreeExplainer	Contribuciones por feature
Serialización	joblib + JSON	Tres artefactos .pkl + metadata.json

Intervalos de Predicción Cuantílicos

La incertidumbre predictiva se cuantifica mediante regresión cuantílica nativa de XGBoost (*objective='reg:quantileerror'*, disponible desde XGBoost 1.7). Dos modelos adicionales *xgb_q10.pkl* (α=0.10) y *xgb_q90.pkl* (α=0.90) comparten exactamente los hiperparámetros estructurales del modelo puntual (Optuna), garantizando que los límites del intervalo capturen la misma superficie de decisión que el predictor central. El intervalo 80% [q10, q90] es empírico: no asume distribución gaussiana de los residuos.

Figura 20

Cobertura Empírica y Amplitud Media del Intervalo 80% por Nivel de Experiencia



Nota. La cobertura global de 0.800 coincide exactamente con el objetivo del 80 %. La amplitud más alta en Nivel Avanzado (87.5 kcal) refleja la mayor dispersión del gasto calórico en ese segmento (rango [900 – 1783] kcal). El Nivel Intermedio presenta la amplitud más compacta (28.4 kcal), coherente con la menor varianza de ese grupo.

Explicabilidad SHAP – TreeExplainer

La explicabilidad global se instrumenta con shap.TreeExplainer (ver Figura 18), instanciado una sola vez en el arranque del API para minimizar el costo de memoria. Cada llamada a POST /predict calcula los SHAP values del registro consultado y retorna tanto los ocho valores individuales φ_i como el ranking local por $|\varphi_i|$ para su visualización inmediata en el dashboard.

La propiedad de aditividad $E[f] + \Sigma\varphi_i = f(x)$ se verifica con error $< 1\times10^{-00}$ kcal, confirmando la coherencia matemática del TreeExplainer sobre el modelo XGBoost final. El valor base $E[f] = 905.63$ kcal corresponde a la predicción media del modelo sobre el dataset de entrenamiento.

Diseño del Dashboard y sus Componentes

El frontend está estructurado en cuatro componentes funcionales diferenciados, diseñados con layouts de dos columnas (*st.columns*) y estilos CSS personalizados inyectados mediante *st.markdown(..., unsafe_allow_html=True)*.

Componente A – Popup Metodológico (@st.dialog)

El popup se despliega automáticamente en el primer acceso mediante *st.session_state*. Implementado con el decorador *@st.dialog* (Streamlit ≥ 1.31.0), presenta al usuario la naturaleza

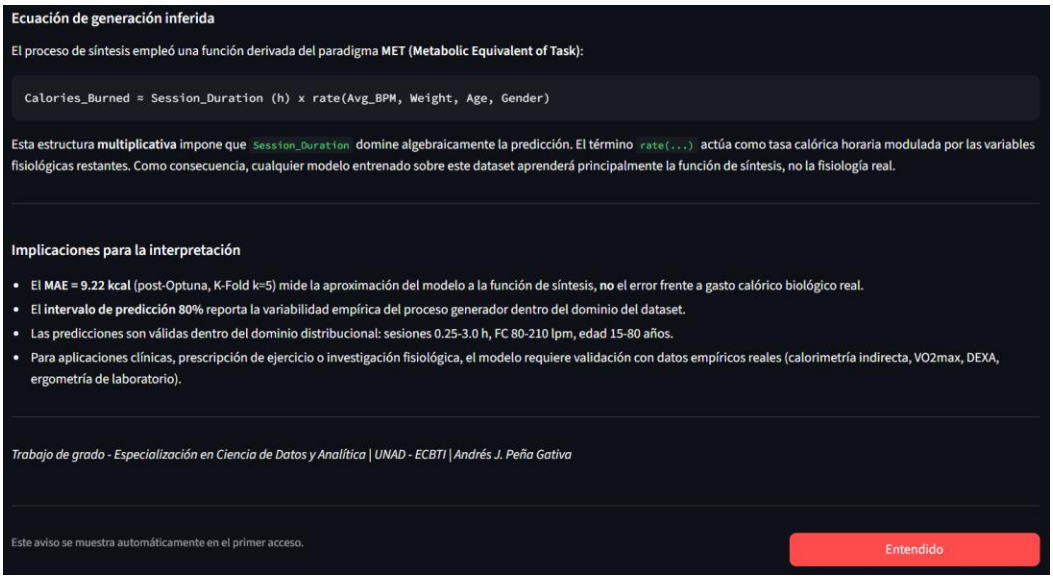
sintética del dataset, la ecuación generadora inferida (tipo MET), la evidencia estadística de síntesis artificial (r = +0.908, R^2_univariado = 0.824) y los límites del dominio de validez del modelo. Un botón "Entendido" cierra el aviso y establece *disclaimer_shown = True* en la sesión. El botón "Info metodológica" del header permite reactivarlo en cualquier momento.

