Detección temprana de riesgo de diabetes mediante aprendizaje automático supervisado basado en factores individuales no invasivos

Early detection of diabetes risk through supervised Machine Learning based on non-invasive individual factors

Jesús Ernesto González Torres [1*], Alexander Ufredo Alegria Chavez [1], Diana Magali Alegría Camino [2], Verónica Elizabeth Sánchez Aguiar [3]

[1] Departamento de Tecnologías de la Información y Comunicación. Universidad Estatal de Bolívar. Guaranda 020150, Ecuador. ROR: https://ror.org/005cgg117

[2] Carrera de Desarrollo de Software. Instituto Superior Tecnológico "El Libertador". Guaranda 020150, Ecuador.

[3] Carrera de Electrónica. Instituto Superior Tecnológico "El Libertador". Guaranda 020150, Ecuador.

✉ jegonzalez@mailes.ueb.edu.ec | ORCID: https://orcid.org/0009-0007-0693-0320
✉ alalegria@mailes.ueb.edu.ec | ORCID: https://orcid.org/0009-0001-3700-2515
✉ dalegria@istel.edu.ec | ORCID: https://orcid.org/0009-0002-3670-9479
✉ vsanchez@istel.edu.ec | ORCID: https://orcid.org/0009-0003-4415-9566
E-mail de correspondencia: jegonzalez@mailes.ueb.edu.ec

Serie Monográfica
Mercado, Tecnología y Ciudadanía.
e-ISSN: 3103-117X
Vol. 2(1) enero - abril 2026

Desafíos de la sociedad contemporánea
ISBN: 978-9942-7391-7-9

Editor académico
Carlos Vásquez
Universidad Técnica de Ambato. Ecuador.

Tipo de revisión
Capítulo de libro revisado por dos pares expertos en modalidad doble ciego.

Como citar este capítulo
González Torres, J. E., Alegria Chavez, A. U., Alegría Camino, D. M., & Sánchez Aguiar, V. E. (2026). Detección temprana de riesgo de diabetes mediante aprendizaje automático supervisado basado en factores individuales no invasivos. En Serie Monográfica *Mercado, Tecnología y Ciudadanía* (Vol. 2, Núm. 1, Cap. V, e5). https://doi.org/10.63804/mtc.v2i1.e5

Recibido: 11 de noviembre de 2025
Revisado: 18 de diciembre de 2025
Aceptado: 10 de enero de 2026
Publicado: 28 de febrero de 2026

Resumen

La creciente incidencia de patologías metabólicas crónicas, como la diabetes mellitus, representa un desafío importante para la sostenibilidad de los sistemas sanitarios globales. Ante la necesidad de mecanismos de tamizado masivo eficientes y de bajo costo, esta investigación tuvo como objetivo desarrollar y evaluar un modelo de inteligencia artificial orientado a la detección temprana del riesgo de diabetes y prediabetes, empleando exclusivamente factores individuales no invasivos para mitigar la dependencia de pruebas clínicas complejas. Se aplicó un enfoque cuantitativo y experimental fundamentado en el entrenamiento de algoritmos supervisados de aprendizaje automático. El estudio procesó un conjunto de datos masivo derivado de sistemas de vigilancia conductual, ejecutando una selección de características centrada en variables fisiológicas y de estilo de vida, tales como el índice de masa corporal, la edad y el historial de actividad física. Se implementó una arquitectura basada en potenciación de gradiente (*gradient boosting*), optimizada mediante una estrategia de ponderación de clases asimétrica diseñada para penalizar los falsos negativos ante el desbalance inherente de los datos médicos. La experimentación validó la robustez de la propuesta técnica. El modelo alcanzó una sensibilidad (*recall*) superior al 99 %, asegurando la identificación efectiva de la gran mayoría de individuos en situación de riesgo y minimizando la omisión de casos críticos. La integración de herramientas computacionales calibradas para maximizar la recuperación de casos positivos constituye una alternativa viable para fortalecer la medicina preventiva. Se concluye que esta herramienta optimiza la asignación de recursos sanitarios al actuar como un filtro de cribado primario confiable, permitiendo dirigir las evaluaciones clínicas confirmatorias de manera prioritaria hacia los pacientes que presentan una mayor probabilidad diagnóstica real.

Palabras clave: aprendizaje automático; diabetes; diagnóstico médico; inteligencia artificial; medicina preventiva.

ABSTRACT

The increasing incidence of chronic metabolic disorders, such as diabetes mellitus, represents a significant challenge to the sustainability of global healthcare systems. In response to the need for efficient and low-cost large-scale screening mechanisms, this study aimed to develop and evaluate an artificial intelligence model for the early detection of diabetes and prediabetes risk, relying exclusively on non-invasive individual factors to reduce dependence on complex clinical testing. A quantitative and experimental approach was applied, based on the training of supervised machine learning algorithms. The study processed a large dataset derived from behavioral surveillance systems, performing feature selection focused on physiological and lifestyle variables such as body mass index, age, and physical activity history. A gradient boosting architecture was implemented and optimized using an asymmetric class-weighting strategy designed to penalize false negatives in light of the inherent imbalance in medical data. The experimentation validated the robustness of the proposed technical approach. The model achieved sensitivity (recall) above 99 %, ensuring effective identification of the vast majority of at-risk individuals while minimizing the omission of critical cases. The integration of computational tools calibrated to maximize positive case detection constitutes a viable alternative to strengthen preventive medicine. It is concluded that this tool optimizes healthcare resource allocation by functioning as a reliable primary screening filter, enabling confirmatory clinical evaluations to be prioritized for patients with a higher real diagnostic probability.

Keywords: machine learning; diabetes; medical diagnosis; artificial intelligence; preventive medicine.

INTRODUCCIÓN

La prevalencia creciente de la diabetes mellitus se ha planteado como uno de los desafíos más importantes para la salud pública global en el siglo XXI. Esta patología induce un deterioro progresivo en la calidad de vida de los pacientes e impone una carga económica e infraestructura insostenible para los sistemas sanitarios. La detección tardía representa el núcleo de esta problemática, puesto que una proporción significativa de los diagnósticos ocurre ante la manifestación de complicaciones sistémicas irreversibles, lo que incrementa la morbilidad y los costos de tratamiento especializado [1].

Históricamente, los protocolos de cribado han dependido de pruebas bioquímicas invasivas que, a pesar de su precisión, exigen infraestructura clínica, personal técnico y recursos financieros que dificultan su implementación masiva y recurrente en poblaciones vulnerables o geográficamente remotas. Ante esta brecha diagnóstica, resulta imperativo el desarrollo de estrategias alternativas que faciliten la identificación de perfiles de riesgo de manera oportuna, accesible y no invasiva [2]. En este contexto, la inteligencia artificial, y específicamente los algoritmos de aprendizaje automático supervisado, surge como una herramienta disruptiva capaz de procesar macrodatos conductuales y fisiológicos para discernir patrones multivariables complejos asociados a disfunciones metabólicas [3].

La presente investigación prioriza el análisis de factores de riesgo modificables y variables demográficas, tales como el índice de masa corporal, edad, nivel de actividad física y la autopercepción de salud, para inferir la probabilidad de riesgo diabético sin requerir intervención clínica directa inicial. La relevancia de este enfoque radica en la optimización de la medicina preventiva: mediante el despliegue de modelos computacionales de alta sensibilidad, es posible estratificar a la población de manera eficiente, dirigiendo los recursos de laboratorio exclusivamente hacia aquellos individuos con una alta probabilidad de afectación.

