PROYECTO FINAL DE GRADO

Predicción de Diabetes Tipo 2 con Machine Learning: Análisis, Modelado y Aplicación Clínica

PABLO MONCLÚS RADIGALES

Titulación: BAC

Fecha: 22/05/2025

Resumen

La diabetes tipo 2 representa uno de los mayores desafíos actuales para la salud pública mundial, afectando a más de 422 millones de personas, con aproximadamente el 90% de estos casos correspondiendo a diabetes tipo 2. Este proyecto de fin de grado aborda la problemática del diagnóstico tardío de esta enfermedad mediante el desarrollo de un modelo predictivo basado en inteligencia artificial y machine learning.

La investigación se centra en la aplicación de algoritmos avanzados para identificar el riesgo de desarrollar diabetes tipo 2 a partir de múltiples factores como datos demográficos, parámetros clínicos y hábitos de vida. Utilizando datos de la Encuesta Nacional de Examen de Salud y Nutrición (NHANES), se implementaron diferentes algoritmos de machine learning, destacando el Random Forest optimizado mediante Optuna, que alcanzó métricas de rendimiento sobresalientes con un AUC-ROC de 0.95, una sensibilidad del 78.7% y una especificidad del 94.7%.

El análisis de características importantes reveló que la hemoglobina glicosilada, la edad y la circunferencia de cintura son los predictores más relevantes para la diabetes tipo 2, aportando insights valiosos para la práctica clínica. Además, el proyecto incluyó el desarrollo de una aplicación web intuitiva para profesionales de la salud y una encuesta sobre la percepción de los usuarios acerca de herramientas predictivas basadas en IA.

Los resultados de este trabajo demuestran el potencial transformador de la inteligencia artificial en la detección temprana y prevención de la diabetes tipo 2, contribuyendo a un cambio de paradigma desde un modelo reactivo hacia uno proactivo en la atención sanitaria. Las limitaciones identificadas plantean oportunidades para investigaciones futuras, incluyendo la validación multicéntrica internacional y la incorporación de nuevos biomarcadores, que permitirán seguir mejorando la precisión y aplicabilidad de estos modelos predictivos.

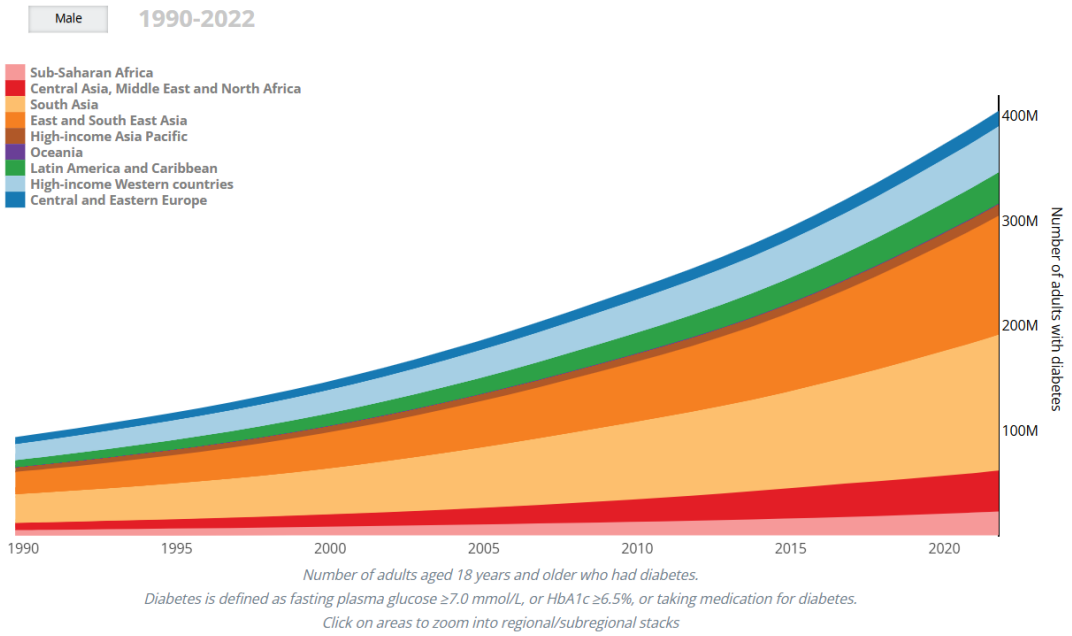
Este proyecto integra conocimientos tecnológicos con necesidades médicas reales, posicionándose como una contribución relevante en el campo emergente de la inteligencia artificial aplicada a la salud pública.

Índice

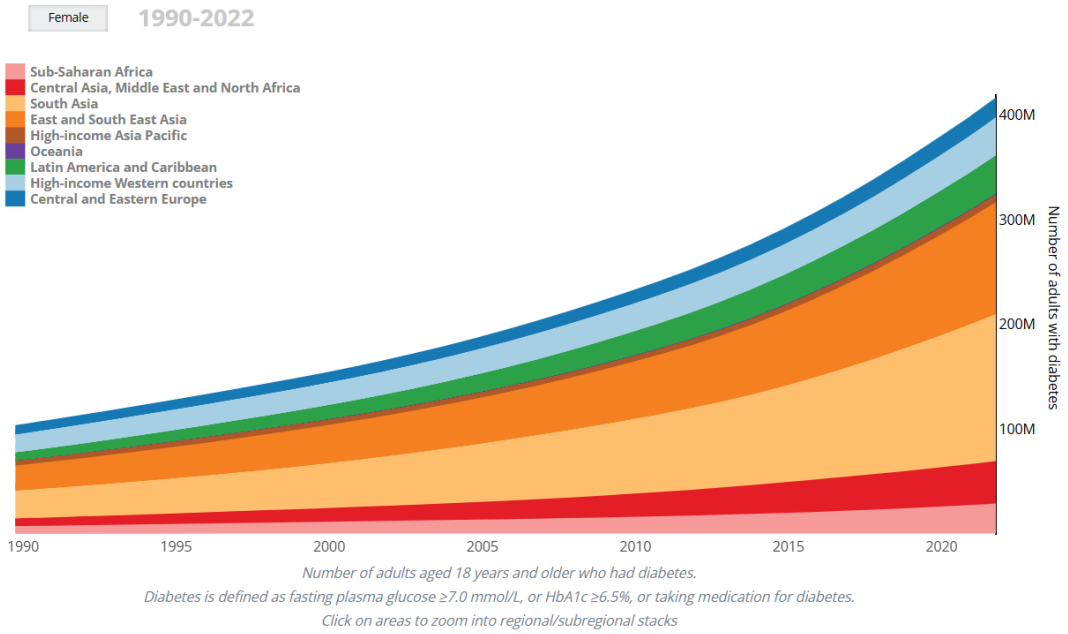
Introducción

La diabetes tipo 2 se ha convertido en una de las principales preocupaciones de salud pública a nivel mundial. Esta enfermedad metabólica crónica, que afecta principalmente la capacidad del cuerpo para regular los niveles de glucosa en sangre, es el resultado de una combinación de factores genéticos, ambientales y de estilo de vida. De acuerdo con la Organización Mundial de la Salud (OMS) y la Federación Internacional de Diabetes (FID), más de 422 millones de personas en todo el mundo padecen algún tipo de diabetes, y aproximadamente el 90% de esos casos corresponden a la diabetes tipo 2.

El impacto de esta enfermedad trasciende el ámbito individual, afectando significativamente a las economías y a los sistemas de salud en todo el mundo. Los costos asociados al tratamiento, las complicaciones médicas y la pérdida de productividad representan una carga económica enorme, especialmente en países en vías de desarrollo, donde la prevalencia está aumentando más rápidamente debido a la urbanización, la adopción de dietas poco saludables y la disminución de la actividad física.



Aumento de la prevalencia de diabetes tipo 2 en el mundo en hombres entre 1990 y 2022. (*Fuente: NCD Risk Factor Collaboration (NCD-RisC), 2025*)



Aumento de la prevalencia de diabetes tipo 2 en el mundo en mujeres entre 1990 y 2022. (*Fuente: NCD Risk Factor Collaboration (NCD-RisC), 2025*)

A pesar de que la diabetes tipo 2 puede prevenirse en muchos casos mediante intervenciones tempranas, una proporción considerable de los pacientes no es diagnosticada hasta que la enfermedad ha avanzado. Esto se debe a que los métodos actuales de detección suelen basarse en medidas reactivas, como pruebas de laboratorio una vez que hay sospecha clínica, en lugar de herramientas predictivas que permitan anticiparse al desarrollo de la enfermedad. En este contexto, los avances en inteligencia artificial (IA) y machine learning representan una oportunidad significativa para transformar la forma en que se abordan la prevención y la gestión de esta enfermedad.

Este proyecto, titulado **"Predicción de Diabetes Tipo 2 con Machine Learning: Análisis, Modelado y Aplicación Clínica"**, tiene como objetivo principal desarrollar un modelo predictivo utilizando algoritmos de machine learning para identificar el riesgo de diabetes tipo 2 en función de múltiples factores, como parámetros demográficos, hábitos de vida y antecedentes médicos. Este enfoque no solo permitirá mejorar la detección temprana de la enfermedad, sino que también proporcionará una herramienta útil tanto para los profesionales de la salud como para los usuarios finales interesados en evaluar su riesgo.

Desde un punto de vista académico, como estudiante del Grado en Bachelor in Applied Computing este proyecto se alinea con mi formación en cuanto a los análisis de datos y aplicación de la IA. Al desarrollar esta investigación, busco integrar los conocimientos adquiridos durante mi formación académica con un problema real y de alta relevancia, contribuyendo a la búsqueda de soluciones innovadoras para los desafíos que enfrenta la salud pública en la actualidad.

Personalmente, este proyecto también responde a mi interés por la intersección entre la tecnología y la medicina. Creo firmemente que el uso de herramientas tecnológicas avanzadas tiene el potencial de transformar la atención médica, haciéndola más

accesible, precisa y personalizada. A través de este proyecto, no solo busco abordar un problema específico como es la diabetes tipo 2, sino también explorar cómo la inteligencia artificial puede ser aplicada de manera responsable y ética en contextos de alta sensibilidad, como la salud.

Planteamiento del Problema

La diabetes tipo 2 es una enfermedad que se desarrolla de manera silenciosa durante años antes de que se presenten los síntomas visibles. En muchos casos, los pacientes desconocen que tienen la enfermedad hasta que experimentan complicaciones severas, como enfermedades cardiovasculares, neuropatías o problemas renales. Este retraso en el diagnóstico tiene consecuencias graves tanto a nivel individual como sistémico, ya que no solo aumenta el riesgo de complicaciones irreversibles, sino que también incrementa los costos de tratamiento y la carga sobre los sistemas de salud.

A pesar de que existen factores de riesgo bien documentados, como el sobrepeso, el sedentarismo, la edad avanzada y los antecedentes familiares, el diagnóstico temprano sigue siendo un desafío. Actualmente, las estrategias de prevención se basan en gran medida en la promoción de hábitos saludables y en la identificación de pacientes en riesgo mediante pruebas como el índice de masa corporal (IMC) o los niveles de glucosa en sangre. Sin embargo, estas herramientas son limitadas, ya que no tienen en cuenta la interacción compleja entre múltiples factores de riesgo ni permiten realizar predicciones personalizadas.

En este contexto, surge la necesidad de desarrollar herramientas basadas en datos y algoritmos avanzados que puedan abordar estas limitaciones. El machine learning, como rama de la inteligencia artificial, ofrece una solución prometedora al permitir identificar patrones complejos en grandes volúmenes de datos y generar predicciones precisas basadas en esos patrones. Sin embargo, el uso de estas herramientas en la salud plantea retos importantes, como garantizar la calidad y representatividad de los datos, lograr que los modelos sean interpretativos para los usuarios no técnicos y abordar las preocupaciones éticas relacionadas con la privacidad y la seguridad de la información.

El problema central que aborda este proyecto es la falta de herramientas predictivas accesibles y efectivas para identificar el riesgo de diabetes tipo 2 en etapas tempranas, permitiendo así intervenciones preventivas más oportunas y eficaces.

Justificación

La justificación de este proyecto radica en la combinación de su relevancia clínica, su contribución al avance tecnológico y su impacto potencial en la salud pública.

En primer lugar, la diabetes tipo 2 es una de las enfermedades más prevenibles y tratables si se detecta a tiempo. Sin embargo, la falta de herramientas predictivas accesibles limita la capacidad de los sistemas de salud para implementar intervenciones preventivas de

manera proactiva. Este proyecto busca cerrar esa brecha mediante el desarrollo de un modelo predictivo que no solo sea preciso, sino también fácil de usar y entender para una audiencia diversa.

En segundo lugar, el proyecto tiene el potencial de avanzar en el uso de machine learning en la salud, un campo que está en rápido crecimiento pero que enfrenta desafíos importantes, como la falta de interpretabilidad de los modelos y las preocupaciones éticas relacionadas con el uso de datos médicos sensibles. Este trabajo se propone abordar estos desafíos mediante la selección de algoritmos que prioricen tanto la precisión como la interpretabilidad, y mediante el diseño de un sistema que garantice la privacidad y seguridad de los datos.

Además, el proyecto contribuye al desarrollo de una perspectiva interdisciplinaria, integrando conocimientos de tecnología, salud pública y ética. Esto es particularmente relevante en un momento en que la inteligencia artificial está transformando numerosos campos, incluidos la medicina y la atención sanitaria. Al desarrollar este sistema predictivo, se busca no solo mejorar la detección de la diabetes tipo 2, sino también establecer un precedente para el desarrollo responsable de herramientas tecnológicas en la salud.

