IEEE ComSoc IEEE Communications Society | IEEE COMPUTER SOCIETY | IEEE Signal Processing Society | VTS | IEEE Transactions on Machine Learning in Communications and Networking

Regresión ordinal aplicada a la estimación de niveles de obesidad con datos comportamentales y físicos

Juan José García Álvarez[1], Valentina Buelvas Martínez[1], y Tomás Cadavid Martínez[1]

[1]Universidad de Antioquia, Facultad de Ingeniería, Ingeniería de Sistemas, Medellín, Colombia

Autor de correspondencia: Tomás Cadavid Martínez (e-mail: tomas.cadavid@udea.edu.co)

Este trabajo fue desarrollado con apoyo del curso Modelos II del programa de Ingeniería de Sistemas de la Universidad de Antioquia.

ABSTRACT Obesity is a growing global public health challenge associated with unhealthy eating behaviors and physical inactivity. This study aims to estimate obesity levels using demographic, lifestyle, and physical activity data through supervised machine learning techniques. The dataset employed is the "Estimation of Obesity Levels Based on Eating Habits and Physical Condition" from the UCI Machine Learning Repository, consisting of 2,111 records and 17 features. An exploratory data analysis is conducted to understand the distribution of the variables, identify relevant patterns, and prepare the data for modeling. Several classification algorithms are evaluated to distinguish between multiple obesity categories, ranging from Insufficient Weight to Obesity Type III. The objective is to develop an accurate and interpretable ordinal regression model capable of supporting early detection and prevention of obesity. The results of this research may contribute to the application of intelligent systems in public health and in the design of personalized interventions for healthier lifestyles.

INDEX TERMS Eating habits, Lifestyle data, Machine learning, Obesity ordinal regression, Obesity prediction, Public health, Risk assessment.

I. INTRODUCCIÓN

La obesidad es uno de los principales problemas de salud pública a nivel mundial y está fuertemente asociada a enfermedades crónicas como la diabetes tipo 2, la hipertensión y las afecciones cardiovasculares. Su prevalencia continúa en aumento debido a estilos de vida poco saludables, caracterizados por hábitos alimenticios inadecuados y bajos niveles de actividad física. En este contexto, el uso de técnicas de Machine Learning (ML) se ha consolidado como una herramienta eficaz para analizar grandes volúmenes de datos y apoyar la estimación de los niveles de obesidad a partir de características comportamentales y condiciones físicas de los individuos.

La aplicación de modelos de clasificación basados en ML permite automatizar el análisis y reconocer patrones relevantes que contribuyen a la identificación temprana de riesgos, facilitando el diseño de estrategias personalizadas de prevención y atención en salud. Además, estas técnicas pueden fortalecer la toma de decisiones en el ámbito de la salud pública mediante la generación de conocimiento a partir de datos.

Este estudio emplea el conjunto de datos "Estimation of Obesity Levels Based on Eating Habits and Physical Condition" del UCI Machine Learning Repository, compuesto por 2.111 registros y múltiples atributos relacionados con hábitos de vida. El objetivo principal es construir un modelo de regresión ordinal capaz de categorizar a los sujetos en niveles que van desde Insufficient Weight hasta Obesity Type III. El trabajo presentado se centra en la etapa de análisis exploratorio de datos y en la definición de la aproximación metodológica para el desarrollo y evaluación del modelo.

El resto de este artículo se organiza de la siguiente manera: la Sección II describe el conjunto de datos y el proceso de preprocesamiento; la Sección III detalla la estrategia de modelado y los algoritmos utilizados; la Sección IV presenta los resultados preliminares y su discusión; finalmente, la Sección V expone las conclusiones y posibles trabajos futuros.

II. DESCRIPCIÓN DEL PROBLEMA

A. CONTEXTO Y UTILIDAD DEL PROBLEMA

El problema abordado consiste en estimar el nivel de obesidad de una persona a partir de características relacionadas

con su alimentación, actividad física y hábitos cotidianos. La utilidad de una solución basada en Machine Learning radica en la capacidad que tiene para aprender patrones complejos entre múltiples variables, lo cual puede apoyar tanto la investigación médica como la implementación de sistemas inteligentes de apoyo al diagnóstico y la prevención de esta enfermedad.