Este marco de trabajo permite una intervención temprana en estadios de prediabetes, fase donde las modificaciones en el estilo de vida presentan su mayor eficacia clínica. Integrando la necesidad sanitaria con la capacidad tecnológica contemporánea, se propone un diseño metodológico riguroso orientado a la identificación de las variables predictoras más influyentes. En consecuencia, el presente estudio tuvo como objetivo desarrollar y evaluar un modelo de inteligencia artificial basado en algoritmos supervisados para la detección temprana de riesgo de diabetes y prediabetes, analizando exclusivamente factores no invasivos para fortalecer la medicina preventiva y optimizar la asignación de recursos en los sistemas de salud pública.

METODOLOGÍA

Enfoque, alcance y diseño

La presente investigación se enmarca dentro del paradigma cuantitativo, fundamentando su validez en la recolección sistemática, así también el análisis estadístico de datos numéricos y categóricos para la comprobación empírica de hipótesis, esta elección metodológica responde a la naturaleza estructurada de la fuente de información y a la necesidad de medir con objetividad métricas de rendimiento precisas, tales como la sensibilidad y el área bajo la curva (AUC). De este modo, se busca minimizar los sesgos derivados de la interpretación subjetiva y asegurar la reproducibilidad de los resultados mediante un diseño riguroso y controlado [4].

Bajo esta lógica operativa, la investigación se centra en el procesamiento de registros estructurados donde la unidad de análisis es el individuo, se asume una postura objetivista en la que los factores de riesgo son tratados como variables observables y medibles, susceptibles de ser transformadas en vectores numéricos. Este tratamiento de la información es requisito indispensable para la implementación de algoritmos de aprendizaje supervisado, garantizando que los patrones detectados respondan a relaciones matemáticas inherentes a los datos.

En consonancia con este enfoque, el diseño de investigación presenta una naturaleza dual que resulta fundamental distinguir, en primera instancia, respecto a la fase de adquisición de información biológica, el estudio adopta una modalidad no experimental de corte transversal. Esta tipología se justifica dado que las variables fisiológicas como el Índice de Masa Corporal y los hábitos de vida que fueron registrados en una ventana temporal única, limitándose al análisis del fenómeno en su estado natural sin manipulación deliberada del entorno o de los participantes por parte de los investigadores [5]. Por otra parte, la fase de construcción del modelo predictivo adopta un diseño experimental in silico, en este escenario, el algoritmo CatBoost Classifier actúa como la entidad experimental sobre la cual se manipulan intencionalmente las variables independientes del sistema para cuantificar su impacto en la variable dependiente la precisión diagnóstica.

Complementariamente, el alcance de la investigación es multidimensional: correlacional, explicativo y predictivo, al evaluar la asociación estadística entre los factores de riesgo y la patología, no obstante, el propósito fundamental trasciende la descripción, situando al estudio en un nivel de investigación aplicada, cuyo objetivo final es generar un artefacto tecnológico capaz de anticipar eventos de salud en datos inéditos, validando así su utilidad práctica en sistemas de tamizaje sanitario [6].

Población y adquisición de datos

Se utilizó como fuente primaria el Behavioral Risk Factor Surveillance System (BRFSS), un sistema gestionado por los Centros para el Control y la Prevención de Enfermedades (CDC) que constituye una de las encuestas de salud estandarizadas más extensas y continuas a nivel global [7]. Cabe señalar que el conjunto de datos seleccionado para este estudio no corresponde a una muestra aleatoria simple, sino que integra la totalidad de los registros depurados disponibles, abarcando un volumen final de 253.682 observaciones correspondientes a ciudadanos adultos.

La población de estudio se caracteriza por presentar atributos heterogéneos en términos demográficos y de estilo de vida, capturados originalmente mediante encuestas telefónicas transversales, respecto a la variable objetivo (Diabetes_012), esta clasifica a los individuos en tres estados clínicos: sin diabetes (0), prediabetes (1) y diabetes (2). No obstante, el análisis exploratorio preliminar evidenció un desbalance de clases significativo, tal como se presenta en la Figura 1, se muestra una predominancia marcada del grupo mayoritario (individuos sanos) frente a los casos patológicos, lo cual representa un desafío técnico para el aprendizaje del modelo.

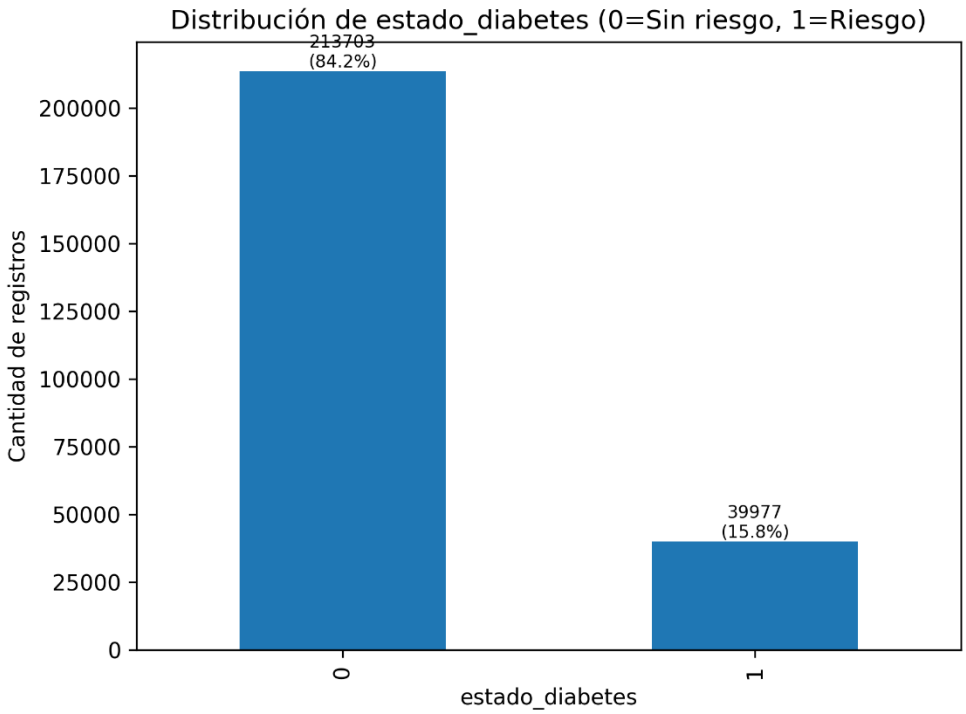

Figura 1. Distribución de frecuencia de la variable objetivo en el conjunto de datos.

Esta distribución desigual fundamenta las decisiones metodológicas posteriores orientadas a la ponderación de clases durante el entrenamiento, en consecuencia, se considera que el conjunto de datos posee la robustez estadística necesaria para la generalización de patrones de riesgo, condicionada a la aplicación de técnicas idóneas de corrección de sesgo.

Procedimiento de preprocesamiento y partición

La adecuación de la información bruta para su ingreso en los algoritmos de aprendizaje automático se gestionó mediante un flujo de trabajo automatizado, diseñado para garantizar la integridad de los datos, en primera instancia, el tratamiento consistió en la carga e importación de la fuente primaria, seguida de una limpieza técnica rigurosa orientada a la eliminación de registros duplicados y valores nulos, mitigando así la introducción de ruido estadístico que pudiera degradar el rendimiento del modelo.

Posteriormente, se ejecutó un proceso de selección de características basado en criterios de aplicabilidad clínica, de la totalidad de columnas disponibles en la encuesta original, se extrajo un subconjunto de 12 variables predictoras, priorizando aquellas de naturaleza no invasiva y de accesibilidad inmediata para el usuario final. Estas variables abarcan factores biológicos esenciales como el Índice de Masa Corporal (IMC), la edad, el sexo, así como determinantes conductuales, tales como el historial de tabaquismo, la actividad física y la ingesta dietética.