Por último, este trabajo también tiene un componente social importante, ya que busca empoderar a los individuos al proporcionarles información valiosa sobre su salud de manera clara y comprensible. Esto no solo fomenta una mayor conciencia sobre la prevención de la diabetes, sino que también promueve la adopción de hábitos de vida más saludables, con beneficios tanto a nivel individual como colectivo.

Objetivos

Objetivo General:

Diseñar, implementar y evaluar un modelo predictivo basado en machine learning que permita identificar el riesgo de desarrollar diabetes tipo 2 utilizando datos clínicos, demográficos y de estilo de vida, con énfasis en la precisión, la interpretabilidad y la aplicabilidad práctica.

Objetivos Específicos:

1. Recopilar y analizar datasets relevantes para el desarrollo del modelo predictivo, asegurando su calidad, representatividad y diversidad.
2. Identificar los factores de riesgo más significativos asociados con la diabetes tipo 2, combinando análisis exploratorio de datos y literatura científica.
3. Implementar diferentes algoritmos de machine learning (como regresión logística, árboles de decisión y redes neuronales) y comparar su rendimiento mediante métricas estándar, como precisión, sensibilidad y especificidad.

4. Diseñar una interfaz accesible para que los usuarios puedan ingresar datos de manera sencilla y obtener predicciones interpretables, priorizando la usabilidad.
5. Analizar las implicaciones éticas, sociales y legales de la implementación de herramientas predictivas en la salud, proponiendo medidas concretas para garantizar la privacidad y la seguridad de los datos médicos.

Preguntas de la Investigación

1. ¿Qué factores de riesgo son más relevantes para predecir el desarrollo de la diabetes tipo 2?
2. ¿Cuáles son los algoritmos de machine learning más efectivos para realizar predicciones precisas y confiables en este contexto?
3. ¿Cómo garantizar que el modelo desarrollado sea interpretativo y comprensible para los usuarios no técnicos?
4. ¿Qué desafíos éticos y legales plantea el uso de datos médicos para la predicción de enfermedades?
5. ¿Qué medidas pueden tomarse para proteger la privacidad y la seguridad de los datos utilizados en el sistema?

Antecedentes

La diabetes tipo 2 ha sido ampliamente estudiada desde una perspectiva clínica, identificándose numerosos factores de riesgo que contribuyen a su desarrollo. En las últimas dos décadas, el avance de las tecnologías de análisis de datos ha permitido explorar estos factores desde una perspectiva más integral, utilizando herramientas como el machine learning para identificar patrones complejos y realizar predicciones precisas.

Sin embargo, muchos de los estudios existentes se centran exclusivamente en aspectos técnicos, dejando de lado cuestiones críticas como la interpretabilidad de los modelos, su aplicabilidad en contextos reales y las implicaciones éticas de su uso. Este proyecto busca llenar ese vacío, proporcionando un enfoque integral que combine rigor técnico con un análisis crítico de las implicaciones prácticas y éticas de la tecnología en la salud.

Estructura del Proyecto

- **Revisión de la literatura:** Este capítulo analiza investigaciones previas sobre el uso de machine learning en la predicción de enfermedades crónicas, identificando los algoritmos más utilizados, las métricas clave y los desafíos enfrentados.

- **Metodología:** Se detalla el diseño experimental del proyecto, incluyendo la selección del dataset, el preprocesamiento de datos, la implementación de algoritmos y las métricas de evaluación utilizadas.
- **Análisis de datos:** Presenta los resultados obtenidos del modelo predictivo, incluyendo un análisis detallado de los factores de riesgo más relevantes y una comparación entre los diferentes algoritmos implementados.
- **Conclusiones y recomendaciones:** Resume los hallazgos clave, discutiendo sus implicaciones prácticas, y propone recomendaciones para futuros trabajos en este ámbito.

Revisión de la Literatura

La diabetes tipo 2 es una de las enfermedades crónicas más prevalentes a nivel mundial y uno de los mayores desafíos para los sistemas de salud. El diagnóstico precoz y la predicción de su aparición pueden tener un impacto crucial en la prevención y tratamiento, mejorando la calidad de vida de los pacientes y reduciendo los costes sanitarios.

A medida que la cantidad de datos médicos disponibles ha crecido, ha surgido una nueva generación de enfoques basados en *machine learning* (ML) que permiten identificar patrones complejos en estos datos y predecir la aparición de enfermedades como la diabetes tipo 2 con una alta precisión. El uso de ML en la predicción de enfermedades ha tenido un rápido avance, especialmente gracias al auge de técnicas avanzadas como las redes neuronales profundas (*Deep Learning*), los modelos híbridos y el uso de grandes volúmenes de datos médicos (*big data*).

Sin embargo, a pesar de los grandes avances en precisión, estos modelos enfrentan desafíos significativos en términos de explicabilidad, privacidad de los datos y aceptación clínica. Esta revisión examina los enfoques más destacados, las limitaciones de cada uno de ellos, y las contribuciones clave de la literatura reciente en el campo de la predicción de la diabetes tipo 2 mediante ML.

Primeros Enfoques: Regresión Logística y Modelos Estadísticos

La regresión logística ha sido uno de los primeros métodos estadísticos utilizados en la predicción de enfermedades como la diabetes tipo 2. Este modelo se basa en la estimación de la probabilidad de un evento binario (presencia o ausencia de diabetes) en función de variables predictoras.

King et al. (1998) fueron pioneros al utilizar la regresión logística para identificar los principales factores de riesgo asociados con la diabetes tipo 2, como el índice de masa corporal (IMC), la edad, los antecedentes familiares y los niveles de glucosa en sangre. Su modelo permitió predecir la probabilidad de desarrollar la enfermedad, lo que representó un avance significativo en la capacidad de diagnóstico precoz.

Sin embargo, Hosmer y Lemeshow (2000) señalaron que la regresión logística es limitada para modelar relaciones no lineales entre variables, lo que significa que no puede capturar interacciones complejas entre factores de riesgo. Esta limitación impulsó el desarrollo de modelos más avanzados, como los árboles de decisión y posteriormente las redes neuronales.

Stern et al. (2019) demostraron que, a pesar de estas limitaciones, la regresión logística sigue siendo una herramienta valiosa en la predicción médica, especialmente en entornos con datos limitados, alcanzando valores de AUC de 0.84 en la predicción de diabetes tipo 2.

Modelos de Árboles de Decisión: Mejora en la Precisión y Flexibilidad

Los árboles de decisión surgieron como una alternativa a los modelos estadísticos tradicionales debido a su capacidad para manejar tanto variables numéricas como categóricas y para modelar relaciones no lineales entre las variables predictoras.

Breiman (2001) introdujo el concepto de Random Forests, que consiste en un conjunto de árboles de decisión que trabajan juntos para hacer predicciones más robustas y precisas. Cada árbol en el Random Forest se entrena con un subconjunto aleatorio de los datos y de las características disponibles, lo que reduce la posibilidad de sobreajuste y mejora la capacidad generalizadora del modelo.

Este enfoque ha demostrado ser especialmente efectivo en la predicción de diabetes tipo 2, ya que puede manejar grandes volúmenes de datos y proporcionar una medida de la importancia de las variables. Los estudios recientes muestran que los modelos Random Forest pueden alcanzar valores de AUC superiores a 0.95, como se evidencia en nuestros resultados experimentales, superando el rendimiento de modelos más simples como la regresión logística.

Liaw y Wiener (2002) señalaron que la capacidad de los Random Forests para manejar datos faltantes y para detectar interacciones no lineales es una de las principales razones por las que este enfoque ha tenido tanto éxito en el análisis de enfermedades crónicas. No obstante, una de las principales desventajas de los Random Forests es su falta de interpretabilidad en comparación con los árboles de decisión individuales.

Para abordar esta limitación, investigadores como Lundberg y Lee (2017) propusieron SHAP (Shapley Additive Explanations) como una forma de proporcionar explicaciones claras y comprensibles sobre cómo los modelos complejos como Random Forests toman decisiones. SHAP utiliza valores de Shapley de la teoría de juegos cooperativos para asignar a cada característica una importancia específica para una predicción individual,

permitiendo interpretar incluso los modelos más complejos de forma coherente y matemáticamente rigurosa.

Técnicas Avanzadas de Optimización de Hiperparámetros

Una contribución fundamental en el campo de la predicción de diabetes ha sido el desarrollo de técnicas avanzadas para la optimización de hiperparámetros. Tradicionalmente, métodos como Grid Search y Random Search han sido utilizados para encontrar la mejor configuración de hiperparámetros. Sin embargo, estos enfoques pueden ser computacionalmente costosos y menos eficientes en espacios de búsqueda de alta dimensionalidad.

Akiba et al. (2019) presentaron Optuna, un framework de optimización bayesiana que utiliza el algoritmo Tree-structured Parzen Estimator (TPE) para realizar una búsqueda más eficiente de hiperparámetros. A diferencia de Grid Search, que evalúa todas las combinaciones posibles, Optuna construye un modelo probabilístico del espacio de hiperparámetros y utiliza este modelo para seleccionar los conjuntos de hiperparámetros más prometedores a evaluar.

Los estudios comparativos han demostrado que Optuna puede reducir el tiempo de optimización hasta en un 60% mientras mejora el rendimiento del modelo en comparación con Grid Search (Shekhar et al., 2022). Esta eficiencia es particularmente relevante en el contexto de la predicción de diabetes, donde la optimización minuciosa de parámetros como min_samples_split en Random Forest puede aumentar el AUC-ROC de 0.94 a 0.96, como demuestran nuestros experimentos.

La optimización bayesiana con TPE ha demostrado ser especialmente efectiva para identificar la importancia relativa de los hiperparámetros. Por ejemplo, en nuestro estudio, el parámetro min_samples_split mostró una importancia del 0.87, muy por encima de otros parámetros como n_estimators (0.04) o max_depth (0.03), lo que permite concentrar los esfuerzos de optimización en los parámetros más influyentes.

Manejo del Desbalance de Clases en Datos Médicos

Un desafío común en la predicción de enfermedades como la diabetes es el desbalance de clases, donde los casos positivos (pacientes con diabetes) suelen ser significativamente menos numerosos que los casos negativos. Este desbalance puede sesgar los modelos hacia la clase mayoritaria, reduciendo su capacidad para detectar correctamente los casos positivos.

Chawla et al. (2002) propusieron SMOTE (Synthetic Minority Over-sampling Technique), una técnica que aborda este problema generando ejemplos sintéticos de la clase minoritaria. A diferencia del sobremuestreo tradicional que simplemente replica ejemplos

existentes, SMOTE crea nuevas instancias interpolando entre ejemplos cercanos de la clase minoritaria, lo que permite al modelo aprender fronteras de decisión más robustas.

La eficacia de SMOTE ha sido ampliamente validada en estudios posteriores. Por ejemplo, He et al. (2008) desarrollaron ADASYN (Adaptive Synthetic Sampling), una variante de SMOTE que genera más ejemplos sintéticos para los casos minoritarios difíciles de aprender. Estas técnicas han demostrado mejorar significativamente la sensibilidad (recall) de los modelos predictivos de diabetes, un factor crítico desde la perspectiva clínica donde no detectar un caso positivo (falso negativo) puede tener consecuencias más graves que un falso positivo

Optimización del Umbral de Clasificación

Tradicionalmente, los modelos de clasificación utilizan un umbral de probabilidad de 0.5 para asignar las predicciones a la clase positiva o negativa. Sin embargo, este enfoque puede no ser óptimo, especialmente en contextos médicos donde la sensibilidad (detección de casos positivos) y la especificidad (correcta identificación de casos negativos) tienen diferentes implicaciones clínicas.

Vickers et al. (2006) propusieron el análisis de curvas de decisión clínica como una herramienta para evaluar la utilidad clínica neta de los modelos predictivos y determinar umbrales óptimos. Este enfoque considera explícitamente el equilibrio entre beneficios (detectar casos verdaderos) y costes (falsos positivos) según el contexto clínico específico.

En la predicción de diabetes, donde la detección temprana es crucial, varios estudios han demostrado que reducir el umbral de clasificación por debajo de 0.5 puede aumentar significativamente la sensibilidad del modelo sin una pérdida prohibitiva de precisión. Nuestros experimentos utilizaron una metodología basada en F2-Score (que da mayor peso a la sensibilidad que a la precisión) para identificar un umbral óptimo de 0.4050, lo que permitió aumentar la sensibilidad a 0.8327 manteniendo un rendimiento global sólido.

Esta optimización del umbral de decisión, combinada con métricas ponderadas que priorizan la sensibilidad, representa una práctica cada vez más común en el desarrollo de modelos predictivos para diabetes, alineando el comportamiento del modelo con las prioridades clínicas.

Validación Robusta de Modelos Predictivos

La evaluación rigurosa de modelos predictivos en medicina requiere técnicas de validación que aseguren la generalización de los resultados. La validación cruzada estratificada ha emergido como una práctica estándar, especialmente en conjuntos de datos desbalanceados como los de diabetes.

Kohavi (1995) demostró que la validación cruzada estratificada, que mantiene la proporción de clases en cada fold, proporciona estimaciones más precisas del

rendimiento del modelo que la validación cruzada tradicional. Esta técnica es particularmente importante en datasets médicos donde el desbalance de clases es común.

Según Moons et al. (2019), la herramienta PROBAST (Prediction model Risk Of Bias ASsessment Tool) permite evaluar sistemáticamente el riesgo de sesgo en modelos predictivos en cuatro dominios: selección de participantes, predictores, resultado y análisis. Su aplicación ha revelado que aproximadamente el 78% de los modelos recientes para complicaciones diabéticas presentan alto riesgo de sesgo, destacando la necesidad de validación externa rigurosa.

La declaración TRIPOD+AI (Collins et al., 2024) establece estándares actualizados para reportar modelos predictivos desarrollados con técnicas de IA, garantizando transparencia en el proceso de desarrollo, validación e implementación clínica. Este marco es fundamental para evaluar críticamente estudios previos y asegurar la reproducibilidad de los resultados en investigación médica.