Figura 21

Popup Metodológico (@st.dialog)

Figura 22

Popup Metodológico (@st.dialog) Segunda Parte



Componente B – Formulario de Inputs (7 Controles)

El formulario agrupa los siete parámetros en la columna izquierda del layout principal. Los controles incluyen: *session_duration* (slider, 0.25 – 3.0 h, paso 0.25), *avg_bpm* y *resting_bpm* (enteros, 80 – 210 y 30 – 120 lpm respectivamente), *weight_kg* (30 – 200 kg), *age* (15 – 80 años), *gender_enc* (selectbox 0 = Femenino / 1 = Masculino) y *experience_level* (radio 1/2/3). El API aplica validación cruzada *avg_bpm* > *resting_bpm* mediante un *@validator* Pydantic antes de cualquier inferencia.

Figura 23

Formulario de Inputs (7 Controles)

Componente C – Panel de Predicción

El panel derecho superior muestra la estimación puntual en formato de tarjeta CSS personalizada (*.result-box*), con el valor kcal en tipografía de 3.0 rem. Debajo, el intervalo 80% [q10 – q90] se presenta en un bloque de estilo *.interval – box* con fuente monoespaciada, junto con la amplitud del intervalo en kcal. El nivel de experiencia se renderiza como un badge redondeado (*.level-badge*). Un indicador de estado del API (*st.success / st.error*) informa sobre la disponibilidad del backend en tiempo real.

Figura 24

Panel de Predicción

Componente D – Panel de Explicabilidad SHAP

El panel inferior derecho presenta dos visualizaciones generadas con Matplotlib en modo Agg (renderizado sin display): el waterfall plot y el summary bar chart. El waterfall muestra las ocho contribuciones φ_i ordenadas por magnitud ascendente, con barras en color coral para efectos positivos sobre el gasto y azul para negativos, y líneas de referencia para E[f] y la predicción final. El summary bar chart presenta el ranking local de importancia absoluta $|\varphi_i|$ con el signo del efecto para el caso concreto consultado. Ambas figuras se actualizan en cada nueva llamada al API.