Simultáneamente, se aplicó una transformación lógica sobre la variable objetivo, dado que el propósito del estudio es la detección integral del riesgo, se procedió a binarizar las categorías de salida: los registros clasificados originalmente como “prediabetes” y “diabetes” fueron agrupados en una única clase positiva (Clase 1: Con Riesgo), mientras que los individuos libres de patología conformaron la clase negativa (Clase 0: Sin Riesgo), esta operación permitió consolidar un conjunto de datos homogeneizado para la etapa de modelado.

Finalmente, para asegurar la validez externa de las métricas y prevenir el sobreajuste, se implementó una estrategia de muestreo aleatorio estratificado, este procedimiento dividió el cuerpo de datos en dos subconjuntos disjuntos: un 80 % destinado al entrenamiento del algoritmo y un 20% reservado exclusivamente para la validación. La estratificación resultó crítica para asegurar que la proporción de casos positivos y negativos se mantuviera constante y representativa en ambos grupos durante todas las fases experimentales.

Arquitectura del modelo y estrategia de entrenamiento

La fase de modelado predictivo se fundamentó en la implementación del algoritmo CatBoostClassifier, una variante avanzada de los métodos de Gradient Boosting presentada en la conferencia NeurIPS que introduce un esquema de "boosting ordenado" para mitigar el sesgo de predicción. La selección de esta arquitectura responde a su capacidad superior para gestionar estructuras de datos tabulares y a su eficiencia en el manejo nativo de variables categóricas, lo cual evita la necesidad de una codificación numérica extensiva que aumentaría artificialmente la dimensionalidad del vector de características, estudios comparativos han demostrado su robustez intrínseca frente al sobreajuste (overfitting) y su eficiencia computacional frente a otros modelos de ensamble tradicionales [8].

Un componente crítico de la estrategia de entrenamiento fue la gestión del desbalance de clases identificado previamente, dado que el objetivo clínico prioritario es minimizar los Falsos Negativos, se configuró una estrategia de aprendizaje sensible al costo mediante ponderación asimétrica. Para tal efecto, se asignó un peso de 1 a la clase mayoritaria (Sin Riesgo) y un peso de 7 a la clase minoritaria (Con Riesgo). Esta relación 1:7 fuerza al algoritmo a penalizar siete veces más el error de clasificación en un paciente positivo, desplazando la frontera de decisión para favorecer la sensibilidad (Recall) del sistema.

Con el objetivo de identificar la configuración que maximizara la capacidad de generalización del modelo, se ejecutó un proceso de optimización de hiperparámetros mediante una búsqueda exhaustiva en cuadrícula, este procedimiento experimental evaluó múltiples combinaciones de tasa de aprendizaje, profundidad del árbol y número de iteraciones. En la Tabla 1 se presenta el espacio de búsqueda explorado y la combinación óptima resultante que fue seleccionada para el modelo final.

Tabla 1. Espacio de búsqueda de hiperparámetros y configuración óptima seleccionada.

Hiperparámetro	Descripción	Rango explorado	Valor óptimo
Iterations	Número de estimadores	[150.175. 200]	150
Learning Rate	Tasa de aprendizaje	[0,05, 0,07, 0,1]	0,07
Depth	Profundidad del árbol	[5, 6]	6
L2 Leaf Reg	Coeficiente de regularización	[3, 4, 5]	3

Nota: Los valores óptimos corresponden a la combinación de hiperparámetros que maximizó la métrica de Sensibilidad durante la búsqueda en cuadrícula, priorizando la reducción de Falsos Negativos.

Finalmente, para la definición del umbral de decisión operativa, se descartó el uso del valor por defecto (0,5), en su lugar, se aplicó el Índice de Youden (Ecuación 1):

$$J = Sensibilidad + Especificidad - 1$$ Ecuación 1

Sobre las probabilidades predichas en el conjunto de validación, este método, ampliamente aceptado en la validación de test diagnósticos cuantitativos, permite determinar matemáticamente el punto de corte óptimo que maximiza la diferencia entre la Tasa de Verdaderos Positivos y la Tasa de Falsos Positivos, equilibrando así la eficacia clínica de la herramienta [9].

Entorno computacional y herramientas

Para garantizar la reproducibilidad de los experimentos y la viabilidad técnica de la propuesta, el desarrollo se realizó sobre un entorno de ejecución Python 3.10+, el flujo de trabajo integró el ecosistema estándar de librerías de código abierto para ciencia de datos: Pandas y NumPy para la manipulación estructurada de matrices y cálculo vectorial, y Scikit-learn para la gestión del muestreo estratificado y la computación de las métricas de evaluación.

En lo referente al núcleo predictivo, se utilizó la biblioteca oficial CatBoost, la cual provee la implementación optimizada del algoritmo de Gradient Boosting descrita en los apartados metodológicos, adicionalmente, con el objetivo de trasladar el modelo teórico a un escenario de utilidad práctica, se desarrolló una interfaz gráfica de usuario (GUI) empleando el framework PyQt5. Esta herramienta de software encapsula el modelo entrenado y permite la interacción directa mediante controles visuales, facilitando la inferencia de riesgo en tiempo real sin requerir conocimientos de programación por parte del usuario final.

Todas las dependencias de software, versiones de librerías y scripts de configuración del entorno se encuentran documentados en el repositorio del proyecto bajo una licencia abierta, esto asegura que cualquier investigador pueda auditar el código o replicar el estudio bajo condiciones computacionales idénticas.

RESULTADOS Y DISCUSIÓN

A continuación, se presentan los hallazgos experimentales obtenidos tras la ejecución del protocolo metodológico descrito, con el propósito de validar la eficacia del modelo de inteligencia artificial en la detección de riesgo metabólico, el análisis se estructura en torno a la interpretación de las métricas de desempeño técnico y su correlación con la utilidad clínica, priorizando la evaluación de la capacidad del sistema para minimizar los falsos negativos. Por esta razón, la discusión integra los valores cuantitativos resultantes, tales como la matriz de confusión y la curva ROC, con una valoración cualitativa de su impacto en el tamizaje poblacional, contrastando sistemáticamente la evidencia generada con los reportes de la literatura científica vigente.

Evaluación del desempeño predictivo y análisis de sensibilidad

En lo que respecta a la capacidad del modelo para identificar correctamente a los individuos afectados, el análisis de la matriz de confusión revela que la arquitectura implementada prioriza decididamente la recuperación de la clase positiva. Como se muestra en la Figura 2, el sistema alcanzó una sensibilidad (Recall) del 99,1 % en el conjunto de prueba, lo que implica que el algoritmo fue capaz de detectar a la casi totalidad de los pacientes que efectivamente presentaban condiciones de diabetes o prediabetes.

