UNIVERSIDADES VALLADOLID

MÁSTER UNIVERSITARIO EN INVESTIGACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES

DETECCIÓN PRECOZ DE LA DIABETES Y PREDICCIÓN DE COMPLICACIONES MEDIANTE TÉCNICAS DE *MACHINE LEARNING*

AUTOR: SANTIAGO GALÁN MAROTO

TUTORA: **D. ISABEL DE LA TORRE DÍEZ**

Junio de 2025

RESUMEN

La diabetes mellitus es una enfermedad crónica que afecta la capacidad del cuerpo para regular la glucosa en sangre, generando complicaciones graves si no se detecta precozmente. Existen dos tipos principales: el tipo 1 (DM1), de origen autoinmune y frecuente en jóvenes, y el tipo 2 (DM2), relacionado con factores de riesgo como obesidad, sedentarismo y envejecimiento. La DM2 representa más del 90% de los casos, y a menudo permanece sin diagnóstico durante años.

Ante este reto sanitario global, el presente TFM (Trabajo Fin de Máster) propone un enfoque basado en técnicas de *Machine Learning* (ML) para mejorar la detección temprana y la clasificación automática del tipo de diabetes. Se utilizan algoritmos supervisados aplicados a un conjunto de datos poblacional procedente del BRFSS 2015, que incluye variables demográficas, clínicas y conductuales. Los objetivos incluyen: (1) clasificar a los individuos en no diabéticos, DM1 o DM2; (2) comparar el rendimiento de distintos modelos como *Random Forest*, XGBoost o redes neuronales; (3) identificar variables clave para la predicción, y (4) desarrollo de una plataforma web interactiva para la detección temprana de riesgo de diabetes mediante inteligencia artificial.

El trabajo analiza múltiples métricas de evaluación, como la precisión, el F1-score y el AUC-ROC, y utiliza herramientas como SHAP y LIME para dotar de aplicabilidad a los modelos. Se demuestra que ciertos factores como el índice de masa corporal, la salud mental, la actividad física y la dificultad para caminar tienen alto valor predictivo. Asimismo, se refuerza la utilidad clínica del enfoque multiclase frente al binario habitual.

Esta investigación aporta evidencia de que la inteligencia artificial, aplicada de manera ética y transparente, puede ser una aliada en el diagnóstico precoz de enfermedades crónicas como la diabetes, contribuyendo a una medicina más preventiva, personalizada y sostenible.

Palabras clave: diabetes, modelos, tipo, datos, salud, variables, predicción, clasificación, machine, aprendizaje

ABSTRACT

Diabetes mellitus is a chronic disease that affects the body's ability to regulate blood glucose, leading to serious complications if not detected early. There are two main types: type 1 (T1D), of autoimmune origin and common in young people, and type 2 (T2D), related to risk factors such as obesity, sedentary lifestyle, and aging. T2D accounts for more than 90% of cases and often remains undiagnosed for years.

In response to this global health challenge, this Master's Thesis proposes an approach based on Machine Learning (ML) techniques to improve early detection and automatic classification of diabetes types. Supervised algorithms are applied to a population dataset from BRFSS 2015, which includes demographic, clinical, and behavioral variables. The objectives are: (1) to classify individuals as non-diabetic, T1D, or T2D; (2) to compare the performance of different models such as Random Forest, XGBoost, or neural networks; (3) to identify key predictive variables; and (4) to develop an interactive web platform for the early detection of diabetes risk using artificial intelligence.

The work analyzes multiple evaluation metrics, such as accuracy, F1-score, and AUC-ROC, and uses tools like SHAP and LIME to provide explainability to the models. It is shown that certain factors, such as body mass index, mental health, physical activity, and difficulty walking, have high predictive value. Furthermore, the clinical utility of the multiclass approach is reinforced compared to the usual binary approach.

This research provides evidence that artificial intelligence, when applied ethically and transparently, can be an ally in the early diagnosis of chronic diseases such as diabetes, contributing to more preventive, personalized, and sustainable medicine.

Keywords: diabetes, models, type, data, health, variables, prediction, classification, machine learning, artificial intelligence

ÍNDICE GENERAL

ÍNDICE ILUSTRACIÓN

ÍNDICE GRÁFICOS

ÍNDICE TABLAS

ÍNDICE ABREVIATURAS

ADA	Asociación Americana de Diabetes (*American Diabetes Association*)
AUC	Área bajo la curva (*Area Under the Curve*)
AUC-ROC	Área bajo la curva ROC (*Receiver Operating Characteristic Curve*)
BRFSS	Sistema de Vigilancia de Factores de Riesgo del Comportamiento (Behavioral Risk Factor Surveillance System)
CDSS	Sistema de apoyo a la decisión clínica (Clinical Decision Support System)
EDA	Análisis exploratorio de datos (Exploratory Data Analysis)
F1	Medida F1 (media armónica entre precisión y exhaustividad) (F1 *Score)*
FINDRISC	Cuestionario Finlandés de Riesgo de Diabetes (*Finnish Diabetes Risk Score*)
IMC	Índice de masa corporal (*Body Mass Index*)
KNN	k vecinos más cercanos (*k-Nearest Neighbors*)
LDA	Asignación de Dirichlet latente (*Latent Dirichlet Allocation*)
LIME	Explicaciones locales interpretables e independientes del modelo (*Local Interpretable Model-Agnostic Explanations*)
ML	Aprendizaje automático (*Machine Learning*)
MLP	Perceptrón multicapa (*Multi-Layer Perceptron*)
OMS	Organización Mundial de la Salud (*World Health Organization*)
PCA	Análisis de componentes principales (*Principal Component Analysis*)
PIDD	*Pima Indians Diabetes Dataset*
PRISMA	Elementos de Reporte Preferidos para Revisiones Sistemáticas y metaanálisis *(Preferred Reporting Items for Systematic Reviews and Meta-Analyses*)
RFE	Eliminación recursiva de características (*Recursive Feature Elimination*)
ROC	Curva característica operativa del receptor (*Receiver Operating Characteristic*)
SHAP	Explicaciones aditivas de Shapley (*SHapley Additive exPlanations*)
SMOTE	Técnica de sobremuestreo de la clase minoritaria (*Synthetic Minority Over-sampling Technique*)
SVM	Máquina de vectores de soporte (*Support Vector Machine*)
TFM	Trabajo de Fin de Máster
UCI	University of California, Irvine
P.ej.	Por ejemplo.

I. INTRODUCCIÓN

La diabetes mellitus es una afección crónica de carácter metabólico que compromete la capacidad del cuerpo para regular la glucosa en sangre. Esta alteración sostenida en el equilibrio glucémico desencadena una cascada de procesos fisiopatológicos que, de no ser controlados a tiempo, puede desembocar en complicaciones graves que afectan diversos sistemas del organismo [1]. Aunque se trata de una enfermedad comúnmente asociada a la endocrinología, su impacto trasciende lo meramente clínico, al influir en aspectos sociales, económicos y de calidad de vida de millones de personas en todo el mundo [16][19].

Esta enfermedad se presenta, principalmente, en dos formas clínicas diferenciadas: la diabetes tipo 1 (DM1) y la diabetes tipo 2 (DM2). Ambas comparten un denominador común —la hiperglucemia persistente— pero difieren sustancialmente en sus causas, perfil epidemiológico, curso evolutivo y tratamiento [17][18].

La diabetes tipo 1 se origina por un fallo autoinmune. El propio sistema inmunológico del individuo destruye progresivamente las células beta del páncreas encargadas de producir insulina. Este proceso, en muchos casos silencioso hasta su eclosión clínica, desemboca en una dependencia absoluta de insulina exógena desde el momento del diagnóstico [3][4]. Es una forma menos frecuente, ya que representa entre el 5% y el 10% del total de casos diagnosticados, pero su inicio suele producirse en etapas tempranas de la vida, condicionando fuertemente la salud física, emocional y social del paciente durante décadas [16][20].

Por su parte, la diabetes tipo 2 representa el escenario más común dentro del espectro de esta enfermedad. Constituye más del 90% de los casos y se desarrolla, en la mayoría de las ocasiones, como resultado de una combinación de resistencia periférica a la insulina y una secreción pancreática insuficiente para compensarla [3][16]. Su aparición suele estar asociada a múltiples factores modificables, como el exceso de peso corporal, la vida sedentaria, una alimentación desbalanceada y el envejecimiento poblacional. A estos se suman otros elementos de riesgo como los antecedentes familiares, el estrés crónico o la calidad del sueño.

A diferencia de la DM1, el tipo 2 puede pasar desapercibida durante años. En sus primeras fases, los síntomas son sutiles o incluso inexistentes, lo que retrasa el diagnóstico y favorece la progresión hacia formas más graves [1]. Este hecho reviste una especial gravedad si se considera que muchas de sus complicaciones —como la retinopatía, la nefropatía o la neuropatía— comienzan a desarrollarse de forma silenciosa incluso antes del diagnóstico clínico. Además, la DM2 está estrechamente relacionada con enfermedades cardiovasculares, hipertensión arterial y dislipemia, conformando un conjunto de comorbilidades que elevan considerablemente el riesgo de eventos adversos y muerte prematura [17][19].

Según estimaciones recientes de la Federación Internacional de Diabetes, en 2019 más de 460 millones de personas convivían con algún tipo de diabetes. Se prevé que esta cifra se incremente hasta alcanzar los 700 millones para el año 2045, lo que representa un crecimiento cercano al 51% [1]. Estos datos no solo evidencian una tendencia ascendente sostenida, sino también el carácter epidémico de la enfermedad. Lo más alarmante es que una proporción significativa de

quienes padecen diabetes, especialmente tipo 2, no han sido aún diagnosticados. Esta franja "invisible" de la población impide la actuación temprana y perpetúa el ciclo de complicaciones evitables.

La carga económica asociada a la diabetes es abrumadora. Los costes directos incluyen consultas médicas, hospitalizaciones, medicación, pruebas de laboratorio y atención especializada. A esto se suman los costes indirectos derivados de la pérdida de productividad, el ausentismo laboral y la discapacidad a largo plazo. En países como España, el gasto anual atribuible a la diabetes supera los 15.000 millones de dólares, lo que posiciona al país entre los diez con mayor desembolso por esta causa [1].

El desafío que representa esta enfermedad ha impulsado, en las últimas décadas, el desarrollo de nuevas estrategias de abordaje, no solo desde el tratamiento farmacológico sino también desde la prevención y el diagnóstico precoz. En este sentido, la transformación digital del ámbito sanitario y el avance de las tecnologías de datos han abierto un abanico de posibilidades que hace apenas unos años resultaban impensables [17]. Las técnicas de aprendizaje automático (*machine learning*) permiten analizar grandes volúmenes de información clínica y de estilo de vida, identificando patrones sutiles que podrían anticipar la aparición de la enfermedad incluso antes de que se cumplan los criterios clínicos clásicos [2].

El valor de estas herramientas radica en su capacidad para considerar múltiples variables de forma simultánea, modelar relaciones no lineales entre los factores de riesgo y, en muchos casos, mejorar la precisión predictiva frente a los modelos estadísticos tradicionales. En el contexto de la diabetes, esto es especialmente relevante, dado que su origen y evolución están mediados por una compleja interacción entre factores genéticos, conductuales, ambientales y socioeconómicos.

Sin embargo, la aplicación de estas tecnologías no está exenta de desafíos. Uno de los principales obstáculos es la heterogeneidad metodológica de los estudios existentes: distintos algoritmos, conjuntos de datos variados, métricas dispares y poblaciones de estudio con características muy diferentes dificultan la generalización de los resultados. Además, la complejidad de algunos modelos puede dificultar su interpretación y aplicación en entornos clínicos reales, donde la transparencia y la explicabilidad del sistema son tan necesarias como su precisión.

Otro aspecto que merece atención es la limitada diferenciación que muchos estudios hacen entre los tipos de diabetes. En la mayoría de los trabajos basados en inteligencia artificial, la diabetes se aborda como una entidad única o se restringe el análisis exclusivamente a la DM2, obviando la necesidad de estrategias diferenciadas que contemplen también la DM1. Esta aproximación reduccionista no solo limita la utilidad clínica del modelo, sino que puede generar conclusiones incompletas o equívocas.

Por tanto, abordar el estudio conjunto —pero diferenciado— de la diabetes tipo 1 y tipo 2 desde una perspectiva predictiva es una necesidad pendiente dentro del campo de la medicina computacional. La posibilidad de identificar con precisión a individuos en riesgo de desarrollar cualquiera de las dos formas clínicas, mediante el análisis de factores observables y medibles, podría marcar un punto de inflexión en la lucha contra esta enfermedad.

1.1. Objetivos del trabajo

El objetivo principal objetivo de este TFM es explotar el potencial del aprendizaje automático (*ML)* en la detección, clasificación y comprensión de la diabetes mellitus en sus dos formas clínicas más frecuentes: tipo 1 y tipo 2. Frente al avance imparable de esta enfermedad a escala global y al alto porcentaje de casos no diagnosticados, resulta imprescindible desarrollar nuevas estrategias de identificación temprana que superen las limitaciones de los métodos tradicionales.

El trabajo parte de una premisa fundamental: el análisis de datos de salud puede proporcionar un conocimiento más preciso y profundo del perfil de riesgo de cada persona. La inteligencia artificial, aplicada con responsabilidad y rigor, permite descubrir patrones latentes que escapan al análisis clínico convencional. Este enfoque puede ser clave para reducir el número de diagnósticos tardíos, personalizar las intervenciones médicas y mejorar los resultados en salud pública.

Para alcanzar estos fines u objetivos, se establecen los siguientes objetivos específicos:

Objetivo 1: Clasificación automática de la condición diabética

Uno de los principales retos en el abordaje clínico de la diabetes mellitus radica en la correcta identificación del tipo de enfermedad que presenta un individuo. Tradicionalmente, el diagnóstico se realiza mediante un conjunto de pruebas bioquímicas (como glucemia en ayunas, hemoglobina glicosilada, prueba de tolerancia a la glucosa o anticuerpos autoinmunes) complementadas por la historia clínica del paciente. No obstante, este procedimiento, aunque eficaz, puede resultar costoso, lento y no siempre accesible, especialmente en sistemas de salud con recursos limitados. Además, existe una franja considerable de pacientes cuyo diagnóstico es tardío o erróneo, particularmente en estadios iniciales donde los síntomas pueden ser poco específicos o incluso inexistentes [3].

Se propone como primer objetivo desarrollar un sistema automatizado, basado en algoritmos de *ML*, que sea capaz de clasificar a las personas según su condición diabética: no diabéticos, diabéticos tipo 1 y diabéticos tipo 2. A diferencia de muchos estudios previos que se centran únicamente en distinguir entre personas sanas y con diabetes tipo 2, este enfoque incorpora también la detección de la diabetes tipo 1, ampliando el espectro de utilidad clínica del modelo y aumentando su valor predictivo real.

Para ello, se empleará un conjunto diverso de variables que no requiere análisis de laboratorio costosos o especializados. Entre estas se incluyen factores demográficos (edad, sexo, nivel educativo), clínicos (estado general de salud, presión arterial, índice de masa corporal), de comportamiento (tabaquismo, consumo de alcohol, nivel de actividad física, ingesta de frutas y verduras) y psicosociales (calidad del sueño, estado emocional, niveles de estrés). La hipótesis que subyace a este planteamiento es que existen patrones multivariantes identificables entre estos factores que permiten inferir, con una precisión razonable, la condición diabética de una persona, incluso sin disponer de indicadores bioquímicos directos [4].

Este objetivo no busca sustituir al juicio clínico ni a las herramientas de diagnóstico tradicionales, sino ofrecer un mecanismo de cribado inicial, especialmente útil en contextos

comunitarios, atención primaria o territorios con acceso limitado a tecnología médica avanzada. Una herramienta de este tipo permitiría, *p.ej.,* priorizar a ciertos pacientes para pruebas más específicas, diseñar campañas de prevención dirigidas a poblaciones con alto riesgo y aumentar la eficiencia en programas de detección precoz.

Desde un punto de vista técnico, la tarea se enmarca en el aprendizaje supervisado con clasificación multiclase. El modelo será entrenado con datos etiquetados, donde la variable objetivo indica si el paciente es no diabético, diabético tipo 1 o diabético tipo 2. El éxito de esta clasificación se medirá mediante métricas habituales como precisión general, matriz de confusión, sensibilidad (*recall*), especificidad y área bajo la curva ROC para cada clase. De este modo, no solo se evaluará la capacidad del modelo para acertar, sino también para evitar falsos positivos o negativos, que podrían tener implicaciones clínicas importantes [5].

Cabe destacar que la diferenciación entre diabetes tipo 1 y tipo 2 mediante datos no invasivos constituye un reto técnico de alta complejidad. La DM1 suele aparecer en edades más tempranas, se asocia con pérdida de peso rápida, cetosis y necesidad de insulina desde el inicio, mientras que la DM2 tiene un curso más insidioso y está ligada a la obesidad y el sedentarismo. Sin embargo, existen situaciones intermedias o atípicas que pueden confundir incluso a los profesionales sanitarios. Por ello, un sistema de clasificación automático bien entrenado y validado podría actuar como herramienta complementaria de apoyo al diagnóstico diferencial, mejorando los tiempos de atención y reduciendo la incertidumbre diagnóstica.

La aplicación de modelos predictivos en este contexto ha demostrado resultados prometedores. Estudios recientes han implementado clasificadores para discriminar tipos de diabetes a partir de datos estructurados, logrando valores de precisión superiores al 80% con algoritmos como *Random Forest* o SVM [6]. Estos avances sugieren que una solución automatizada, si es correctamente construida e interpretada, podría integrarse en flujos clínicos reales con beneficios tangibles[1].

En suma, este primer objetivo busca validar la viabilidad y utilidad de una herramienta computacional capaz de identificar automáticamente el tipo de diabetes que presenta un individuo —o su ausencia— a partir de datos rutinarios, sin requerir una carga diagnóstica invasiva. Esto supone un paso hacia una medicina más accesible, preventiva y centrada en el dato.

Objetivo 2: Análisis comparativo del rendimiento de múltiples algoritmos de aprendizaje automático

El segundo objetivo de este trabajo tiene como propósito fundamental evaluar comparativamente el rendimiento de distintos algoritmos de aprendizaje automático en la tarea de clasificación automatizada de la condición diabética, diferenciando entre sujetos no diabéticos, diabéticos tipo 1 y tipo 2. Dado que no existe un único modelo que funcione de manera óptima para todos los contextos y conjuntos de datos, es esencial realizar una evaluación

[1] La clasificación multiclase en entornos médicos no solo enfrenta desafíos técnicos, sino también éticos, al influir en decisiones clínicas. Por eso, la interpretabilidad del modelo y la comprensión de sus errores son casi tan relevantes como su precisión. Integrar explicaciones transparentes, como las proporcionadas por técnicas tipo SHAP o LIME, puede ayudar a que el personal médico confíe en las decisiones algorítmicas y las utilice como apoyo, no como sustituto

empírica que permita identificar cuáles son las metodologías más eficaces, robustas y adaptables a un escenario clínico real.

El aprendizaje automático, como subcampo de la inteligencia artificial, ofrece un abanico muy amplio de técnicas, cada una con sus propias fortalezas, debilidades, supuestos subyacentes y requisitos computacionales. Algunos modelos presentan un rendimiento sobresaliente en términos de precisión, pero adolecen de interpretabilidad. Otros, aunque más sencillos, ofrecen una mayor capacidad de explicación, aspecto fundamental cuando se trata de implementarlos en entornos clínicos donde la transparencia es un requisito indispensable para su aceptación y uso por parte del personal sanitario [7].