B. DESCRIPCIÓN DEL CONJUNTO DE DATOS

El conjunto de datos "Estimation of Obesity Levels Based on Eating Habits and Physical Condition" proviene del UCI Machine Learning Repository. Contiene 2,111 registros y 17 variables, entre las cuales se incluyen datos demográficos (género, edad, altura, peso), hábitos alimenticios (consumo de calorías, frecuencia de comidas, consumo de agua y alcohol), y actividad física (frecuencia de ejercicio, uso de transporte activo). La variable objetivo corresponde al nivel de obesidad, categorizado en siete clases: Insufficient_Weight, Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Obesity_Type_II y Obesity_Type_III.

Las variables se encuentran codificadas en distintos formatos: numéricos (edad, altura, peso) y categóricos (hábitos o respuestas de "sí" o "no"). No se reportan valores faltantes en el dataset, por lo que no se requiere imputación. Sin embargo, las variables categóricas deben ser transformadas mediante codificación one-hot o label encoding para que puedan ser procesadas por los modelos de aprendizaje supervisado.

C. PARADIGMA DE APRENDIZAJE Y JUSTIFICACIÓN

Dado que la variable objetivo representa niveles de obesidad con categorías discretas y ordenadas, se adopta un enfoque de aprendizaje supervisado basado en regresión ordinal en lugar de una clasificación multiclase tradicional. Este paradigma resulta más adecuado, ya que incorpora explícitamente la relación de orden entre las categorías, evitando tratar los niveles como clases independientes y permitiendo que el modelo capture adecuadamente la progresión entre ellos. Los modelos candidatos considerados son K-Nearest Neighbors (K-NN), Random Forest, Gradient Boosting, Perceptrón Multicapa (MLP) y Support Vector Machines (SVM), adaptados a la naturaleza ordinal del problema.

III. ESTADO DEL ARTE

Los modelos de aprendizaje supervisado aplicados a la estimación de niveles de obesidad han sido ampliamente estudiados en los últimos años, utilizando información relacionada con hábitos alimentarios, actividad física y variables sociodemográficas. Entre las propuestas recientes se encuentra el trabajo de Yagin et al. [1], quienes emplean un conjunto de datos de 498 individuos de Colombia, Perú y México para clasificar siete niveles de obesidad mediante una red neuronal multicapa (MLP) con una sola capa oculta. El estudio evalúa tres técnicas de selección de características (chi-square, F-Classify y mutual information) e incorpora optimización bayesiana para el ajuste automático de hiperparámetros. La validación utiliza una partición 75/25 con un proceso adicional de validación interna, repetido diez veces para estimar la variabilidad del desempeño. Aunque el modelo con todas las características presenta la mayor exactitud, las clases minoritarias, particularmente "Obesity Type I" y "Obesity Type II", mantienen un rendimiento inferior debido al desbalance del conjunto de datos.

Otro enfoque relevante es el modelo híbrido por votación mayoritaria propuesto por Galli et al. [2], donde se comparan varios clasificadores tradicionales, incluyendo SVM, GaussianNB, k-NN, Árboles de Decisión, Random Forest, Gradient Boosting, XGBoost y MLP. Tras la evaluación, Gradient Boosting, XGBoost y MLP son seleccionados como clasificadores base para el ensamble. La predicción final corresponde a la clase mayoritaria entre los modelos, obteniéndose un accuracy de 97.16%, superior al de los clasificadores individuales. No obstante, el estudio se centra únicamente en exactitud global, lo cual limita la interpretación del desempeño por categoría e impide identificar efectos del desbalance entre clases.

De forma complementaria, Santisteban Quiroz [3] analiza un conjunto mayor de 2.111 muestras procedentes de los mismos países e incorpora múltiples algoritmos supervisados, entre ellos LightGBM, XGBoost, Random Forest, Extra Trees y Regresión Logística. La validación se realiza mediante partición 80/20 y un proceso de entrenamiento repetido cien veces para reducir la variabilidad. LightGBM obtiene el mayor rendimiento, con un AUC del 99.90% y un accuracy del 97.45%. A pesar de estos resultados, el autor señala que las clases menos representadas continúan generando inestabilidad en las predicciones, lo que evidencia la persistencia del problema de desbalance.