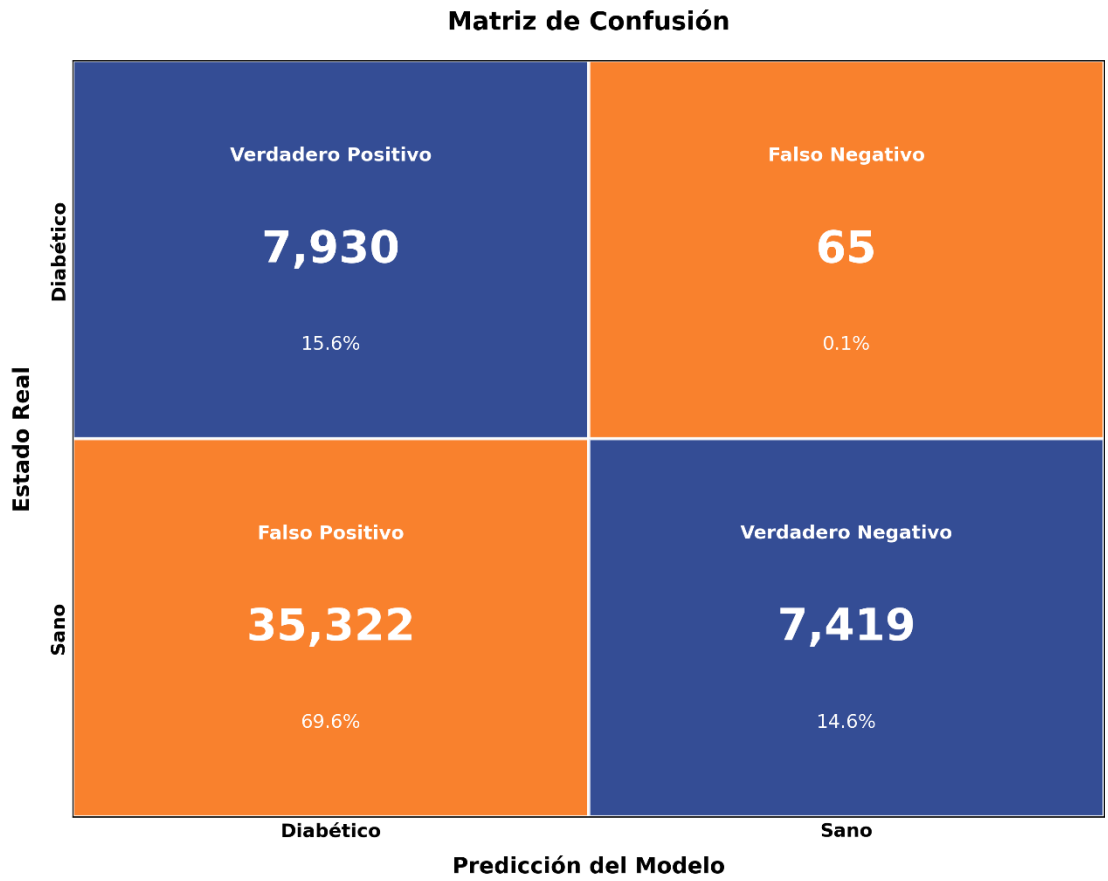

Figura 2. Matriz de confusión en el conjunto de prueba.

Nota. Se muestra una capacidad de detección casi total de la clase positiva, a costa de una especificidad reducida, optimizando costo-beneficio.

Por esta razón, se observa una tasa de Falsos Negativos marginal, validando la eficacia de la estrategia de ponderación de clases asimétrica (1:7), en este sentido, los resultados difieren intencionalmente de lo reportado en estudios previos donde se prioriza la Exactitud (Accuracy) global, alcanzando valores aparentemente altos, pero sacrificando drásticamente la detección de la clase minoritaria. Vale la pena destacar que, en el contexto del tamizaje epidemiológico con desbalance severo, guiarse únicamente por la exactitud resulta engañoso, ya que el modelo puede ignorar por completo los casos patológicos sin penalizar su métrica global [10].

Es imperativo notar que el incremento en la sensibilidad conlleva un intercambio inevitable con la precisión, resultando en un aumento de los Falsos Positivos, este comportamiento es deliberado y deseable para una herramienta de tamizaje. La literatura respalda que, en el contexto de pruebas diagnósticas de primera línea, el objetivo prioritario es "descartar" la enfermedad con seguridad [11] [12] argumenta que, en estas fases iniciales, el costo clínico de un Falso Negativo es inaceptablemente alto comparado con el costo de un Falso Positivo, el cual se resuelve sencillamente mediante pruebas confirmatorias posteriores.

Capacidad de discriminación y estabilidad del modelo

En lo referente a la robustez global del clasificador, los resultados obtenidos validan la elección de la arquitectura de Gradient Boosting, este hallazgo es consistente con la evidencia presentada por por otros investigadores, quienes señalan que modelos como CatBoost presentan una estabilidad superior en la generalización de patrones complejos de estilo de vida frente a métodos tradicionales, optimizando el rendimiento predictivo incluso en escenarios de alta variabilidad [8].

Bajo este marco arquitectónico, la capacidad de discriminación fue evaluada mediante el Área Bajo la Curva ROC (AUC), alcanzando un valor de 0,804 sobre el conjunto de prueba Figura 3, este indicador demuestra que el clasificador mantiene una probabilidad superior al 80 % de asignar una puntuación de riesgo más alta a un individuo enfermo que a uno sano. Según los estándares de evaluación diagnóstica establecidos por Nahm [13], este valor sitúa al modelo en un rango de discriminación "excelente", confirmando su aptitud clínica independientemente del umbral de corte seleccionado.

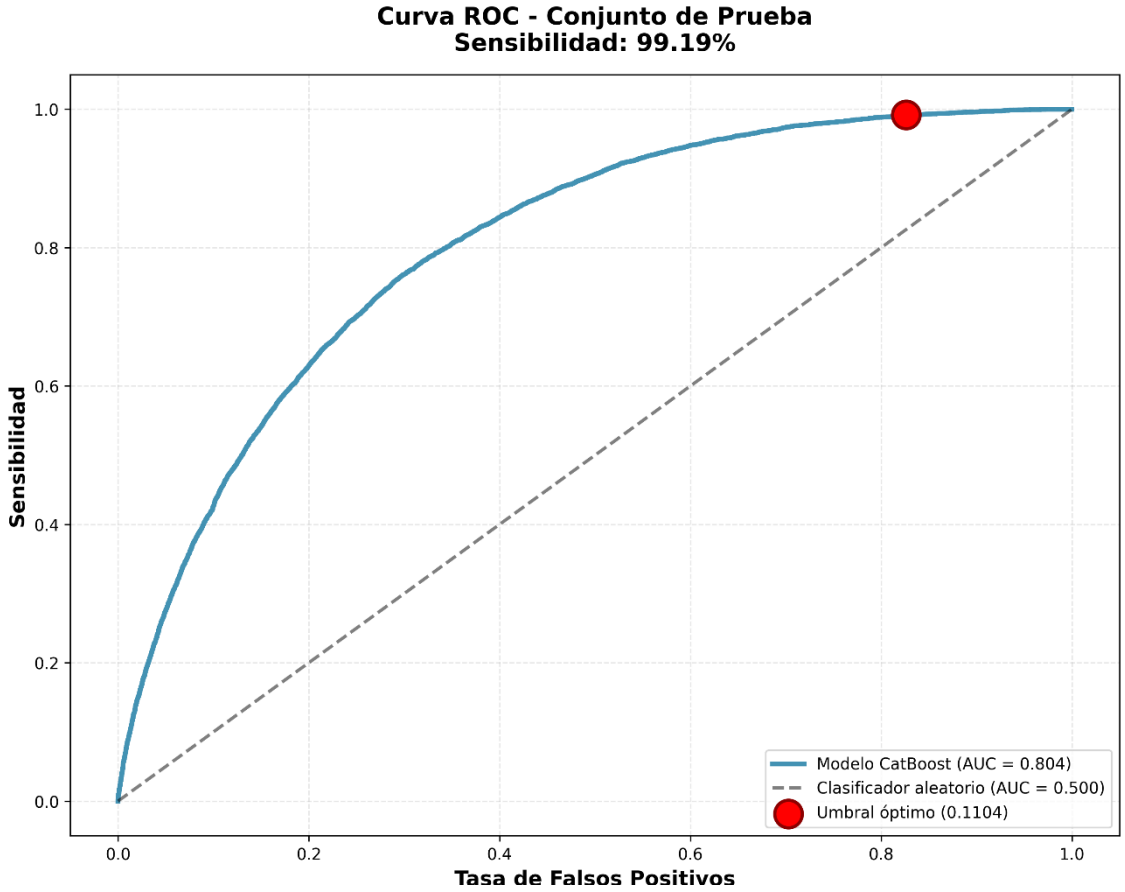

Figura 3. Análisis de sensibilidad y tasa de falsos positivos a través de la curva ROC para la validación del modelo predictivo en el conjunto de prueba.