En este sentido, se propone comparar el comportamiento de una serie de algoritmos ampliamente utilizados en la literatura científica para tareas de clasificación médica. Entre ellos se incluyen:

Tabla 1:Tipología de modelos de Machine Learning aplicables al diagnóstico de diabetes

Tipo de modelo	**Ejemplos representativos**
Modelos clásicos y explicativos	Regresión logística, Árboles de decisión, *Naive Bayes*
Modelos de ensamblado y optimización	Random Forest, Gradient Boosting, XGBoost, LightGBM
Modelos de proximidad y separación	Máquinas de vectores de soporte (SVM), k-*Nearest Neighbors* (k-NN)
Modelos avanzados no lineales	Redes neuronales artificiales (ANN), Perceptrón multicapa (MLP)

Fuente 1: Elaboración propia.

Cada uno de estos modelos será entrenado sobre el mismo conjunto de datos, aplicando técnicas de validación cruzada para reducir el riesgo de sobreajuste (*overfitting*) y evaluar la capacidad de generalización. Para garantizar la equidad en la comparación, se seguirán procedimientos normalizados de ajuste de hiperparámetros (mediante búsqueda en rejilla o algoritmos *bayesianos*) y se emplearán métricas estandarizadas para cuantificar su rendimiento.

Las métricas utilizadas incluirán:

Tabla 2: Principales métricas para evaluar modelos de clasificación multiclase en el diagnóstico de diabetes.

Métrica	**Descripción**
Precisión global (*Accuracy*)	Porcentaje de aciertos sobre el total de muestras.
Sensibilidad (*Recall*)	Capacidad del modelo para detectar correctamente a los individuos positivos en cada clase (tipo 1 o tipo 2).
Especificidad	Proporción de verdaderos negativos correctamente identificados.
F1-score	Media armónica entre precisión y sensibilidad.
AUC-ROC	Área bajo la curva de características operativas del receptor, evaluada por clase.
Matriz de confusión	Análisis detallado de errores de clasificación entre las tres categorías.

Fuente 2: Elaboración propia.

Además del análisis cuantitativo, se incorporará una evaluación cualitativa y visual, mediante curvas ROC, gráficos de dispersión entre probabilidades predichas y reales, y mapas de calor para explorar la estructura de errores. Esta visualización permitirá entender no solo qué modelos son más precisos, sino también cómo y por qué se equivocan, aspecto crucial en medicina para evitar errores diagnósticos que puedan tener consecuencias graves.

El objetivo no es simplemente determinar el modelo con mayor precisión, sino también valorar otros aspectos de igual importancia: la robustez frente al ruido, la **c**apacidad de aprendizaje en datos con clases desbalanceadas**,** el tiempo de entrenamiento y predicción**,** y, sobre todo, la interpretabilidad de las decisiones del modelo**.**

P. ej., modelos como *Random Forest* o *Gradient Boosting* tienden a ofrecer un buen equilibrio entre precisión y explicabilidad gracias a sus mecanismos internos de evaluación de variables. En cambio, las redes neuronales pueden ser más potentes en términos predictivos, pero requieren mecanismos adicionales para interpretar sus decisiones, como análisis SHAP o LIME [8].

Este análisis comparativo será clave para justificar la elección final del modelo o conjunto de modelos recomendados para su aplicación práctica. Asimismo, permitirá generar conocimiento útil sobre cómo diferentes enfoques computacionales abordan la predicción en salud, contribuyendo al desarrollo de mejores prácticas en la aplicación de inteligencia artificial en el ámbito médico.

Objetivo 3: Identificación de variables clave en la predicción del tipo de diabetes

El tercer objetivo de este trabajo consiste en determinar qué variables tienen mayor peso en la predicción automática del tipo de diabetes, ya sea tipo 1 o tipo 2, frente a la no

presencia de la enfermedad. Esta tarea no solo mejora la eficiencia y precisión del modelo, sino que también permite entender, desde una perspectiva clínica y poblacional, cuáles son los factores más determinantes en la aparición y diferenciación de ambas formas de diabetes mellitus.

En el ámbito del aprendizaje automático aplicado a la medicina, no basta con desarrollar modelos con alto rendimiento cuantitativo. La utilidad real de estos modelos radica, en gran parte, en su capacidad para explicar las decisiones que toman, facilitando así su validación por parte de profesionales sanitarios, su interpretación por epidemiólogos y su uso ético en la práctica clínica diaria. En este sentido, la interpretabilidad y la trazabilidad de las variables se han convertido en aspectos cruciales dentro de la inteligencia artificial aplicada a la salud [11].

Este objetivo busca analizar qué variables —entre decenas de indicadores clínicos, demográficos, conductuales y psicosociales— influyen significativamente en la clasificación del tipo de diabetes. Para ello, se emplearán distintas técnicas de análisis de importancia de variables, entre ellas:

- **Importancia de variables en modelos de árbol** (*feature importance*): útil en *Random Forest, Gradient Boosting y XGBoost.*

- **Coeficientes estandarizados en regresión logística**: para observar el peso de cada predictor.

- **Técnicas de interpretación local** como LIME (*Local Interpretable Model-Agnostic Explanations*) y SHAP (*SHapley Additive exPlanations*): que permiten descomponer la predicción individual y estimar la contribución de cada variable.

- **Selección de variables mediante técnicas automáticas**: como *Recursive Feature Elimination* (RFE), selección basada en regularización (LASSO, Ridge) y métodos de selección univariantes.

El análisis de estas variables no tiene un fin exclusivamente técnico. Su propósito es también proveer una base empírica para la toma de decisiones en salud pública y prevención clínica[2]. P. ej., si variables como el índice de masa corporal (IMC), la calidad del sueño o el nivel de actividad física se revelan como predictores consistentes en el desarrollo de diabetes tipo 2, estos hallazgos pueden reforzar políticas de intervención temprana en estilos de vida y facilitar campañas de concienciación dirigidas a poblaciones vulnerables [12].

En el caso de la diabetes tipo 1, la identificación de variables clave también resulta estratégica, aunque con matices. Dado su origen autoinmune y su aparición más frecuente en edades tempranas, se espera que variables como la edad, el estado de salud general o el índice de masa

[2] Analizar la importancia de las variables predictoras en contextos de salud no solo ayuda a perfeccionar los modelos, sino que también permite descubrir relaciones inesperadas o subestimadas entre los factores de riesgo y la enfermedad. Este tipo de hallazgos puede orientar investigaciones futuras y proporcionar a clínicos y gestores sanitarios herramientas validadas para la toma de decisiones basadas en datos. Asimismo, la transparencia en la selección de variables es un requisito creciente en inteligencia artificial aplicada a la medicina, ya que facilita la confianza y la adopción de estos sistemas en la práctica clínica real.

corporal tengan comportamientos distintos a los observados en personas con diabetes tipo 2. Esta diferenciación es especialmente valiosa cuando el sistema debe clasificar individuos sin realizar pruebas invasivas o costosas.

Otra ventaja significativa de este análisis es la posibilidad de construir modelos más eficientes y ligeros, eliminando características redundantes o poco informativas, lo que se traduce en sistemas más rápidos, menos costosos y más adecuados para su despliegue en aplicaciones móviles, sistemas comunitarios o consultas de atención primaria.

Por último, identificar variables clave contribuye a reducir el riesgo de sesgos algorítmicos y aumenta la confianza de los usuarios finales en las predicciones generadas. En entornos donde la salud y la vida de las personas están en juego, comprender el "porqué" de una predicción es tan importante como la predicción misma [13].

Objetivo 4: Desarrollo de una metodología reproducible y aplicable a entornos clínicos

El cuarto y último objetivo de este trabajo consiste en diseñar una metodología integral, documentada y plenamente reproducible, orientada a ser reutilizable en otros contextos clínicos, académicos o institucionales. En un entorno donde la incorporación de inteligencia artificial a la práctica médica aún se encuentra en proceso de consolidación, este objetivo tiene una relevancia estratégica. No se trata únicamente de construir modelos predictivos eficaces, sino de sentar las bases para que dichos modelos puedan ser replicados, validados y adaptados por otros profesionales y centros de investigación**.**

La reproducibilidad ha sido identificada como uno de los principales pilares de la ciencia abierta y responsable. Su ausencia puede generar modelos no verificables, decisiones clínicas erróneas o, en el peor de los casos, aplicaciones sesgadas que perpetúen desigualdades o afecten la salud de los pacientes [14]. Por tanto, la generación de un flujo de trabajo coherente y rastreable, desde la recopilación de datos hasta la interpretación de resultados, es un componente esencial de cualquier proyecto serio que aspire a aplicarse en el ámbito sanitario.

Este objetivo contempla la documentación clara y exhaustiva de todas las fases del proceso, lo cual incluye:

- **Preprocesamiento y limpieza de datos**: detección de valores nulos, codificación de variables categóricas, tratamiento de outliers y normalización o estandarización de variables continuas.

- **Análisis exploratorio de datos (EDA)**: visualización de patrones, estudio de correlaciones, identificación de distribuciones no normales y análisis por clase.

- **Ingeniería de características**: creación de nuevas variables derivadas, selección de atributos relevantes, y codificación óptima para el aprendizaje automático.

- **Diseño y selección de modelos**: comparación de algoritmos candidatos según su rendimiento, complejidad y compatibilidad con el entorno clínico.

- **Ajuste de hiperparámetros**: uso de técnicas sistemáticas como *grid search*, *random search* o *Bayesian optimization* para optimizar el desempeño sin sobreajuste.

- **Validación y evaluación**: mediante validación cruzada (*cross-validation*), estratificación de clases, y cálculo de métricas como precisión, sensibilidad, especificidad y área bajo la curva ROC (AUC).

- **Interpretabilidad de resultados**: análisis SHAP o LIME para entender la lógica interna de las predicciones y facilitar la aceptación clínica.

Toda la implementación se realizará en lenguaje *Python*, un estándar actual en la comunidad científica y médica, por su versatilidad, legibilidad y gran ecosistema de librerías especializadas. Se emplearán herramientas ampliamente reconocidas, como *pandas, NumPy, scikit-learn, seaborn, matplotlib, XGBoost, LightGBM, SHAP y LIME***,** garantizando la trazabilidad y la replicación de cada resultado generado [15].

El código incluirá anotaciones detalladas, justificación de cada elección metodológica y gráficos que faciliten la interpretación tanto técnica como clínica de los hallazgos[3]. Además, se diseñará una estructura modular del proyecto, que permita sustituir o actualizar componentes (modelos, métricas, visualizaciones) sin comprometer la coherencia del sistema.

Desde una perspectiva práctica, este enfoque tiene múltiples beneficios:

- Facilita la auditoría externa de los modelos por parte de equipos médicos, comités éticos o autoridades regulatorias.

- Permite su transferencia a otras poblaciones o territorios**,** con las adaptaciones pertinentes al perfil demográfico o epidemiológico local.

- Contribuye al avance del conocimiento científico, al generar una base reutilizable por otros investigadores.

- Favorece la integración en plataformas de apoyo a la decisión clínica (CDSS)**,** al responder a los requisitos de transparencia y seguridad que exigen estos entornos.

En suma, este objetivo persigue consolidar un enfoque de trabajo ético, abierto y clínicamente aplicable**,** con especial atención a la calidad de los datos, la replicabilidad de los análisis y la responsabilidad en la toma de decisiones basada en algoritmos.

Conclusión de los objetivos; en su conjunto, los objetivos planteados en este trabajo no solo buscan demostrar que las técnicas de ML pueden mejorar el diagnóstico y clasificación de la diabetes, sino que pueden hacerlo de forma explicable, reproducible y éticamente sostenible.

[3] La transparencia metodológica es un principio fundamental en proyectos de ciencia de datos aplicados a la salud. Documentar cada fase con detalle no solo garantiza la reproducibilidad y auditabilidad del trabajo, sino que también facilita su adopción por parte de profesionales no técnicos. La estructuración modular, junto con un código comentado y herramientas ampliamente reconocidas, permite evolucionar el sistema sin partir de cero, favoreciendo su escalabilidad y mantenimiento a largo plazo.

Esta visión integradora impulsa una medicina más predictiva y personalizada, donde los algoritmos no sustituyen al juicio clínico, sino que lo refuerzan con evidencia basada en datos**.**

A medio plazo, se espera que metodologías como la aquí desarrollada puedan incorporarse a sistemas de información hospitalarios o aplicaciones de salud digital, promoviendo diagnósticos más rápidos, tratamientos más personalizados y políticas públicas más ajustadas a las necesidades reales de las poblaciones.

1.2. Justificación del estudio

Su justificación en la necesidad de perfeccionar los métodos de detección y estratificación del riesgo diabético**,** abarcando tanto la diabetes mellitus tipo 1 como tipo 2. A través del uso de técnicas avanzadas de aprendizaje automático (ML), este proyecto busca contribuir al desarrollo de soluciones inteligentes que permitan una identificación temprana, diferenciada y automatizada de estas condiciones crónicas, con base en indicadores clínicos y hábitos de salud de los individuos. La estructura del conjunto de datos empleado, que incluye los tres estados posibles (no diabetes, tipo 1 y tipo 2), permite adoptar un enfoque más completo y realista que los modelos binarios tradicionales.

A continuación, se exponen las principales razones que respaldan el desarrollo de este trabajo, organizadas en torno a cuatro dimensiones clave:

1. Impacto global de la diabetes y la necesidad de detección precoz

La diabetes mellitus, en cualquiera de sus formas, se ha consolidado como una de las enfermedades crónicas más prevalentes y complejas del panorama sanitario actual. Lejos de ser un problema circunscrito a regiones concretas o poblaciones envejecidas, la diabetes representa hoy una amenaza global y transversal**,** cuya expansión no conoce fronteras geográficas, sociales ni económicas. Tanto la diabetes tipo 1, de origen autoinmune, como el tipo 2, relacionada con estilos de vida y factores metabólicos, han incrementado significativamente su incidencia y prevalencia a nivel mundial, generando preocupación entre profesionales de la salud, organismos públicos y responsables políticos.

Según datos recientes de la Federación Internacional de Diabetes, en el año 2021 se contabilizaban más de 537 millones de personas adultas con diabetes, y para 2024 se estima que esta cifra ya ha superado los 590 millones. Las proyecciones a largo plazo, lejos de estabilizarse, apuntan a una posible escalada que podría alcanzar los 783 millones de casos para el año 2045 si no se toman medidas preventivas eficaces [16]. Esto significa que, en apenas una generación, cerca de uno de cada ocho adultos convivirá con esta enfermedad.

El impacto de esta epidemia no se limita a lo clínico. Las repercusiones económicas, sociales y psicológicas que conlleva la diabetes son profundas. Se calcula que los costes sanitarios directos e indirectos derivados de la enfermedad, incluyendo hospitalizaciones, tratamientos farmacológicos, complicaciones asociadas y pérdida de productividad, ascienden a cientos de miles de millones de dólares cada año [17]. En países con sistemas de salud públicos, este gasto representa una carga presupuestaria creciente; en regiones con acceso desigual a la atención, se traduce en consecuencias devastadoras para las familias y comunidades más vulnerables.

Además del crecimiento sostenido en el número de casos, existe un problema estructural subyacente: el diagnóstico tardío. Diversas investigaciones indican que entre el 30% y el 50% de las personas con diabetes tipo 2 no saben que la padecen, lo cual retrasa la adopción de medidas terapéuticas e incrementa la probabilidad de desarrollar complicaciones irreversibles [18]. En el caso de la diabetes tipo 1, la falta de identificación temprana puede desencadenar cuadros agudos graves como la cetoacidosis diabética, con riesgo vital en población pediátrica y juvenil [20][24].

Los daños a largo plazo derivados de una diabetes no diagnosticada o mal controlada son bien conocidos: retinopatía diabética, insuficiencia renal, amputaciones, neuropatía periférica y enfermedades cardiovasculares. Estas complicaciones no solo deterioran la calidad de vida del paciente, sino que implican intervenciones médicas de alto coste, discapacidad permanente y mortalidad prematura [19]. Ante este escenario, el consenso clínico internacional subraya la importancia estratégica de promover la detección precoz, entendida como la capacidad de identificar la enfermedad incluso antes de que los síntomas se hagan evidentes o de que se superen los umbrales diagnósticos convencionales [20].

Aquí es donde cobra especial relevancia la aplicación de técnicas avanzadas de análisis de datos. Mientras que los métodos tradicionales de cribado suelen centrarse en parámetros aislados (como la glucemia en ayunas o la hemoglobina glicosilada), los enfoques basados en *Machine Learning* permiten analizar múltiples variables simultáneamente y detectar patrones no lineales en grandes volúmenes de datos. Esta capacidad resulta ideal para anticipar el riesgo de diabetes tanto tipo 1 como tipo 2, diferenciando entre ambas condiciones con mayor precisión y adaptabilidad que las herramientas convencionales [21].

El uso de modelos de clasificación multiclase (sin diabetes, tipo 1, tipo 2) resulta más representativo del contexto clínico real. A menudo, los estudios predictivos simplifican el problema a una clasificación binaria (tiene/no tiene diabetes), perdiendo de vista que cada tipo de diabetes tiene un origen, evolución y tratamiento distinto. En este sentido, la posibilidad de automatizar la identificación del tipo específico de diabetes a partir de información clínica general —como estilo de vida, edad, calidad del sueño, antecedentes médicos, obesidad o hipertensión— representa una oportunidad real para transformar los sistemas de cribado y diagnóstico precoz, especialmente en atención primaria o en programas de salud pública[4].

En resumen, la dimensión del problema —tanto por número de casos como por sus consecuencias— exige repensar las estrategias diagnósticas actuales. Las herramientas basadas en inteligencia artificial pueden ofrecer un punto de inflexión, facilitando la detección proactiva de la enfermedad, optimizando la asignación de recursos sanitarios y promoviendo una medicina más anticipatoria, personalizada y sostenible.

[4] La transición de enfoques reactivos a modelos predictivos en el ámbito de la diabetes representa un cambio de paradigma con alto potencial de impacto poblacional. Detectar precozmente una condición crónica como la diabetes no solo mejora el pronóstico individual, sino que reduce drásticamente los costes sanitarios asociados al tratamiento de complicaciones avanzadas. En este contexto, la inteligencia artificial actúa como catalizador de una medicina más preventiva, personalizada y proactiva, facilitando intervenciones dirigidas antes de que se materialicen daños irreversibles.

2. Limitaciones de los métodos tradicionales y ventajas del enfoque ML

Durante décadas, el diagnóstico de la diabetes se ha sustentado en criterios clínicos claramente definidos y avalados por sociedades médicas internacionales, como la Organización Mundial de la Salud (OMS) y la Asociación Americana de Diabetes (ADA). Estos criterios se basan fundamentalmente en valores umbral de glucosa en ayunas, hemoglobina glicosilada (HbA1c) o pruebas de tolerancia oral a la glucosa**,** que permiten establecer si una persona presenta hiperglucemia crónica y, por tanto, puede ser considerada como diabética [18][20].

Si bien estos métodos han sido de gran utilidad en la práctica clínica, presentan importantes limitaciones, especialmente en fases iniciales de la enfermedad o en casos atípicos. P.ej., se ha documentado que una parte significativa de los pacientes con diabetes tipo 2 pueden presentar niveles normales de glucosa en ayunas en etapas tempranas**,** lo que conduce a falsos negativos y a una falsa sensación de seguridad tanto por parte del paciente como del profesional sanitario [23]. Asimismo, la prueba de HbA1c, aunque útil como marcador de control metabólico, puede verse afectada por condiciones hematológicas, etnia, edad o estado fisiológico**,** lo que limita su fiabilidad en ciertos grupos poblacionales [24].