En la misma línea, Gozukara Bag et al. [4] emplean el conjunto de datos público correspondiente a hábitos alimentarios y actividad física de Colombia, Perú y México. Su propuesta integra SMOTE-NC para el tratamiento del desbalance, selección de características mediante RFE y optimización bayesiana. Tres modelos supervisados (Regresión Logística, Random Forest y XGBoost) son evaluados en este esquema. Los resultados muestran que la Regresión Logística alcanza precisiones cercanas al 99% tanto con todas las características como con el subconjunto óptimo, destacando que una estrategia adecuada de preprocesamiento puede superar a modelos de mayor complejidad sin incrementar los costos computacionales.

En conjunto, estos estudios evidencian avances significativos en la clasificación de niveles de obesidad mediante técnicas supervisadas. Sin embargo, también muestran limitaciones persistentes relacionadas principalmente con el desbalance de clases y la dependencia excesiva de métricas globales como la exactitud, las cuales dificultan una evaluación exhaustiva por categoría. En este trabajo se busca abordar estas limitaciones mediante un análisis centrado en

el rendimiento por clase y en la mejora del aprendizaje en categorías minoritarias.

IV. ENTRENAMIENTO Y EVALUACIÓN DE LOS MODELOS

A. CONFIGURACIÓN EXPERIMENTAL

1) Metodología de Validación y Preprocesamiento

Para garantizar la robustez de los resultados y mitigar el sesgo de selección, se empleó una estrategia de Validación Cruzada Estratificada con $k = 5$ particiones (Stratified K-Fold). Esta técnica asegura que la distribución de las clases de obesidad se mantenga constante en cada fold de entrenamiento y validación. El conjunto de datos se dividió previamente en un 70% para entrenamiento/validación y un 30% reservado estrictamente para la prueba final(Test).

2) Manejo del Desbalance de Clases

A diferencia de técnicas de oversampling sintético (como SMOTE), en este experimento se optó por un enfoque algorítmico mediante el ajuste de pesos. En los modelos LogisticRegression, RandomForest y SVM, se configuró el hiperparámetro class_weight='balanced', el cual penaliza los errores de clasificación en las clases minoritarias de manera inversamente proporcional a su frecuencia. Los datos de entrada fueron estandarizados utilizando StandardScaler dentro de un Pipeline para evitar fugas de información (data leakage) entre los conjuntos de entrenamiento y validación [5], [6].

3) Modelos e Hiperparámetros

Se evaluaron cinco familias de modelos para cubrir diferentes enfoques de aprendizaje (paramétrico, basado en instancias, ensambles, redes neuronales y vectores de soporte). La optimización de hiperparámetros se realizó mediante una búsqueda exhaustiva (GridSearchCV). El espacio de búsqueda se detalla en la TABLA 1.

4) Justificación de Métricas

Dado que la variable objetivo (Nivel de Obesidad) es categórica pero posee un orden intrínseco (Ordinal), la métrica principal seleccionada para la optimización fue el Error Absoluto Medio (MAE) sobre las etiquetas codificadas numéricamente.

- **MAE (Mean Absolute Error):** Se priorizó sobre la exactitud porque penaliza la distancia del error. Confundir "Peso Normal" con "Obesidad Tipo III" es un error mayor que confundirlo con "Sobrepeso Nivel I".
- **Accuracy y Spearman:** Métricas secundarias usadas para comparabilidad con la literatura y para medir correlación de rango.

TABLA 1. Espacio de Búsqueda de Hiperparámetros

Tipo de Modelo	Algoritmo	Hiperparámetros Explorados (Grid)
Regresión (Paramétrico)	Regresión Logística	$C \in \{0.01, 0.1, 1, 10\}$ Solver: `lbfgs` (Multinomial)
No Paramétrico	k-Nearest Neighbors (k-NN)	Vecinos (k) $\in \{3, 5, 7\}$ Ponderación: {uniform, distance}
Ensamble	Random Forest	Árboles $\in \{100, 200\}$ Max Depth $\in \{\text{None}, 10\}$ Min Samples Leaf $\in \{1, 2\}$
Boosting	Gradient Boosting	Árboles $\in \{100, 200\}$ Tasa de aprendizaje $\in \{0.05, 0.1\}$ Max Depth $\in \{3, 5\}$
Red Neuronal	Perceptrón Multicapa (MLP)	Capas Ocultas $\in \{(50,), (100,), (50, 50)\}$ Alpha (L2) $\in \{0.0001, 0.001\}$
Máquina de Soporte	SVM	$C \in \{0.1, 1, 10\}$ Kernel $\in \{\text{linear}, \text{rbf}\}$