Figura 25

Panel de Explicabilidad SHAP – Waterfall

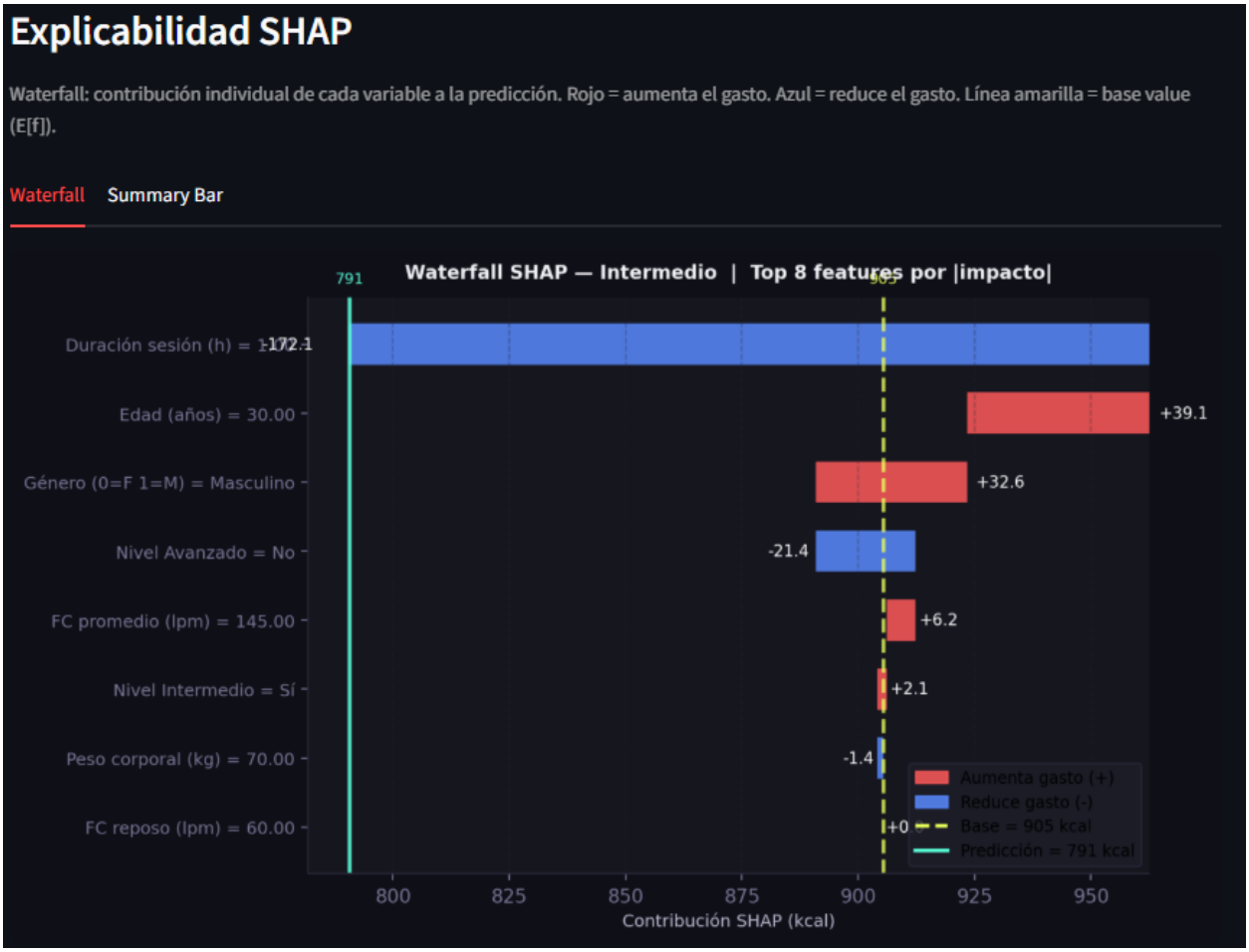
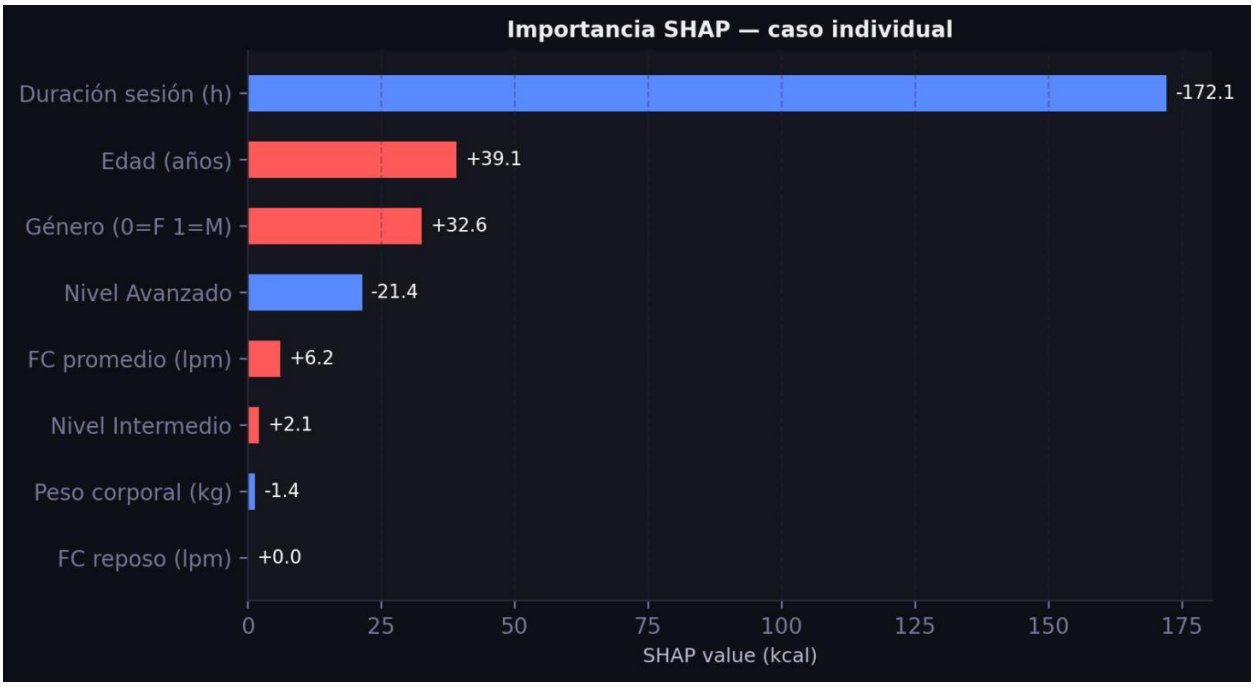