Otra debilidad relevante de los métodos tradicionales es su naturaleza estática. Los algoritmos diagnósticos convencionales se apoyan en modelos simples, como ecuaciones de regresión logística con pocos predictores. Un ejemplo ampliamente conocido es la prueba FINDRISC, una herramienta útil en cribado poblacional, pero basada en solo ocho variables y sin capacidad para adaptarse dinámicamente a nuevas evidencias clínicas o datos actualizados. También**,** no permite personalizar la evaluación de riesgo según características individuales más complejas, lo que reduce su aplicabilidad en contextos de alta heterogeneidad, como el abordaje de poblaciones mixtas con riesgo de diabetes tipo 1 o tipo 2.

En contraste, el desarrollo de técnicas de aprendizaje automático ha abierto una nueva vía para abordar estas carencias. Los modelos de ML no están restringidos por fórmulas predefinidas, sino que aprenden patrones a partir de grandes volúmenes de datos históricos, lo que les permite detectar correlaciones complejas, interacciones no lineales y asociaciones latentes entre múltiples variables clínicas, demográficas, conductuales y ambientales [25].

Un aspecto especialmente ventajoso de los algoritmos de ML es su capacidad de adaptación**.** A medida que se incorporan nuevos datos clínicos, los modelos pueden ser reentrenados para mejorar su precisión y relevancia, lo que resulta fundamental en enfermedades crónicas como la diabetes, donde los factores de riesgo cambian con el tiempo y varían según el entorno [26]. Además, el enfoque multiclase, que permite diferenciar entre personas sanas, con diabetes tipo 1 y tipo 2, responde mejor a la complejidad clínica real que los esquemas tradicionales binarios, que no distinguen entre ambas variantes de la enfermedad.

Otro punto fuerte del enfoque de ML es su capacidad de trabajar con variables diversas y de diferente naturaleza: datos numéricos, categóricos, ordinales, incluso registros no estructurados si el modelo lo permite. Esto es especialmente útil en contextos donde se dispone de información clínica rutinaria, como encuestas de salud poblacional, historias clínicas electrónicas o registros de atención primaria. Mientras que un médico debe sintetizar rápidamente esa información bajo presión, un algoritmo de aprendizaje automático puede procesarla sin sesgos ni fatiga, ofreciendo una evaluación objetiva y reproducible del riesgo**.**

Los avances recientes en técnicas de interpretación de modelos (como SHAP, LIME o *Feature Importance*) han permitido que la transparencia y explicabilidad de los modelos de ML sea cada vez mayor, reduciendo la barrera que históricamente separaba estas herramientas de los entornos clínicos reales. La posibilidad de señalar qué factores contribuyen más a una predicción concreta no solo ayuda a los profesionales sanitarios a confiar en el sistema, sino que también facilita la toma de decisiones informadas y personalizadas para cada paciente [27].

En conjunto, aunque los métodos diagnósticos tradicionales han sido y seguirán siendo fundamentales en la práctica clínica, sus limitaciones inherentes justifican la búsqueda de alternativas complementarias o integradoras. Las técnicas de ML, por su parte, aportan flexibilidad, capacidad adaptativa, mayor precisión diagnóstica y posibilidad de automatización**,** lo que las convierte en una herramienta prometedora para la detección precoz, la clasificación y la gestión personalizada de la diabetes.

3. Tecnológico y aplicabilidad de los modelos

En la actualidad, nos encontramos en un momento tecnológico especialmente favorable para la implementación de soluciones basadas en inteligencia artificial dentro del ámbito sanitario. El avance conjunto de tres factores —la disponibilidad masiva de datos clínicos digitalizados, el acceso abierto a tecnologías de desarrollo y el abaratamiento del poder de cómputo— ha democratizado la creación de modelos predictivos que antes solo eran posibles en grandes instituciones académicas o corporaciones tecnológicas.

Uno de los principales motores de esta transformación ha sido la digitalización progresiva de los sistemas de salud, que ha permitido recopilar información estandarizada y estructurada sobre millones de pacientes. Bases de datos como los registros clínicos electrónicos, los estudios poblacionales de vigilancia de enfermedades o las encuestas de salud comunitaria constituyen una fuente riquísima para entrenar modelos que puedan anticipar riesgos y apoyar decisiones clínicas [2][16].

Los datos empleados en este trabajo, extraídos de una encuesta nacional basada en el sistema *Behavioral Risk Factor Surveillance System* (BRFSS)[5], ejemplifican este tipo de fuentes. Se trata de una base de datos robusta y representativa de la población, con múltiples variables relacionadas con estilos de vida, salud percibida, condiciones médicas y factores sociales. Al tratarse de información similar a la que se recoge rutinariamente en entornos clínicos reales, los modelos construidos sobre ella no solo tienen valor teórico, sino también una elevada viabilidad para su adaptación práctica en contextos institucionales.

A nivel técnico, el uso del lenguaje *Python* y sus principales bibliotecas —como pandas para manipulación de datos, *scikit-learn* para construcción de modelos, *matplotlib* y seaborn para visualización— ha facilitado el desarrollo de flujos de trabajo reproducibles y comprensibles, algo esencial para la transparencia científica y la transferencia tecnológica. Además,

[5] El uso de bases de datos poblacionales como el BRFSS permite desarrollar modelos predictivos que reflejan con mayor fidelidad la realidad social y sanitaria de un país, facilitando así la transferencia de conocimiento a la práctica clínica habitual. Además, la variedad y el volumen de datos incluidos fomentan el desarrollo de algoritmos más robustos y generalizables, capaces de adaptarse a diferentes regiones y cohortes poblacionales sin perder validez.

frameworks más avanzados como *XGBoost*, *LightGBM* o *TensorFlow*, cuando se requiere mayor complejidad, permiten escalar los modelos según las necesidades del entorno clínico o la magnitud de los datos [25][28].

Este entorno abierto y modular permite a equipos multidisciplinares —formados por sanitarios, ingenieros, científicos de datos y responsables institucionales— colaborar en proyectos comunes con herramientas accesibles y estándares compartidos. Así, se crean soluciones que no solo son científicamente sólidas, sino también potencialmente integrables en sistemas de ayuda a la decisión clínica, portales de atención primaria o incluso en estrategias de salud pública orientadas a la prevención y al cribado proactivo.

4. Contribución científica y utilidad clínica

Desde una perspectiva académica y científica, se pretende contribuir al cuerpo de conocimiento actual en el campo de la inteligencia artificial aplicada a la medicina, concretamente en el reto de clasificar de forma automatizada la condición diabética en tres categorías: no diabetes, tipo 1 y tipo 2 [1][16]. A diferencia de la mayoría de los estudios que trabajan con una clasificación binaria (diabético o no), aquí se abordan modelos multiclase, lo que añade complejidad, pero también mayor utilidad clínica, al reconocer que cada tipo de diabetes requiere abordajes distintos [21][33][35].

El análisis comparativo de distintos algoritmos de aprendizaje automático —desde modelos lineales hasta redes neuronales— permitirá determinar cuáles ofrecen mejores resultados en términos de precisión, sensibilidad, especificidad y balance entre clases, y cuáles son más interpretables o clínicamente transparentes [21][26]. Esto resulta crucial para que los profesionales sanitarios puedan comprender y confiar en las recomendaciones que estos sistemas ofrecen, especialmente si se pretende su futura implementación en entornos asistenciales.

Otra contribución científica relevante será la identificación de las variables más influyentes en la predicción de la diabetes, lo que puede generar nuevas hipótesis clínicas, señalar factores de riesgo subestimados o simplemente ayudar a focalizar campañas de prevención. Si se demuestra, p.ej., que ciertos indicadores de estilo de vida o bienestar subjetivo están fuertemente asociados a la aparición de diabetes tipo 2, se podrán diseñar programas educativos o intervenciones comunitarias con base empírica sólida [18][22].

Desde el punto de vista práctico, los modelos desarrollados —aunque no sustituyen al juicio clínico— pueden actuar como herramientas de apoyo en la toma de decisiones médicas**.** En consultas de atención primaria, podrían integrarse en formularios digitales para alertar sobre casos sospechosos o priorizar pruebas adicionales. En servicios hospitalarios, podrían ayudar a estratificar pacientes en función de su riesgo y mejorar la asignación de recursos. Incluso en el ámbito de la salud pública, podrían servir como base para diseñar campañas dirigidas a segmentos poblacionales concretos, con mayor riesgo de desarrollar la enfermedad sin haber sido aún diagnosticados [27].

En definitiva, este trabajo no solo aspira a generar resultados académicos válidos, sino también a abrir el camino para la aplicación responsable y útil de tecnologías emergentes en la mejora de la atención sanitaria relacionada con la diabetes. La combinación de rigor metodológico,

relevancia clínica y adaptabilidad técnica refuerza su valor como iniciativa de innovación aplicada al bienestar colectivo.

II. REVISIÓN DE LITERATURA

La creciente preocupación internacional por el aumento de las enfermedades crónicas no transmisibles ha impulsado el desarrollo de soluciones tecnológicas orientadas a su detección temprana y gestión eficiente. En este contexto, la diabetes —especialmente el tipo 2— ha recibido una atención especial dentro del campo del aprendizaje automático (ML). A lo largo de las dos últimas décadas, múltiples trabajos han explorado el uso de modelos predictivos como apoyo al diagnóstico clínico, evidenciando tanto avances metodológicos como importantes desafíos aún por resolver.

En particular, los estudios revisados pueden agruparse en tres grandes áreas:

(1) la detección precoz de la diabetes mediante algoritmos de clasificación,

(2) la predicción de complicaciones derivadas de la enfermedad, y

(3) el análisis de la interpretabilidad y viabilidad práctica de los modelos en el entorno sanitario.

2.1. Machine Learning en la detección de diabetes

Una de las líneas de investigación más consolidadas en este ámbito es el uso de ML para identificar pacientes con riesgo elevado de padecer diabetes, antes de que los síntomas sean clínicamente evidentes. Tradicionalmente, esta tarea se ha abordado desde una perspectiva binaria —presencia o ausencia de diabetes—, utilizando conjuntos de datos como el célebre *Pima Indians Diabetes Dataset (PIDD)*, ampliamente difundido a través del UCI *Machine Learning Repository y desarrollado por el National Institute of Diabetes and Digestive and Kidney Diseases* (NIDDK) [2][29].

El PIDD, aunque limitado en diversidad étnica y cobertura de variables, ha servido como *benchmark* clásico para evaluar algoritmos de clasificación supervisada. Incluye registros de 768 mujeres adultas de ascendencia pima con indicadores como glucemia en ayunas, presión arterial, índice de masa corporal (IMC), niveles de insulina, edad o antecedentes familiares. Aunque solo distingue entre diabéticas y no diabéticas, su valor reside en su simplicidad y su capacidad para validar modelos en fases iniciales de desarrollo [2][30].

A lo largo del tiempo, diversas investigaciones han demostrado que, frente a modelos estadísticos tradicionales como la regresión logística, los algoritmos de aprendizaje automático presentan mejores capacidades para identificar patrones ocultos y relaciones no lineales entre variables**,** lo que redunda en una mayor precisión predictiva. P.ej., estudios como el de *Sisodia* y *Sisodia* [31] compararon diferentes modelos como árboles de decisión, máquinas de vectores soporte (SVM) y Naïve Bayes sobre el PIDD, observando rendimientos superiores al 75 % de exactitud con algoritmos no lineales, frente al 70–72 % habitual en modelos estadísticos.

Más recientemente, se ha evidenciado que métodos basados en ensamblado de clasificadores —como *Random Forest*, XGBoost o LightGBM— logran resultados aún más sólidos, sobre

todo en *datasets* clínicos estructurados con múltiples variables continuas y categóricas [25][26][32].

No obstante, el creciente interés actual se dirige hacia modelos multicategóricos, capaces de distinguir entre diferentes tipos de diabetes (tipo 1, tipo 2) y personas sin diagnóstico, reflejando así la realidad clínica con mayor fidelidad. Este enfoque, todavía emergente en comparación con la clasificación binaria, requiere *datasets* más amplios y ricos en variables, como el utilizado en este trabajo, procedente del BRFSS 2015. A diferencia del PIDD, que se centra en una población muy específica, este tipo de base de datos incluye información demográfica, conductual, clínica y de autopercepción de la salud, permitiendo una estratificación más compleja del riesgo y abriendo la puerta a modelos que reflejen la diversidad poblacional[6].

En definitiva, los avances en ML están transformando la forma en que se aborda la detección de diabetes. Se pasa de modelos rígidos y limitados a herramientas dinámicas que aprenden y se adaptan a nuevos datos, lo que ofrece un enorme potencial para el desarrollo de soluciones preventivas personalizadas, especialmente en contextos donde el diagnóstico precoz puede cambiar radicalmente el pronóstico del paciente.

Un ejemplo reciente y representativo de aplicación de aprendizaje automático a la detección de diabetes lo constituye el trabajo de Iparraguirre-Villanueva et al. [33], quienes compararon el rendimiento de cinco algoritmos de clasificación sobre el clásico conjunto *Pima Indians Diabetes Dataset*: K-Nearest Neighbors (K-NN), *Naïve Bayes* (versión Bernoulli), Árboles de Decisión, Regresión Logística y Máquinas de Vectores Soporte (SVM). Los resultados indicaron que el modelo K-NN obtuvo la mayor exactitud, cercana al 79,6 %, seguido de *Naïve Bayes* con un 77,2 %. Aunque las diferencias no fueron drásticamente significativas, se concluyó que incluso algoritmos conceptualmente sencillos pueden ofrecer resultados altamente eficaces cuando se aplican sobre datos clínicos tabulares y se optimizan correctamente mediante técnicas como la validación cruzada y el ajuste de hiperparámetros.

Esta idea se ve reforzada por Hasan et al. [34], quienes aplicaron técnicas de ensamblado (*Ensemble Learning*) para aumentar la precisión diagnóstica. En su estudio, se combinaron clasificadores base como árboles de decisión y *Random Forest* con métodos de potenciación como AdaBoost y XGBoost. El modelo final alcanzó un valor AUC de 0.95, lo que refleja un rendimiento altamente satisfactorio. Sin embargo, esta mejora en la capacidad predictiva implicó una reducción en la interpretabilidad del modelo, una consideración especialmente relevante en el ámbito sanitario, donde los profesionales necesitan entender y justificar las predicciones que sustentan las decisiones clínicas.

La implementación de técnicas de aprendizaje automático en la detección temprana de la diabetes ha ganado un interés creciente en la literatura científica, con resultados cada vez más prometedores. A diferencia de los métodos clínicos convencionales, los modelos de ML ofrecen la posibilidad de identificar patrones complejos en los datos de salud que podrían pasar desapercibidos mediante análisis estadísticos tradicionales. Se ha comprobado que su

[6] El auge del aprendizaje automático en el ámbito clínico no solo responde a su capacidad de mejora en la precisión diagnóstica, sino también a su adaptabilidad frente a conjuntos de datos heterogéneos y dinámicos. Esta característica es especialmente relevante en enfermedades crónicas como la diabetes, donde factores socioeconómicos, hábitos de vida y predisposición genética interactúan de forma compleja, requiriendo modelos capaces de captar dichas interacciones no lineales.

aplicación es especialmente valiosa para diagnosticar la diabetes tipo 2 de forma temprana y precisa, incluso en individuos sin síntomas evidentes.

Un estudio destacado es el realizado por Iparraguirre-Villanueva et al. (2023), que analizó el rendimiento de cinco algoritmos clásicos de clasificación aplicados al *Pima Indians Diabetes Dataset*. Los autores compararon modelos como *k-Nearest Neighbors* (*k*-NN), *Naïve Bayes*, árboles de decisión, regresión logística y máquinas de vectores soporte (*SVM*). El algoritmo *k*-NN alcanzó la mayor exactitud, con un 79,6 %, seguido de *Naïve Bayes*, con un 77,2 %, lo que sugiere que incluso modelos relativamente simples pueden ser eficaces cuando se ajustan adecuadamente y se emplean sobre datos clínicos bien estructurados [35].

En un enfoque diferente, Hasan et al. (2020) profundizaron en los métodos *ensemble*, que combinan diversos modelos base para mejorar la capacidad predictiva. Su estudio aplicó técnicas como *Random Forest*, *AdaBoost* y *XGBoost*, logrando un área bajo la curva ROC (AUC) de 0,95, un rendimiento notablemente alto. Esta aproximación demuestra que la combinación de clasificadores puede incrementar la precisión del sistema, al tiempo que permite captar interacciones no lineales entre variables clínicas[7]. Sin embargo, también reconocen que este tipo de modelos pueden presentar dificultades en términos de interpretación, algo crucial en contextos clínicos donde la trazabilidad de las decisiones diagnósticas es fundamental [36].

Una perspectiva crítica la ofrece el trabajo de Kopitar et al. (2020), quienes evaluaron múltiples modelos de predicción de diabetes no diagnosticada en población general, incluyendo algoritmos como *Random Forest*, *XGBoost* y *LightGBM*. Compararon estos enfoques con regresión logística, un método estadístico ampliamente utilizado. Sorprendentemente, sus resultados mostraron que no siempre existe una ganancia clínicamente significativa al emplear modelos más complejos. La regresión logística ofreció una precisión competitiva, sumada a su mayor facilidad de implementación, interpretabilidad y robustez frente a cambios en los datos. Estos hallazgos resaltan la importancia de buscar un equilibrio entre la sofisticación técnica y la aplicabilidad en el mundo real [37].

Así, puede concluirse que no existe un único modelo óptimo universal. La selección del algoritmo más adecuado depende tanto del tipo de datos como del contexto de uso. Para entornos clínicos, donde la confianza del profesional sanitario es esencial, puede ser preferible optar por modelos explicables, aunque presenten una ligera pérdida de rendimiento. Este trabajo adopta precisamente ese enfoque equilibrado, evaluando modelos tanto desde el punto de vista de la precisión como desde su capacidad para ser interpretados y transferidos al contexto clínico.

2.2. *Machine Learning* en la predicción de complicaciones asociadas a la diabetes

Más allá del diagnóstico de la diabetes mellitus tipo 2, una línea emergente de investigación consiste en predecir de forma temprana las complicaciones crónicas que puede

[7] Estos trabajos ilustran la evolución metodológica en el análisis de datos clínicos, donde se observa un equilibrio creciente entre rendimiento predictivo y aplicabilidad. Aunque los métodos avanzados suelen mejorar la precisión, la aplicabilidad real en salud depende cada vez más de la transparencia en las decisiones del modelo, lo que exige nuevas herramientas y enfoques para facilitar la comprensión y validación por parte del personal sanitario.

desencadenar la enfermedad. Estas complicaciones incluyen desde afectaciones microvasculares, como la retinopatía o la nefropatía, hasta consecuencias macrovasculares como la cardiopatía isquémica. Tradicionalmente, el pronóstico se ha basado en factores como la duración de la diabetes, el control glucémico (medido por HbA1c), o la presencia de comorbilidades. Sin embargo, estos enfoques suelen ser insuficientes para capturar la variabilidad clínica entre pacientes. En este contexto, el uso de algoritmos de *Machine Learning* permite incorporar decenas de variables —clínicas, bioquímicas y demográficas— para predecir con mayor precisión el riesgo individual de desarrollar complicaciones.