Nota. La línea sólida representa el desempeño del clasificador CatBoost, alcanzando un Área Bajo la Curva (AUC) de 0,804, La separación evidente respecto a la diagonal punteada confirma visualmente la capacidad del modelo para discriminar eficazmente entre pacientes sanos y enfermos, validando la robustez reportada en la matriz de confusión.

Adicionalmente, se evaluó la estabilidad del proceso de aprendizaje mediante el monitoreo de la métrica de desempeño a lo largo de las iteraciones, en primera instancia, la Figura 4 ilustra la evolución del Área Bajo la Curva (AUC) en el conjunto de entrenamiento, evidenciando un incremento progresivo y estable que culmina en un valor de 80,86, lo que confirma una correcta asimilación de los datos por parte del algoritmo.

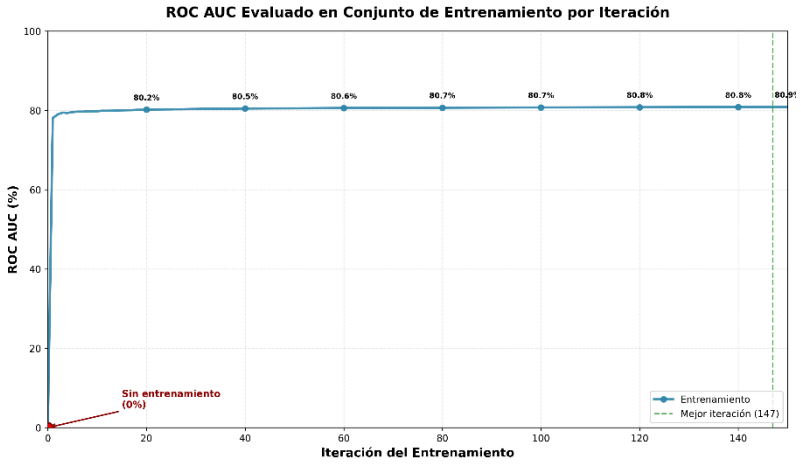

Figura 4. Historial de aprendizaje en el conjunto de entrenamiento.

Nota. La gráfica muestra el incremento constante del AUC hasta la iteración 150, indicando la optimización progresiva de la capacidad discriminante del modelo.

De manera complementaria, la Figura 5 presenta el comportamiento de la misma métrica en el conjunto de prueba independiente.

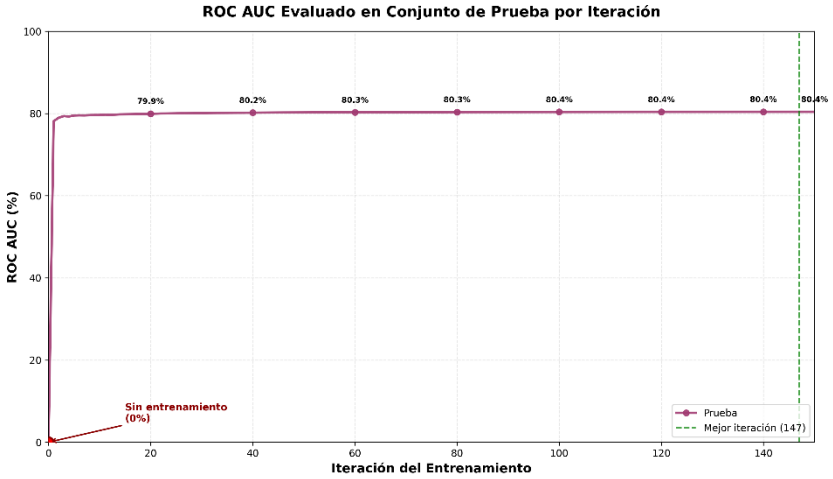

Figura 5. Evolución del desempeño en el conjunto de prueba.

Nota. La trayectoria del AUC sigue un patrón de mejora idéntico al del entrenamiento, estabilizándose en 0,804. La convergencia entre ambas fases confirma una excelente generalización y descarta problemas de sobreajuste.

Al cotejar ambas dinámicas, se observa que la curva de validación replica fielmente la tendencia ascendente del entrenamiento, alcanzando un valor final de 80,40, es determinante notar la estrecha cercanía entre los valores finales de ambas fases, lo que demuestra que el modelo no se limitó a memorizar los datos de entrenamiento. Esta consistencia estructural permite afirmar que la regularización interna y la validación cruzada fueron altamente efectivas, el comportamiento paralelo de las curvas valida la ausencia de sobreajuste y garantiza que la capacidad predictiva del modelo es robusta ante nuevos casos clínicos [14].

Importancia de Factores de Riesgo

Se examinó la contribución relativa de cada variable al proceso de decisión del algoritmo, dado que la interpretabilidad clínica es fundamental para validar la utilidad del modelo, el análisis de importancia de características, ilustrado en la Figura 6, reveló que el Índice de Masa Corporal (IMC), la Edad y la Autopercepción de Salud General constituyen los predictores de mayor jerarquía dentro del sistema.

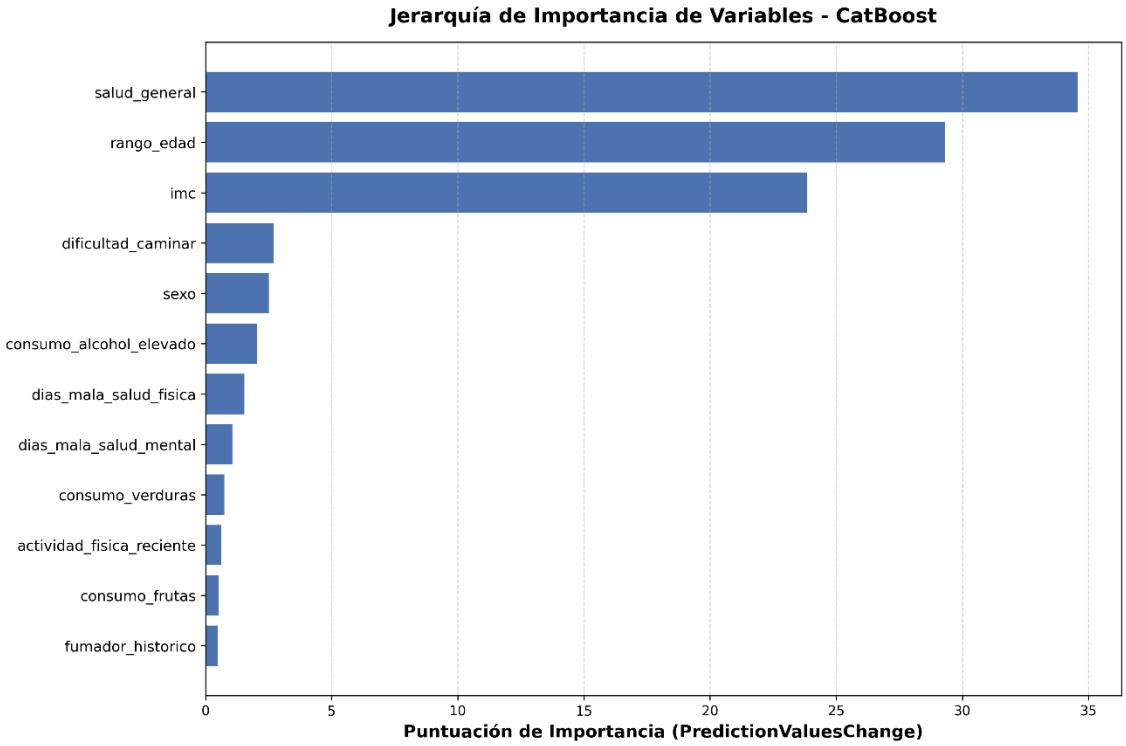

Figura 6. Jerarquía de importancia de las variables predictoras.