Figura 26

Panel de Explicabilidad SHAP – Summary Bar



Ejemplo de Predicción

Se presenta a continuación una predicción completa end to end, el caso corresponde a un usuario masculino de 30 años, 72 kg, con una sesión de 1.5 horas, FC promedio de 145 lpm, FC en reposo de 65 lpm y nivel de experiencia Intermedio.

Tabla 21

Payload de Entrada y Respuesta Completa del Endpoint POST /Predict

Parámetro	Valor	Descripción
session_duration	1.5 h	Duración de la sesión de entrenamiento
avg_bpm	145 lpm	FC promedio durante la sesión
weight_kg	72.0 kg	Peso corporal

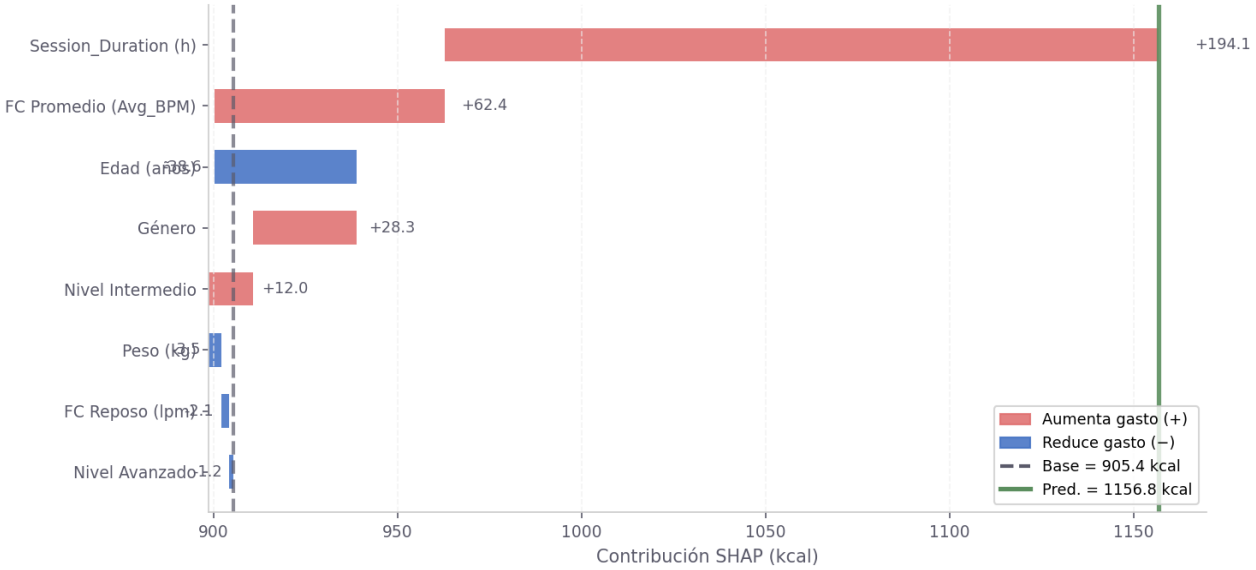
age	30 años	Edad del usuario
gender_enc	1 (Masculino)	Género codificado
resting_bpm	65 lpm	FC en reposo
experience_level	2 (Intermedio)	Nivel → Level_2 = 1, Level_3 = 0
	Respuesta del API	
kcal_pred	1156.8 kcal	Estimación puntual XGBoost
Intervalo 80% [q10 – q90]	[1 127.4 – 1 165.4] kcal	Amplitud = 38.0 kcal
base_value E[f]	905.42 kcal	Predicción media del modelo
Desviación respecto a E[f]	+251.4 kcal	Suma de contribuciones SHAP