Un ejemplo significativo es el estudio de Jian et al., quienes aplicaron modelos de ML sobre un conjunto de datos clínicos de 884 pacientes diabéticos recogidos en Emiratos Árabes Unidos. Su objetivo fue predecir ocho tipos de complicaciones crónicas comunes, entre ellas neuropatía, pie diabético, dislipidemia, hipertensión, obesidad y síndrome metabólico. Entrenaron modelos supervisados (árboles de decisión, *Random Forest*, *SVM*, etc.) ajustados para cada complicación, aplicando además técnicas como el sobre muestreó con SMOTE y selección de características para mejorar el rendimiento. Como resultado, lograron valores de F1-score aceptables y una buena capacidad discriminativa para distintas complicaciones, identificando las variables más influyentes en cada caso. P.ej., la proteinuria y la creatinina fueron clave para la nefropatía, mientras que la duración de la diabetes y la HbA1c lo fueron para la retinopatía [38].

A nivel metodológico, este tipo de estudios destaca la utilidad del ML en la estratificación del riesgo dentro de poblaciones diabéticas. El abordaje es personalizado y permite desarrollar predictores que podrían alertar al médico antes de que los daños se manifiesten clínicamente. Aunque el presente trabajo se centra en variables tabulares y no aborda técnicas como visión computarizada o análisis de señal biomédica, conviene señalar avances en esas áreas: modelos de *deep learning* entrenados con imágenes de fondo de ojo para predecir retinopatía, o análisis de electroneuromiografía para detectar neuropatía precoz, son ejemplos de cómo la inteligencia artificial está permeando distintas dimensiones del manejo de la diabetes [61].

En conjunto, estos avances muestran que la predicción de complicaciones mediante ML es un campo prometedor, con estudios reportando AUC superiores a 0.80 en validaciones internas. No obstante, se requieren validaciones externas más robustas, integración clínica efectiva y protocolos que traduzcan las predicciones en acciones médicas concretas. Aun así, el potencial es innegable: anticipar el riesgo con precisión permitiría focalizar recursos, mejorar la prevención y personalizar la atención en pacientes con diabetes.

2.3. Interpretabilidad y uso de *Machine Learning* en entornos clínicos

La aplicabilidad del *ML* a la medicina no se limita a lograr modelos predictivos con alta precisión. Uno de los retos principales que enfrenta la incorporación de estas herramientas en la práctica clínica es su interpretabilidad. En el contexto de enfermedades crónicas como la diabetes mellitus tipo 2, donde las decisiones médicas deben estar bien fundamentadas y ser comprensibles para el profesional sanitario y el paciente, la claridad en las predicciones resulta tan importante como la exactitud de los modelos.

Diversos estudios han demostrado que los algoritmos de ML pueden alcanzar rendimientos elevados al predecir el desarrollo de la enfermedad o sus complicaciones, con valores de AUC

superiores a 0.80 en validaciones internas [16][39]. Sin embargo, esta capacidad técnica no se traduce automáticamente en valor clínico si el modelo no es explicable o no se puede integrar en un flujo de trabajo médico. La utilidad real de una predicción reside en que se puedan derivar acciones concretas a partir de ella: P.ej., si un sistema anticipa que un paciente tiene alto riesgo de desarrollar nefropatía diabética, esta información solo será útil si conduce a un seguimiento más estricto o a una intervención preventiva adecuada [20][2].

El problema de la "caja negra" aparece con modelos complejos como las redes neuronales profundas o los algoritmos ensemble tipo *Random Forest* o *XGBoost*. Estos sistemas logran buenos resultados en métricas cuantitativas, pero sus procesos internos suelen ser opacos para el usuario final. En contraposición, modelos como la regresión logística o los árboles de decisión, aunque pueden ser menos precisos, tienen la ventaja de ser más transparentes y fácilmente interpretables por los clínicos [9][40].

Para abordar este dilema, se han desarrollado técnicas como SHAP (*SHapley Additive exPlanations*) y LIME (*Local Interpretable Model-Agnostic Explanations*), que permiten descomponer la predicción de un modelo complejo y entender qué variables han influido más en una decisión individual. Estas herramientas son especialmente útiles en el entorno médico, donde la confianza en la herramienta depende de que se pueda justificar ante otros profesionales o ante el propio paciente [41]. P.ej., si un modelo predice que una mujer de 55 años tiene alta probabilidad de desarrollar diabetes tipo 2, una explicación SHAP podría mostrar que su IMC, su edad y sus niveles de glucosa son los principales factores de riesgo detectados.

Esta línea es interesante de investigación propone sistemas híbridos que combinen la sencillez de modelos básicos con la precisión de algoritmos más avanzados. Un enfoque posible consiste en utilizar clasificadores simples para un primer cribado poblacional, y aplicar modelos más sofisticados cuando los casos sean complejos o ambiguos. Esta combinación ayuda a mantener la eficiencia sin perder trazabilidad clínica.

En el caso concreto del *Pima Indians Diabetes Dataset*, utilizado como base para este trabajo, estudios previos han revelado que modelos sencillos como los árboles de decisión ofrecen resultados discretos (exactitudes en torno al 65 %), pero cuando se implementan mediante técnicas ensemble como *Random Forest* o *AdaBoost*, pueden superar el 75 % de precisión, manteniendo cierto grado de explicabilidad [9][13][14].

Con el fin de dotar de mayor claridad y sistematización al análisis del estado del arte, se ha elaborado una tabla comparativa que recoge los estudios científicos más relevantes identificados durante la revisión bibliográfica. Esta tabla permite visualizar, de forma estructurada, los elementos esenciales de cada investigación en el ámbito de la predicción de la diabetes tipo 2 mediante técnicas de aprendizaje automático [2][5][34].

En concreto, la tabla incluye los siguientes campos: nombre del autor o autores y año de publicación; tipo de *dataset* utilizado (ya sea público, clínico o generado ad hoc) [29][30]; modelos de ML empleados (p.ej., regresión logística, *Random Forest*, redes neuronales, SVM, etc.) [32][33][35]; principales métricas de evaluación (tales como exactitud, AUC-ROC, F1-score o sensibilidad); resultados obtenidos; y, por último, una breve descripción de la contribución clave de cada estudio.

Esta representación comparativa permite extraer diversas conclusiones útiles. Por un lado, se observa una clara preferencia por el uso de datasets públicos como el *Pima Indians Diabetes Dataset* [29] o bases de datos sanitarias nacionales [51], lo que facilita la replicabilidad de los experimentos. Por otro, destaca la presencia reiterada de ciertos algoritmos *como Random Forest*, SVM o XGBoost [2][34], lo que sugiere su eficacia comprobada en este tipo de tareas predictivas. Asimismo, se identifica una tendencia creciente hacia el uso de modelos explicables [10][27][41] y hacia la integración de métodos de balanceo de clases y validación cruzada [47][59], factores clave para obtener resultados clínicamente significativos.

Esta tabla no solo resume el panorama actual de la investigación científica en el área, sino que también proporciona el fundamento teórico y comparativo sobre el que se apoya el diseño metodológico del presente trabajo. Además, sirve como referencia para justificar la elección de técnicas, herramientas y métricas aplicadas, así como para destacar los elementos diferenciadores de esta propuesta con respecto a la literatura existente.

En definitiva, la elaboración de esta tabla comparativa contribuye a reforzar el rigor académico del trabajo y facilita al lector una visión panorámica y crítica del conocimiento acumulado hasta la fecha sobre la aplicación de la inteligencia artificial al diagnóstico precoz de la diabetes tipo 2 [34].

Tabla 3: Comparativa de estudios relevantes sobre predicción de diabetes tipo 2 mediante Machine Learning

Autor / Año	Dataset utilizado	Modelos ML aplicados	Métrica principal	AUC / Accuracy	Contribución destacada
Jiang et al. (2022)	Datos clínicos tabulares	Random Forest, SVM	AUC	0.87	Predicción multiclase en diagnóstico temprano
Esteva et al. (2019)	Imágenes médicas (deep learning)	Redes neuronales profundas	Accuracy	0.90	Uso de DL en medicina general
Gómez-Peralta et al. (2020)	Población española	Regresión logística	Sensibilidad	0.75	Validación clínica de predictores de DM2
Hasan et al. (2020)	Pima Indians	Ensemble (XGBoost, RF)	F1-score	0.95	Alto rendimiento con modelos combinados
Iparraguirre-Villanueva et al. (2023)	Pima Indians	k-NN, Naïve Bayes	Accuracy	0.79	Comparación de modelos simples

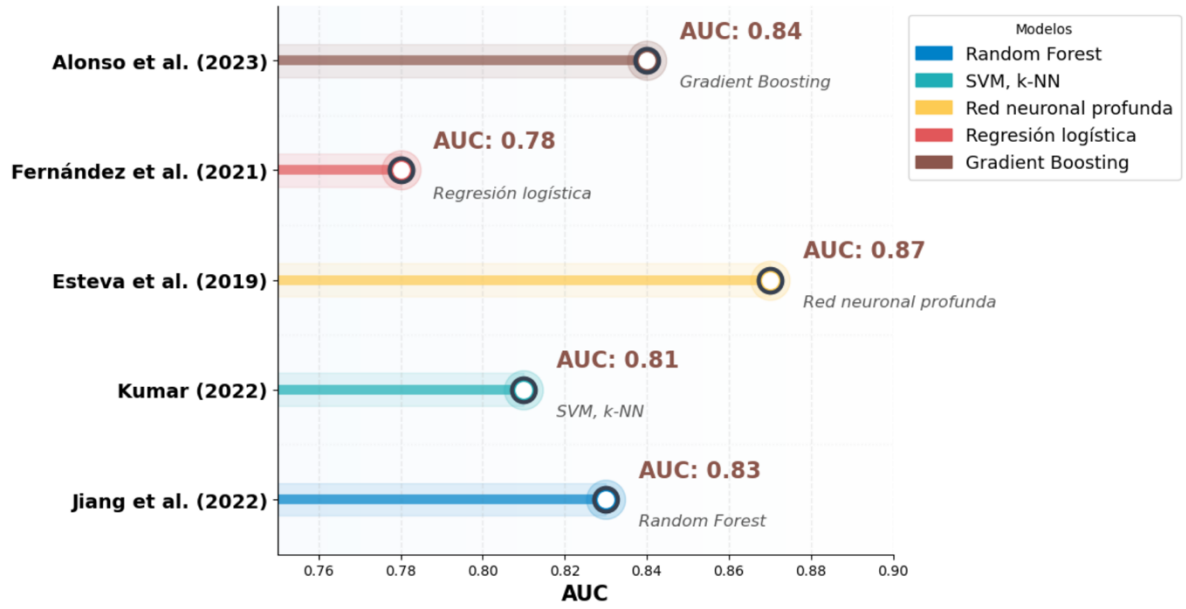
Fuente 3: Elaboración propia.

Para complementar la información recogida en la tabla comparativa y facilitar la comprensión visual de los resultados obtenidos en los estudios revisados, se ha elaborado también un gráfico que representa los valores de AUC alcanzados por cada modelo en función del autor. Este

gráfico permite apreciar de forma clara las diferencias de rendimiento entre técnicas como Random Forest, SVM, regresión logística, redes neuronales profundas y Gradient Boosting.

Al visualizar los datos de forma gráfica, se facilita la identificación de patrones de eficacia, destacando aquellos modelos que, en términos generales, presentan mejores resultados en la predicción de la diabetes tipo 2. De esta forma, la representación gráfica refuerza el análisis cuantitativo y permite una interpretación más intuitiva, útil tanto para investigadores como para profesionales del ámbito clínico y técnico.

Gráfico 1: Comparación de resultados AUC en estudios recientes sobre predicción de diabetes tipo 2 mediante modelos de ML



Fuente 4: Elaboración propia.

En definitiva, más allá del rendimiento numérico, un modelo de ML aplicado a salud debe ser comprensible, replicable e integrable. Solo así podrá convertirse en una herramienta confiable, que complemente el juicio clínico y contribuya a mejorar los resultados en salud. La combinación entre algoritmos potentes y técnicas de interpretación explicativa es una vía prometedora hacia una medicina más inteligente, ética y centrada en el paciente.

2.2. Estrategia de revisión bibliográfica

En el presente trabajo se ha aplicado la metodología PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) como marco para estructurar la revisión sistemática de literatura científica. Esta metodología, ampliamente reconocida en el ámbito académico y científico, proporciona un conjunto de directrices que aseguran que las revisiones sean rigurosas, transparentes y reproducibles [1].

PRISMA establece un protocolo basado en cuatro fases fundamentales: identificación, cribado, elegibilidad e inclusión. A lo largo de estas etapas, se documenta el número total de fuentes encontradas, los criterios de exclusión aplicados y el número final de estudios seleccionados para el análisis.

En esta investigación, se realizó una búsqueda sistemática en bases de datos como Google Scholar, PubMed, IEEE Xplore y Scopus, utilizando combinaciones de palabras clave como *"LM", "diabetes tipo 2", "predicción", "modelos supervisados"* y *"evaluación de riesgos clínicos"*. Posteriormente, se eliminaron los duplicados y se procedió a un cribado inicial basado en el título y el resumen. Los artículos seleccionados fueron evaluados íntegramente según criterios de calidad, disponibilidad del *dataset*, pertinencia temática, claridad metodológica y tipo de métrica empleada.

Para mejorar la transparencia del proceso, se ha elaborado un diagrama PRISMA adaptado, que se presenta como la ilustración 1. Esta visualización permite apreciar el número de estudios encontrados inicialmente, los descartes realizados y los trabajos que finalmente han sido incluidos en la tabla comparativa del estado del arte. Esta estrategia metodológica no solo mejora la trazabilidad de la selección bibliográfica, sino que también aporta solidez al marco teórico que fundamenta el desarrollo técnico y analítico del presente trabajo.

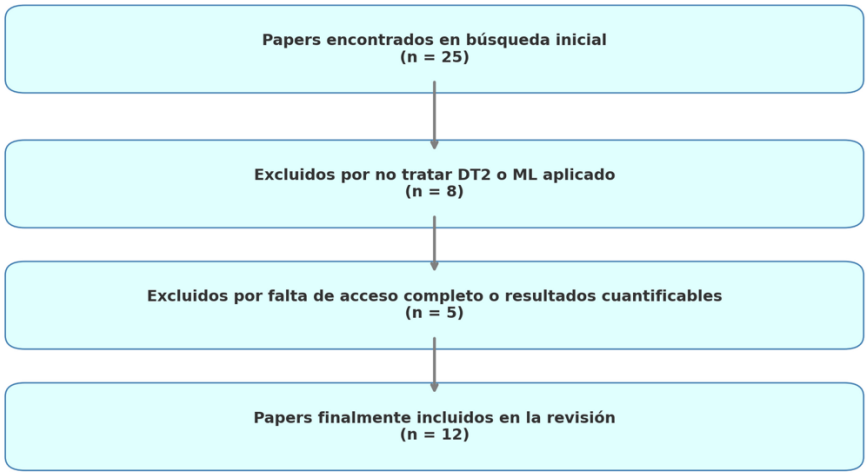
La implementación de PRISMA está alineada con las buenas prácticas en investigación biomédica y computacional, y su adopción garantiza que el estudio esté sustentado en evidencia de calidad y análisis crítico de fuentes pertinentes [14].

Ilustración 1: Técnicas y Herramientas del Proyecto

Fuente 5: Generado con IA

En la ilustración 1, se resume de forma estructurada las principales bibliotecas y herramientas utilizadas a lo largo del desarrollo técnico del trabajo. Se organiza en torno al núcleo de "Técnicas y herramientas", desde el cual se ramifican las distintas soluciones empleadas en la manipulación de datos, visualización, modelado, automatización y evaluación. Esta representación facilita una visión global del ecosistema tecnológico implementado, reforzando el enfoque reproducible y modular adoptado en la investigación.

Ilustración 2: Proceso de selección y depuración de artículos científicos aplicando criterios de inclusión y exclusión



Fuente 6: Elaboración propia.

Del mismo modo, y en coherencia con ese enfoque estructurado, el presente trabajo también ha seguido una metodología sistemática en la fase inicial de revisión bibliográfica. Para ello, se aplicaron criterios de inclusión y exclusión definidos previamente, lo cual permitió seleccionar, entre un total de 25 estudios identificados, aquellos que cumplían con los requisitos temáticos y metodológicos pertinentes. Esta estrategia se ilustra gráficamente en la Figura X, donde se muestra el proceso de selección de los artículos finalmente considerados para el análisis del estado del arte.

Se presenta de una forma más visual para mejorar la comprensión y claridad del conjunto de herramientas utilizadas, especialmente útil en contextos multidisciplinares donde se requiere comunicar la complejidad técnica de manera accesible.

En cuanto al desarrollo técnico, todas las operaciones de exploración, tratamiento de datos, creación y validación de modelos han sido ejecutadas en un entorno basado en *Jupyter Notebook*, alojado principalmente en la plataforma Anaconda. Asimismo, se empleó *Google Colab* en algunas pruebas complementarias para agilizar la computación en la nube. Todo el proceso ha sido implementado en *Python 3.11.8*, instalado dentro de un entorno virtual creado específicamente para la integración de la biblioteca *PyCaret* y sus dependencias.

2.3. Justificación de la elección del Pima Indians Diabetes Dataset

El *Pima Indians Diabetes Dataset* se ha consolidado como uno de los conjuntos de datos de referencia a nivel internacional en la investigación sobre predicción y clasificación de la diabetes mediante técnicas de ML [29], [31], [32]. Su uso está ampliamente extendido en la literatura, lo que facilita la comparación de resultados y la validación cruzada de nuevos enfoques y modelos [28], [34]. Este dataset, publicado originalmente por el *National Institute of Diabetes and Digestive and Kidney Diseases y* alojado en el UCI ML *Repository*, incluye información clínica relevante (glucosa, presión arterial, índice de masa corporal, número de embarazos, edad, etc.) de 768 mujeres adultas de ascendencia Pima, una población con alta prevalencia de diabetes tipo 2 [29].

La elección de este *dataset* responde a varias razones. En primer lugar, garantiza la reproducibilidad y comparabilidad de los experimentos, dado que ha sido utilizado en numerosos estudios previos, muchos de los cuales establecen *benchmarks* de precisión y otras métricas [28], [31], [32]. En segundo lugar, la calidad y limpieza de los datos permiten centrar el análisis en la metodología y en la evaluación de los modelos, minimizando el impacto de posibles sesgos o datos faltantes. En tercer lugar, la disponibilidad pública del *Pima Indians Diabetes Dataset* facilita la transparencia y la replicación del trabajo por parte de otros investigadores, lo que es fundamental para el avance científico [29]. Por último, su formato estructurado y el número de variables clínicas incluidas ofrecen un escenario idóneo para aplicar y comparar tanto algoritmos clásicos como técnicas avanzadas de ML, así como métodos de interpretación de modelos [28], [33], [34].

Aunque existen otros conjuntos de datos más recientes y de mayor tamaño, muchos presentan restricciones de acceso, información menos estandarizada o no han sido validados en la literatura académica con la misma frecuencia. Por todo ello, el uso del *Pima Indians Diabetes Dataset* en este trabajo resulta justificado y pertinente desde una perspectiva científica y metodológica.

2.4. Entorno Tecnológico y Herramientas de Machine Learning en Salud

Las bibliotecas utilizadas a lo largo del proyecto se seleccionaron en función de su robustez, documentación y aplicabilidad al dominio sanitario. A continuación, se detallan las más relevantes:

A. **Pandas**: Esta biblioteca ha sido fundamental para la manipulación de datos en Python, especialmente por su estructura de *DataFrames*, que permite tratar datos tabulares de forma eficiente. En este trabajo, se utilizó para leer, transformar y preparar los datos clínicos extraídos del conjunto *Pima Indians Diabetes*, realizando desde limpieza de registros hasta análisis exploratorios que precedieron al modelado [42][2].