B. RESULTADOS DEL ENTRENAMIENTO DE MODELOS

1) Análisis Comparativo de Desempeño

Los resultados cuantitativos obtenidos tras la validación cruzada y la evaluación en el conjunto de test se resumen en la TABLA 2. El modelo Gradient Boosting demostró el desempeño superior, alcanzando un MAE en el conjunto de prueba de 0.046 y una exactitud (Accuracy) del 97.8%. Esto indica que el modelo casi nunca equivoca la categoría, y cuando lo hace, es típicamente hacia una clase adyacente.

2) Análisis de Sobreajuste (Overfitting)

- **k-NN:** Mostró un claro sobreajuste. Obtuvo un Accuracy perfecto (1.0) y MAE (0.0) en entrenamiento, pero el desempeño cayó drásticamente en validación y test, indicando memorización del ruido en los datos.
- **Gradient Boosting y SVM:** Presentaron las curvas de aprendizaje más estables (Ver FIGURA 1), con una brecha pequeña entre las métricas de entrenamiento y validación, lo que sugiere una buena capacidad de generalización.

3) Intervalos de Confianza

Para asegurar la significancia estadística de los resultados, se calcularon intervalos de confianza (al 95%) utilizando los percentiles 2.5 y 97.5 de los resultados de los 5 folds de validación cruzada [7], [8].

- El MAE promedio de validación para Gradient Boosting fue de 0.045 con un intervalo de confianza de $(0.028, 0.063)$.

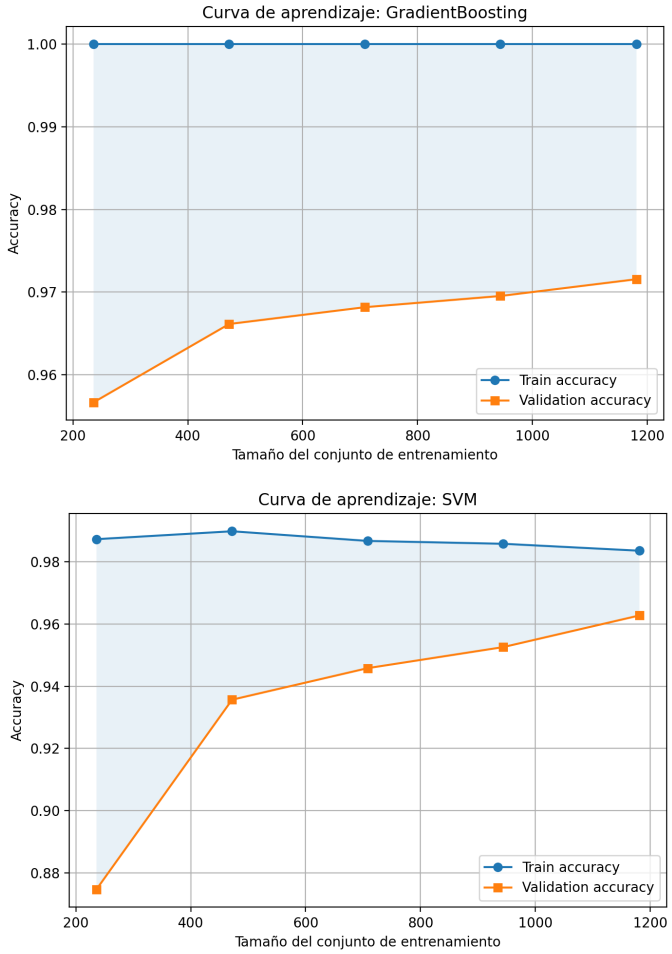

FIGURA 1. Curvas de aprendizaje para Gradient Boosting y SVM. La convergencia entre entrenamiento y validación indica buena generalización.

- Para SVM, el intervalo de confianza de Accuracy fue $(95.0\%, 98.1\%)$, demostrando una alta estabilidad ante variaciones en los datos de entrenamiento.

4) Importancia de Variables

Mediante el análisis de importancia por permutación (aplicado sobre el conjunto de Test), se determinó que el Índice de Masa Corporal (BMI) es la variable dominante para todos los modelos, con una importancia superior al 60% en la toma de decisiones del Gradient Boosting y SVM. Variables como el Género y la Edad ocupan un segundo nivel de relevancia, mientras que los hábitos de transporte (MTRANS) tuvieron un impacto marginal.