Nota. El gráfico de barras muestra el peso asignado por el algoritmo CatBoost a cada factor de riesgo.

Este perfil de riesgo se alinea rigurosamente con la evidencia fisiopatológica establecida, la preeminencia del IMC corrobora los estándares de la American Diabetes Association [15], [16], que identifican al exceso de peso y la obesidad visceral como los determinantes modificables más críticos para el desarrollo de la resistencia a la insulina. La alta relevancia asignada a la edad es consistente con la biología del envejecimiento, la cual documenta un deterioro progresivo en la función de las células beta pancreáticas y una menor sensibilidad a la insulina a medida que avanza la edad biológica [17]. Es notable destacar que el modelo fue capaz de jerarquizar estas relaciones de manera autónoma.

Existe una diferencia sustancial respecto a la literatura previa; estudios recientes demuestran que, para alcanzar precisiones superiores al 80 %, los algoritmos tradicionales suelen depender críticamente de la glucosa plasmática. En contraste, el presente modelo logra un AUC competitivo de 0,804 utilizando exclusivamente variables no invasivas, lo que posibilitó un tamizaje masivo y de bajo costo.

Finalmente, variables de comportamiento como la actividad física y la dieta, aunque presentaron una contribución individual menor en comparación con el IMC, mostraron un efecto protector

acumulativo. Este hallazgo replica los resultados del histórico Diabetes Prevention Program [18], el cual probó definitivamente que la modificación combinada del estilo de vida es más efectiva que la intervención farmacológica para reducir la incidencia de diabetes tipo 2.

Validación de la herramienta, limitaciones y trabajo futuro

La utilidad real de un modelo de aprendizaje automático no reside únicamente en su capacidad predictiva numérica, sino en su viabilidad para ser integrado en el flujo de trabajo de la salud pública, desde un ejercicio de transparencia científica, se discuten las restricciones del alcance actual y se proponen estrategias concretas para evolucionar el sistema hacia una medicina de precisión.

Implementación y pruebas de la interfaz

En lo concerniente a la aplicabilidad tecnológica de la propuesta, la integración del modelo predictivo en una interfaz gráfica de usuario demostró ser funcional y eficiente, como se muestra en la Figura 7, la herramienta permite procesar las 12 variables de entrada y generar una estimación de riesgo en tiempo real mediante un flujo de trabajo intuitivo.

Las pruebas de usabilidad indican que el sistema opera sin requerir conexión a internet ni infraestructura de cómputo de alto rendimiento, esto representa una ventaja significativa frente a las arquitecturas basadas puramente en la nube; como señalan en un informe de la Organización Panamericana de la Salud (OPS), la brecha digital y la desigualdad en la infraestructura de conectividad en América Latina siguen siendo barreras críticas que pueden excluir a poblaciones vulnerables [19]. Esta herramienta, al ser de ejecución local, garantiza su funcionalidad en campañas de salud en zonas rurales o desconectadas.

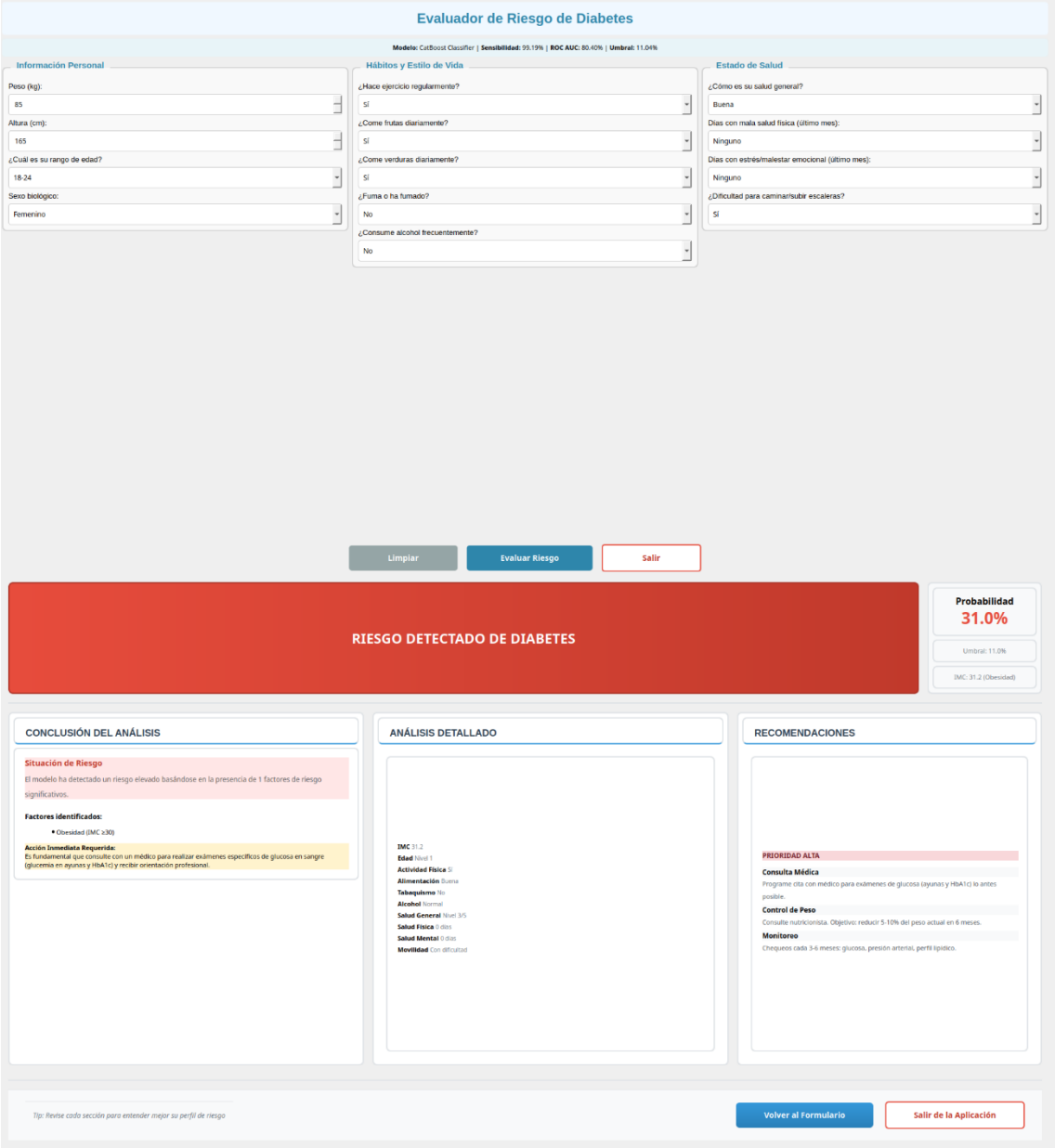

Figura 7. Interfaz gráfica de usuario de la herramienta de predicción de riesgo. Módulo de ingreso de datos (arriba). Módulo de resultados (abajo).

Nota. Despliega la estimación de riesgo generada por el modelo y una alerta visual para la toma de decisiones.

La aplicación fue implementada utilizando la librería PyQt5, lo que permite su ejecución nativa en ordenadores de escritorio sin depender de navegadores web.

Limitaciones y perspectivas futuras

Es imperativo reconocer las restricciones inherentes al presente estudio, la principal limitación radica en la naturaleza de la fuente de datos primaria, la cual se basa en información auto-reportada. Estudios validados [20], [21], han demostrado estadísticamente que existe una discrepancia entre los datos declarados y los reales, observándose una tendencia a subestimar el Índice de Masa Corporal (IMC) y sobreestimar la estatura, lo cual podría introducir un sesgo en la predicción.