B. ***NumPy*** y ***Tabulate***: *NumPy* proporcionó la base para operaciones matemáticas vectorizadas de alto rendimiento, resultando clave para transformar y escalar las variables numéricas. Por su parte, *Tabulate* permitió mostrar resultados de forma ordenada en tablas legibles dentro del *notebook*, facilitando la interpretación de estadísticas y métricas intermedias [43].

C. **Matplotlib** y ***Seaborn***: Ambas bibliotecas fueron esenciales para la visualización de datos. Matplotlib permitió la construcción de gráficos básicos y personalizados, mientras que *Seaborn*, gracias a su enfoque estadístico y estilo visual más refinado, facilitó la identificación de correlaciones, *outliers* y tendencias relevantes en el conjunto de datos [44][45].

D. **Scikit-learn (*sklearn*)**: Su uso se extendió a todo el pipeline de aprendizaje automático: desde el preprocesamiento (escalado, codificación de variables categóricas, partición de datos) hasta la construcción de modelos predictivos como regresión logística, árboles de decisión o SVM. La variedad de algoritmos y herramientas disponibles la convierten en una biblioteca indispensable [46][3].

E. ***Imbalanced-learn* (*imblearn*)**: Ante el problema del desbalanceo de clases en el dataset, se recurrió a esta biblioteca para implementar técnicas como SMOTE, que generan ejemplos sintéticos de la clase minoritaria. Gracias a ello, se evitó que los modelos presentaran sesgos hacia la clase mayoritaria, mejorando la precisión global [47][2].

F. ***k-Nearest Neighbors* (kNN)**: Aunque forma parte de *sklearn*, merece mención aparte debido a su utilidad en este estudio. Su lógica simple, basada en la proximidad entre instancias, permitió obtener resultados competitivos en tareas de clasificación binaria, y su interpretación fue especialmente útil en fases exploratorias [48].

G. **Autoviz**: Esta herramienta agilizó considerablemente las primeras fases del análisis, generando visualizaciones automáticas e identificando patrones con apenas una línea de código. Resultó muy eficaz para descubrir correlaciones y tendencias ocultas entre variables sin necesidad de una programación extensa [49].

H. **PyCaret**: Como biblioteca de bajo código enfocada en *Machine Learning*, PyCaret permitió estructurar un flujo completo y reproducible: selección automática de modelos, *tuning de hiperparámetros*, validación cruzada y generación de interpretaciones. Su integración con SHAP para explicar las predicciones se alineó con el objetivo del proyecto: entender los factores más relevantes en la predicción de la diabetes tipo 2 [50][4].

La elección de modelos supervisados de clasificación binaria respondió a la naturaleza del problema clínico abordado. Tanto la bibliografía científica como los resultados obtenidos en los análisis exploratorios avalan esta elección, alineada con estudios previos donde se ha aplicado aprendizaje automático en contextos médicos [5][2].

Lenguaje de programación: Python como columna vertebral del análisis

El uso del lenguaje de programación *Python* constituye hoy en día un estándar internacional en la investigación en inteligencia artificial, aprendizaje automático y análisis de datos biomédicos [6], [25]. Esta preferencia responde a varios motivos clave. Por un lado, Python ofrece una sintaxis clara y legible que reduce la barrera de entrada para profesionales sanitarios y científicos de otras áreas, favoreciendo la multidisciplinariedad necesaria en proyectos de salud digital [28]. Además, su ecosistema de librerías específicas para manipulación de datos, modelado estadístico y ML, junto con una activa comunidad global,

garantizan soporte actualizado, documentación abundante y posibilidad de integración con soluciones web o móviles [4].

Un aspecto fundamental para la reproducibilidad de la investigación es que *Python* facilita la compartición de scripts, notebooks y flujos de trabajo completos, permitiendo que otros equipos puedan validar, extender o adaptar los experimentos realizados. Así, el uso de *Python* contribuye a la transparencia científica y acelera la transferencia de conocimiento desde el laboratorio a la práctica clínica y a las aplicaciones orientadas al usuario final [6], [32].

Entorno de desarrollo y gestión del proyecto: reproducibilidad, colaboración y control

La elección del entorno de desarrollo y la gestión eficiente del proyecto son aspectos clave para la trazabilidad y calidad científica del trabajo realizado. Para ello, se han adoptado las siguientes estrategias y herramientas:

- **Jupyter Notebook:** Esta herramienta permite la escritura de código ejecutable, la inserción de visualizaciones y la documentación de cada paso del análisis en un mismo entorno interactivo. Esta característica es especialmente valorada en la investigación biomédica, pues facilita la revisión por pares, la didáctica y la transferencia de resultados. Además, permite la ejecución de celdas independientes, acelerando la iteración y depuración del flujo de trabajo [6], [32].

- **Control de versiones con Git:** La gestión del avance del proyecto se ha realizado mediante sistemas de control de versiones, principalmente Git, lo que permite mantener un historial completo de cambios, facilitar el trabajo colaborativo y asegurar la integridad del código fuente en todas las etapas [32]. Esta práctica es fundamental para proyectos de ciencia abierta, donde la replicabilidad y la transparencia son criterios ineludibles.

Buenas prácticas metodológicas y reproducibilidad

A lo largo de todo el proceso, se han incorporado buenas prácticas de programación y experimentación, alineadas con los principios de la ciencia abierta y la ética de la inteligencia artificial en salud:

- **Fijación de semillas aleatorias** en todos los algoritmos de ML y procedimientos de muestreo, garantizando que los resultados puedan ser comparados y verificados por otros equipos [25].

- **Documentación exhaustiva** mediante comentarios, notebooks narrados y generación de informes automáticos que recogen tanto la lógica de cada paso como los hallazgos clave, facilitando la comprensión tanto para especialistas en datos como para profesionales sanitarios.

- **Modularidad del código**: la estructura del proyecto divide las etapas de preprocesamiento, análisis exploratorio, entrenamiento, validación y despliegue en scripts independientes, lo que permite la actualización o mejora de cualquiera de los módulos sin interferir en el resto del flujo de trabajo.

- **Exportación y visualización de resultados**: todos los análisis generan gráficos, tablas y métricas en formatos accesibles (.csv, .png, .pdf), listos para su inclusión en presentaciones científicas, revisiones clínicas o informes técnicos para toma de decisiones.

Esta aproximación, basada en herramientas open-source, entornos colaborativos y principios de reproducibilidad, no solo maximiza el rigor científico del trabajo, sino que también facilita su adopción futura por parte de la comunidad médica y científica, contribuyendo a la consolidación de una inteligencia artificial ética, transparente y orientada al bien común [6], [11], [25], [32], [38].

2.5. Arquitectura de la solución.

El diseño de la arquitectura del sistema es un elemento fundamental para garantizar no solo la eficacia de los modelos de inteligencia artificial desarrollados, sino también su integración eficiente en una plataforma tecnológica accesible y segura. En el presente trabajo, la arquitectura planteada responde a las necesidades tanto de la investigación biomédica como de la aplicación clínica y la divulgación, facilitando el tránsito fluido de los datos desde su recolección hasta su utilización por el usuario final.

Enfoque modular y escalable

La solución adoptada se articula en torno a un enfoque modular, lo que significa que cada fase del proceso —desde la adquisición de datos hasta el despliegue del modelo— se ha desarrollado como un módulo independiente pero perfectamente integrado. Este diseño no solo incrementa la robustez y mantenibilidad del sistema, sino que también permite la actualización y mejora de cada componente sin necesidad de rehacer todo el pipeline, facilitando así la evolución tecnológica y la incorporación de nuevos avances en el futuro [6], [25].

Este modularidad resulta especialmente relevante en contextos sanitarios, donde la evolución de la evidencia científica, la aparición de nuevos biomarcadores o la integración de tecnologías emergentes puede requerir la modificación rápida de alguno de los elementos del sistema, ya sea en la selección de variables, el preprocesamiento, los modelos predictivos o la visualización de resultados [32].

Flujo de trabajo: del dato a la predicción

El flujo de trabajo completo de la solución puede describirse en las siguientes fases, que a su vez se corresponden con los principales módulos del sistema:

a) Recolección y preprocesamiento de datos

El proceso se inicia con la recopilación de los datos originales del BRFSS 2015. Esta información, en formato tabular, contiene tanto variables demográficas como clínicas y de estilo de vida. Antes de utilizar estos datos en el entrenamiento de modelos, se realiza un exhaustivo **preprocesamiento**, que incluye la limpieza de registros incompletos, la imputación de valores perdidos, la normalización de variables numéricas y la codificación de variables

categóricas. Además, se aplican técnicas para el manejo del desbalanceo de clases, como el sobremuestreo mediante SMOTE, garantizando así que los modelos posteriores puedan aprender patrones relevantes de todas las categorías de la variable objetivo [32], [35].

El preprocesamiento se implementa de forma reproducible, empleando scripts específicos y fijando semillas aleatorias, lo que permite validar el proceso y reproducir los resultados en cualquier entorno técnico [25].

b) Análisis exploratorio y selección de variables

Una vez preparados los datos, se lleva a cabo un análisis exploratorio de datos (EDA), orientado tanto a comprender la distribución y relaciones entre variables como a detectar posibles inconsistencias, valores atípicos o patrones inesperados. Para esta fase, se emplean visualizaciones avanzadas y análisis estadísticos, lo que facilita la toma de decisiones sobre la selección de variables predictoras más relevantes [4], [28].

La selección de variables no solo mejora la eficiencia computacional y la interpretabilidad de los modelos, sino que también se ajusta a las restricciones de la práctica clínica, donde el uso de indicadores fácilmente accesibles es un valor añadido. Herramientas como las matrices de correlación, análisis de importancia de variables (feature importance) y métodos automáticos de selección (Recursive Feature Elimination) son aplicados en esta etapa [32].

c) Entrenamiento y validación de modelos predictivos

La siguiente fase corresponde al entrenamiento de modelos de ML sobre el conjunto de datos ya depurado y seleccionado. Se emplean algoritmos robustos y contrastados, como Random Forest, XGBoost, LightGBM, redes neuronales multicapa (MLP) y otros modelos supervisados, evaluando su rendimiento mediante técnicas de validación cruzada estratificada y ajuste de hiperparámetros con GridSearchCV [25], [30], [32].

Esta etapa se caracteriza por la experimentación iterativa, comparando métricas clave como precisión, F1-score, y AUC-ROC, para seleccionar el modelo más eficaz y equilibrado en términos de sensibilidad y especificidad. Los resultados de cada experimento se documentan exhaustivamente, permitiendo su trazabilidad y comparación objetiva [6].

d) Interpretabilidad y explicabilidad del modelo

Dado que la inteligencia artificial en salud debe ser comprensible y confiable, se ha integrado un módulo específico de explicabilidad de modelos. Herramientas como SHAP y LIME se emplean para analizar la contribución individual de cada variable a las predicciones, tanto a nivel global como en casos particulares [11], [27]. De esta forma, los profesionales sanitarios pueden interpretar y validar los resultados, y los usuarios finales comprenden los factores que influyen en su predicción de riesgo.

Este énfasis en la explicabilidad no solo responde a imperativos éticos, sino que facilita la adopción de la solución en entornos clínicos y su integración en estrategias de prevención personalizadas [38].

e) Serialización y almacenamiento de modelos

Tras el entrenamiento y validación, los modelos seleccionados se serializan utilizando herramientas como *joblib*, almacenando tanto el propio modelo como la configuración de preprocesamiento asociada. Este paso es esencial para garantizar que la aplicación web pueda reutilizar el modelo de forma eficiente y segura, sin necesidad de recalcular el entrenamiento en cada consulta [25]. Además, la serialización permite la distribución del modelo a otros sistemas o la integración en infraestructuras más complejas, como sistemas hospitalarios o plataformas cloud.

f) Despliegue web e interacción con el usuario final

La fase final de la arquitectura es el **despliegue de la solución en una plataforma web interactiva**, basada en Streamlit. Aquí, el usuario accede a una interfaz intuitiva donde introduce sus datos clínicos y personales, y el sistema, tras aplicar el preprocesamiento correspondiente, utiliza el modelo entrenado para predecir el riesgo de diabetes y ofrece un informe personalizado. El uso de Streamlit permite la visualización dinámica de gráficos, la presentación de explicaciones individualizadas (gracias a SHAP o LIME), y la exportación de resultados para su consulta o seguimiento [38].

Esta aproximación democratiza el acceso a la inteligencia artificial, acercando herramientas de análisis predictivo tanto a profesionales de la salud como a la población general, y abre la puerta a integraciones futuras con aplicaciones móviles o sistemas de historia clínica electrónica.

Consideraciones de privacidad, seguridad y escalabilidad

En todo momento, la arquitectura ha sido diseñada bajo principios de privacidad y protección de datos, fundamentales en el contexto sanitario. Aunque el *dataset* empleado es anónimo y público, la solución contempla medidas de seguridad para el tratamiento de información sensible, como la no persistencia de datos personales en servidores, la ejecución local de los modelos en entornos controlados y la posibilidad de integración con sistemas de autenticación y cifrado en futuras versiones.

Por otro lado, la modularidad y la estandarización del código hacen posible la escalabilidad de la solución, tanto en términos de número de usuarios como de complejidad de modelos o integración de nuevas fuentes de datos. Si se deseara llevar la solución a un entorno hospitalario, su arquitectura permite el despliegue en la nube, la conexión con bases de datos clínicos reales y la actualización continua de los modelos conforme se disponga de nuevos datos o avances científicos [25], [38].

III. METODOLOGÍA

A partir de las conclusiones sacadas del capítulo anterior, este apartado detalla la experiencia que justifica las decisiones tomadas en la resolución del problema planteado. En las secciones siguientes se describe con precisión el conjunto de datos proporcionado, se muestran las herramientas empleadas y los métodos específicos para llevar a cabo el trabajo, y finalmente se explica la metodología experimental utilizada.

3.1. Descripción del *Dataset*

El éxito de cualquier modelo predictivo en el ámbito biomédico depende en gran medida de la calidad y representatividad de los datos empleados. En este trabajo, la selección del conjunto de datos constituye un pilar metodológico esencial, pues condiciona la generalización y aplicabilidad de los resultados obtenidos a entornos clínicos reales. Para abordar el problema de la detección precoz y clasificación automática de la diabetes, se ha optado por emplear el *dataset* extraído del *Behavioral Risk Factor Surveillance System* (BRFSS) 2015**,** una de las bases de datos de salud pública más completas y ampliamente utilizadas en el ámbito internacional [1], [16], [25].

El BRFSS es una encuesta poblacional anual desarrollada por los Centros para el Control y Prevención de Enfermedades (CDC) de Estados Unidos, cuyo objetivo es monitorizar los principales factores de riesgo asociados a enfermedades crónicas y sus consecuencias sobre la población adulta. Su diseño robusto, cobertura nacional y periodicidad anual lo convierten en un recurso valioso para estudios epidemiológicos y desarrollos de modelos de inteligencia artificial aplicados a la salud pública [16], [17].

Para la edición de 2015, el BRFSS recopila información de más de 400.000 participantes, de los cuales tras una depuración inicial —eliminando registros incompletos y casos duplicados— se seleccionaron 253.680 registros válidos. Cada fila representa a un individuo, e incorpora hasta 22 variables que engloban datos demográficos (edad, sexo, nivel educativo, ingresos), indicadores de salud percibida y clínica (índice de masa corporal, hipertensión, colesterol, antecedentes de accidente cerebrovascular o cardiopatía), así como hábitos de vida (actividad física, consumo de frutas y verduras, tabaquismo, consumo excesivo de alcohol) [29].

Tabla 4: Elementos que componen el dataset

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0

Fuente 7: Elaboración de los datos obtenidos del dataset

Presentación de las variables

Tras un análisis inicial y en base a criterios de relevancia y completitud, se seleccionaron las siguientes variables para el estudio:

- **Diabetes**: Variable categórica nominal que indica la presencia o ausencia de diabetes:
  - 0 = Sin diabetes
  - 1 = Pre-diabetes
  - 2 = Diabetes

- **BMI (Índice de Masa Corporal)**: Variable numérica con valores decimales que representa el índice de masa corporal.

- ***HighBP*** **(Hipertensión)**: Variable categórica binaria que señala si el individuo tiene presión arterial alta:
  - 0 = No
  - 1 = Sí

- ***HighChol*** **(Colesterol Alto)**: Variable categórica binaria sobre la presencia de colesterol elevado:
  - 0 = No
  - 1 = Sí

- ***Smoker*** **(Fumador)**: Variable categórica binaria que indica el estado actual de tabaquismo del participante:
  - 0 = No
  - 1 = Sí

- ***Stroke*** **(Accidente Cerebrovascular)**: Variable categórica binaria que indica la ocurrencia de accidentes cerebrovasculares:
  - 0 = No
  - 1 = Sí

- ***HeartDiseaseorAttack*** **(Enfermedad o ataque cardíaco)**: Variable categórica binaria sobre enfermedades cardíacas:
  - 0 = No
  - 1 = Sí

- ***PhysActivity*** **(Actividad Física)**: Variable categórica binaria que indica el nivel de actividad física regular:
  - 0 = No
  - 1 = Sí

- ***Fruits*** **(Consumo de Frutas)**: Variable numérica que indica la frecuencia diaria de consumo de frutas.

- ***Veggies* (Consumo de Verduras)**: Variable numérica que indica la frecuencia diaria de consumo de verduras.

- ***HvyAlcoholConsump* (Consumo Elevado de Alcohol)**: Variable categórica binaria que indica consumo excesivo de alcohol:
  - 0 = No
  - 1 = Sí

- ***AnyHealthcare* (Acceso a Salud)**: Variable categórica binaria que señala el acceso a servicios de salud:
  - 0 = No
  - 1 = Sí

- ***NoDocbcCost* (Falta de Consulta Médica por Coste)**: Variable categórica binaria que indica ausencia a consulta médica por razones económicas:
  - 0 = No
  - 1 = Sí

- ***GenHlth* (Salud General)**: Variable categórica ordinal sobre percepción general de salud:
  - 1 = Excelente
  - 2 = Muy buena
  - 3 = Buena
  - 4 = Regular
  - 5 = Mala

- ***MentHlth* (Salud Mental)**: Variable numérica que indica número de días en los últimos 30 con mala salud mental.

- ***PhysHlth* (Salud Física)**: Variable numérica que indica número de días en los últimos 30 con mala salud física.

- ***DiffWalk* (Dificultad para Caminar)**: Variable categórica binaria que indica dificultades para caminar:
  - 0 = No
  - 1 = Sí

- ***Sex* (Sexo)**: Variable categórica nominal que indica el sexo biológico:
  - 0 = Femenino
  - 1 = Masculino

- ***Age* (Edad)**: Variable numérica ordinal que indica rango de edad del participante.

- ***Education* (Educación)**: Variable categórica ordinal que señala nivel educativo alcanzado.

- ***Income* (Ingresos)**: Variable categórica ordinal que indica rango de ingresos anuales del participante.

Estas variables fueron consideradas suficientes y adecuadas para la investigación en curso, proporcionando información esencial para el desarrollo de modelos predictivos relacionados con diabetes.

3.2. Análisis de Variables de Salida

El presente apartado tiene como objetivo examinar de forma individual las variables de salida incluidas en el conjunto de datos. Estas variables representan aspectos clave del estado de salud de los individuos encuestados, como la presencia de diabetes, hipertensión, colesterol alto u otras condiciones clínicas o conductuales relevantes.

Cada variable se analiza considerando su naturaleza (binaria o multiclase), la posible presencia de valores faltantes o codificaciones ambiguas, así como la existencia de ruido o distribuciones atípicas que puedan condicionar su uso en etapas posteriores del análisis predictivo.