V. REDUCCIÓN DE DIMENSIÓN

En esta sección se analizaron individualmente las variables, se aplicaron métodos lineales y no lineales, y se evaluó el efecto de estas transformaciones sobre los dos mejores modelos encontrados en la Sección 4 [9], [10], [11].

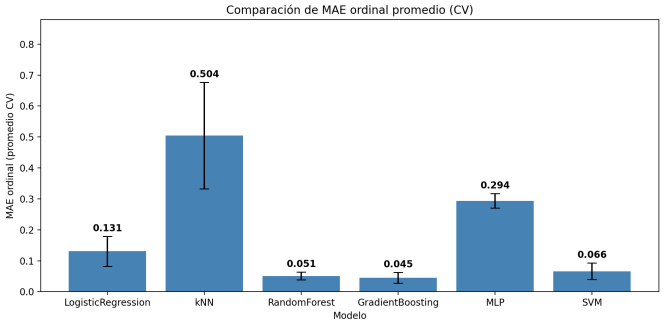

FIGURA 2. Error Absoluto Medio (MAE) en validación cruzada con intervalos de confianza del 95%.

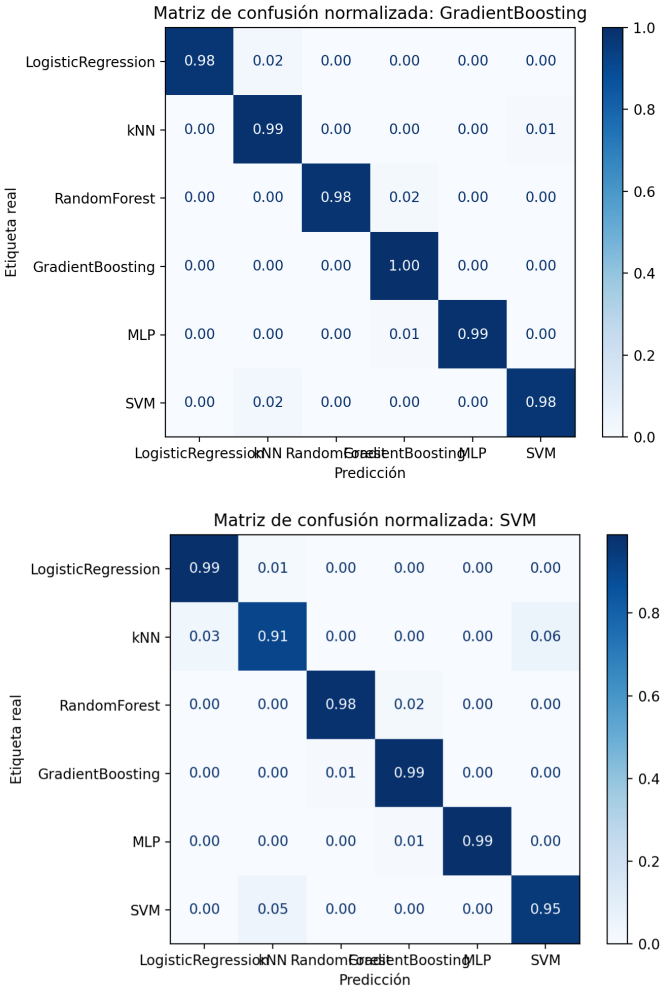

FIGURA 3. Matrices de confusión normalizadas para Gradient Boosting y SVM sobre el conjunto de prueba.

TABLA 2. Resultados Detallados del Mejor Modelo (Validación Cruzada 5-Fold y Métrica de Test)