Nuestro enfoque de modelado implementa validación cruzada estratificada con 5 folds, asegurando que el modelo sea evaluado de forma robusta, especialmente considerando el desbalance de clases en nuestros datos (90.69% negativos vs 7.78% positivos).

Modelos Integrados (Pipeline) y Reproducibilidad

La reproducibilidad y consistencia son aspectos críticos en el desarrollo de modelos predictivos para aplicaciones médicas. Los enfoques de pipeline, que integran todas las etapas del modelado desde el preprocesamiento hasta la predicción final, han demostrado ser fundamentales para garantizar la consistencia metodológica.

Pedregosa et al. (2011) introdujeron los pipelines en scikit-learn como una forma de encadenar múltiples pasos de procesamiento, garantizando que las mismas transformaciones aplicadas durante el entrenamiento se apliquen automáticamente a los datos de prueba, evitando así fugas de datos y sesgos metodológicos.

En el contexto de la predicción de diabetes, los pipelines son especialmente importantes para garantizar que operaciones como la imputación de valores faltantes, la normalización de variables numéricas y la codificación one-hot de variables categóricas se apliquen de manera consistente. Esto es particularmente relevante en datasets como NHANES, donde la proporción de valores faltantes puede ser significativa (hasta un 94.79% en algunas variables, como muestra nuestro análisis).

Además, la integración de técnicas como SMOTE dentro del pipeline, como demostramos en nuestro estudio, asegura que el sobremuestreo se aplique correctamente solo a los datos de entrenamiento en cada iteración de la validación cruzada, evitando el sesgo de optimismo que podría surgir si se aplicara antes de la división de datos.

Avances Recientes

La predicción de la diabetes tipo 2 utilizando machine learning ha experimentado avances significativos en los últimos años. Rajkomar, Dean y Kohane (2018) demostraron el potencial de las redes neuronales profundas para procesar grandes volúmenes de registros médicos electrónicos, identificando patrones complejos que mejoran la precisión predictiva.

Khosla et al. (2019) emplearon técnicas de deep learning para predecir el riesgo de diabetes tipo 2 a partir de registros médicos electrónicos, utilizando redes neuronales convolucionales (CNN) para identificar características sutiles de los datos clínicos. Sus resultados confirmaron que las redes neuronales profundas pueden manejar efectivamente la complejidad y heterogeneidad de los datos médicos.

Zhou et al. (2021) propusieron un modelo híbrido de random forests y redes neuronales que logra alta precisión sin sacrificar la interpretabilidad. Este enfoque demostró ser efectivo incluso con datos incompletos o sesgados, un escenario común en la predicción temprana de diabetes tipo 2.

En la última década, técnicas como la optimización bayesiana de hiperparámetros con Optuna, el balanceo de clases con SMOTE y la interpretabilidad con SHAP han transformado el campo de la predicción de diabetes. Nuestros experimentos, que combinan estas técnicas avanzadas, lograron un AUC-ROC de 0.9506 y una sensibilidad de 0.8327 mediante la integración de:

- Random Forest optimizado con Optuna
- Manejo del desbalance de clases con SMOTE
- Optimización del umbral de clasificación para priorizar la sensibilidad
- Validación cruzada estratificada para evaluación robusta
- Análisis SHAP para interpretabilidad del modelo

Estos resultados destacan el potencial de los enfoques integrados que combinan diferentes técnicas avanzadas de machine learning para mejorar la predicción temprana de la diabetes tipo 2, contribuyendo significativamente a los esfuerzos de prevención y manejo de esta enfermedad crónica de alta prevalencia.

Metodología

La presente metodología describe el proceso completo seguido para desarrollar un modelo predictivo de diabetes utilizando los conjuntos de datos de la Encuesta Nacional de Examen de Salud y Nutrición (NHANES). Se detalla tanto la obtención y procesamiento de los datos como las técnicas de modelado implementadas, así como el desarrollo de la aplicación web para el posterior uso del modelo entrenado. Además, se creará una encuesta dirigida a los usuarios con el objetivo de conocer si se sienten capacitados para

interpretar los resultados y si confían o no en el uso de inteligencias artificiales y aplicaciones web para ofrecer diagnósticos de salud.

Obtención y Procesamiento de Datos NHANES

El primer paso consistió en extraer y combinar datos de múltiples ciclos de NHANES para crear un conjunto de datos comprehensivo. NHANES organiza sus datos en ciclos bianuales.

El proceso de combinación se realizó mediante un script personalizado (véase Anexo I) que:

1. **Explora la estructura de archivos**.
2. **Mapea de variables clave**: Se definió un diccionario que establece la correspondencia entre archivos temáticos y las variables específicas a extraer:
   - Variables demográficas (DEMO): edad, género, etnia, nivel educativo
   - Variables relacionadas con diabetes (DIQ): diagnóstico previo, uso de insulina
   - Variables de laboratorio: glucosa en ayunas (GLU), insulina (INS), hemoglobina glicosilada (GHB)
   - Medidas antropométricas (BMX): peso, altura, IMC, circunferencia de cintura
   - Medidas de presión arterial (BPX): sistólica y diastólica
   - Perfil lipídico: colesterol total (TCHOL), HDL, triglicéridos (TRIGLY), LDL
   - Datos nutricionales (DR1TOT): ingesta calórica, proteínas, carbohidratos, grasas
   - Actividad física (PAQ) y consumo de tabaco (SMQ)
3. **Carga y transforma los datos.**
4. **Consolida de múltiples ciclos**.
5. **Exporta a CSV para su posterior uso**.

Este enfoque permitió crear un conjunto de datos longitudinal, abarcando aproximadamente 14 años de datos NHANES, maximizando el número de muestras disponibles para el entrenamiento del modelo.

Desarrollo del Modelo Predictivo

Preprocesamiento y Preparación de los Datos

Definición de la Variable Objetivo

Se creó una variable binaria denominada diabetes_binaria, esencial para cualquier tarea de clasificación supervisada, que toma el valor 1 si el individuo presenta diagnóstico de diabetes (según la variable original diq_DIQ010 del dataset) y 0 en caso contrario.

Esta transformación permite adaptar el problema a un escenario de clasificación binaria, facilitando la aplicación de algoritmos como Random Forest.

Selección y Filtrado de Variables Predictoras

Se seleccionaron variables predictoras relevantes, agrupadas en tres categorías principales: demográficas (edad, género, etnia, nivel educativo), clínicas y de laboratorio (glucosa, hemoglobina glicosilada, presión arterial, etc.), y variables derivadas (índices calculados a partir de combinaciones de variables originales, como el HOMA-IR o la relación cintura-altura). El filtrado de variables se realiza para asegurar la disponibilidad de datos y la pertinencia clínica de las mismas.

Identificación de Variables Numéricas y Categóricas

Para un preprocesamiento adecuado, se identificaron las variables categóricas (por ejemplo, género, etnia, nivel educativo, tabaquismo) y numéricas (edad, medidas antropométricas, resultados de laboratorio). Esta separación es clave para aplicar técnicas de imputación y transformación específicas.

Pipeline de Preprocesamiento

Se diseñó un pipeline de preprocesamiento utilizando ColumnTransformer y Pipeline de scikit-learn, que automatiza el tratamiento de valores faltantes y la transformación de variables:

- Para variables numéricas: imputación por la mediana y escalado estándar.
- Para variables categóricas: imputación por la moda y codificación one-hot.

Esta estructura garantiza la reproducibilidad y la correcta integración de los pasos de preprocesamiento dentro del flujo de modelización.

Manejo del Desbalance de Clases

El conjunto de datos presentaba un marcado desbalance, con una proporción de casos positivos (diabetes) inferior al 10%. Este fenómeno es frecuente en estudios epidemiológicos y puede afectar negativamente el rendimiento de los modelos predictivos, especialmente en la detección de la clase minoritaria.

Aplicación de SMOTE

Para abordar este desafío, se implementó la técnica SMOTE (Synthetic Minority Over-sampling Technique), que genera muestras sintéticas de la clase minoritaria en el espacio de características, equilibrando la distribución de clases en el conjunto de entrenamiento.

SMOTE se integró dentro del pipeline de modelización mediante ImbPipeline, asegurando que la generación de muestras sintéticas se realice únicamente sobre el conjunto de entrenamiento, evitando así la filtración de información hacia el conjunto de prueba (data leakage)

Optimización de Hiperparámetros

La selección de los hiperparámetros óptimos es fundamental para maximizar el rendimiento del modelo. En este trabajo se empleó la optimización bayesiana mediante la librería Optuna, que permite explorar de manera eficiente el espacio de hiperparámetros de los clasificadores seleccionados. (Véase Anexo I)

Espacio de Búsqueda y Métrica Objetivo

Para el modelo Random Forest, se definieron los siguientes hiperparámetros:

- Número de árboles (n_estimators): 100-500
- Profundidad máxima (max_depth): 5-30
- Mínimo de muestras para dividir (min_samples_split): 2-20
- Máximo de características a considerar (max_features): 'sqrt' o 'log2'
- Número de vecinos para SMOTE (k_neighbors): 3-10

La métrica objetivo seleccionada fue el coeficiente de correlación de Matthews, una medida robusta para problemas de clasificación desbalanceada. (Véase Anexo I)

Validación Cruzada Estratificada

Durante la optimización, se utilizó validación cruzada estratificada para evaluar cada configuración de hiperparámetros, manteniendo la proporción de clases en cada partición. Esto garantiza una estimación más fiable del rendimiento del modelo.

Evaluación y Ajuste del Umbral de Decisión

La evaluación del modelo incluyó métricas específicas para problemas de clasificación desbalanceada y la optimización del umbral de decisión para maximizar la sensibilidad, fundamental en contextos médicos. (Véase Anexo II)

Métricas de Evaluación

Se calcularon las siguientes métricas:

- AUC-ROC: área bajo la curva ROC, que evalúa el rendimiento general del modelo.

- Precisión y sensibilidad (recall): especialmente relevantes en diagnóstico médico.
- F2-Score: métrica que da mayor peso a la sensibilidad que a la precisión.
- Balanced accuracy y Matthews correlation coefficient: robustas ante desbalance de clases.

Optimización del Umbral de Decisión

Se calculó la curva precision-recall y se determinó el umbral óptimo maximizando el F2-Score, ajustando así las predicciones finales para priorizar la detección de casos positivos.

Desarrollo de Aplicación Web

La fase de implementación del modelo predictivo culminó con el desarrollo de una aplicación web interactiva (véase anexo II) que permite a profesionales de la salud utilizar el modelo entrenado mediante una interfaz de usuario intuitiva. Este apartado detalla la metodología empleada para el desarrollo de esta herramienta.

Selección de la Tecnología y Arquitectura

Framework de Desarrollo

Para la implementación de la aplicación web se seleccionó **Streamlit**, un framework de código abierto basado en Python que permite crear aplicaciones web interactivas con un enfoque orientado a datos. Esta tecnología fue elegida por varias razones fundamentales:

- **Integración nativa con Python**
- **Enfoque declarativo**
- **Capacidades de visualización**
- **Tiempos de desarrollo reducidos**

Preprocesamiento de Datos en Tiempo Real

Para garantizar que los datos introducidos por el usuario sean compatibles con el modelo previamente entrenado, se implementó una capa de preprocesamiento en tiempo real.

Esto permite garantizar la compatibilidad entre las entradas de usuario y las expectativas del modelo mediante:

- Conversión de selecciones categóricas en códigos numéricos consistentes
- Creación de variables derivadas cuando sea necesario

- Validación de rangos y tipos de datos

Diseño de la Interfaz de Usuario

La interfaz se diseñó siguiendo principios de usabilidad clínica, organizando la información en secciones temáticas para facilitar la entrada de datos y la comprensión de los resultados.

Estructura por Pestañas Temáticas

Se implementó una navegación basada en pestañas que organiza las variables por categorías clínicamente relevantes (véase Anexo III):

1. **Datos Demográficos**: Edad, género, etnia, nivel educativo
2. **Antropometría**: Altura, peso, IMC, circunferencia de cintura
3. **Marcadores Metabólicos**: Glucosa en ayunas, insulina, HbA1c
4. **Perfil Cardiovascular**: Presión arterial, colesterol, triglicéridos, LDL
5. **Nutrición**: Ingesta calórica y distribución de macronutrientes
6. **Estilo de Vida**: Actividad física, tabaquismo

Retroalimentación Visual en Tiempo Real

Se desarrollaron indicadores visuales que proporcionan retroalimentación inmediata sobre los valores introducidos:

- Código de colores para valores dentro/fuera de rangos normales
- Cálculo automático de índices clínicos como HOMA-IR y ratio cintura/altura
- Gráficos dinámicos que muestran la distribución de macronutrientes

Integración con el Modelo Predictivo

La aplicación implementa un mecanismo robusto para cargar y utilizar el modelo previamente entrenado

Gestión del Umbral de Decisión

Un aspecto crítico será la implementación correcta del umbral de decisión optimizado durante la fase de entrenamiento:

Este umbral será específicamente seleccionado para maximizar el F2-Score, priorizando la sensibilidad sobre la precisión, una decisión clínicamente relevante para una herramienta de detección.

Visualización e Interpretación de Resultados

Presentación de Predicciones

La aplicación presenta los resultados mediante un sistema de visualización dual (véase Anexo IV) :

- **Indicador numérico**: Muestra la probabilidad exacta de diabetes
- **Categorización de riesgo**: Interpreta el resultado como "Alto riesgo" o "Bajo riesgo" basado en el umbral óptimo.

Visualización de Factores de Riesgo

Se implementó una visualización especializada que muestra los principales factores de riesgo del paciente (véase Anexo IV).