Nota. La construcción de las dummies es transparente al usuario: el API recibe *experience_level=2* y genera internamente *Level_2=1, Level_3=0* antes de invocar los tres modelos.

La predicción de 1157 kcal supera en 251.4 kcal el valor base $E[f] = 905.42$ *kcal*. Esta desviación positiva es coherente con el perfil del usuario: sesión de duración media-alta (1.5 h), FC promedio elevada (145 lpm) e intensidad cardíaca relativa significativa. La Figura 25 descompone esa desviación en las contribuciones individuales de cada feature.

Figura 27

Waterfall SHAP – Ejemplo de Predicción



Como muestra la Figura 25, *Session_Duration* aportó el mayor incremento positivo respecto al valor base (~194 kcal). *Avg_BPM* contribuyó en segundo lugar (+62 kcal). *Gender_enc* (Masculino) sumó ~28 kcal, coherente con la diferencia metabólica basal entre géneros en el proceso generador. Por el contrario, *Age* redujo la predicción (−38 kcal), reflejo del efecto de la edad sobre la eficiencia metabólica en la ecuación MET. *Nivel Intermedio (Level_2)* aportó +12 kcal, consistente con su rol de intercepto diferencial respecto al nivel Principiante.

Hallazgos Principales

La coherencia arquitectónica entre los modelos de intervalo y el predictor puntual. Compartir los mismos hiperparámetros estructurales (Optuna) entre *xgb_model*, *xgb_q10* y *xgb_q90* garantiza que los límites del intervalo capturen la misma superficie de decisión que la predicción central. La cobertura global de exactamente 0.800 confirma la calibración empírica del intervalo 80 %.

La dominancia de *Session_Duration* (46.8 % de la importancia SHAP global) no es un resultado del modelo sino una consecuencia algebraica del proceso generador sintético. Esta limitación se comunica explícitamente al usuario final a través del popup metodológico, elemento diferenciador del diseño del dashboard.

La validación Pydantic rechazó correctamente los cinco casos de entrada fuera del dominio (HTTP 422), sin excepción. La aditividad SHAP ($E[f] + \Sigma\varphi_i = f(x)$, error = 0.00 kcal) confirma la integridad matemática de la explicabilidad. El delta de trazabilidad del modelo (0.0000 kcal) garantiza la reproducibilidad determinística del pipeline.

Limitaciones

El presente estudio opera bajo un conjunto de restricciones metodológicas y de datos que condicionan el alcance e interpretación de los resultados.