Modelo	Mejores Parámetros	CV MAE (Mean)	CV MAE CI (95%)	CV Acc (Mean)	CV Acc CI (95%)	CV Spearman (Mean)	CV Spearman CI (95%)	Train Acc (MAE)	Test Acc (MAE)	Test Acc (Spearman)
LogReg	C: 10	0.13065	[0.084, 0.180]	0.93568	[0.919, 0.958]	0.94765	[0.921, 0.967]	0.05552	0.22240	0.92429
kNN	k: 5, 'distance'	0.50455	[0.370, 0.714]	0.81379	[0.769, 0.846]	0.75483	[0.631, 0.846]	0.00000	0.59937	0.78076
RF	Max D: None, Min SL: 1	0.05077	[0.041, 0.066]	**0.97833**	[**0.973**, **0.985**]	**0.97661**	[**0.965**, **0.984**]	0.00203	0.12618	**0.96530**
GB	LR: 0.1, Max D: 3	**0.04467**	[**0.028**, **0.062**]	0.97428	[0.966, 0.975]	0.98367	[0.975, 0.989]	0.00000	0.04574	0.97792
MLP	Alpha: 0.001, HL: (50,)	0.29382	[0.268, 0.314]	0.88153	[0.863, 0.899]	0.86353	[0.851, 0.875]	0.03792	0.42271	0.86278
SVM	C: 10, 'rbf'	0.06633	[0.030, 0.084]	0.96277	[0.950, 0.976]	0.97614	[0.967, 0.990]	0.03114	0.13565	0.94953

A. ANÁLISIS INDIVIDUAL DE VARIABLES

Para determinar la capacidad discriminativa de cada característica se utilizó el método de permutation importance aplicado a los dos mejores modelos identificados previamente: Gradient Boosting y SVM. Este análisis evalúa cuánto aumenta el error cuando una variable es permutada aleatoriamente, manteniendo el resto constantes [10], [12]. Las variables con incrementos mínimos en MAE o accuracy fueron marcadas como posibles candidatas a eliminación.

Los resultados muestran que características como MTRANS, CH2O y algunas subcategorías de actividad física tienen impacto marginal en las métricas del modelo, mientras que atributos como BMI, Age, Family History With Overweight, CALC y NCP contribuyen de manera significativa a la reducción del error [12]. No obstante, dado que el impacto negativo de eliminar variables de baja importancia fue leve pero presente, se decidió mantener el conjunto completo de características para los experimentos de reducción de dimensión mediante métodos de extracción de componentes [10].

1) Conclusión Análisis Individual de Variables

Existen variables con capacidad discriminativa limitada; sin embargo, su eliminación no mejora las métricas y puede introducir pérdida de información, por lo que la reducción se aborda mediante técnicas de extracción de características en lugar de selección directa [10], [13].

B. EXTRACCIÓN DE CARACTERÍSTICAS LINEAL (PCA)

Se aplicó Análisis de Componentes Principales (PCA) para obtener una representación linealmente reducida del conjunto original de 25 variables. El criterio adoptado fue mantener un $\geq$ 95% de la varianza explicada acumulada, lo cual produjo 13 componentes principales, equivalente a una reducción del 48% en dimensionalidad.

Las visualizaciones de los dos primeros componentes evidencian una fuerte superposición entre las clases ordinales, lo cual sugiere que el problema no presenta estructura lineal fácilmente separable en bajo número de dimensiones. A pesar de ello, se evaluó el impacto del embedding PCA en los dos mejores modelos de la Sección 4.

Los resultados indican que PCA conserva prácticamente el mismo MAE y accuracy que los modelos entrenados con los datos originales, con una reducción significativa en dimensionalidad y sin degradación sustancial del desempeño.

1) Conclusión Extracción de Características Lineal

PCA reduce la complejidad del espacio de entrada casi a la mitad manteniendo el rendimiento, por lo que constituye una alternativa viable si se busca eficiencia computacional o interpretabilidad estructural.

C. EXTRACCIÓN DE CARACTERÍSTICAS NO LINEAL (UMAP)

Se aplicó UMAP para obtener un embedding no lineal del conjunto original. Se seleccionó n_components = 2, criterio justificado por:

- La necesidad de comparar con t-SNE y PCA, cuyas proyecciones se interpretan usualmente en 2D.
- La naturaleza visual de UMAP orientada a exploración en espacios reducidos.
- El interés en evaluar el impacto de una reducción extrema (92%) sobre el desempeño de los modelos finales.

Las proyecciones en 2D mostraron nubes altamente solapadas y ausencia de fronteras claras entre clases, lo cual es consistente con la complejidad ordinal y la distribución continua de las muestras. Cuando el embedding UMAP fue utilizado para entrenar nuevamente los modelos Gradient Boosting y SVM, las métricas se deterioraron de manera significativa, lo que indica pérdida de información relevante al comprimir a solo dos dimensiones.