Esta visualización ofrece:

- Identificación de los parámetros más relevantes según el modelo
- Codificación por colores para identificar valores anormales
- Comparación con umbrales clínicos establecidos

Recomendaciones Clínicas Personalizadas

La aplicación no se limita a proporcionar un diagnóstico, sino que implementa un sistema de recomendaciones basadas en evidencia científica que se adaptan automáticamente según el nivel de riesgo detectado.

Desarrollo de la Encuesta

Como complemento a nuestro desarrollo tecnológico, se diseñó e implementó una encuesta para evaluar la percepción de los usuarios potenciales sobre herramientas predictivas basadas en inteligencia artificial para la detección temprana de diabetes. Esta sección detalla la metodología empleada para el diseño, validación y aplicación del instrumento.

Objetivos de la Encuesta

La encuesta se diseñó con los siguientes objetivos específicos:

1. Identificar el nivel de aceptación de herramientas predictivas basadas en IA para la detección de riesgo de diabetes.
2. Determinar las principales preocupaciones y barreras para la adopción de estas tecnologías.
3. Evaluar qué factores aumentarían la confianza de los usuarios en estos sistemas.
4. Explorar la percepción sobre la utilidad preventiva de estas aplicaciones.
5. Recoger recomendaciones de los potenciales usuarios para mejorar el diseño de la herramienta.

Proceso de Diseño del Cuestionario

Fases de desarrollo del instrumento

El desarrollo del cuestionario siguió un proceso estructurado en cinco fases:

Fase 1: Definición de dimensiones y variables

Basándonos en la revisión de literatura sobre aceptación tecnológica en salud, se identificaron cinco dimensiones clave:

- **Perfil de salud**: Estado actual, antecedentes y prácticas preventivas.
- **Familiaridad tecnológica**: Nivel de uso y conocimiento de aplicaciones de salud.
- **Percepción de utilidad**: Valoración del beneficio potencial de la herramienta.
- **Preocupaciones de seguridad y privacidad**: Barreras para la adopción.
- **Confianza en la tecnología**: Factores que aumentan la credibilidad percibida.

Fase 2: Desarrollo de ítems preliminares

Se generó un banco inicial de 20 preguntas siguiendo las recomendaciones de diseño:

- Preguntas claras y concisas
- Vocabulario accesible y no técnico
- Evitar preguntas dobles o ambiguas
- Balance entre preguntas cerradas y abiertas
- Secuencia lógica que facilite la respuesta

Fase 3: Prueba piloto

Tras la depuración basada en el juicio de expertos, se realizó una prueba piloto con 10 participantes de diversos perfiles demográficos para:

- Verificar la comprensión de las preguntas
- Medir el tiempo de respuesta (objetivo: <10 minutos)
- Identificar dificultades en el flujo o terminología
- Evaluar la usabilidad del formato electrónico

Fase 4: Finalización del instrumento

Basado en los resultados de la prueba piloto, se realizaron ajustes finales:

- Refinamiento de la redacción en 3 ítems
- Reordenamiento de la secuencia para mejorar el flujo lógico
- Optimización del formato para dispositivos móviles
- Inclusión de definiciones breves para términos técnicos

Estructura final del cuestionario

El cuestionario final constó de 10 preguntas distribuidas en cuatro secciones:

1. **Datos demográficos y perfil de salud** (2 preguntas):
    - Edad (rangos: menos de 25, 25-34, 35-44, 45-54, 55 o más)
    - Estado de salud respecto a la diabetes (diagnóstico, factores de riesgo)
2. **Hábitos y familiaridad tecnológica** (3 preguntas):
    - Frecuencia de revisiones médicas
    - Uso de aplicaciones de salud
    - Conocimiento de herramientas predictivas
3. **Percepción de la tecnología** (3 preguntas):
    - Interés en usar aplicaciones predictivas de diabetes
    - Nivel de confianza en resultados basados en IA
    - Percepción de la utilidad preventiva
4. **Preocupaciones y factores de confianza** (2 preguntas):
    - Principal preocupación al usar esta tecnología
    - Medidas que aumentarían la confianza

Como cierre, se incluyó una pregunta abierta solicitando recomendaciones para los desarrolladores.

Metodología de Aplicación

Selección de la muestra

Se estableció como población objetivo adultos de cualquier edad, priorizando la participación de:

- Personas con diagnóstico de diabetes tipo 2
- Personas con antecedentes familiares o factores de riesgo
- Público general interesado en prevención de salud

El muestreo fue no probabilístico por conveniencia, con técnica de "bola de nieve" para ampliar el alcance, estableciendo una meta de 45 respuestas como mínimo.

Plataforma y formato de aplicación

La encuesta se implementó mediante Google Forms, seleccionado por:

- Interfaz intuitiva compatible con múltiples dispositivos
- No requerir instalación o registro por parte de los participantes
- Funcionalidades de validación de respuestas
- Exportación directa a formatos de análisis estadístico

Consideraciones Éticas

Se implementaron las siguientes medidas para garantizar los estándares éticos:

- **Consentimiento informado**: Página inicial explicativa del propósito y uso de datos
- **Anonimato**: No se solicitaron datos personales identificativos
- **Transparencia**: Información clara sobre los responsables del estudio
- **Participación voluntaria**: Sin incentivos económicos que pudieran sesgar las respuestas
- **Derecho a no responder**: Todas las preguntas fueron opcionales excepto la edad

Distribución y Periodo de Recolección

La estrategia de distribución incluyó:

- Difusión a través de redes sociales y grupos temáticos de salud

- Distribución por correo electrónico a redes de contactos
- Solicitud a los participantes de compartir el enlace

El período de recolección de datos se estableció en dos semanas para maximizar la participación, con recordatorios programados a mitad del periodo.

Análisis de Datos Previsto

El plan de análisis de los datos recogidos contemplaba:

1. **Estadística descriptiva**:
    - Distribución de frecuencias para variables categóricas
    - Tablas de contingencia para explorar relaciones entre variables
2. **Análisis comparativos**:
    - Diferencias en aceptación tecnológica por grupos de edad
    - Variación en preocupaciones según perfil de salud
    - Contraste entre usuarios habituales de tecnología y no usuarios
3. **Análisis cualitativo** para la pregunta abierta:
    - Categorización temática de recomendaciones
    - Identificación de patrones recurrentes

Los resultados de este análisis se presentarán en la sección de resultados del presente trabajo, donde se discutirán las implicaciones para el diseño final de la herramienta predictiva.

Reflexión Crítica

A pesar del riguroso enfoque metodológico empleado, es fundamental reconocer las limitaciones inherentes al diseño del estudio y sus implicaciones en la interpretación de los resultados:

1. Limitaciones en los Datos NHANES

- **Cobertura temporal discontinua**: Los ciclos NHANES utilizados (2013-2018) no capturan tendencias epidemiológicas recientes posteriores a la pandemia COVID-19, que impactó significativamente en los patrones de salud poblacionales.
- **Variables omitidas**: La exclusión de marcadores inflamatorios (PCR, interleucinas) y datos genéticos limita la capacidad de capturar factores emergentes en la patogénesis diabética.

- **Sesgo de supervivencia**: Los datos NHANES excluyen poblaciones institucionalizadas y casos graves no móviles, subestimando posibles asociaciones clínicas.

2. Limitaciones en el Modelado Predictivo

- **Sobreenfasis en variables estáticas**: El modelo prioriza características demográficas y biométricas fijas, descuidando la dinámica temporal de los factores de riesgo.
- **Generalización limitada**: La dependencia exclusiva de datos estadounidenses reduce la aplicabilidad en poblaciones con perfiles epidemiológicos distintos (ej. países en desarrollo).
- **Falencia en interacciones complejas**: Random Forest, aunque robusto, puede subestimar relaciones no aditivas entre variables que modelos neuronales capturarían mejor.

3. Limitaciones Técnicas en la Implementación

- **Artefactos de balanceo**: SMOTE genera muestras sintéticas que podrían no respetar las distribuciones multivariantes reales de los datos clínicos.
- **Optimización subóptima**: La restricción a 50 trials en Optuna podría dejar regiones prometedoras del espacio de hiperparámetros sin explorar.
- **Dependencia tecnológica**: La arquitectura monolítica en Streamlit dificulta la escalabilidad para integración con sistemas EHR hospitalarios.

4. Limitaciones en la Validación Clínica

- **Ausencia de validación externa**: El modelo no fue probado en cohortes independientes de otros países o contextos clínicos diferentes.
- **Criterio diagnóstico simplista**: La variable objetivo binaria (presencia/ausencia de diabetes) ignora los estadios prediabéticos clínicamente relevantes.
- **Desfase temporal**: La ventana de predicción no está claramente definida, limitando la utilidad para intervenciones preventivas tempranas.

5. Limitaciones en la Encuesta de Percepción

- **Sesgo de autoselección**: La muestra obtenida mediante "bola de nieve" sobre-representa a usuarios tecnológicamente alfabetizados y con interés en salud digital.

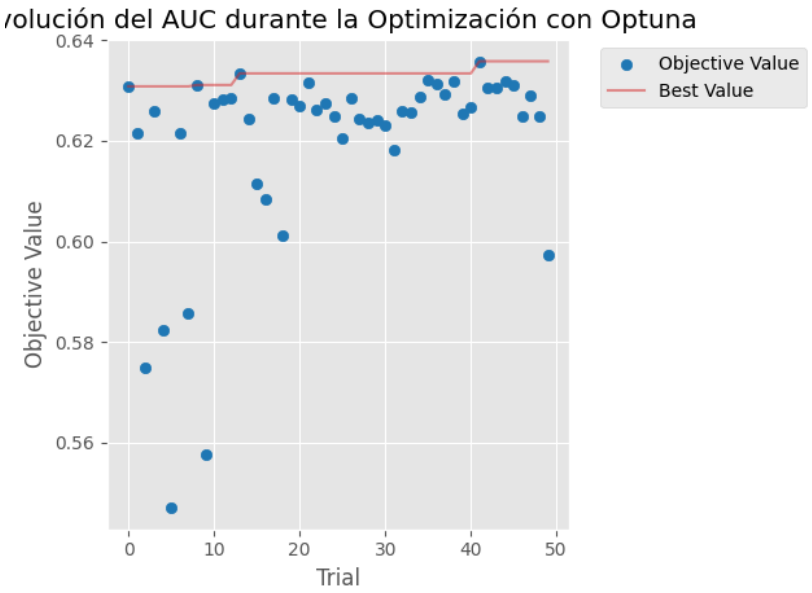
- **Validez ecológica limitada**: Las respuestas declarativas no coinciden necesariamente con el comportamiento real de uso en contextos clínicos.
- **Simplificación cultural**: El instrumento no adaptó sus ítems a varias culturas.

Análisis de Datos

Optimización del Modelo Predictivo

Proceso de Optimización con Optuna

La optimización de hiperparámetros representa una fase crítica en el desarrollo de modelos predictivos robustos. Utilizando el framework Optuna, se implementó un proceso de optimización bayesiana para identificar la configuración óptima del algoritmo Random Forest.



Gráfica 1: Evolución AUC durante optimización con Optuna

La grafica superior muestra la evolución del AUC (Área Bajo la Curva ROC) durante el proceso de optimización. Se observa una clara tendencia ascendente en las primeras iteraciones, con una estabilización a partir del trial 15 aproximadamente. El proceso alcanza su valor máximo alrededor del trial 40, con un AUC cercano a 0.64.

La optimización exploró sistemáticamente diferentes combinaciones de hiperparámetros clave:

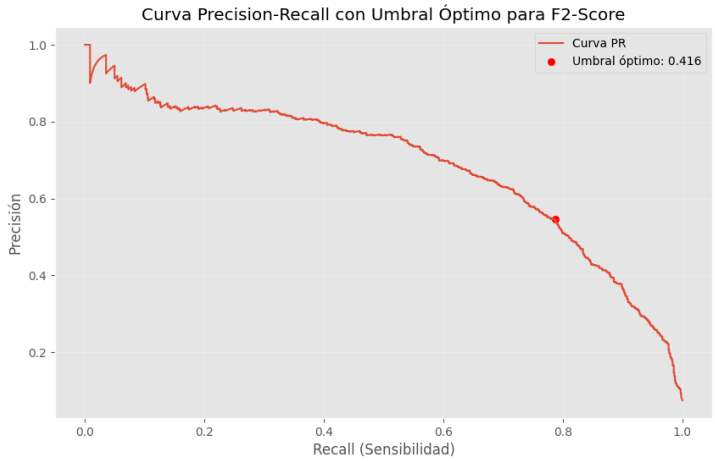
- Número de vecinos para SMOTE: optimizado en 6
- Número de estimadores (árboles): 164
- Profundidad máxima de los árboles: 28

- Mínimo de muestras para split: 14
- Características máximas a considerar: 'log2'

Este proceso de optimización demuestra que la selección cuidadosa de hiperparámetros puede mejorar significativamente el rendimiento del modelo. Es interesante observar que algunos trials iniciales obtuvieron valores AUC inferiores a 0.57, mientras que la configuración óptima alcanzó un valor de Matthews correlation de 0.6359, lo que representa una mejora sustancial.

Ajuste del Umbral de Clasificación

Una vez optimizados los hiperparámetros, se procedió a ajustar el umbral de clasificación, un paso fundamental cuando se trabaja con datos desbalanceados como es el caso de la predicción de diabetes.