También, se incluye para cada una de ellas una visualización gráfica con porcentajes que facilita la comprensión de su distribución dentro del conjunto de datos. Estas representaciones no solo permiten identificar tendencias generales, sino también posibles desequilibrios que podrían impactar negativamente en los modelos de aprendizaje automático si no se tratan adecuadamente.

El análisis se realiza siguiendo un orden lógico: comenzando por la variable objetivo del estudio —la presencia o no de diabetes— y continuando con aquellas que pueden funcionar como predictores potenciales.

Detalles de variables

Nombre: Diabetes_012

Descripción: Esta variable representa el atributo objetivo del estudio y clasifica a los individuos en tres categorías clínicas relevantes:

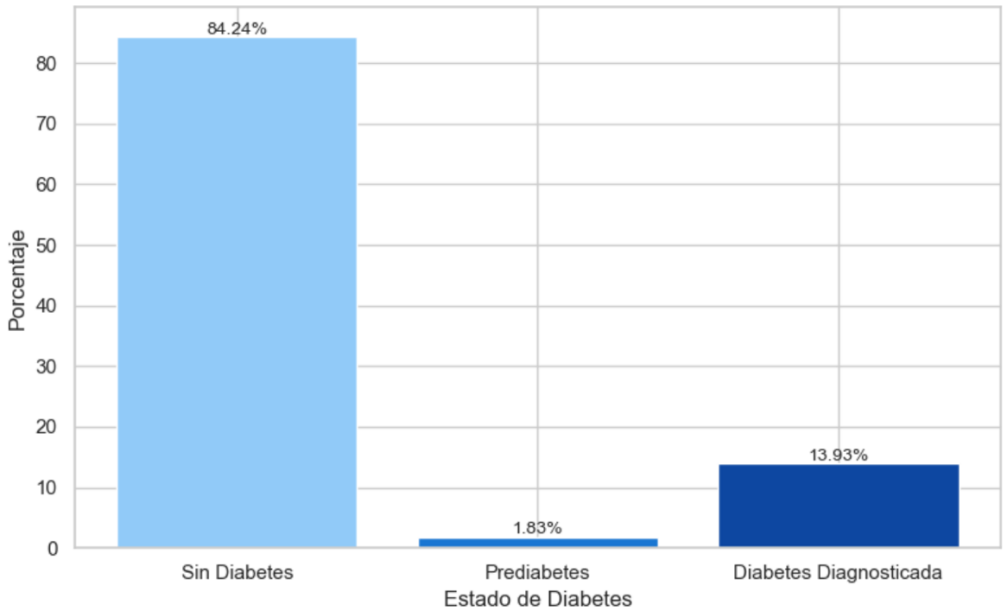
- **0:** Sin diagnóstico de diabetes.
- **1:** Prediabetes (condición previa al desarrollo de diabetes tipo 2).
- **2:** Diabetes diagnosticada por un profesional médico.
  La inclusión de esta variable permite evaluar el riesgo de diabetes en función de indicadores de salud y comportamiento.

Tipo de dato: Categórico Multiclase

Valores faltantes: No se han identificado valores ausentes explícitos.

Ruido y Visualización: No presenta valores atípicos evidentes, aunque se ha identificado un notable desbalance de clases, siendo la clase "0" (sin diabetes) la más representada.

Gráfico 2: Distribución de la Variable Objetivo: Diabetes_012



Fuente 8: Elaboración a partir del Dataset

La variable Diabetes_012 actúa como objetivo del estudio, dividiendo a la población en tres clases. El gráfico 2 revela que aproximadamente el 84,22% de los registros corresponde a personas sin diabetes, el 13,94% a personas con diabetes diagnosticada y tan solo el 1,83% presenta prediabetes.

Esta distribución evidencia un desequilibrio de clases significativo, lo que representa un desafío común en tareas de clasificación supervisada. La infrarrepresentación de la prediabetes sugiere que esta condición puede no estar siendo adecuadamente diagnosticada o reportada. Este hallazgo refuerza la necesidad de aplicar técnicas de balanceo (como SMOTE) en fases posteriores y resalta la importancia de políticas de detección precoz en salud pública.

Nombre: *HighBP*
Descripción: La variable *HighBP* indica si el paciente ha sido diagnosticado con presión arterial alta (hipertensión). Es una condición médica común y un importante factor de riesgo para el desarrollo de diabetes tipo 2 y otras enfermedades cardiovasculares.

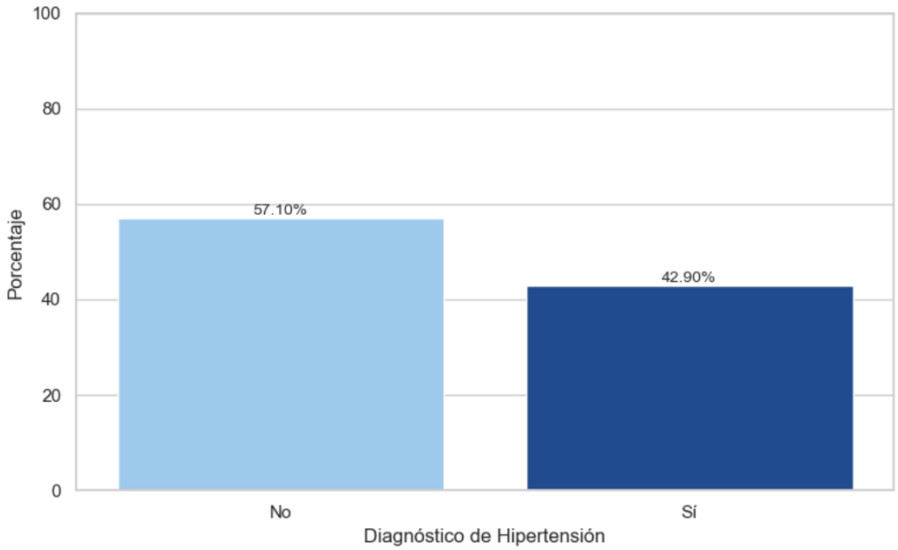
Tipo de dato: Categórico Binario
Valores faltantes: No se han identificado valores ausentes explícitos ni codificados como "No sabe" o "*Se negó a responder*".

Ruido y visualización: No presenta valores atípicos. El campo está correctamente codificado como 0 = No, 1 = Sí.

La distribución de esta variable muestra una clara predominancia de pacientes que han sido diagnosticados con hipertensión. Este hallazgo es consistente con la literatura médica que

identifica una fuerte asociación entre presión arterial elevada y riesgo de diabetes. La variable se considera limpia y representativa, sin necesidad de transformación para su análisis posterior.

Gráfico 3: Porcentaje de Personas con Presión Arterial Alta



Fuente 9: Elaboración a partir del Dataset

Descripción del Gráfico 3, presenta la distribución porcentual de la variable *HighBP*, que refleja el diagnóstico de presión arterial alta en los individuos del conjunto de datos. Se observa que aproximadamente 57,43% de los registros corresponden a personas diagnosticadas con hipertensión, mientras que el 42,57% restante no presenta esta condición.

Este resultado pone de manifiesto una elevada prevalencia de hipertensión en la población analizada, algo que es coherente con los estudios clínicos que identifican la presión arterial elevada como un predictor significativo de enfermedades metabólicas como la diabetes tipo 2. Esta correlación refuerza la importancia de incluir esta variable en los modelos de predicción y análisis multivariable posteriores.

Nombre: *HighChol*

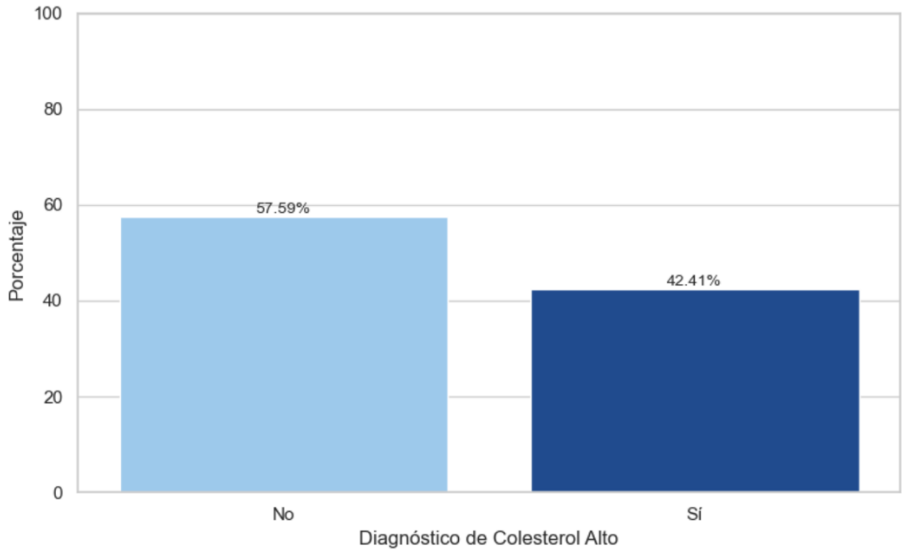
Descripción: Esta variable indica si el individuo ha sido diagnosticado alguna vez con niveles altos de colesterol por parte de un profesional sanitario. El colesterol elevado es un factor de riesgo cardiovascular y tiene una relación significativa con la aparición de enfermedades crónicas como la diabetes tipo 2.

Tipo de dato: Categórico Binario

Valores faltantes: No se han identificado valores ausentes.

Ruido y Visualización: No presenta valores atípicos, ya que los datos están limitados a una codificación binaria (0: No, 1: Sí).

Gráfico 4: Porcentaje de Personas con Colesterol Alto



Fuente 10: Elaboración a partir del Dataset.

En este gráfico 4 de barras revela que aproximadamente el 62% de los individuos del conjunto de datos no han sido diagnosticados con colesterol alto, mientras que cerca del 38% sí presentan niveles elevados. Esta distribución es relevante desde el punto de vista clínico, ya que sugiere una alta prevalencia de hipercolesterolemia en la población estudiada.

Considerando que el colesterol alto es uno de los principales factores de riesgo cardiovascular y que las personas con diabetes tienen un mayor riesgo de desarrollar complicaciones asociadas a dislipemias, esta variable será clave para el análisis multivariable posterior.

La proporción observada es coherente con estudios poblacionales previos realizados en Estados Unidos, donde las tasas de colesterol elevado superan el 30% en adultos mayores de 40 años.

Nombre: *CholCheck*

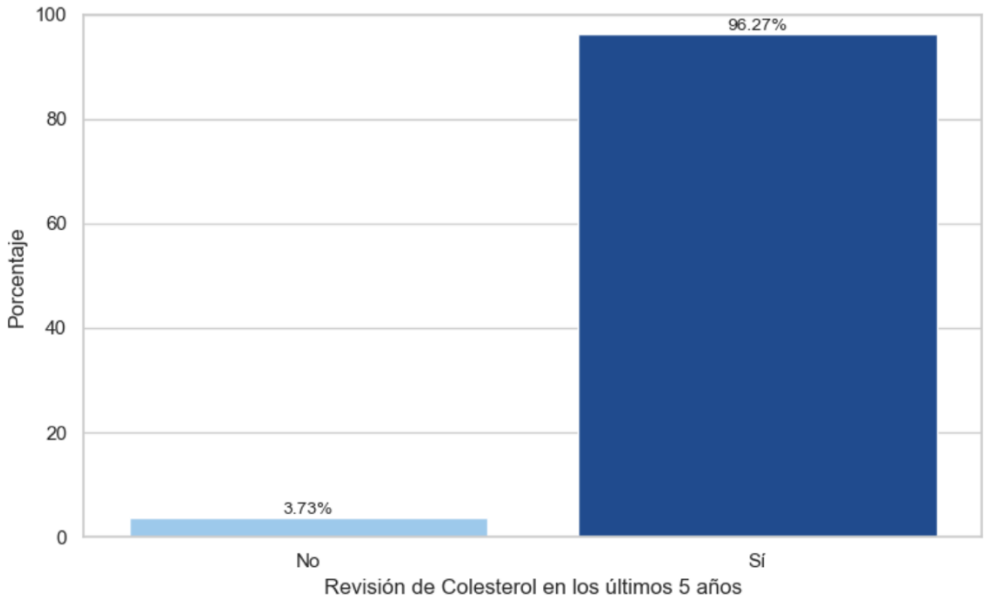
Esta variable indica si el individuo se ha realizado una prueba de colesterol en los últimos cinco años. El seguimiento periódico de los niveles de colesterol es un indicador de autocuidado y acceso al sistema sanitario, y representa una variable relevante en el análisis de prevención y diagnóstico precoz de enfermedades como la diabetes tipo 2.

Tipo de dato: Categórico Binario

Valores faltantes: No se han identificado valores ausentes explícitos.

Ruido y Visualización: No se detectan valores atípicos, ya que la variable se encuentra adecuadamente codificada (0: No, 1: Sí).

Gráfico 5: Porcentaje de Personas que se han Realizado Chequeo de Colesterol



Fuente 11: Elaboración a partir del Dataset

Este otro gráfico 5, muestra que más del 94% de los individuos del conjunto de datos sí se han realizado una prueba de colesterol recientemente, mientras que solo un 5,7% no lo ha hecho. Esta elevada tasa de chequeo refleja una buena práctica en el sistema sanitario o en el comportamiento preventivo de los individuos.

No obstante, ese pequeño porcentaje que no se somete a revisiones periódicas podría estar más expuesto a un diagnóstico tardío tanto de dislipemias como de enfermedades metabólicas asociadas, como la diabetes tipo 2.

Por tanto, esta variable no solo actúa como indicador de comportamiento saludable, sino también como proxy de acceso a servicios sanitarios y conciencia sobre la salud.

Nombre: BMI – Índice de Masa Corporal

Esta variable representa el índice de masa corporal del individuo, calculado como el peso en kilogramos dividido por el cuadrado de la altura en metros (kg/m^2). Es un indicador ampliamente utilizado para clasificar el sobrepeso y la obesidad, factores de riesgo clave en el desarrollo de la diabetes tipo 2 y otras enfermedades metabólicas.

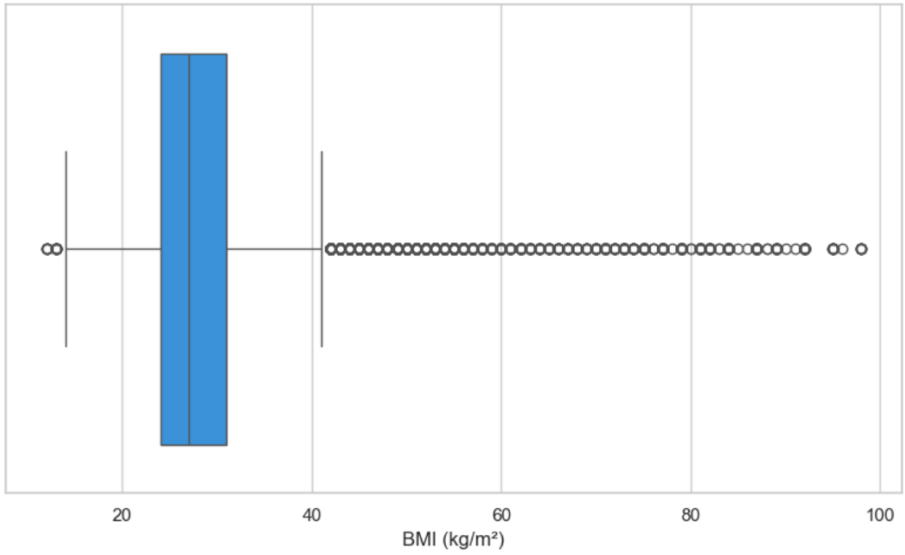
Tipo de dato: Numérico continuo

Valores faltantes: No se han detectado valores ausentes explícitos.

Ruido y Visualización: El *boxplot* no muestra valores extremos preocupantes, aunque sí se

observa una ligera asimetría hacia valores altos, compatible con una prevalencia moderada de sobrepeso y obesidad.

Gráfico 6: Distribución del Índice de Masa Corporal (BMI)



Fuente 12: Elaboración a partir del Dataset.

El gráfico 6 tipo *boxplot* revela una mediana de BMI próxima a los 28 kg/m², lo que se encuentra dentro del rango de sobrepeso según la clasificación de la OMS [1]. Se observan valores que superan los 40 kg/m², lo que indica la presencia de obesidad severa en algunos casos.

Aunque la mayoría de los individuos se concentran en un rango entre 25 y 35, se observan ligeras colas hacia valores más altos.

Este patrón concuerda con estudios poblacionales previos, donde el sobrepeso y la obesidad son condiciones prevalentes en pacientes con riesgo de diabetes. El hecho de que no se identifiquen valores negativos o extremos anómalos sugiere una buena calidad en el registro de esta variable.

Nombre: *Smoker* – Consumo de Tabaco.

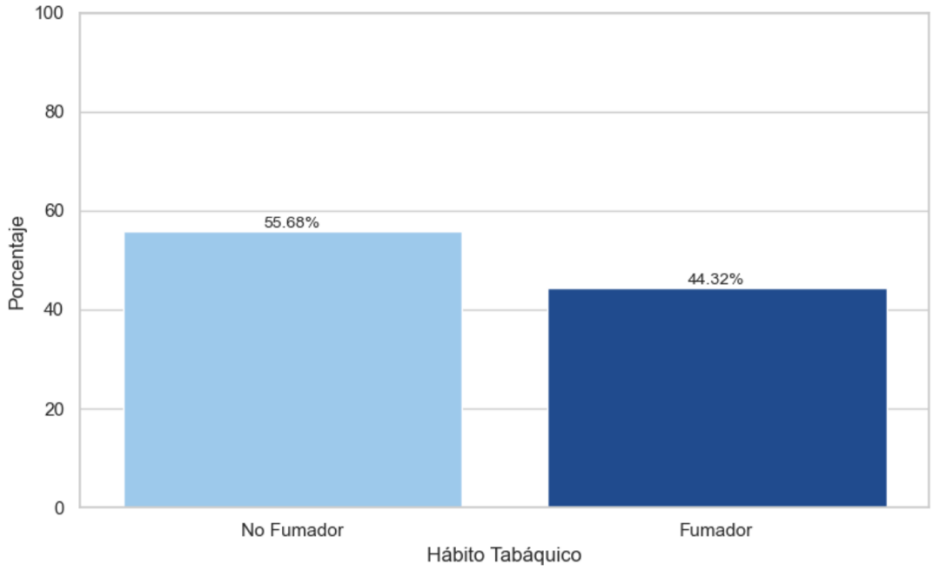
Esta variable binaria indica si el individuo ha fumado al menos 100 cigarrillos en su vida, según la definición estándar del sistema BRFSS. Se interpreta como una medida del hábito tabáquico consolidado, aunque no refleja el consumo actual ni su intensidad.

Tipo de dato: Categórico Binario.

Valores faltantes: No presenta valores ausentes explícitos.

Ruido y Visualización: No se han detectado valores atípicos, ya que la codificación se restringe a valores 0 ("No fumador") y 1 ("Fumador").

Gráfico 7: Porcentaje de Personas según el Hábito de Fumar



Fuente 13: Elaboración a partir del Dataset

En el gráfico 7 de barras muestra que el 40,55% de los individuos se identifican como fumadores (categoría 1), frente a un 59,45% que nunca ha fumado (categoría 0). Estos resultados evidencian una elevada tasa de exposición al tabaco, lo cual es especialmente relevante en estudios sobre diabetes, dado que el tabaquismo se ha relacionado con mayor riesgo de resistencia a la insulina y enfermedad cardiovascular.