1) Conclusión Extracción de Características No Lineal

Aunque UMAP permite capturar relaciones no lineales locales, la reducción a dos dimensiones destruye información discriminativa necesaria para el problema, por lo que no se recomienda utilizar esta representación en el modelo final.

D. DISCUSIÓN Y CONCLUSIONES DE LA REDUCCIÓN DE DIMENSIÓN

La evaluación completa del sistema permitió identificar qué modelos funcionan mejor, qué variables aportan más información y cómo afecta la reducción de dimensionalidad al rendimiento final. En general, los resultados muestran que el problema tiene categorías muy cercanas entre sí y un comportamiento claramente ordinal, lo que hace necesario usar métricas más específicas que la exactitud.

En el entrenamiento de modelos, SVM y Gradient Boosting fueron los que obtuvieron el mejor desempeño, logrando

los menores errores ordinales (MAE) y mayores correlaciones de Spearman. Además, mostraron estabilidad en validación cruzada estratificada, lo que indica buenos niveles de generalización. Esto contrasta con algunos trabajos del estado del arte donde los mejores resultados se obtienen con MLP o con ensambles más complejos. La diferencia principal es que nuestro estudio evalúa el problema con métricas orientadas al orden de las clases, mientras que muchos estudios previos se basan únicamente en precisión global.

Del mismo modo que lo reporta la literatura, los resultados confirman que existe un desbalance importante entre categorías. Las matrices de confusión muestran que las clases intermedias son las más difíciles de predecir, lo cual coincide con los hallazgos de otros autores. El uso de MAE ordinal ayudó a identificar mejor estas confusiones y a entender en qué transiciones entre niveles se presentan los mayores errores.

En el análisis de características, las variables BMI, Age, Family History With Overweight, NCP y CALC resultaron ser las más relevantes para la predicción, lo cual coincide con lo reportado en los estudios revisados, donde los factores físicos y de hábitos tienen mayor peso. A diferencia de trabajos que aplican eliminación de características, en nuestro caso quitar variables no mejoró los resultados, por lo que se optó por aplicar técnicas de extracción de características (PCA y UMAP).

En reducción de dimensionalidad, PCA logró mantener el desempeño de los modelos reduciendo casi la mitad de las variables, mientras que UMAP afectó negativamente las métricas al comprimir demasiado la información. Aunque ninguno de los métodos generó grupos de clases claramente separados, PCA ofreció una representación más estable y útil para el modelo.

En conjunto, los resultados muestran que el objetivo del proyecto se cumplió: se identificaron las variables más relevantes, se evaluó el impacto de la reducción de dimensión y se compararon distintos modelos bajo condiciones apropiadas para datos ordinales y desbalanceados. Se concluye que SVM y Gradient Boosting son modelos confiables para este tipo de clasificación, que PCA es una opción válida para simplificar el espacio de características, y que el problema sigue siendo complejo debido al solapamiento entre categorías y al desbalance, tal como lo indica el estado del arte.

TABLA 3. **Comparación de Modelos con Reducción de Dimensionalidad**

Reducción	GradientBoosting		SVM	
Dimensionalidad	Accuracy	MAE	Accuracy	MAE
Original	0.9716	0.0473	0.8785	0.3438
PCA	0.8391	0.4148	0.8691	0.3785
UMAP	0.6640	0.9685	0.5410	1.3328