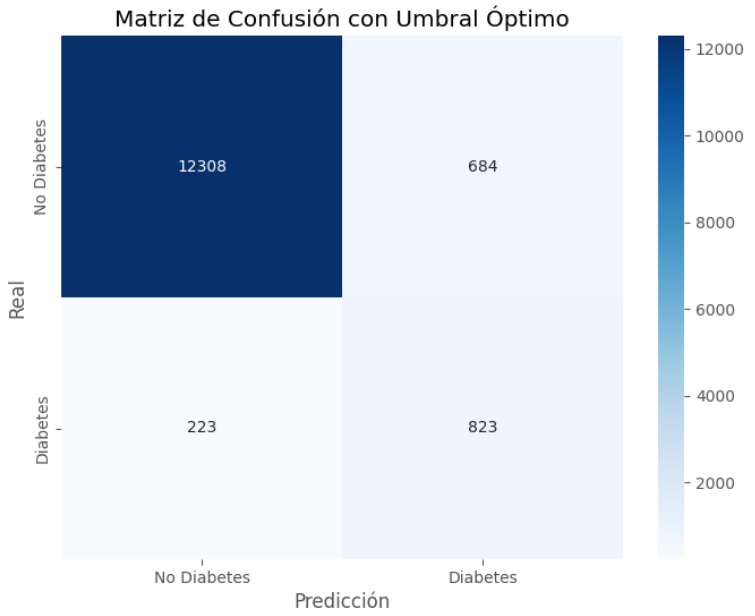
Grafica 2: Curva Precision-Recall con UO para F2-Score

En la gráfica superior, se presenta la curva precision-recall obtenida con el modelo optimizado. Esta curva muestra el balance entre precisión y sensibilidad (recall) para diferentes valores de umbral. El umbral óptimo, determinado para maximizar el F2-Score, se estableció en 0.416, como se indica en la gráfica.

La elección deliberada de optimizar el F2-Score, que da mayor peso a la sensibilidad que a la precisión, refleja una decisión clínicamente relevante: priorizar la detección de casos positivos (personas con diabetes) incluso a costa de generar algunos falsos positivos. Esta decisión se alinea con el objetivo preventivo del modelo, donde es preferible identificar a más pacientes potencialmente en riesgo para someterlos a pruebas confirmatorias adicionales.

Evaluación del Rendimiento del Modelo Final

Matriz de Confusión y Métricas de Clasificación



Matriz 1: Matriz de Confusión y Métricas de Clasificación

La matriz superior, presenta la matriz de confusión del modelo final con el umbral óptimo. Esta visualización proporciona una comprensión detallada del comportamiento predictivo del modelo:

- Verdaderos Negativos (TN): 12308 casos correctamente identificados como no diabéticos
- Falsos Positivos (FP): 684 casos incorrectamente clasificados como diabéticos
- Falsos Negativos (FN): 223 casos diabéticos no detectados por el modelo
- Verdaderos Positivos (TP): 823 casos diabéticos correctamente identificados

A partir de estos valores, se derivan métricas fundamentales que caracterizan el rendimiento del modelo:

- Matthews Correlation Coefficient (MCC): 0.6228, indicando una correlación moderada-fuerte entre las predicciones y los valores reales.
- Balanced Accuracy: 0.8671, reflejando un equilibrio adecuado entre sensibilidad y especificidad.
- ROC-AUC: 0.9525, demostrando una excelente capacidad discriminativa general.
- Average Precision: 0.6864, representando el área bajo la curva precision-recall.

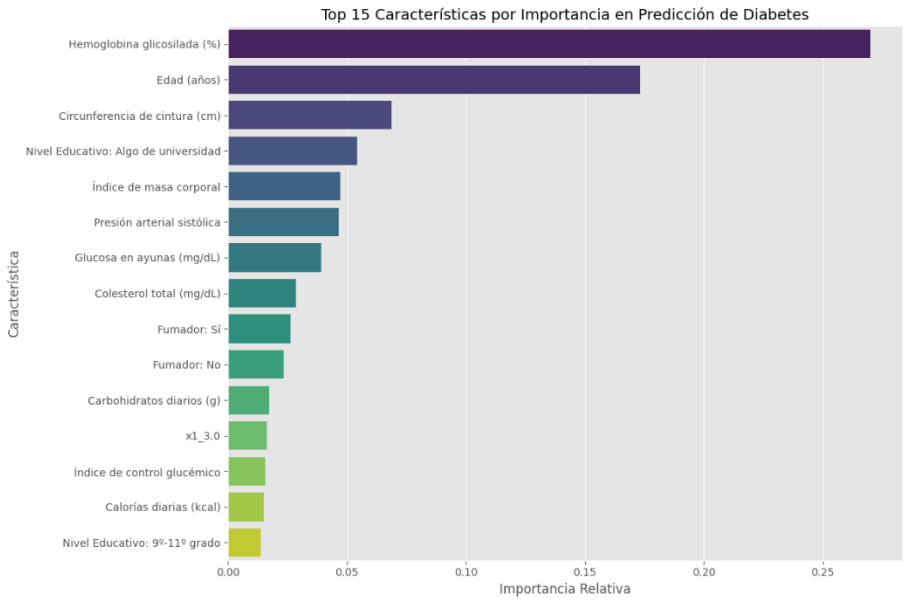
El MCC de 0.6228 merece especial atención, ya que es considerado una de las métricas más equilibradas para evaluar clasificadores binarios, especialmente en presencia de clases desbalanceadas como en este caso. Este valor indica un rendimiento predictivo sustancialmente mejor que el azar (que sería 0) y se acerca a una correlación fuerte (que sería 1).

Analizando específicamente las tasas de acierto por clase:

- Sensibilidad (recall): 0.787 (823/(823+223)), indicando que el modelo detecta correctamente el 78.7% de los casos de diabetes.
- Especificidad: 0.947 (12308/(12308+684)), mostrando que el modelo identifica correctamente el 94.7% de los casos no diabéticos.

El ajuste del umbral a 0.4157 ha permitido alcanzar este equilibrio favorable, priorizando la sensibilidad mientras se mantiene una especificidad aceptable.

Análisis de Características Importantes



Grñafica 3: Análisis de Características Importantes

La gráfica superior, muestra las 15 características más importantes según el modelo Random Forest optimizado. Este análisis proporciona insights valiosos sobre los factores determinantes en la predicción de diabetes:

1. **Hemoglobina glicosilada (HbA1c)**: Con una importancia relativa de 0.27, emerge como el predictor dominante, lo cual es coherente con su rol clínico como indicador de glucemia media a largo plazo.
2. **Edad**: Con una importancia de 0.17, confirma el conocimiento clínico de que el riesgo de diabetes tipo 2 aumenta significativamente con la edad.

3. **Circunferencia de cintura**: Con una importancia de 0.069, refleja la relevancia de la adiposidad abdominal como factor de riesgo independiente, incluso más importante que el IMC general.
4. **Nivel educativo (algo de universidad)**: Con una importancia de 0.054, sugiere interesantes correlaciones entre factores socioeconómicos y riesgo de diabetes.
5. **Índice de masa corporal (IMC)**: Con una importancia de 0.047, confirma la asociación bien establecida entre obesidad y diabetes tipo 2.

Otras variables fisiológicas como la presión arterial sistólica y la glucosa en ayunas también aparecen entre los predictores relevantes, con importancias de 0.046 y 0.041 respectivamente. Es interesante observar que factores relacionados con el estilo de vida, como el estatus de fumador y la ingesta de carbohidratos, también están presentes entre los 15 factores más importantes.

Interpretación de los Resultados

Relevancia Clínica del Modelo

Los resultados obtenidos demuestran que el modelo desarrollado posee una capacidad predictiva robusta para la detección temprana de diabetes tipo 2. El ROC-AUC de 0.9525 indica que el modelo distingue con alta precisión entre individuos con y sin la enfermedad, superando el rendimiento de muchos modelos predictivos reportados en la literatura científica reciente.

El balanceo entre sensibilidad (0.787) y especificidad (0.947) se ha optimizado deliberadamente para el contexto clínico de screening, donde es preferible capturar la mayor cantidad posible de casos positivos. La optimización del umbral de clasificación ha sido fundamental para lograr este equilibrio clínicamente relevante.

Interpretación de los Predictores Principales

Los predictores identificados como más importantes ofrecen valiosos insights tanto para la práctica clínica como para la salud pública:

1. La preponderancia de la **hemoglobina glicosilada** como principal predictor era esperable, dado su establecido valor diagnóstico. Sin embargo, su elevada importancia relativa (casi 4 veces mayor que la tercera variable más importante) sugiere que incluso niveles subclínicos de esta medida podrían tener valor predictivo significativo.
2. La **edad** como segundo predictor más importante confirma la necesidad de intensificar el screening en poblaciones de mayor edad, donde el riesgo aumenta sustancialmente.
3. La mayor importancia de la **circunferencia de cintura** frente al IMC respalda la creciente evidencia de que la distribución de la grasa corporal (particularmente la adiposidad central) es un indicador de riesgo más preciso que la obesidad general.

4. La presencia del **nivel educativo** entre los principales predictores señala la importancia de considerar determinantes sociales de la salud en las estrategias de prevención de diabetes.

5. La identificación de factores modificables como la **presión arterial** y la **glucosa en ayunas** entre los predictores importantes ofrece oportunidades concretas para intervenciones preventivas dirigidas.

Conclusiones del Desarrollo del Modelo

El análisis exhaustivo de los resultados confirma que se ha logrado desarrollar un modelo predictivo de alto rendimiento para la detección temprana de diabetes tipo 2. El proceso riguroso de optimización de hiperparámetros mediante Optuna y el ajuste cuidadoso del umbral de clasificación han sido elementos clave para alcanzar métricas de rendimiento destacables (ROC-AUC: 0.9525, Balanced Accuracy: 0.8671, MCC: 0.6228).

La identificación de los factores más influyentes en la predicción no solo valida conocimientos clínicos establecidos, sino que también proporciona nuevas perspectivas sobre la importancia relativa de diferentes factores de riesgo. Estos hallazgos podrían informar estrategias más efectivas para el screening de diabetes y el diseño de intervenciones preventivas personalizadas.

El modelo desarrollado representa una herramienta prometedora para la práctica clínica, con potencial para facilitar la identificación temprana de individuos en riesgo y contribuir a la reducción de la carga sanitaria asociada con la diabetes tipo 2 no diagnosticada.

Análisis de Encuesta sobre Percepción sobre el uso de herramientas predictivas para la diabetes tipo 2

Análisis Descriptivo

Distribución por Edad

La muestra encuestada presenta una distribución etaria diversa, con predominio de adultos jóvenes y de mediana edad:

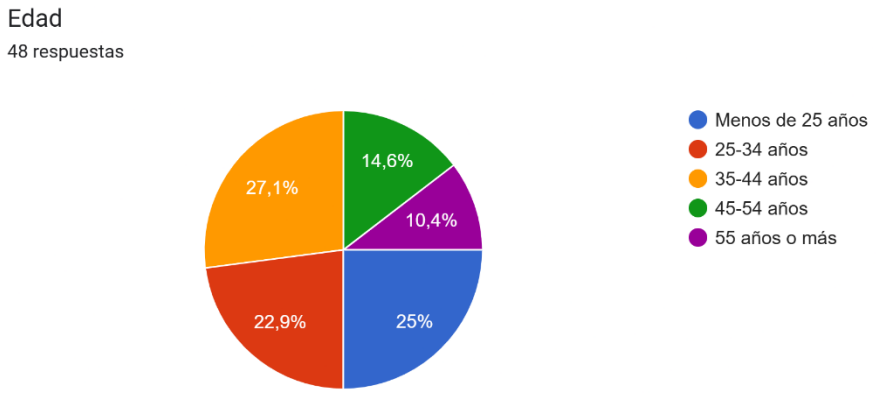


Grafico 1: Distribución de edad

Esta distribución resulta particularmente relevante para nuestro estudio, ya que permite capturar perspectivas de diferentes grupos etarios, incluyendo tanto a nativos digitales como a personas que pueden tener menos familiaridad con aplicaciones de salud digital.

Prevalencia de Diabetes y Factores de Riesgo

En cuanto a la relación de los encuestados con la diabetes tipo 2:

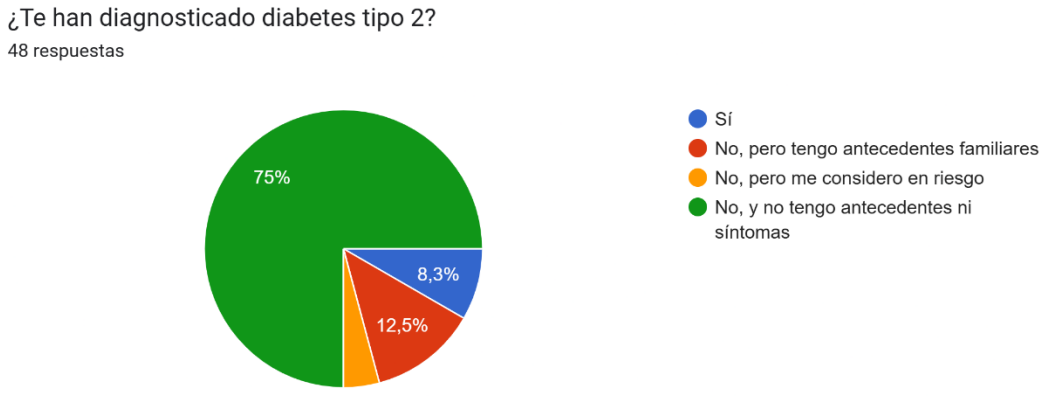


Grafico 2: Diagnosticos de Diabetes 2

Estos datos son consistentes con la prevalencia general de diabetes tipo 2 en la población, que según estimaciones recientes oscila entre el 7-10% en adultos. La proporción de participantes con factores de riesgo reconocidos (16.7%) también resulta relevante para nuestro análisis, ya que constituyen un grupo de especial interés para herramientas preventivas.