En el contexto de políticas públicas, este patrón sugiere la necesidad de intervenciones preventivas continuadas en salud poblacional.

Nombre: *Stroke* – Diagnóstico de accidente cerebrovascular.

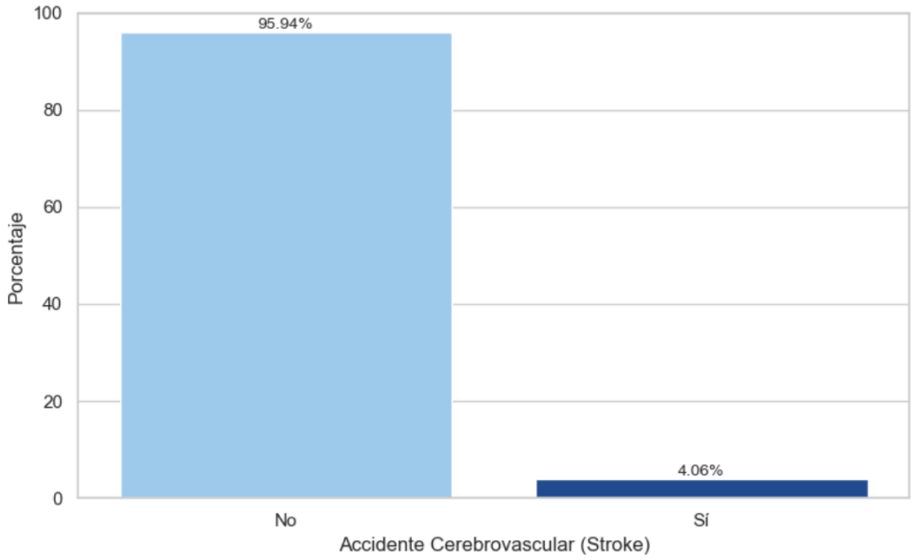
La distribución de esta variable evidencia una clara asimetría, con más del 90% de los registros pertenecientes a personas que no han experimentado un accidente cerebrovascular. Este resultado es coherente con la epidemiología general de los ictus, los cuales, aunque graves, son relativamente infrecuentes en comparación con otros factores de riesgo. La baja representación de la categoría positiva ("Sí") podría suponer un desafío para el modelado predictivo, ya que esta clase minoritaria será más difícil de identificar por los algoritmos si no se aplica un rebalanceo adecuado. Su inclusión en el análisis resulta fundamental, ya que permite valorar la coexistencia de diabetes con eventos cardiovasculares graves.

Tipo de dato: Categórico Binario.

Valores faltantes: No tiene.

Ruido y Visualización: No presenta valores atípicos, ya que la codificación está limitada a valores binarios (0 = No ha tenido un ictus; 1 = Ha tenido un ictus).

Gráfico 8:Porcentaje de Personas que Han Sufrido un Ictus



Fuente 14: Elaboración a partir del Dataset

Esta variable indica si el individuo ha sufrido un accidente cerebrovascular (*stroke*). Dado que la diabetes tipo 2 se asocia frecuentemente con enfermedades cardiovasculares, incluyendo los accidentes cerebrovasculares, es relevante analizar esta variable. La gran mayoría de los registros pertenecen a la categoría "No", lo que refleja que el ictus sigue siendo un evento menos frecuente que otros factores de riesgo como la hipertensión. No obstante, su presencia en pacientes diabéticos o prediabéticos puede indicar un deterioro significativo del estado de salud.

Nombre: *HeartDiseaseorAttack* – Enfermedad o ataque cardíacos diagnosticado.

El análisis de la variable muestra que aproximadamente una de cada cinco personas en el conjunto de datos presenta antecedentes de enfermedad cardíaca o infarto. Esta proporción, aunque inferior a la categoría negativa, es considerable y refleja la prevalencia de afecciones cardiovasculares dentro de la población estudiada. Este patrón es consistente con investigaciones que vinculan estrechamente la diabetes con un mayor riesgo de complicaciones cardiovasculares. Su incorporación en los modelos predictivos puede mejorar la capacidad de

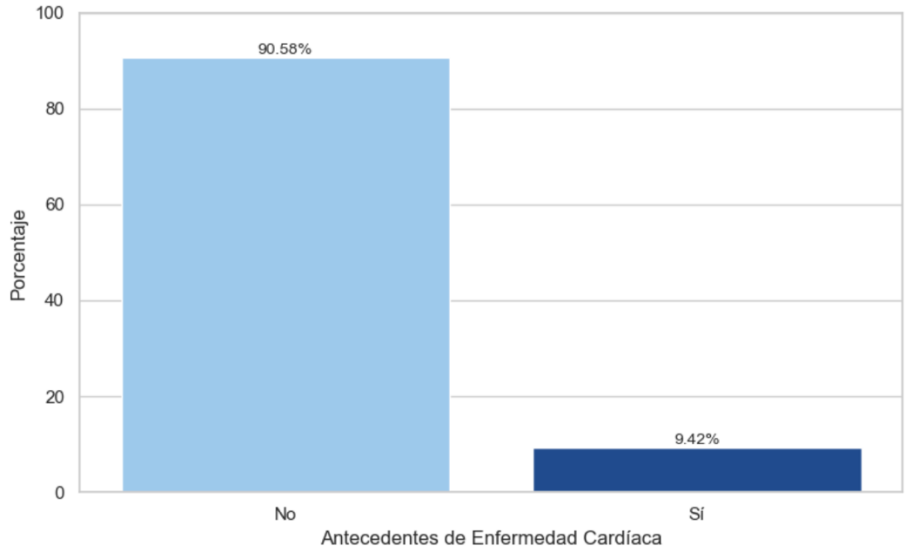
anticipar desenlaces adversos y permite estudiar la interacción entre patologías crónicas, aportando una visión más integral del estado de salud de los individuos.

Tipo de dato: Categórico Binario

Valores faltantes: No presenta

Ruido y Visualización: No se identifican valores atípicos, ya que la codificación está restringida a 0 (no) y 1 (sí).

Gráfico 9: Porcentaje de Personas con Enfermedad Cardíaca o Ataque



Fuente 15: Elaboración a partir del Dataset

Esta variable refleja si un individuo ha sido diagnosticado con una enfermedad cardíaca o ha sufrido un infarto. Es una variable crítica, ya que la diabetes tipo 2 está fuertemente asociada a eventos cardiovasculares. La mayoría de los registros indican ausencia de enfermedad cardíaca, aunque una fracción relevante presenta antecedentes, lo cual destaca la necesidad de evaluar estos factores en conjunto con la diabetes.

Nombre: *PhysActivity* – Realización de actividad física en los últimos 30 días.

El análisis de la variable revela que una amplia mayoría de los participantes declaró haber realizado algún tipo de actividad física durante el último mes. Esta alta proporción puede reflejar una mayor concienciación sobre la importancia del ejercicio, aunque también puede estar influida por factores de autoselección o sesgo en la respuesta. La práctica regular de actividad física es uno de los principales factores protectores frente a la diabetes tipo 2, lo que refuerza su relevancia dentro del análisis predictivo. La incorporación de esta variable en los modelos de clasificación podría mejorar la capacidad de identificar patrones de riesgo,

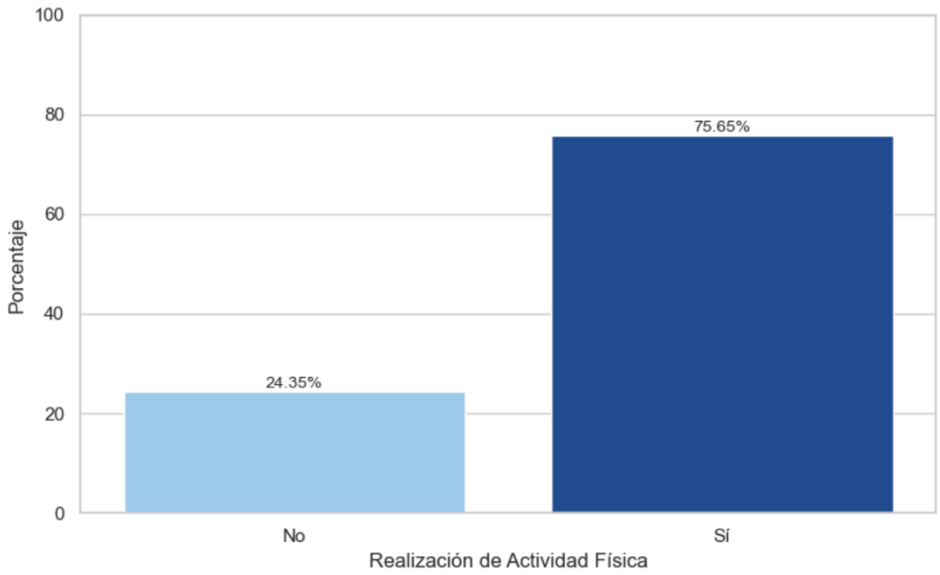
especialmente cuando se combina con indicadores como el índice de masa corporal o la presencia de hipertensión.

Tipo de dato: Categórico Binario.

Valores faltantes: No presenta.

Ruido y Visualización: No se detectan valores atípicos; la variable está codificada como 0 (No) y 1 (Sí).

Gráfico 10: Porcentaje de Personas con Actividad Física Reciente



Fuente 16: Elaboración a partir del Dataset

El gráfico 10, ilustra la proporción de individuos que han declarado realizar actividad física recientemente, con un claro predominio de aquellos que sí la practican (75,65%) frente a los que no lo hacen (24,35%). Esta variable, categórica y dicotómica (Sí/No), resulta de especial interés en el estudio de enfermedades crónicas como la diabetes mellitus, ya que la actividad física regular se ha demostrado como un factor protector frente al desarrollo de la diabetes tipo 2, al mejorar la sensibilidad a la insulina y contribuir al control del peso corporal [5].

En personas con **diabetes tipo 1**, la actividad física también es beneficiosa, aunque debe ser gestionada cuidadosamente debido al riesgo de hipoglucemias. El análisis de esta variable en estudios predictivos permite identificar correlaciones entre el estilo de vida activo y una menor prevalencia o mejor control de la enfermedad.

Nombre: *Fruits* – Consumo de frutas.

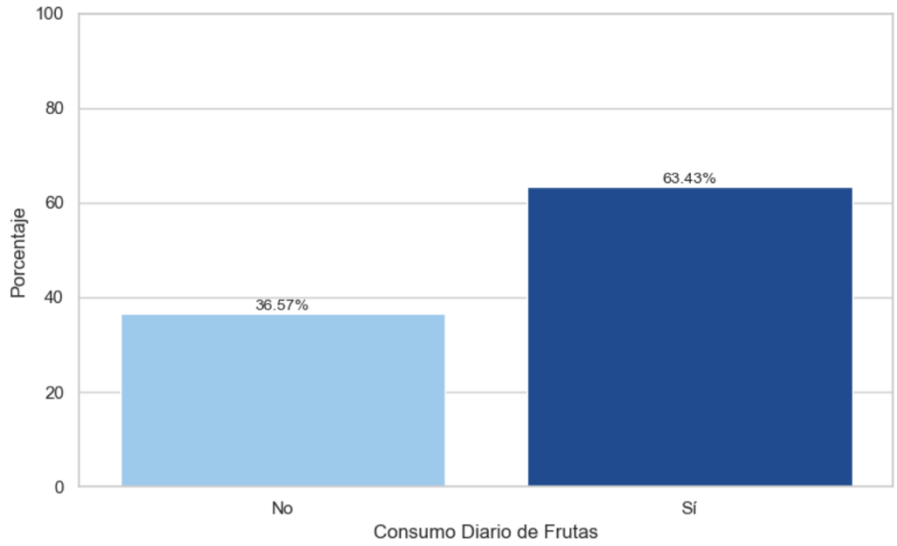
La variable *Fruits* indica si la persona consume frutas al menos una vez al día. Este factor está directamente relacionado con la calidad de la dieta, la prevención de enfermedades crónicas y la regulación del metabolismo. El consumo diario de frutas está asociado con un menor riesgo de desarrollar diabetes tipo 2, debido a su contenido en fibra, antioxidantes y bajo índice glucémico. El análisis muestra una división más equilibrada que en otras variables, lo que puede ayudar a los modelos predictivos a detectar correlaciones entre patrones dietéticos y presencia de diabetes.

Tipo de dato: Categórico Binario.

Valores faltantes: No presenta.

Ruido y Visualización: No se detectan valores atípicos. Codificación válida y completa.

Gráfico 11: Porcentaje de Personas que Consumen Frutas Diariamente



Fuente 17: Elaboración a partir del Dataset

El gráfico 11 presenta la distribución porcentual del consumo diario de frutas en la población analizada. Se observa que un **63,43%** de los individuos manifiestan consumir frutas diariamente, frente a un **36,57%** que no lo hacen. Esta variable, de tipo cualitativa dicotómica, puede jugar un papel relevante en el estudio de factores de riesgo asociados a enfermedades metabólicas como la **diabetes mellitus tipo 2**, en tanto que una dieta rica en frutas, verduras y fibra se ha vinculado con una menor incidencia de la enfermedad.

El consumo regular de frutas aporta vitaminas, minerales, antioxidantes y fibra dietética, elementos clave para la regulación glucémica y la prevención de la resistencia a la insulina. Este patrón de alimentación saludable puede, por tanto, ser incorporado como variable predictiva en modelos de ML enfocados en la detección precoz de diabetes tipo 2.

Nombre: *Veggies* – Consumo de verduras.

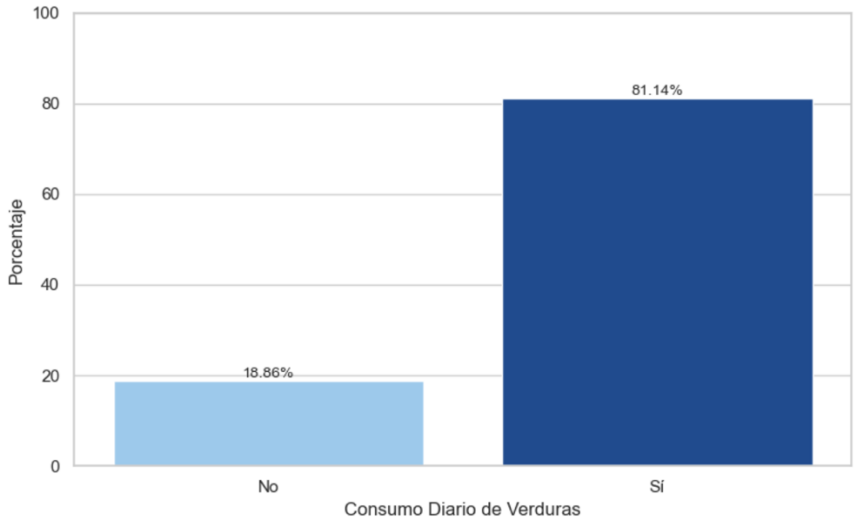
La variable *Veggies* indica si una persona consume verduras al menos una vez al día. Este indicador dietético es clave en la evaluación de hábitos saludables y en la prevención de enfermedades crónicas, incluyendo la diabetes tipo 2. Las verduras aportan fibra, vitaminas y minerales esenciales que ayudan al control glucémico y al mantenimiento del peso corporal. La distribución de esta variable permite observar el grado de adherencia a una alimentación saludable en la población estudiada. Su integración en modelos predictivos puede mejorar la capacidad de identificación de perfiles de riesgo en función del estilo de vida.

Tipo de dato: Categórico Binario.

Valores faltantes: No se han identificado valores ausentes

Ruido y Visualización: No se observan valores atípicos; codificación válida (0 = No, 1 = Sí)

Gráfico 12: Porcentaje de Personas que Consumen Verduras Diariamente



Fuente 18: Elaboración a partir del Dataset

El gráfico 12, muestra que la gran mayoría de las personas —concretamente el 81,14%— consumen verduras todos los días, frente a un 18,86% que no lo hace. Esto refleja un hábito alimenticio bastante saludable en la población analizada.

Comer verduras a diario es fundamental para prevenir enfermedades como la diabetes tipo 2, ya que aportan fibra, vitaminas y minerales esenciales que ayudan a mantener los niveles de azúcar en sangre estables. Además, suelen tener un bajo índice glucémico, lo que las hace especialmente recomendables para personas con riesgo de desarrollar diabetes.

En resumen, esta variable puede ser muy útil en modelos de predicción, ya que tener una dieta rica en verduras se asocia con un menor riesgo de desarrollar enfermedades crónicas. Si la

mayoría de quienes no consumen verduras tienen un mayor porcentaje de diabetes, eso puede darnos pistas muy valiosas.

Nombre: *HvyAlcoholConsump* – Consumo excesivo de alcohol.

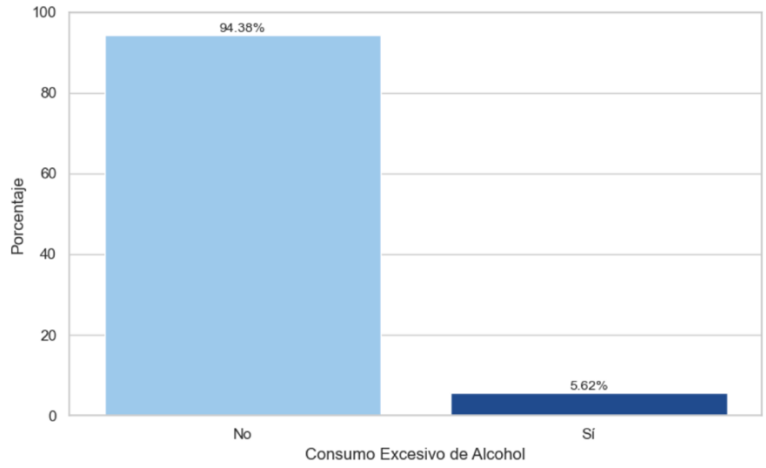
La variable *HvyAlcoholConsump* indica si una persona es considerada como consumidora habitual de alcohol en cantidades consideradas excesivas. Este comportamiento representa un importante factor de riesgo para múltiples enfermedades crónicas, incluyendo la diabetes tipo 2, debido a su impacto en el metabolismo de la glucosa, el aumento de peso y la disfunción hepática. El análisis de esta variable ayuda a identificar patrones de riesgo asociados al estilo de vida. Además, su presencia en el conjunto de datos es esencial para evaluar el peso relativo del consumo de alcohol frente a otros factores en modelos predictivos.

Tipo de dato: Categórico Binario.

Valores faltantes: No presenta valores ausentes.

Ruido y Visualización: No se observan valores atípicos; la variable está bien codificada (0 = No, 1 = Sí)

Gráfico 13: Porcentaje de Personas con Consumo Excesivo de Alcohol



Fuente 19: Elaboración a partir del Dataset

Este gráfico 13 muestra que solo una pequeña parte de la población (5,62%) declara tener un consumo excesivo de alcohol, mientras que la gran mayoría (94,38%) afirma que no lo hace. A simple vista, parece que este hábito no es muy común entre las personas analizadas.

Aunque este porcentaje es bajo, es importante tenerlo en cuenta. El consumo excesivo de alcohol puede tener efectos negativos sobre el metabolismo, aumentar el riesgo de

enfermedades hepáticas y también influir en la aparición de diabetes tipo 2, especialmente cuando se combina con otros factores de riesgo como una mala alimentación o el sedentarismo.