Adicionalmente, aunque la alta sensibilidad del modelo (99,1 %) es ideal para el cribado, la tasa de Falsos Positivos implica que algunos usuarios sanos serán alertados, por ello, esta herramienta debe actuar como un "filtro primario" y no como sustituto del diagnóstico clínico. Finalmente, el trabajo futuro debe orientarse hacia la validación clínica prospectiva y la incorporación de biomarcadores genéticos.

Tal como discuten otros investigadores [21], la transición hacia una "medicina de precisión" en diabetes permitirá estratificar el riesgo considerando componentes étnicos y genéticos específicos de nuestra población, mejorando la especificidad de los algoritmos predictivos sin sacrificar su accesibilidad.

CONCLUSIONES

En primer lugar, se desarrolló y validó con éxito un modelo predictivo basado en el algoritmo CatBoost, capaz de estratificar el riesgo de diabetes y prediabetes mediante el análisis exclusivo de 12 variables no invasivas. Se concluye que la integración de determinantes demográficos, conductuales y de estilo de vida en una arquitectura de aprendizaje supervisado es suficiente para inferir la probabilidad de alteración metabólica con alta fidelidad. Este resultado es disruptivo, ya que permite prescindir de la dependencia absoluta de pruebas hematológicas invasivas en las fases iniciales del cribado poblacional, reduciendo barreras de acceso y costos operativos.

En segundo lugar, la evidencia experimental demuestra que el empleo de estrategias de ponderación asimétrica de clases, en conjunto con la optimización del umbral de decisión (punto de corte de 0,1104), constituye una solución técnica robusta ante el desbalance inherente de los registros sanitarios. El modelo alcanzó una sensibilidad (recall) del 99,1 %, validando la hipótesis de que es técnica y clínicamente posible maximizar la recuperación de pacientes en riesgo, minimizando los falsos negativos que representan un peligro crítico en el diagnóstico preventivo.

Asimismo, el análisis de importancia de características identificó al Índice de Masa Corporal (IMC), la Edad y la Autopercepción de Salud como los predictores biológicos de mayor peso relativo, resultados que mantienen una estricta coherencia con la fisiopatología metabólica establecida. En tercer lugar, se demostró la viabilidad tecnológica del sistema mediante el despliegue de una interfaz gráfica de ejecución local. A diferencia de las soluciones basadas en la nube, esta arquitectura garantiza la autonomía funcional en entornos con conectividad limitada, asegurando la soberanía de los datos y la privacidad del paciente.

La validación del software confirmó que el motor de inferencia opera eficientemente en hardware convencional, democratizando el acceso a herramientas de inteligencia artificial sin requerir infraestructura costosa. En resumen, la principal contribución de este trabajo a la medicina preventiva radica en la provisión de un artefacto tecnológico de bajo costo que optimiza la gestión de recursos. La herramienta actúa como un filtro primario, permitiendo a las instituciones de salud focalizar sus intervenciones clínicas en segmentos alertados por el modelo, facilitando así la transición de un modelo sanitario reactivo hacia uno proactivo y preventivo, centrado en la detección temprana antes del desarrollo de complicaciones crónicas sistémicas.

Contribución y autoría

J.G. y **A.A.:** Conceptualización, metodología, software, análisis formal, investigación, curación de datos y redacción: preparación del borrador original. **D.A.** y **V.S.:** Validación, recursos, redacción: revisión y edición, visualización y supervisión. Todos los autores han leído y aceptado la versión publicada del manuscrito.

J.G. and **A.A.:** Conceptualization, methodology, software, formal analysis, investigation, data curation, and writing: original draft preparation. **D.A.** and **V.S.:** Validation, resources, writing: review and editing, visualization, and supervision. All authors have read and agreed to the published version of the manuscript.

Financiamiento

Este trabajo fue financiado por las universidades patrocinantes en el marco de sus actividades de investigación.

Declaración ética

Se declara que la presente investigación se fundamenta en un análisis secundario in silico de datos provenientes del Behavioral Risk Factor Surveillance System (BRFSS), específicamente el conjunto de datos correspondiente al año 2015 [Ref]. Esta es una base de datos de acceso público gestionada por los Centros para el Control y la Prevención de Enfermedades (CDC), disponible para la comunidad científica bajo protocolos de datos abiertos.

Por esta razón, este estudio no involucró experimentación directa ni intervención clínica con seres humanos ni con animales por parte de los autores, por lo que no se requirió la aprobación específica de un comité de ética institucional para la ejecución del modelado computacional. Adicionalmente, se garantiza que el conjunto de datos utilizado se encuentra totalmente anonimizado, cumpliendo con los estándares de confidencialidad y protección de datos personales, ya que no contiene información que permita la identificación individual de los sujetos encuestados. Cabe destacar que la recolección original de los datos por parte del CDC se realizó bajo estrictos protocolos de consentimiento informado y aprobación ética vigentes en su momento.

Uso de inteligencia artificial

La concepción del estudio, el diseño experimental, el análisis e interpretación de los resultados, así como la redacción y revisión crítica del manuscrito, fueron realizados de manera autónoma por los autores, quienes asumen la responsabilidad plena por el contenido final del artículo. Los autores declaran que se utilizaron herramientas de inteligencia artificial generativa de forma asistida exclusivamente para apoyar tareas técnicas de programación.

Disponibilidad de datos

Los datos y los artefactos de investigación que respaldan los resultados de este estudio están disponibles de forma abierta en el repositorio Zenodo (DOI: 110.5281/zenodo.18452614). Asimismo, con el objetivo de garantizar la reproducibilidad computacional, el código fuente completo, los *scripts* de preprocesamiento, configuración del entrenamiento y módulos de despliegue han sido depositados en el repositorio público de GitHub: github.com/cxborjas/Model-for-Early-Detection-of-Diabetes-Risk.

Conflicto de interés

Los autores declaran no tener ningún conflicto de interés de carácter financiero, académico o personal en relación con la realización, interpretación o publicación del presente trabajo de investigación.

Agradecimientos

Se expresa un agradecimiento especial a Claudio Borja Saltos por su colaboración técnica en la revisión de la arquitectura del modelo predictivo y sus observaciones críticas, las cuales contribuyeron significativamente a la validación de los resultados presentados.

REFERENCIAS

[1] P. Arokiasamy, S. Salvi, y Y. Selvamani, «Global Burden of Diabetes Mellitus», en *Handbook of Global Health*, I. Kickbusch, D. Ganten, y M. Moeti, Eds., Cham: Springer International Publishing, 2021, pp. 1-44. doi: 10.1007/978-3-030-05325-3_28-2.

[2] S. Laha, A. Rajput, S. S. Laha, y R. Jadhav, «A Concise and Systematic Review on Non-Invasive Glucose Monitoring for Potential Diabetes Management», *Biosensors*, vol. 12, n.º 11, p. 965, nov. 2022, doi: 10.3390/bios12110965.

[3] E. K. Oikonomou y R. Khera, «Machine learning in precision diabetes care and cardiovascular risk prediction», *Cardiovasc Diabetol*, vol. 22, n.º 1, p. 259, sep. 2023, doi: 10.1186/s12933-023-01985-3.

[4] J. Torales y I. Barrios, «Diseño de investigaciones: algoritmo de clasificación y características esenciales», *Medicina Clínica y Social*, vol. 7, n.º 3, pp. 210-235, sep. 2023, doi: 10.52379/mcs.v7i3.349.