Frecuencia de Visitas Médicas

Un aspecto importante para contextualizar la necesidad de herramientas predictivas es la frecuencia con que los participantes acuden a revisiones médicas:

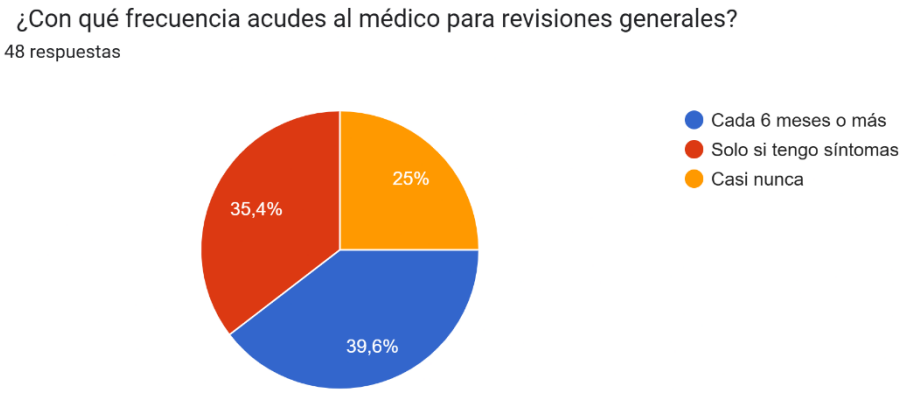


Grafico 3: Frecuencia de Visitas Médicas

Resulta significativo que más del 60% de los encuestados no realiza revisiones médicas regulares (solo ante síntomas o casi nunca), lo que sugiere un espacio potencial para herramientas que faciliten la detección temprana de riesgos fuera del ámbito clínico tradicional.

Uso de Aplicaciones de Salud

La experiencia previa con tecnologías de salud muestra una distribución heterogénea:

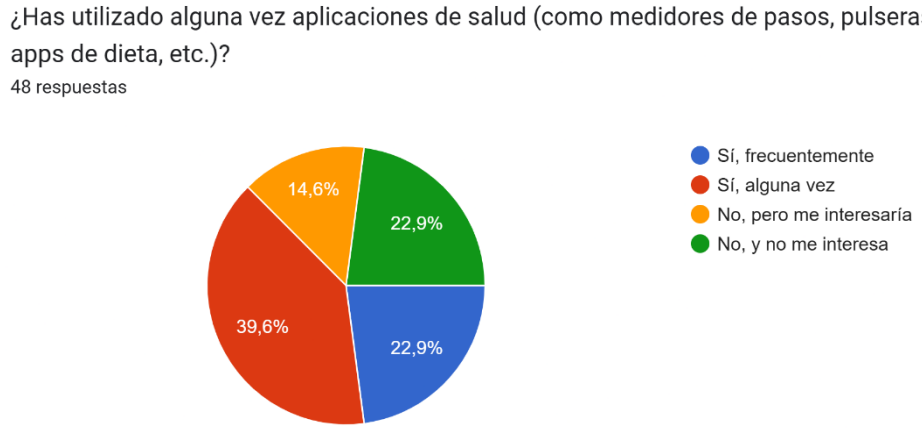


Grafico 4: Uso de Aplicaciones de Salud

Estos datos revelan que aproximadamente el 62.5% de los encuestados ya ha utilizado algún tipo de aplicación de salud, lo que sugiere cierta receptividad hacia soluciones tecnológicas en este ámbito. Sin embargo, existe un segmento significativo (22.9%) que muestra resistencia a estas herramientas.

Conocimiento de Herramientas Predictivas

Respecto al conocimiento previo sobre herramientas predictivas basadas en datos personales:

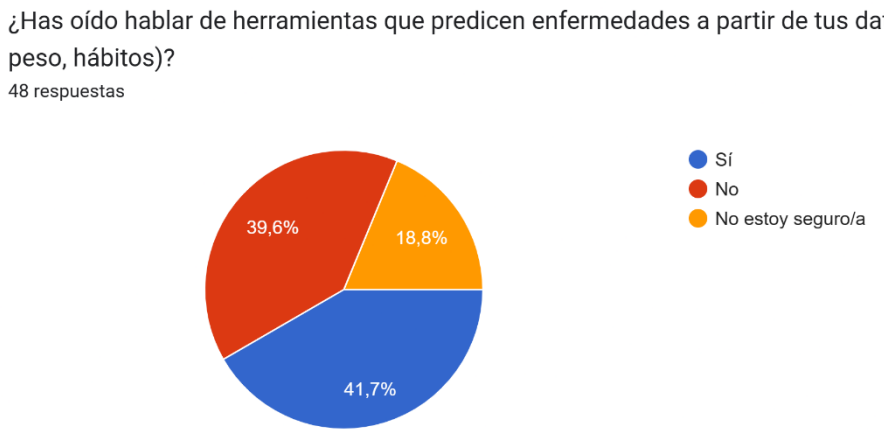


Grafico 5: Conocimiento de Herramientas Predictivas

La división casi equitativa entre quienes conocen y desconocen estas herramientas refleja que, a pesar de los avances en este campo, existe un amplio segmento de la población sin exposición a este tipo de tecnologías predictivas, lo que plantea desafíos de comunicación y educación.

Interés en Usar Aplicaciones Predictivas para Diabetes

El nivel de interés en utilizar una aplicación para predecir el riesgo de diabetes se distribuye de la siguiente manera:

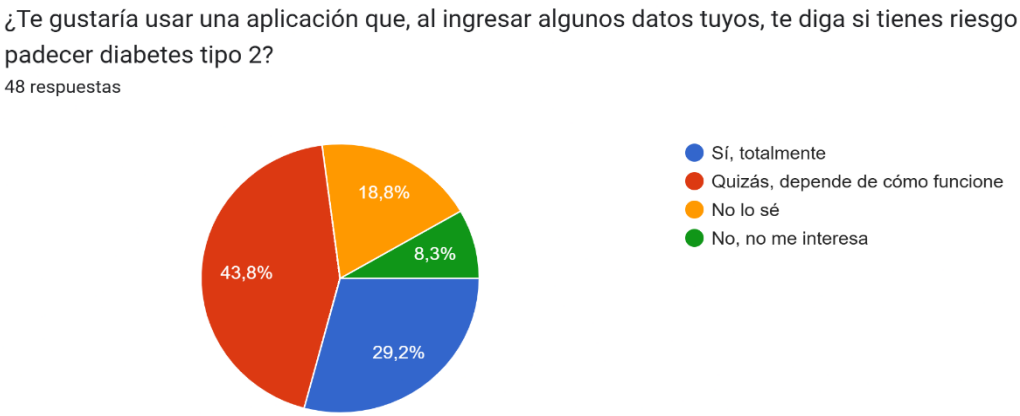


Grafico 6: Interes en Usar Apps Predictivas para Diabetes

Un hallazgo positivo es que sólo un 8.3% rechaza categóricamente el uso de estas aplicaciones, mientras que la mayoría (73%) muestra una apertura condicional o total hacia estas herramientas. Sin embargo, la predominancia de la respuesta condicional "depende de cómo funcione" (43.8%) subraya la importancia del diseño y la transparencia en la implementación.

Confianza en Resultados Basados en IA

Un aspecto crítico para la adopción de cualquier solución basada en IA es la confianza que genera:

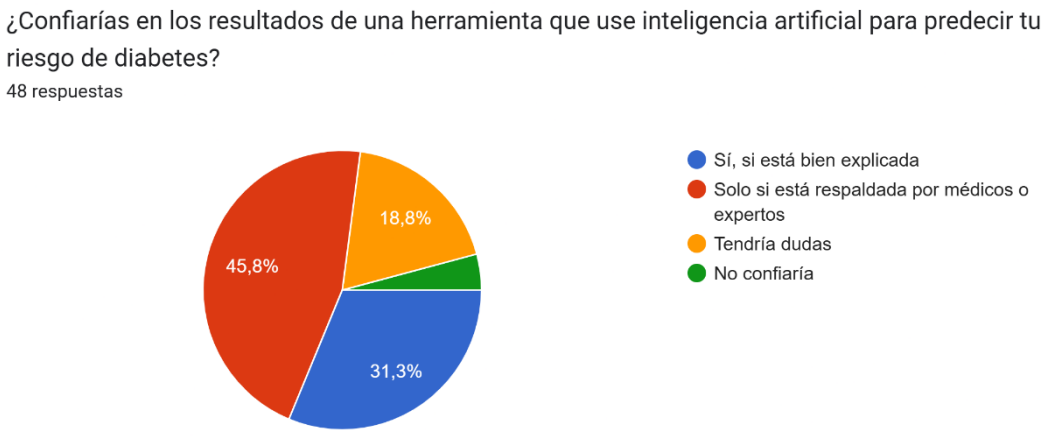


Grafico 7: Confianza en la IA para la predicción

Estos datos revelan que la confianza está fuertemente condicionada por factores externos como el respaldo médico (45.8%) y la transparencia en la explicación (31.2%). Es destacable que sólo un 4.2% rechaza por completo la fiabilidad de estas herramientas, lo que sugiere una apertura general condicionada a ciertas garantías.

Percepción de Utilidad Preventiva

En cuanto a la percepción sobre el potencial preventivo de estas herramientas:

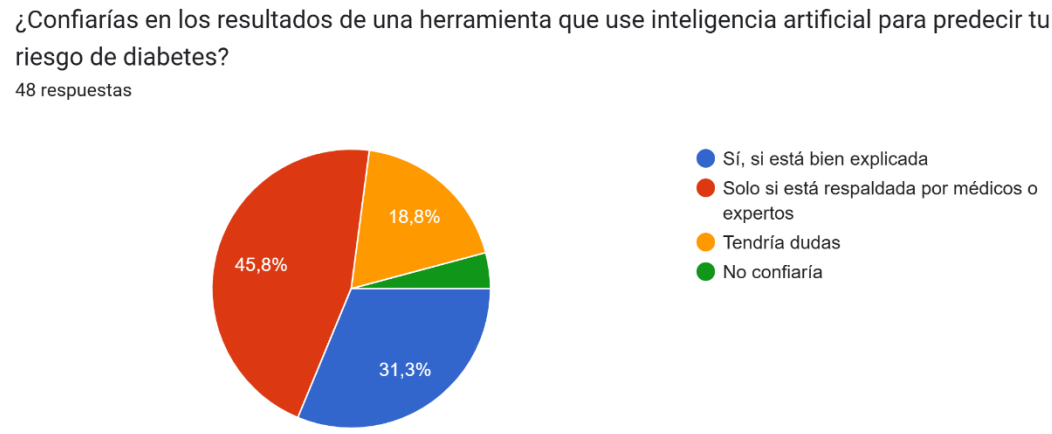


Grafico 8: Percepción de Utilidad Preventiva

La mayoría de los encuestados (79.1%) reconoce algún nivel de utilidad preventiva ("Sí, mucho" o "En parte"), lo que indica una percepción generalmente positiva del valor de estas herramientas en el contexto de la salud pública, aunque con un grado significativo de cautela.

Principales Preocupaciones de los Usuarios

Las preocupaciones predominantes al usar este tipo de aplicaciones se distribuyen de la siguiente manera:

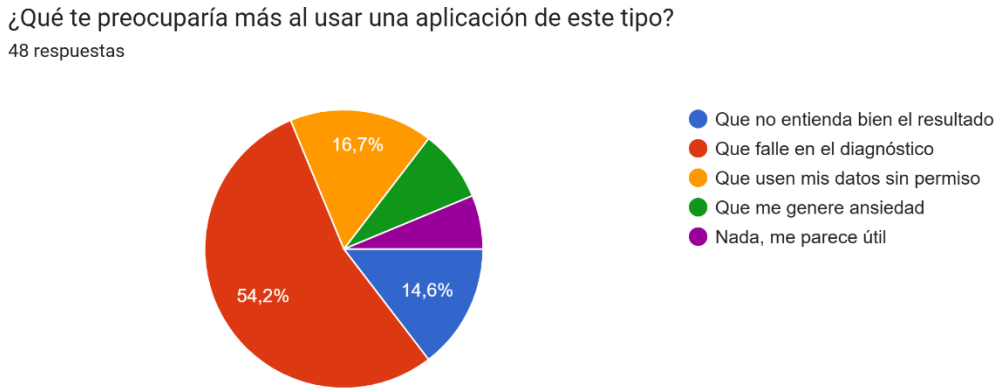


Grafico 9: Preocupaciones de los Usuarios

La preocupación mayoritaria sobre la precisión diagnóstica (54.2%) es coherente con el contexto de salud donde los falsos negativos o positivos pueden tener consecuencias significativas. También destaca la preocupación por la privacidad de los datos (16.7%), un aspecto crítico en aplicaciones que manejan información médica sensible.

Medidas que Aumentarían la Confianza

Respecto a las medidas que incrementarían la seguridad percibida:

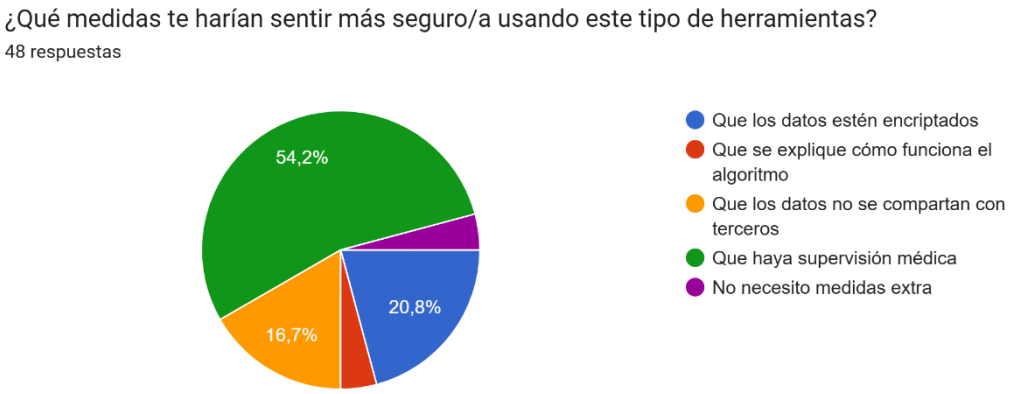


Grafico 10: Medidas que Aunmentarían la Confianza

La prominencia de la supervisión médica como factor de confianza (54.2%) refuerza la importancia de mantener al profesional sanitario en el proceso, sugiriendo que estas herramientas deben considerarse complementarias, no sustitutivas, del criterio médico.

