CLASIFICACION MULTICLASE DE OBESIDAD UTILIZANDO DATOS DE ESTILO DE VIDA Y ACTIVIDAD FÍSICA

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 multiclass classification 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, Multiclass obesity classification, 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 clasificación multiclase 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. CONTEXTO DEL PROBLEMA Y CONJUNTO DE DATOS

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 es categórica y discreta, se selecciona un enfoque de aprendizaje supervisado de tipo clasificación multiclase. Algunos modelos candidatos que pueden ayudar a resolver el problema son algoritmos como Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN) y Neural Networks.

III. ESTADO DEL ARTE

Uno de los enfoques más comunes en la clasificación de obesidad es el uso de modelos supervisados con información sobre hábitos y condiciones físicas. Entre las propuestas recientes se encuentra la de Yagin et al. [1]. En su estudio trabajan con datos de 498 personas de Colombia, Perú y México, buscando distinguir siete clases diferentes de obesidad a partir de una red neuronal MLP de una sola capa oculta. Antes de entrenar el modelo se probaron tres técnicas de selección de características (chi-square, F-Classify y mutual information), para evaluar si reducir el número de variables podía mejorar el desempeño. Además, aplican optimización bayesiana para ajustar los hiperparámetros sin tener que hacerlo manualmente. La validación se realizó con una partición 75 % entrenamiento y 25 % prueba, y dentro del conjunto de entrenamiento se hizo una validación interna. Este proceso se repitió varias veces —diez en total— con el fin de observar cómo variaban los resultados y asegurar que el modelo no estuviera dependiendo del azar. Aun así, algunas clases con menos ejemplos continuaron mostrando dificultades de predicción, lo que confirma que el desbalance en la distribución de datos sigue siendo un reto relevante.

Un planteamiento distinto es el presentado en Diagnostics [2], donde se propone un modelo híbrido basado en votación mayoritaria para la clasificación de niveles de obesidad. En ese trabajo se comparan varios algoritmos clásicos —entre ellos SVM, GaussianNB, k-NN, Regresión Logística, Árboles de Decisión, Random Forest, Gradient Boosting, XGBoost y MLP— y, tras la evaluación, se seleccionan Gradient Boosting, XGBoost y MLP como clasificadores base para el ensamble. La agregación es de tipo un-stage: la clase final corresponde a la más votada por los modelos base (majority voting). La evaluación se realizó con una partición 80/20 (entrenamiento/prueba) y los autores reportan accuracies individuales y del ensamble —por ejemplo, XGBoost 96.06 %, MLP 93.38 % y el ensamble 97.16 %—. Aunque los resultados numéricos son llamativos, la evaluación se apoya principalmente en Accuracy, lo que impide conocer con detalle el comportamiento por clase y dificulta la apreciación del rendimiento en categorías minoritarias.

En general, estos estudios muestran progresos importantes, pero también dejan ver limitaciones que aún afectan su uso real: el desbalance del dataset sigue perjudicando la clasificación de las categorías menos representadas, y basar la evaluación en una sola métrica impide entender qué tan bien responde el modelo en cada nivel de obesidad. Por eso, en este trabajo se busca probar un modelo que tenga en cuenta estos dos puntos: mejorar el aprendizaje en clases minoritarias y analizar el desempeño por categoría, con el fin de obtener resultados más útiles en un entorno clínico.

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, 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, 2610, Aug. 2023, doi: 10.3390/diagnostics13152610.