[5] Centers for Disease Control and Prevention (CDC), «BRFSS 2015 Annual Data». 2015. [En línea]. Disponible en: https://www.cdc.gov/brfss/annual_data/annual_2015.html

[6] M. D'Agostino *et al.*, «Toward a holistic definition for Information Systems for Health in the age of digital interdependence», *Revista Panamericana de Salud Pública / Pan American Journal of Public Health*, 2021, [En línea]. Disponible en: https://iris.paho.org/handle/10665.2/55195

[7] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, y M. Cabanillas-Carbonell, «Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes», *Diagnostics*, vol. 13, n.º 14, p. 2383, jul. 2023, doi: 10.3390/diagnostics13142383.

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, y A. Gulin, «CatBoost: unbiased boosting with categorical features», en *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, y R. Garnett, Eds., Curran Associates, Inc., 2018. [En línea]. Disponible en: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

[9] F. H. Vera-Rivera, «Cognitive complexity points: a metric to evaluate the design of microservices-based applications», *inycomp*, vol. 26, n.º 1, mar. 2024, doi: 10.25100/iyc.v26i1.13145.

[10] A. Vabalas, E. Gowen, E. Poliakoff, y A. J. Casson, «Machine learning algorithm validation with a limited sample size», *PLoS ONE*, vol. 14, n.º 11, p. e0224365, nov. 2019, doi: 10.1371/journal.pone.0224365.

[11] R. Trevethan, «Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice», *Front. Public Health*, vol. 5, p. 307, nov. 2017, doi: 10.3389/fpubh.2017.00307.

[12] D. Chicco y G. Jurman, «The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation», *BMC Genomics*, vol. 21, n.º 1, p. 6, dic. 2020, doi: 10.1186/s12864-019-6413-7.

[13] F. S. Nahm, «Receiver operating characteristic curve: overview and practical use for clinicians», *Korean J Anesthesiol*, vol. 75, n.º 1, pp. 25-36, feb. 2022, doi: 10.4097/kja.21209.

[14] A. Martín Conejero y M. Alonso García, «Epidemiological studies or how we have to design our research», *Angiologia*, 2023, doi: 10.20960/angiologia.00543.

[15] American Diabetes Association Professional Practice Committee *et al.*, «2. Diagnosis and Classification of Diabetes: *Standards of Care in Diabetes—2024*», *Diabetes Care*, vol. 47, n.º Supplement_1, pp. S20-S42, ene. 2024, doi: 10.2337/dc24-S002.

[16] Y. Qin *et al.*, «Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type», *IJERPH*, vol. 19, n.º 22, p. 15027, nov. 2022, doi: 10.3390/ijerph192215027.

[17] M. E. Cerf, «Beta Cell Dysfunction and Insulin Resistance», *Front. Endocrinol.*, vol. 4, 2013, doi: 10.3389/fendo.2013.00037.

[18] «Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin», *N Engl J Med*, vol. 346, n.º 6, pp. 393-403, feb. 2002, doi: 10.1056/NEJMoa012512.

[19] A. Lana Pérez, M. J. García Fernández, y M. L. López González, «Evaluación del proceso de un programa realizado a través de Internet y de la telefonía móvil para promover conductas saludables en estudiantes de educación secundaria de España y México», *Rev. Esp. Salud Publica*, vol. 87, n.º 4, pp. 393-406, ago. 2013, doi: 10.4321/S1135-57272013000400009.

[20] A. Martín Conejero, «Validation of diagnostic tests (part-II). Quantitative tests», *Angiologia*, 2023, doi: 10.20960/angiologia.00475.

[21] V. Quick *et al.*, «Concordance of self-report and measured height and weight of college students», *J Nutr Educ Behav*, vol. 47, n.º 1, pp. 94-98, 2015, doi: 10.1016/j.jneb.2014.08.012.

Biografía de los autores

Jesús Ernesto González Torres

Ingeniero de Software graduado por la Universidad Estatal de Bolívar, cuenta con una sólida experiencia profesional en la industria tecnológica, habiendo desempeñado roles estratégicos como Desarrollador de Software y Líder de Proyectos. En la actualidad, ejerce como desarrollador en el Departamento de Tecnologías de la Información y Comunicación (TIC) de la UEB. Sus áreas de interés técnico y líneas de investigación se centran en la programación avanzada, el diseño y gestión de bases de datos, y la implementación de soluciones de inteligencia artificial aplicada a problemas reales.

Software Engineer graduated from the State University of Bolívar (UEB). He possesses solid professional experience in the technology industry, having held strategic roles such as Software Developer and Project Leader. Currently, he serves as a developer in the Department of Information and Communication Technologies (ICT) at UEB. His technical interests and research lines focus on advanced programming, database design and management, and the implementation of artificial intelligence solutions applied to real-world problems.

Alexander Ufredo Alegria Chaves

Ingeniero de Software titulado por la Universidad Estatal de Bolívar (UEB), Ecuador. Cuenta con experiencia profesional como desarrollador de software y actualmente ejerce como Especialista de Desarrollo de Software en el Departamento de Tecnologías de la Información y Comunicación de la misma universidad. Sus líneas de interés e investigación abarcan la programación, la inteligencia artificial, así como las tecnologías blockchain y criptomonedas.

Software Engineer holding a degree from the State University of Bolívar (UEB), Ecuador. He has professional experience as a software developer and currently serves as a Software Development Specialist in the Department of Information and Communication Technologies (ICT) at the same university. His research interests encompass programming, artificial intelligence, as well as blockchain technologies and cryptocurrencies.

Diana Magali Alegría Camino

Es Licenciada en Ciencias de la Educación con mención en Informática Educativa y Magíster en Sistemas de Información, con especialización en Inteligencia de Negocios y Analítica de Datos Masivos. Su formación se complementa con la certificación en Formación de Formadores, avalada por el Sistema Nacional de Cualificaciones y Capacitación Profesional, lo que respalda su sólida competencia metodológica. Actualmente, se desempeña como docente en la carrera de Desarrollo de Software, donde integra su doble perfil pedagógico y tecnológico. Sus principales áreas de interés e investigación abarcan la programación, la investigación educativa y la aplicación de la inteligencia artificial, enfocándose en el uso de nuevas tecnologías para la innovación académica y el análisis de datos.

She holds a Licentiate degree in Education Sciences with a major in Educational Informatics and a Master's degree in Information Systems, specializing in Business Intelligence and Big Data Analytics. Her background is complemented by the "Training of Trainers" certification, endorsed by the

National System of Qualifications and Professional Training, which underpins her strong methodological competence. Currently, she serves as a professor in the Software Development program, where she integrates her dual pedagogical and technological profile. Her main areas of interest and research include programming, educational research, and the application of artificial intelligence, focusing on the use of emerging technologies for academic innovation and data analysis.

Verónica Elizabeth Sánchez Aguiar

Es Licenciada en Ciencias de la Educación con mención en Informática Educativa y Magíster en Sistemas de Información, con especialización en Inteligencia de Negocios y Analítica de Datos Masivos. Su preparación metodológica está respaldada por la certificación en Formación de Formadores del Sistema Nacional de Cualificaciones y Capacitación Profesional. En la actualidad, ejerce como docente en la carrera de Electrónica, donde aplica su formación interdisciplinaria. Sus principales áreas de interés y desarrollo profesional se centran en las Tecnologías de la Información, integrando herramientas de gestión de datos y sistemas informáticos en el ámbito de la ingeniería y la educación técnica.

She holds a Licentiate degree in Education Sciences with a major in Educational Informatics and a Master's degree in Information Systems, specializing in Business Intelligence and Big Data Analytics. Her methodological preparation is supported by the "Training of Trainers" certification from the National System of Qualifications and Professional Training. Currently, she serves as a professor in the Electronics program, where she applies her interdisciplinary background. Her primary areas of interest and professional development focus on Information Technologies, integrating data management tools and computer systems within the fields of engineering and technical education.

Descargo de responsabilidad