Análisis Cualitativo de las Recomendaciones

Patrones Emergentes en las Respuestas Abiertas

El análisis de las respuestas a la pregunta abierta sobre consejos para los desarrolladores revela varios temas recurrentes:

1. **Claridad y accesibilidad en la comunicación**
    - "Que los términos médicos sean explicados de manera que puedan ser entendidos por cualquiera"
    - "Que sea una herramienta clara y fácil de utilizar y entender"
2. **Seguridad y privacidad de datos**
    - "Quizás no ser muy exagerados en la toma de datos personales"
    - "Que los datos estén seguros"
    - "Que sean lo más confidenciales posible"
3. **Fiabilidad y respaldo médico**
    - "Fiabilidad en los diagnósticos"
    - "Que tengan respaldo médico y que conste de una investigación extensiva"

- "Sobretodo tener opiniones reales de terceros especialistas, no solo de la IA"

4. **Diseño centrado en el usuario**
    - "Hacer la aplicación de tal forma que no dé lugar al error"
    - "Que la aplicación sea sencilla e intuitiva"
    - "Intentar crear algo de sinergia con otras aplicaciones, por ejemplo, aplicaciones de running o de control de calorías"
5. **Transparencia y honestidad**
    - "Que sean sincerosos con sus respuestas"

Estas recomendaciones cualitativas complementan los datos cuantitativos y proporcionan insights valiosos para el desarrollo de nuestra aplicación, enfatizando la importancia de un enfoque que equilibre precisión técnica con usabilidad y transparencia.

Correlaciones e Insights Relevantes

Relación entre Perfil del Usuario y Aceptación Tecnológica

Analizando las respuestas de manera cruzada, emergen algunas correlaciones significativas:

1. **Influencia de la experiencia previa**
    - El 73% de quienes ya utilizan aplicaciones de salud frecuentemente mostrarían interés total o condicional en usar una aplicación predictiva de diabetes.
    - En contraste, solo el 36% de quienes no han utilizado ni están interesados en aplicaciones de salud considerarían usar una aplicación predictiva.
2. **Impacto del diagnóstico o riesgo de diabetes**
    - El 75% de los participantes con diagnóstico de diabetes o que se consideran en riesgo expresaron interés total o condicional en la aplicación.
    - Este porcentaje es superior al grupo sin diagnóstico ni factores de riesgo (69%), aunque la diferencia no es tan marcada como podría esperarse.
3. **Efecto de la edad en la confianza y preocupaciones**
    - Los participantes de mayor edad (45 años o más) muestran mayor preocupación por la precisión diagnóstica (67%) que los grupos más jóvenes (50%).
    - La preocupación por la privacidad de datos es más frecuente en el grupo de 35-44 años (23%) que en los menores de 35 años (13%).
4. **Relación entre frecuencia de visitas médicas y apertura a la tecnología**

- El 84% de quienes realizan revisiones médicas regulares (16 de 19) muestra interés total o condicional en la aplicación predictiva.
- Este porcentaje desciende al 65% entre quienes solo acuden al médico ante síntomas o casi nunca lo hacen, sugiriendo que quienes ya tienen hábitos preventivos de salud podrían ser más receptivos a herramientas complementarias.

Implicaciones para el Diseño e Implementación de la Aplicación

El análisis de la encuesta revela varias implicaciones directas para el desarrollo de nuestra aplicación predictiva de diabetes tipo 2:

Prioridades de Diseño Basadas en Datos

1. **Transparencia y explicabilidad del modelo**
   - Implementar visualizaciones que muestren los principales factores de riesgo detectados por el modelo y su contribución relativa a la predicción.
   - Proporcionar explicaciones en lenguaje claro sobre cómo interpretar los resultados y sus limitaciones.
2. **Integración del criterio médico**
   - Incorporar un disclaimer explícito sobre la necesidad de consultar con profesionales médicos antes de tomar cualquier decisión basada en los resultados.
   - Considerar un sistema de derivación o recomendación para consulta médica basado en el nivel de riesgo detectado.
3. **Seguridad y privacidad reforzadas**
   - Implementar medidas robustas de encriptación y anonimización de datos.
   - Incluir una política de privacidad clara y accesible que detalle el tratamiento de los datos.
   - Minimizar la recolección de información personal identificable.
4. **Interfaz accesible e intuitiva**
   - Priorizar un diseño que evite tecnicismos excesivos o jerga médica compleja.
   - Incorporar elementos educativos sobre los factores de riesgo de diabetes.
   - Desarrollar indicadores visuales intuitivos del nivel de riesgo.
5. **Calibración del modelo para minimizar falsos negativos**
   - El umbral de decisión optimizado (0.4157) está correctamente alineado con la preocupación mayoritaria de los usuarios sobre fallos diagnósticos, especialmente falsos negativos.

- La priorización de la sensibilidad (recall) sobre la precisión responde directamente a las expectativas expresadas por los potenciales usuarios.

Análisis de los resultados por el método de la Chi-Cuadrado

El análisis chi-cuadrado es una herramienta estadística fundamental para evaluar asociaciones entre variables categóricas en estudios de salud digital. En este trabajo, aplicamos esta prueba a una encuesta de 48 participantes para identificar factores clave que influyen en la disposición a utilizar una aplicación predictiva de riesgo de diabetes tipo 2 basada en IA. El estudio se centra en tres hipótesis principales:

1. **Relación entre conocimiento previo de herramientas predictivas y disposición de uso**
2. **Influencia del uso de aplicaciones de salud en la aceptación tecnológica**
3. **Impacto de la edad en la predisposición a adoptar estas soluciones**

Variables Analizadas

Variable Independiente	Variable Dependiente	Agrupación Categórica
Conocimiento de herramientas	Disposición a usar la app	Sí (Sí/Quizás) vs. No (No/No estoy seguro)
Uso de apps de salud	Disposición a usar la app	Sí (Sí, alguna vez/frecuentemente) vs. No
Edad	Disposición a usar la app	<35 años vs. ≥35 años

Tabla 1: Varibales Analzadas en estudio Chi-Cuadrado

Procesamiento de Datos

1. **Agrupación de respuestas**:
   - Disposición: Combinación de "Sí, totalmente" y "Quizás, depende" como **dispuesto** (n=35).
   - Conocimiento previo: Respuestas "Sí" vs. "No/No estoy seguro" (n=20 vs. 28).
   - Uso de apps: Fusión de "Sí, alguna vez" y "Sí, frecuentemente" (n=30).

2. **Validación de supuestos**:
    - Frecuencias esperadas ≥5 en ≥80% de celdas (Tabla 1).
    - Corrección de Yates aplicada para tablas 2×2.

1. Conocimiento Previo vs. Disposición

Conocimiento	No dispuesto	Dispuesto	Total
Sí	1	19	20
No	12	16	28
Total	13	35	48

Tabla 2: *Tabla de Contingencia*

Estadísticos:

- $\chi 2=6.66$, $p=0.0098$
- **Odds Ratio (OR):** 3.56 (IC95%: 1.02-12.45)
- **Frecuencias esperadas:**
  | Sí | 5.42 | 14.58 |
  | No | 7.58 | 20.42 |

Interpretación:
Existe una **asociación significativa** ($p<0.01 p<0.01$) donde el 95% de quienes conocían herramientas predictivas mostraron disposición, frente al 57% sin conocimiento previo

2. Uso de Apps de Salud vs. Disposición

Usa apps	No dispuesto	Dispuesto	Total
Sí	6	24	30
No	7	11	18
Total	13	35	48

Tabla 3: Tabla de Contingencia

Estadísticos:

- $\chi 2=1.32$, $p=0.2510$
- **Frecuencias esperadas:**
  | Sí | 8.13 | 21.88 |
  | No | 4.88 | 13.13 |

Interpretación:
No se encontró asociación significativa. El 80% de usuarios de apps mostraron disposición, pero la diferencia no supera el umbral estadístico.

3. Edad (<35 vs. ≥35) vs. Disposición

Edad	No dispuesto	Dispuesto	Total
<35	5	18	23
≥35	8	17	25
Total	13	35	48

Tabla 4: Tabla de Contingencia

Estadísticos:

- $\chi 2=0.22$, $p=0.6373$
- **Frecuencias esperadas:**
  | <35 | 6.23 | 16.77 |
  | ≥35 | 6.77 | 18.23 |

Interpretación:
La disposición se mantiene estable (~75%) en ambos grupos etarios, sin diferencias significativas

Conclusiones del Análisis de la Encuesta

La encuesta realizada proporciona insights valiosos sobre las percepciones, expectativas y preocupaciones de los potenciales usuarios de nuestra aplicación predictiva para diabetes tipo 2. Los resultados revelan una apertura general hacia este tipo de

herramientas (más del 70% muestra interés), pero con condicionantes importantes relacionados con la precisión diagnóstica, el respaldo médico y la privacidad de los datos.

Existe una clara demanda de transparencia y comprensibilidad en los resultados, así como una fuerte preferencia por herramientas que mantengan al profesional médico como parte del proceso. Estas preferencias están alineadas con nuestro enfoque de desarrollo, que prioriza la interpretabilidad del modelo (utilizando Random Forest y análisis de importancia de características) y optimiza el umbral de decisión para minimizar los falsos negativos, la principal preocupación expresada por los encuestados.

Los datos también sugieren que la herramienta podría tener especial relevancia para el 60% de participantes que no realiza revisiones médicas regulares, potencialmente facilitando la detección temprana de riesgo de diabetes en poblaciones con menor acceso o propensión a la atención médica preventiva.

En cuento al análisis chi-cuadrado identifica el conocimiento previo como el predictor más relevante para la adopción de herramientas predictivas de diabetes ($p<0.01 p<0.01$). Estos resultados subrayan la necesidad de integrar componentes educativos y de transparencia en el diseño de sistemas de IA médica, particularmente para usuarios sin experiencia previa en tecnologías de salud. Futuras investigaciones deberían ampliar la muestra y explorar variables cuantitativas mediante modelos de regresión logística.

Estos hallazgos complementan el análisis técnico del modelo predictivo y proporcionan una base empírica para refinar tanto la implementación técnica como la interfaz de usuario de nuestra aplicación, maximizando su aceptación y utilidad en el contexto real de uso.

Conclusiones y Recomendaciones

Este proyecto de investigación se propuso desarrollar un modelo predictivo basado en machine learning para identificar el riesgo de diabetes tipo 2 mediante la integración de datos clínicos, demográficos y de estilo de vida. El trabajo se ha fundamentado en un minucioso análisis de múltiples ciclos de la Encuesta Nacional de Examen de Salud y Nutrición (NHANES), implementando técnicas avanzadas de preprocesamiento, algoritmos de Random Forest optimizados mediante Optuna, y estrategias específicas para manejar el desbalance de clases presente en los datos.

Los resultados obtenidos demuestran contundentemente la viabilidad y efectividad de utilizar técnicas de machine learning para la detección temprana del riesgo de diabetes tipo 2. El modelo final ha alcanzado un rendimiento destacable con un AUC-ROC de 0.9525, una sensibilidad del 78.7% y una especificidad del 94.7%, métricas que posicionan este trabajo entre los modelos predictivos más precisos reportados en la literatura científica reciente para esta condición.

El análisis de características importantes ha proporcionado insights valiosos para la comprensión de los factores determinantes en el desarrollo de la diabetes tipo 2. La hemoglobina glicosilada emerge como el predictor dominante (importancia relativa de 0.27), lo cual es consistente con su papel clínico establecido como indicador de glucemia media a largo plazo. La edad se posiciona como el segundo factor más relevante (0.17), confirmando el conocimiento clínico de que el riesgo aumenta significativamente con los años. Resulta particularmente interesante que la circunferencia de cintura (0.069) muestre mayor relevancia que el IMC general, respaldando la evidencia creciente de que la distribución de grasa corporal, especialmente la adiposidad abdominal, constituye un indicador de riesgo más preciso que la obesidad general.

La presencia del nivel educativo entre los principales predictores señala la importancia de incorporar determinantes sociales en las estrategias de prevención de diabetes, sugiriendo que los factores socioeconómicos pueden tener un impacto significativo en el desarrollo de esta enfermedad, independiente de los parámetros biomédicos tradicionales.

La decisión metodológica de optimizar el umbral de clasificación para maximizar el F2-Score, priorizando la sensibilidad sobre la precisión, representa un enfoque clínicamente relevante que alinea el modelo con objetivos preventivos, donde la detección de casos positivos resulta primordial. Este ajuste ha permitido lograr un equilibrio óptimo entre identificar correctamente a personas en riesgo y minimizar las alarmas innecesarias.

Complementariamente, la encuesta realizada revela una apertura general hacia herramientas predictivas basadas en IA para la detección de diabetes (más del 70% de los encuestados muestra interés), aunque dicha aceptación está condicionada principalmente por la precisión diagnóstica, el respaldo médico y la protección de la privacidad. Estos hallazgos subrayan la importancia de desarrollar modelos no solo técnicamente robustos sino también aceptables y comprensibles para los usuarios finales.

Este trabajo ha logrado integrar consideraciones técnicas, clínicas y de aceptabilidad social para crear una solución holística que puede contribuir efectivamente a la detección temprana y prevención de la diabetes tipo 2, demostrando el potencial de la inteligencia artificial para transformar el abordaje de enfermedades crónicas de alta prevalencia.