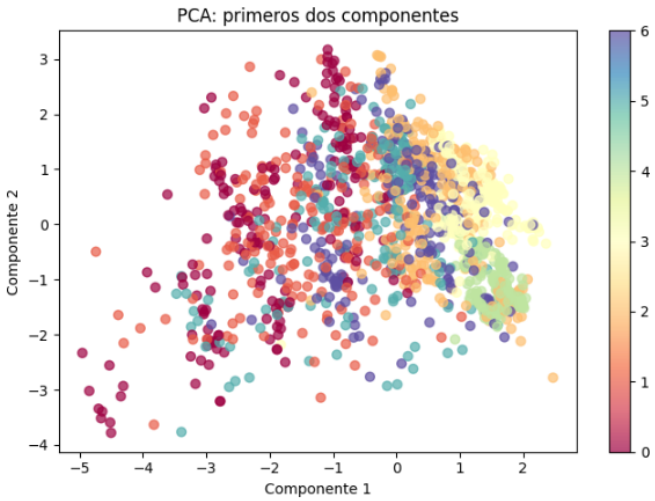

FIGURA 4. **Proyección PCA (Componentes 1 vs 2)**

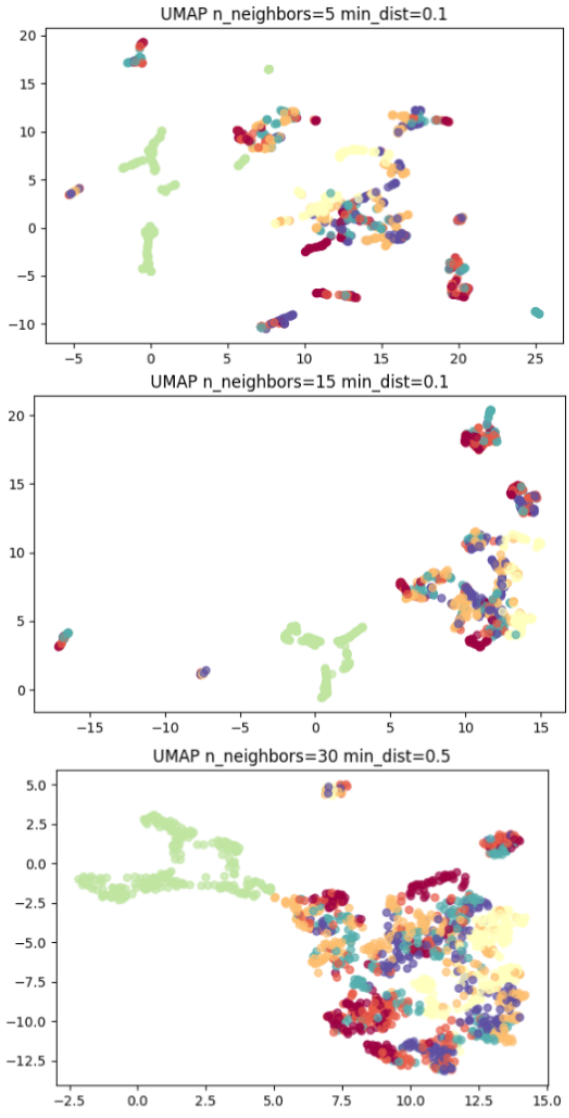

FIGURA 5. **Proyección UMAP (2D)**

REFERENCIAS

[1] F. H. Yagin et al., "Estimation of obesity levels with a trained neural network approach optimized by the Bayesian technique," *Appl. Sci.*, vol. 13, no. 6, p. 3875, Mar. 2023, doi: 10.3390/app13063875.

[2] D. D. Solomon et al., "Hybrid majority voting: Prediction and classification model for obesity," *Diagnostics*, vol. 13, no. 15, p. 2610, Aug. 2023, doi: 10.3390/diagnostics13152610.

[3] J. P. Santisteban Quiroz, "Estimation of obesity levels based on dietary habits and physical condition using computational intelligence," *Inform. Med. Unlocked*, vol. 29, p. 100901, Mar. 2022, doi: 10.1016/j.imu.2022.100901.

[4] H. G. Gozukara Bag et al., "Estimation of obesity levels through the proposed predictive approach based on physical activity and nutritional habits," *Diagnostics*, vol. 13, no. 18, p. 2949, Sep. 2023, doi: 10.3390/diagnostics13182949.

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[7] G. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. O'Reilly Media, 2022.

[8] W. J. Conover, *Practical Nonparametric Statistics*, 3rd ed. John Wiley & Sons, 1999.

[9] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[10] B. Ghojogh, M. Crowley, F. Karray, y A. Ghodsi, *Elements of Dimensionality Reduction and Manifold Learning*. Cham, Switzerland: Springer Nature Switzerland AG, 2023.

[11] Y. Ma y Y. Fu, Eds., *Manifold Learning Theory and Applications*. Boca Raton, FL, USA: CRC Press, 2012.

[12] D. Ryckelynck, F. Casenave, N. Akkari, *Manifold Learning: Model Reduction in Engineering*, SpringerBriefs in Computer Science. Cham, Switzerland, 2024.

[13] B. K. Tripathy, A. Sundareswaran, S. Ghela, *Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization*. CRC Press, 2022.