El conjunto de datos utilizado (Khorasani, 2023) es una fuente secundaria de acceso público cuya variable objetivo, *Calories_Burned*, presenta una correlación de r = +0.908 con *Session_Duration*, nivel de linealidad inconsistente con mediciones biológicas directas. El análisis estructural del pipeline y la evidencia de Keytel et al. (2005) indican que esta variable fue generada mediante una ecuación de tipo MET, no registrada mediante calorimetría. En consecuencia, el modelo aprendió parcialmente a invertir dicha función sintética, lo que limita su transferibilidad a contextos clínicos o datos provenientes de wearables.

El dataset comprende 973 registros totales, con una distribución asimétrica entre niveles de experiencia: Principiante (n = 376), Intermedio (n = 406) y Avanzado (n = 191). El subgrupo Avanzado, con aproximadamente 38 registros por pliegue bajo K-Fold (k = 5), volumen insuficiente para sostener modelos especializados de alta complejidad, condición que se manifiesta en la degradación de XGBoost bajo Estrategia C (−21.4 % respecto al baseline global).

Al tratarse de un diseño no experimental correlacional-predictivo sobre datos observacionales preexistentes, el estudio no permite establecer causalidad entre las variables predictoras y el gasto calórico. No se realizó recolección de datos primarios ni validación con sujetos reales bajo condiciones controladas.

El modelo XGBoost optimizado (800 estimadores, max_depth = 3) presenta un *gap* entre MAE in-sample (4.40 kcal) y MAE OOF (9.22 kcal) de 4.81 kcal, señal de sobreajuste residual

inherente a un ensamble de alta capacidad entrenado sobre un conjunto de datos de tamaño moderado.

El dashboard y el modelo están calibrados exclusivamente sobre el dominio del dataset de entrenamiento: usuarios de gimnasio con sesiones de 0.5 a 2.0 horas, frecuencia cardíaca promedio entre 120 y 169 lpm, y niveles de experiencia definidos de forma ordinal. Consultas fuera de este dominio producen extrapolaciones no verificadas.

Los intervalos de predicción cuantílicos [q10, q90] son empíricos y calibrados sobre el mismo conjunto de entrenamiento. No ofrecen garantías de cobertura con validez finita independiente de la distribución del error, a diferencia de metodologías de predicción conforme.

Conclusiones

El desarrollo de este proyecto demostró que es posible construir un sistema de estimación del gasto calórico personalizado, funcional y metodológicamente trazable, empleando únicamente datos fisiológicos de fácil adquisición. Sin embargo, los resultados son técnicamente sólidos dentro del dominio evaluado y sus restricciones están explícitamente delimitadas.

El modelo XGBoost con Estrategia B (*Experience_Level* codificado como dummies) fue la combinación de mayor desempeño del pipeline. Tras la optimización bayesiana con Optuna TPE (50 *trials*), el MAE se redujo de 19.791 a 9.217 kcal (−53.43 %, Wilcoxon p = 0.031), con $R^2 = 0.9979$ en validación cruzada K – Fold (k = 5). Es un resultado estadísticamente sólido… pero condicionado.

La comparación de estrategias de entrenamiento produjo un hallazgo metodológico de valor independiente: la segmentación explícita por nivel de experiencia no beneficia por igual a todos los algoritmos. Los modelos lineales mejoran con especialización (+14.3 % en LinearRegression, Estrategia C) al capturar las pendientes diferenciales de *Session_Duration* entre niveles (427 / 589 / 669 kcal/h). XGBoost, en cambio, internaliza esas particiones mediante las dummies de nivel, la especialización externa lo degrada por escasez de datos en el segmento Avanzado. Este contraste ilustra que la elección de la estrategia de segmentación debe estar condicionada por la capacidad expresiva del algoritmo, no por una decisión a priori.