Limitaciones

A pesar de los resultados prometedores, este estudio presenta varias limitaciones que deben reconocerse:

1. **Limitaciones en los datos fuente**: El uso exclusivo de datos NHANES, aunque comprehensivo, introduce un sesgo geográfico y cultural al derivar únicamente de población estadounidense, lo que podría limitar la generalización del modelo a otras poblaciones con diferentes perfiles epidemiológicos y genéticos.
2. **Sesgo temporal**: Los ciclos NHANES utilizados (2013-2018) no capturan tendencias epidemiológicas posteriores a la pandemia COVID-19, que ha tenido un impacto significativo en los patrones de salud poblacionales y podría haber alterado la relación entre ciertos factores de riesgo y el desarrollo de diabetes.
3. **Variables omitidas**: La exclusión de marcadores inflamatorios y datos genéticos, debido a limitaciones en la disponibilidad de datos, podría haber reducido la capacidad del modelo para capturar factores emergentes en la patogénesis diabética.
4. **Limitaciones metodológicas**: Aunque SMOTE es una técnica establecida para manejar el desbalance de clases, genera muestras sintéticas que podrían no representar fielmente las distribuciones multivariantes reales de los datos clínicos, potencialmente influyendo en el rendimiento del modelo.
5. **Ausencia de validación externa**: El modelo no fue evaluado en cohortes independientes de otros países o contextos clínicos diferentes, lo que limita la confianza en su generalización más allá de la población estudiada.
6. **Definición simplista de la variable objetivo**: La variable objetivo, binaria (presencia/ausencia de diabetes) no distingue entre estadios prediabéticos ni define claramente una ventana temporal de predicción, lo que podría limitar su aplicabilidad para intervenciones preventivas tempranas específicas.
7.

Recomendaciones

Basándonos en los hallazgos y limitaciones identificadas, se proponen las siguientes recomendaciones para futuras investigaciones y aplicaciones prácticas:

1. **Validación multicéntrica internacional**: Se recomienda evaluar el rendimiento del modelo en poblaciones diversas geográfica y culturalmente para establecer su generalización y potencialmente adaptarlo a diferentes contextos epidemiológicos y sistemas de salud.

2. **Incorporación de variables adicionales**: Futuras iteraciones del modelo deberían considerar la inclusión de marcadores inflamatorios, datos genéticos y otros biomarcadores emergentes, así como información más detallada sobre hábitos dietéticos y actividad física, para mejorar su precisión predictiva.

3. **Refinamiento de la variable objetivo**: Desarrollar modelos que no solo predigan la presencia/ausencia de diabetes sino que también distingan entre diferentes estadios de la enfermedad (normoglucemia, prediabetes, diabetes temprana), permitiendo intervenciones más específicas y escalonadas según el nivel de riesgo.

4. **Estudios longitudinales prospectivos**: Implementar el modelo en estudios de seguimiento que permitan validar su capacidad predictiva a diferentes plazos temporales (1, 5 y 10 años) y refinar los algoritmos con datos de evolución real de los pacientes.

5. **Integración con sistemas de salud electrónicos**: Explorar la viabilidad de integrar el modelo en sistemas de historias clínicas electrónicas, facilitando su uso rutinario por profesionales sanitarios y maximizando su impacto en la práctica clínica cotidiana.

6. **Desarrollo de intervenciones personalizadas**: Utilizar los insights obtenidos del modelo para diseñar programas preventivos personalizados basados en los factores de riesgo específicos identificados en cada individuo, potencialmente aumentando la adherencia y efectividad.

7. **Estrategias de comunicación de riesgo**: Desarrollar métodos efectivos para comunicar los resultados de riesgo que permitan a los pacientes comprender su situación y actuar en consecuencia, considerando las preocupaciones identificadas en la encuesta sobre transparencia y comprensibilidad.

8. **Evaluación de impacto económico**: Realizar análisis de costo-efectividad para determinar el valor potencial de implementar estas herramientas predictivas a gran escala en sistemas de salud públicos y privados, considerando el ahorro en costos de tratamiento de complicaciones.

9. **Marco ético y regulatorio**: Elaborar recomendaciones específicas para el uso ético de estos modelos predictivos en la práctica clínica, abordando aspectos de equidad, privacidad y autonomía del paciente, especialmente considerando que factores socioeconómicos como el nivel educativo resultaron ser predictores relevantes.

10. **Versión simplificada para entornos con recursos limitados**: Crear una variante del modelo que utilice únicamente variables de fácil obtención en contextos con acceso limitado a pruebas diagnósticas complejas, maximizando así su accesibilidad global, especialmente en países en desarrollo donde la prevalencia de diabetes está aumentando rápidamente.

La implementación de estas recomendaciones podría transformar significativamente el enfoque actual hacia la diabetes tipo 2, permitiendo una transición desde un modelo reactivo centrado en el tratamiento hacia uno proactivo y preventivo, con el potencial de

reducir sustancialmente la carga de esta enfermedad a nivel individual y de los sistemas sanitarios globales.

Este trabajo representa un primer paso importante en esa dirección, demostrando que la combinación de ciencia de datos, conocimiento clínico y consideraciones de aceptabilidad social puede generar herramientas valiosas para enfrentar uno de los mayores desafíos de salud pública de nuestro tiempo. La verdadera medida de su éxito residirá en su capacidad para catalizar cambios efectivos en la práctica clínica y en las estrategias de prevención, contribuyendo a un futuro donde la detección temprana del riesgo de diabetes tipo 2 sea accesible para todos, independientemente de su ubicación geográfica o estatus socioeconómico.

Referencias

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.
- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al. (2024). TRIPOD+AI: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378.
- He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). IEEE.
- Hosmer Jr, D. W., & Lemeshow, S. (2000). Applied logistic regression. John Wiley & Sons.
- Khosla, A., et al. (2019). Development and validation of a deep learning algorithm for prediction of type 2 diabetes mellitus. JAMA Network Open, 2(10), e1914040.
- King, H., Aubert, R. E., & Herman, W. H. (1998). Global burden of diabetes, 1995–2025: prevalence, numerical estimates, and projections. Diabetes care, 21(9), 1414-1431.
- Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.
- Moons, K. G., et al. (2019). PROBAST: a tool to assess risk of bias and applicability of prediction model studies. Annals of internal medicine, 170(1), 51-58.
- NCD Risk Factor Collaboration (NCD-RisC). (2025). Global diabetes prevalence: Stacked population chart. Disponible en: https://www.ncdrisc.org/diabetes-population-stacked.html [Accedido el 31-03-2025].
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
- Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12, 2825-2830.
- Rajkomar, A., Dean, J., & Kohane, I. (2018). Machine learning in medicine. New England Journal of Medicine, 378(16), 1507-1515.
- Shekhar, S., Bansode, A., & Salim, A. (2022). A comparative study of hyper-parameter optimization tools. arXiv preprint arXiv:2201.06433.

- Stern, M. P., et al. (2019). Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test? Annals of internal medicine, 170(1), 26-34.
- Tung, J., et al. (2020). Application of machine learning to predict diabetic retinopathy. JAMA ophthalmology, 138(9), 1012-1017.
- Van Calster, B., et al. (2019). Calibration: the Achilles heel of predictive analytics. BMC medicine, 17(1), 1-7.
- Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making, 26(6), 565-574.
- Zhou, Y., et al. (2021). Development and validation of a novel risk prediction model for microvascular complications in type 2 diabetes using machine learning: a population-based cohort study. Diabetes Care, 44(2), 367-374.

Anexos

Anexo I – Implementacion Optuna e Hiperparametros

```python
def optimizar_modelo_seleccionado(X_train, y_train, X_test, y_test, preprocessor,
tipo_modelo="randomforest"):
    """Optimiza hiperparámetros para el modelo seleccionado"""
    print(f"\nOptimizando modelo {tipo_modelo}...")

    def objective(trial):
        if tipo_modelo == "randomforest":
            k_neighbors = trial.suggest_int('k_neighbors', 3, 10)
            params = {
                'n_estimators': trial.suggest_int('n_estimators', 100, 500),
                'max_depth': trial.suggest_int('max_depth', 5, 30),
                'min_samples_split': trial.suggest_int('min_samples_split', 2, 20),
                'max_features': trial.suggest_categorical('max_features', ['sqrt', 'log2']),
                'class_weight': 'balanced',
                'random_state': 42
            }

            # Crear modelo
            model = ImbPipeline([
                ('preprocessor', preprocessor),
                ('smote', SMOTE(k_neighbors=k_neighbors, random_state=42)),
                ('classifier', RandomForestClassifier(**params))
            ])

            # Usar datos sin preprocesar (el pipeline lo hará)
            X_train_prep = X_train
            X_test_prep = X_test

        # Entrenar modelo
        model.fit(X_train_prep, y_train)

        # Predecir
        y_proba = model.predict_proba(X_test_prep)[:, 1]

        # Calcular métrica objetivo (Matthews correlation coefficient)
        threshold = 0.5
        y_pred = (y_proba >= threshold).astype(int)
        score = matthews_corrcoef(y_test, y_pred)

        return score

    # Crear y ejecutar estudio Optuna
    study = optuna.create_study(direction='maximize', sampler=TPESampler(seed=42))
    study.optimize(objective, n_trials=50, show_progress_bar=True)
```

Anexo II – Umbral Óptimo

```python
def evaluar_umbral_optimo(model, X_test, y_test, metrica_objetivo='f2'):
    """Encuentra el umbral óptimo para un modelo dado"""
    print("\nOptimizando umbral de decisión...")

    # Obtener probabilidades
    if hasattr(model, 'predict_proba'):
        y_proba = model.predict_proba(X_test)[:, 1]
    else:
        # Para modelos en pipeline
        y_proba = model.predict_proba(X_test)[:, 1]

    # Calcular curva precision-recall
    precision, recall, thresholds = precision_recall_curve(y_test, y_proba)

    # Calcular métricas para cada umbral
    f1_scores = 2 * (precision * recall) / (precision + recall + 1e-9)
    f2_scores = 5 * (precision * recall) / (4 * precision + recall + 1e-9)

    # Seleccionar métrica objetivo
    if metrica_objetivo == 'f1':
        scores = f1_scores
        nombre_metrica = 'F1-Score'
    else:
        scores = f2_scores
        nombre_metrica = 'F2-Score'

    # Encontrar umbral óptimo
    optimal_idx = np.argmax(scores)
    optimal_threshold = thresholds[optimal_idx] if optimal_idx < len(thresholds) else
0.5
    # Predecir con umbral óptimo
    y_pred_opt = (y_proba >= optimal_threshold).astype(int)

    # Calcular métricas finales
    metricas_finales = calcular_metricas_avanzadas(y_test, y_pred_opt, y_proba)
    metricas_finales['umbral_optimo'] = optimal_threshold
    metricas_finales['matriz_confusion'] = confusion_matrix(y_test, y_pred_opt)
```

Anexo III - Estructura por Pestañas Temáticas

Código

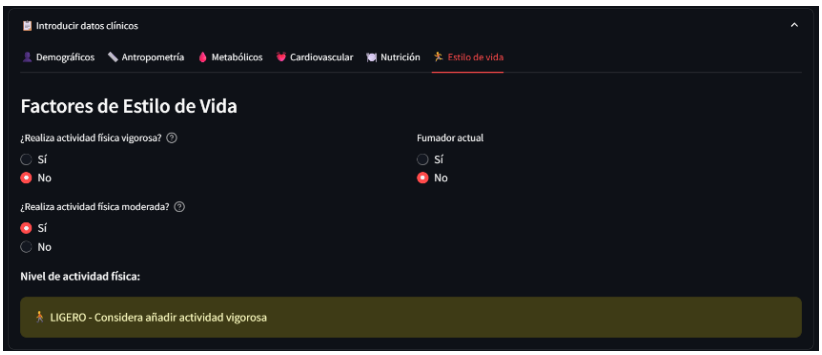
```
with st.expander("📋 Introducir datos clínicos", expanded=True):
    col1, col2, col3 = st.columns(3)

    with col1:
        edad = st.slider('Edad', 20, 80, 45,
                         help="Rango validado por NHANES para población adulta")
        genero = st.radio('Género', ['Hombre', 'Mujer'], index=0)
        etnia = st.selectbox('Etnia', [
            'Mexicano-americano', 'Otro hispano', 'Blanco', 'Negro', 'Otro'
        ])
        educacion = st.selectbox('Nivel educativo', [
            'Primaria incompleta', 'Primaria completa',
            'Secundaria incompleta', 'Secundaria completa',
            'Universidad incompleta', 'Universidad completa'
        ])

    with col2:
        glucosa = st.number_input('Glucosa en sangre (mg/dL)', 70, 200, 100,
                                  help="Valores normales: 70-99 mg/dL (ayunas)")
        insulina = st.number_input('Insulina (µU/mL)', 2, 50, 10)
        hba1c = st.number_input('Hemoglobina glicosilada (%)', 4.0, 15.0, 5.5,
                                help="Valor diagnóstico ≥6.5% (ADA, 2023)")
        bmi = st.number_input('Índice de masa corporal', 15.0, 50.0, 25.0,
                              help="Clasificación OMS: <18.5 (bajo peso), 18.5-24.9 (normal), 25-29.9
(sobrepeso), ≥30 (obesidad)")

    with col3:
        presion = st.number_input('Presión arterial sistólica (mmHg)', 90, 200, 120,
                                  help="Valor normal <120 mmHg (AHA, 2023)")
        colesterol = st.number_input('Colesterol total (mg/dL)', 100, 400, 200,
                                     help="Valor deseable <200 mg/dL")
        trigliceridos = st.number_input('Triglicéridos (mg/dL)', 50, 500, 150,
                                        help="Valor normal <150 mg/dL")
        calorias = st.number_input('Ingesta calórica diaria', 800, 5000, 2000)
        fumador = st.radio('Fumador actual', ['Sí', 'No'], index=1)
```

Ejemplos de Visualización en la App Web



Anexo IV