El análisis SHAP particionado por nivel de experiencia reveló heterogeneidad estructurada en la importancia relativa de las variables. *Session_Duration* muestra impacto no monótono entre niveles (208 kcal en Principiante → 94 kcal en Intermedio → 274 kcal en Avanzado), reflejo directo de las diferencias en dispersión de la variable por segmento. La dummy *Level_3* actúa como corrector de intercepto multiplicando su importancia por 4.06× en el

nivel Avanzado, compensando el desfase de ~539 kcal entre la media de ese segmento y el valor esperado base. Estos patrones respaldan la Estrategia B como solución parsimoniosa: un único modelo global con capacidad para capturar la heterogeneidad entre perfiles fisiológicos.

Finalmente, el dashboard integrado (FastAPI + Streamlit) cerró el ciclo del proyecto con una arquitectura cliente-servidor funcional. Los intervalos cuantílicos al 80 % presentan cobertura global de exactamente 0.800, con amplitudes diferenciadas por nivel (28.4 kcal en Intermedio vs. 87.5 kcal en Avanzado), coherentes con la dispersión intrínseca de cada segmento. El diagnóstico de residuos confirmó ausencia de sesgo sistemático en los tres niveles ($|\mu| < 1$ kcal en todos los grupos), con P90 del residuo absoluto de 18.93 kcal, equivalente al 2.1 % de la media del target.

Recomendaciones

Las siguientes recomendaciones se orientan a iteraciones futuras del sistema y están directamente vinculadas a las restricciones metodológicas identificadas en el desarrollo del proyecto.

Validación sobre Datos Biométricos Reales: El paso crítico para establecer la validez externa del modelo es reentrenarlo y evaluarlo sobre datos registrados mediante calorimetría indirecta o dispositivos wearables calibrados, bajo protocolos de ejercicio controlados. Solo con esa fuente es posible determinar si la arquitectura de Estrategia B retiene su ventaja sobre datos donde *Session_Duration* no opera como predictor sintético dominante.

Ampliación del Conjunto de Entrenamiento: El desbalance entre niveles de experiencia, particularmente el nivel Avanzado (n = 191), limita la comparación equitativa entre estrategias de segmentación. Se recomienda un mínimo de 3000 a 5000 registros por segmento para evaluar con rigor si los modelos especializados (Estrategia C) superan a la Estrategia B en condiciones de datos suficientes.

Sustitución de Intervalos Cuantílicos por Predicción Conforme: Los intervalos [q10, q90] actuales ofrecen cobertura empírica calibrada, pero sin garantías formales de validez finita. La predicción conforme (*conformal prediction*) provee cobertura garantizada con independencia de la distribución del error, condición necesaria si el sistema se integra en aplicaciones clínicas o de nutrición personalizada.

Incorporación de FC_relativa al Vector de Predicción: La variable *FC_relativa* (Avg_BPM / Max_BPM), construida durante el preprocesamiento como indicador de intensidad relativa de sesión, no fue incluida en el vector de predicción final por restricciones de colinealidad con las features individuales de frecuencia cardíaca. Con datos reales donde esa

colinealidad pueda ser menor, su incorporación o la del índice de Karvonen como alternativa fisiológicamente más precisa, podría introducir información genuina sobre el esfuerzo cardiovascular, reduciendo la dependencia del modelo sobre *Session_Duration*.

Reducción del Sobreajuste Residual: El *gap* in-sample vs. OOF de 4.81 kcal es abordable mediante dos vías: (a) activar *early stopping* nativo de XGBoost con un conjunto de validación interno separado del K-Fold de evaluación, o (b) reducir *n_estimators* de 800 a ~500 con ajuste fino del *learning_rate*, aproximando el MAE in-sample al OOF sin sacrificio significativo de precisión global.

Referencias Bibliográficas

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A next-generation hyperparameter optimization framework.* Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

Benton, D., & Young, H. A. (2017). *Reducing calorie intake may not help you lose body weight.* Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 12(5), 703–714. https://doi.org/10.1177/1745691617690878

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). *Algorithms for hyper-parameter optimization.* Advances in Neural Information Processing Systems, *24*, 2546–2554. https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html

Booth, F. W., Roberts, C. K., & Laye, M. J. (2012). *Lack of exercise is a major cause of chronic diseases.* Comprehensive Physiology, 2(2), 1143-1211. https://doi.org/10.1002/cphy.c110025

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Clarinda, K., Deepa, N., Dhamotharan, R., Krishnan, R. S., Raj, J. R. F., & Anchitaalagammai, J. V. (2024). *Leveraging machine learning algorithms to optimize diet maintenance based on user food intake and fitness data.* 2024 International Conference on Sustainable Communication Networks and Application (ICSCNA), 1800–1807. https://doi.org/10.1109/ICSCNA63714.2024.10864099

Garber, C. E., Blissmer, B., Deschenes, M. R., Franklin, B. A., Lamonte, M. J., Lee, I.-M., Nieman, D. C., & Swain, D. P. (2011). *Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: Guidance for prescribing exercise.* Medicine & Science in Sports & Exercise, 43(7), 1334–1359. https://doi.org/10.1249/MSS.0b013e318213fefb

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., & Delp, S. L. (2018). *Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities.* Journal of Biomechanics, 81, 1-11. https://doi.org/10.1016/j.jbiomech.2018.09.009

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd ed.)*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hutter, F., Hoos, H., & Leyton-Brown, K. (2014). *An efficient approach for assessing hyperparameter importance.* Proceedings of the 31st International Conference on Machine Learning, *32*(1), 754–762. https://proceedings.mlr.press/v32/hutter14.html

Keytel, L. R., Goedecke, J. H., Noakes, T. D., Hiiloskorpi, H., Laukkanen, R., van der Merwe, L., & Lambert, E. V. (2005). *Prediction of energy expenditure from heart rate monitoring during submaximal exercise.* Journal of Sports Sciences, 23(3), 289–297. https://doi.org/10.1080/02640410470001730089

Khorasani, V. (2023). *Gym members exercise tracking* [Dataset]. Kaggle. https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset

Lavanya, T. V., & Sivaraman, K. (2024). *A Machine learning approach for predicting physical activity intensity from wearable sensor data.* 2024 7th International Conference on

Circuit Power and Computing Technologies (ICCPCT), 1769–1774. https://doi.org/10.1109/ICCPCT61902.2024.10673260

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* Advances in Neural Information Processing Systems, 30, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Miah, J., Mamun, M., Rahman, M. M., Mahmud, M. I., Ahmad, S., & Nasir, M. H. B. (2022). *MHfit: Mobile health data for predicting athletics fitness using machine learning models.* 2022 2nd International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), 584–589. https://doi.org/10.1109/ISMODE56940.2022.10180967

Priscilla, M., Suriya, A., Srikanth, J., Jagadhishwaran, S., Kumar, M. N., & Yasvanth, D. (2024). *Evolution of artificial intelligence based burned calories prediction system using novel hybrid learning methodology.* 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 1–7. https://doi.org/10.1109/ICONSTEM60960.2024.10568726

Salanke, V. S., & Sathyajeeth, R. (2024). *Advanced calories burn forecasting with XGBoost: A comprehensive predictive analysis.* 2024 Global Conference on Communications and Information Technologies (GCCIT), 1–6. https://doi.org/10.1109/GCCIT63234.2024.10862479

Santhiya, S., Senthamarai, M., Jayadharshini, P., Juber, B., Akshay, J., & Aneesh, S. (2024). *Machine learning-based calorie burn estimation for enhanced physical activity monitoring.* 2024 2nd International Conference on Advances in Computation,

Communication and Information Technology (ICAICCIT), 324–330. https://doi.org/10.1109/ICAICCIT64383.2024.10912335

World Health Organization. (2000). *Obesity: Preventing and managing the global epidemic: Report of a WHO consultation.* World Health Organization. https://iris.who.int/handle/10665/42330

Yi, J. (2024). *Athlete physical fitness assessment and training optimization assisted by artificial intelligence.* 2024 International Conference on Information Technology, Communication Ecosystem and Management (ITCEM), 191–196. https://doi.org/10.1109/ITCEM65710.2024.00043