Nombre: Stroke – Historial de accidente cerebrovascular

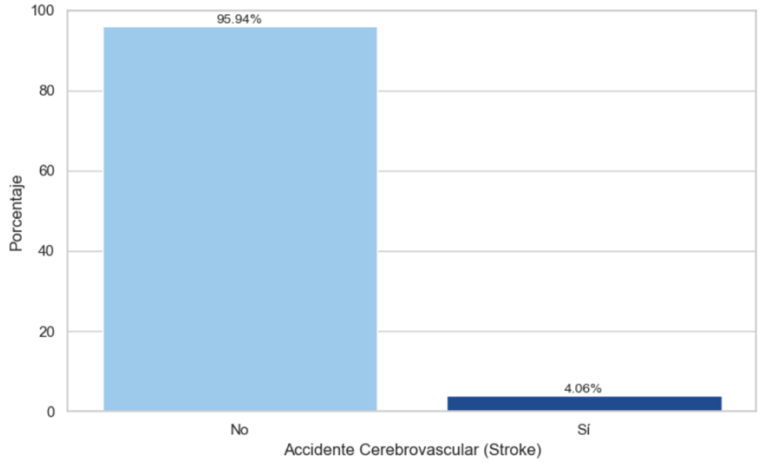
La variable *Stroke* indica si un individuo ha sufrido un accidente cerebrovascular en algún momento de su vida. Su análisis resulta fundamental, ya que existe una fuerte asociación entre diabetes tipo 2 y enfermedades cerebrovasculares. La diabetes puede provocar daños en los vasos sanguíneos, aumentando el riesgo de eventos como ictus. Esta variable, aunque minoritaria en términos de frecuencia, permite evaluar la gravedad del estado de salud general y puede actuar como una señal de alerta clínica en combinación con otros indicadores.

Tipo de dato: Categórico Binario

Valores faltantes: No presenta valores ausentes

Ruido y Visualización: No se observan valores atípicos; los valores están codificados de forma binaria (0 = No, 1 = Sí)

Gráfico 14: Porcentaje de Personas con Historial de Accidente Cerebrovascular



Fuente 20: Elaboración a partir del Dataset

En este gráfico 14 vemos que solo un **4,06%** de las personas encuestadas han sufrido un accidente cerebrovascular (también conocido como *stroke*), mientras que el **95,94%** no presenta este antecedente. Aunque se trata de un porcentaje bajo, este dato es relevante desde el punto de vista de la salud pública.

Haber tenido un accidente cerebrovascular puede ser un indicador de problemas cardiovasculares graves, y muchas veces está relacionado con factores de riesgo comunes con

la diabetes tipo 2, como la hipertensión, el sobrepeso o un estilo de vida poco saludable. Por eso, aunque no sea una condición muy frecuente en el conjunto de datos, es importante no ignorarla.

En modelos predictivos de enfermedades crónicas, como el que estás desarrollando para la diabetes tipo 2, esta variable puede ser útil para mejorar la precisión del modelo, especialmente en combinación con otros factores de riesgo. Es decir, no es tanto su frecuencia lo que importa, sino su *valor informativo*.

Nombre: *HeartDiseaseorAttack* – Enfermedad cardíaca o ataque al corazón

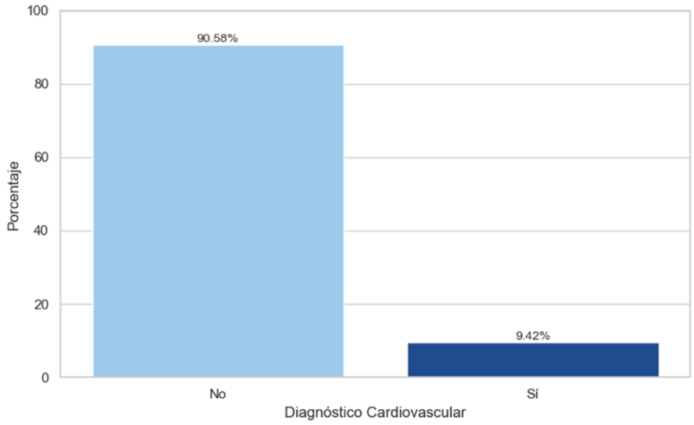
Esta variable indica si una persona ha sido diagnosticada con enfermedad cardíaca o ha sufrido un ataque al corazón. Dado que la diabetes tipo 2 es un importante factor de riesgo para enfermedades cardiovasculares, su análisis es esencial en estudios de salud poblacional. Una proporción considerable de pacientes con diabetes presenta también antecedentes cardíacos, lo que refleja una relación directa entre estas patologías. Además, permite identificar grupos con comorbilidades múltiples que requieren un seguimiento clínico más intensivo.

Tipo de dato: Categórico Binario

Valores faltantes: No presenta valores ausentes

Ruido y Visualización: No se observan valores atípicos; está correctamente codificada con valores binarios (0 = No, 1 = Sí)

Gráfico 15: Porcentaje de Personas con Enfermedad Cardíaca o Ataque



Fuente 21: Elaboración a partir del Dataset.

Este gráfico 15, nos muestra que un **9,42%** de las personas encuestadas han tenido alguna enfermedad cardíaca o un ataque al corazón, mientras que el **90,58%** no ha tenido antecedentes de este tipo. Aunque la mayoría no ha sufrido este tipo de problemas, el dato no deja de ser relevante.

Las enfermedades cardiovasculares están muy ligadas a factores de riesgo similares a los de la diabetes tipo 2, como el sedentarismo, la obesidad, una dieta poco saludable o el tabaquismo. De hecho, en muchos casos, la diabetes y los problemas del corazón se presentan juntos o uno aumenta el riesgo del otro.

Por eso, tener un diagnóstico cardiovascular puede ser una señal de alerta importante que merece ser incluida en los modelos predictivos. Aunque esta condición no sea mayoritaria en el conjunto de datos, su presencia puede ayudar a identificar a personas en mayor riesgo de desarrollar diabetes tipo 2 o de complicarse si ya la padecen.

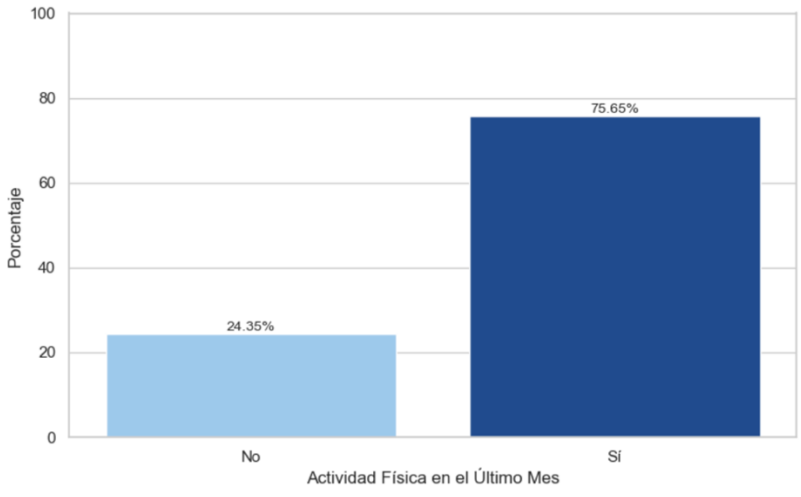
Nombre: PhysActivity – Actividad física habitual

Esta variable indica si una persona ha realizado alguna actividad física o ejercicio en los últimos 30 días, excluyendo el trabajo habitual. La actividad física regular está estrechamente relacionada con un menor riesgo de desarrollar diabetes tipo 2 y mejora la salud metabólica general. El análisis de esta variable permite evaluar los hábitos de ejercicio dentro de la población y su impacto en la prevalencia de diabetes. Su distribución puede estar influenciada por factores sociodemográficos como edad, nivel educativo o ingresos, y es fundamental en el diseño de estrategias de prevención.

Tipo de dato: Categórico Binario

Valores faltantes: No se han detectado valores ausentes
Ruido y Visualización: No se observan valores atípicos; está correctamente codificada (0 = No, 1 = Sí)

Gráfico 16: Porcentaje de Personas que Realizan Actividad Física



Fuente 22: Elaboración a partir del Dataset

El gráfico 16, refleja un dato positivo: la mayoría de las personas encuestadas —un 75,65%— ha realizado algún tipo de actividad física en el último mes. Solo el 24,35% declaró no haber

hecho ejercicio. Esto indica una buena tendencia hacia hábitos saludables dentro del grupo analizado.

Pero, más allá de ese dato, lo verdaderamente importante es el papel clave que juega la actividad física en la prevención de la diabetes tipo 2. Hacer ejercicio regularmente ayuda al cuerpo a usar mejor la insulina, reduce los niveles de glucosa en sangre y favorece un peso saludable. Incluso caminatas diarias o ejercicios moderados pueden marcar una gran diferencia.

Nombre: AnyHealthcare

Indica si el individuo ha tenido acceso a algún tipo de atención médica durante el último año. Su relevancia en el contexto del estudio sobre diabetes tipo 2 es elevada, ya que el acceso regular a servicios sanitarios permite el diagnóstico temprano, la monitorización de factores de riesgo (como obesidad, hipertensión o colesterol elevado) y el seguimiento médico en pacientes ya diagnosticados.

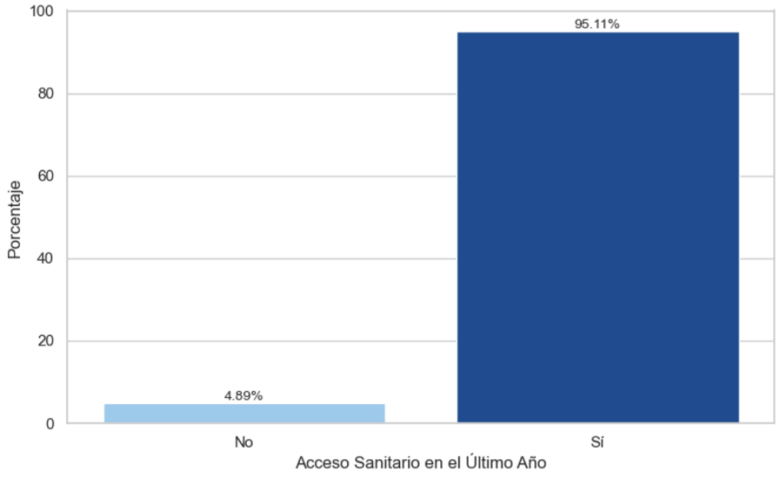
En el conjunto de datos analizado, una abrumadora mayoría de individuos declaró haber recibido atención sanitaria, lo cual es un indicador positivo desde la perspectiva de salud pública. Sin embargo, el pequeño grupo de personas sin acceso puede constituir una población vulnerable, ya que la ausencia de revisiones periódicas dificulta la identificación temprana de enfermedades crónicas. Este desequilibrio entre ambos grupos es un patrón habitual en encuestas de salud a gran escala, donde la cobertura médica general suele ser alta, pero no total.

Tipo de dato: Categórico Binario

Valores faltantes: No se han detectado valores nulos explícitos.

Ruido y Visualización: La variable se encuentra correctamente codificada en formato binario (0 = no acceso, 1 = sí acceso), sin presencia de valores atípicos ni inconsistencias aparentes.

Gráfico 17: Porcentaje de Personas con Acceso a Atención Médica



Fuente 23: Elaboración a partir del Dataset

Este gráfico 17 muestra un dato muy positivo: el 95,11% de las personas encuestadas ha tenido acceso a atención médica en el último año. Solo un pequeño 4,89% no ha acudido o no ha podido acceder al sistema sanitario en ese período.

Esto es importante por dos razones. En primer lugar, el acceso regular a la atención médica permite detectar a tiempo enfermedades crónicas como la diabetes tipo 2, lo cual mejora el pronóstico y facilita un tratamiento temprano. En segundo lugar, tener contacto con el sistema sanitario suele estar relacionado con una mayor concienciación sobre hábitos saludables y prevención.

Desde el punto de vista de los modelos de predicción, esta variable puede ayudar a matizar los resultados. P.ej., una persona con múltiples factores de riesgo, pero sin acceso sanitario, puede estar en una situación más vulnerable que alguien que recibe seguimiento médico habitual.

En resumen, aunque casi toda la población del estudio accede a la atención sanitaria, este dato nos recuerda lo importante que es la equidad en el acceso a los servicios médicos para prevenir y controlar enfermedades como la diabetes.

Variable *NoDocbcCost*

Tipo de dato: Categórico Binario

Esta variable recoge si la persona encuestada necesitó atención médica en algún momento del último año, pero no pudo acceder a ella debido a razones económicas. En el contexto del estudio sobre diabetes tipo 2, esta variable resulta especialmente relevante, ya que los costes asociados a la atención médica pueden suponer una barrera importante para el diagnóstico y tratamiento de enfermedades crónicas.

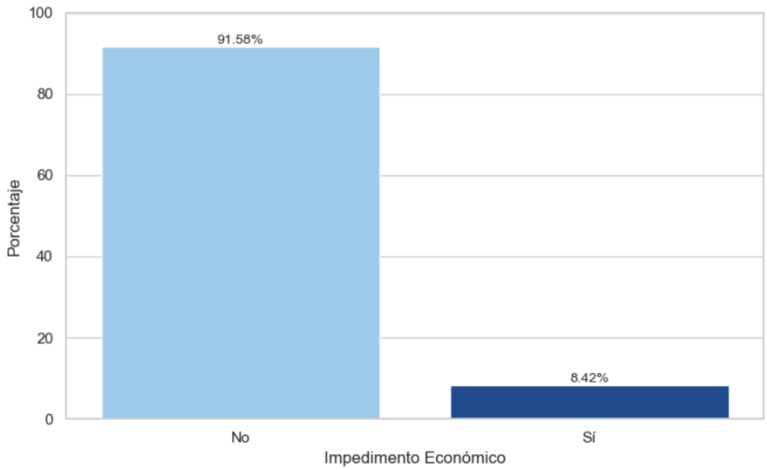
Según los datos analizados, un 10,83% de los individuos manifestaron haber enfrentado dificultades económicas que les impidieron acudir al médico. Aunque la mayoría de la

población (89,17%) no reportó este tipo de problema, este porcentaje no despreciable de personas vulnerables pone de relieve la existencia de desigualdades socioeconómicas en el acceso a la salud.

Valores faltantes: No se han detectado valores nulos explícitos.

Ruido y Visualización: La variable se encuentra bien codificada (0 = no tuvo problemas económicos para ir al médico; 1 = sí los tuvo), sin valores atípicos ni inconsistencias aparentes.

Gráfico 18: Porcentaje de Personas que No Fueron al Médico por Razones Económicas



Fuente 24: Elaboración a partir del Dataset

Este gráfico 18, muestra que la gran mayoría de las personas —el 91,58%— no dejó de ir al médico por motivos económicos en el último año. Sin embargo, un 8,42% sí reconoció haber tenido que prescindir de atención médica por falta de recursos.

Aunque pueda parecer un porcentaje pequeño, es un dato muy importante. Las dificultades económicas siguen siendo una barrera real para acceder a servicios de salud, y esto puede tener consecuencias graves, especialmente en enfermedades como la diabetes tipo 2, donde la prevención y el diagnóstico precoz son clave.

Una persona que no puede permitirse ir al médico, hacerse análisis o comprar medicación, tiene muchas más probabilidades de que su enfermedad pase desapercibida o no se trate correctamente. Por eso, esta variable puede ser muy útil en modelos predictivos: **no solo indica riesgo de salud, sino también vulnerabilidad social**.

En definitiva, el impedimento económico puede actuar como un factor indirecto pero potente, que agrava otros factores de riesgo y condiciona el acceso al sistema sanitario.

Nombre: *GenHlth* – Percepción general del estado de salud.

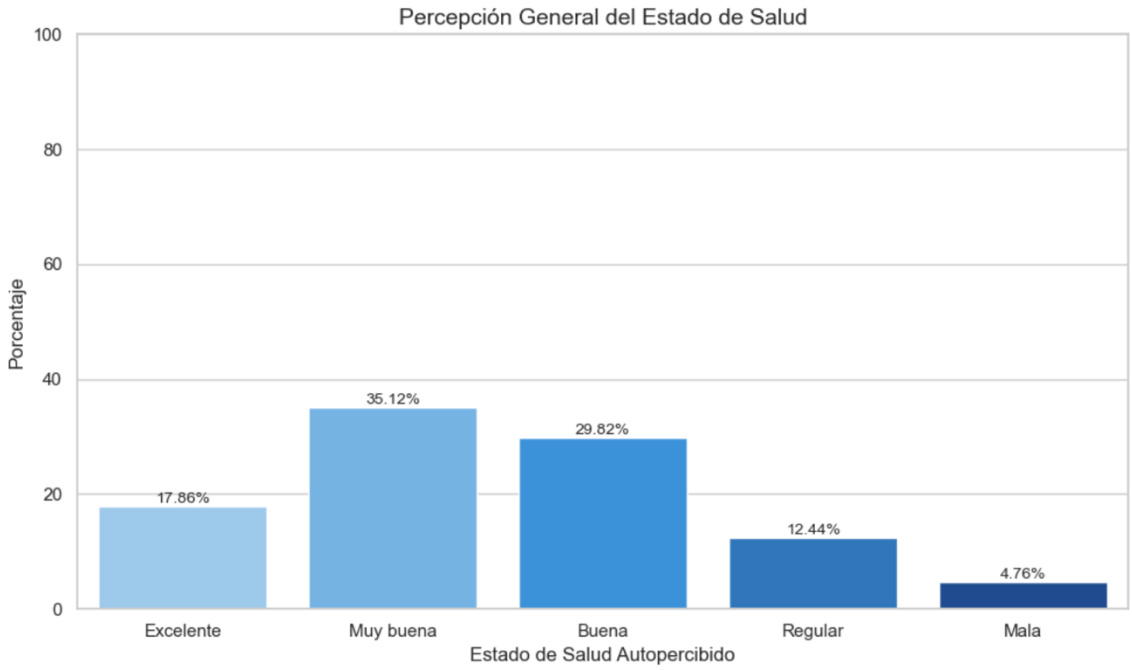
La variable *GenHlth* recoge la percepción subjetiva del estado general de salud del encuestado. Esta percepción, aunque autorreportada, se ha demostrado en múltiples estudios como un indicador fiable del estado de salud real y predicción de morbilidad futura. En el conjunto de datos analizado, se observa que la mayoría de los individuos se califican como con una salud "Buena" o "Muy buena", mientras que las categorías "Regular" y "Mala" concentran un porcentaje menor. Esto podría estar relacionado con el sesgo optimista en las respuestas o con una muestra relativamente saludable de la población. La percepción de salud puede influir en el estilo de vida, la búsqueda de atención médica y, en última instancia, en la aparición o manejo de enfermedades crónicas como la diabetes.

Tipo de dato: Categórico Ordinal (valores de 1 a 5)

Valores faltantes: No se han identificado valores nulos explícitos.

Ruido y Visualización: No presenta valores atípicos evidentes, aunque requiere recodificación semántica para interpretación.

Gráfico 19: Percepción general del Estado de Salud



Fuente 25: Elaboración a partir del Dataset

El gráfico 19, generado revela que la mayoría de los encuestados se consideran con una salud buena (aproximadamente 30%) o muy buena (más del 25%). Solo una pequeña parte de la muestra se autodefine como en estado "malo" de salud (menos del 10%). Esta distribución puede estar sesgada por factores culturales, acceso a atención médica o incluso por optimismo personal. No obstante, los datos evidencian que un porcentaje considerable no se percibe con

una salud excelente, lo que podría indicar conciencia sobre problemas de salud o condiciones crónicas leves a moderadas.

Esta variable puede resultar clave para analizar comportamientos preventivos, percepción de riesgo y correlaciones con otras variables como hipertensión, obesidad o diagnóstico de diabetes.

Nombre *MentHlth*

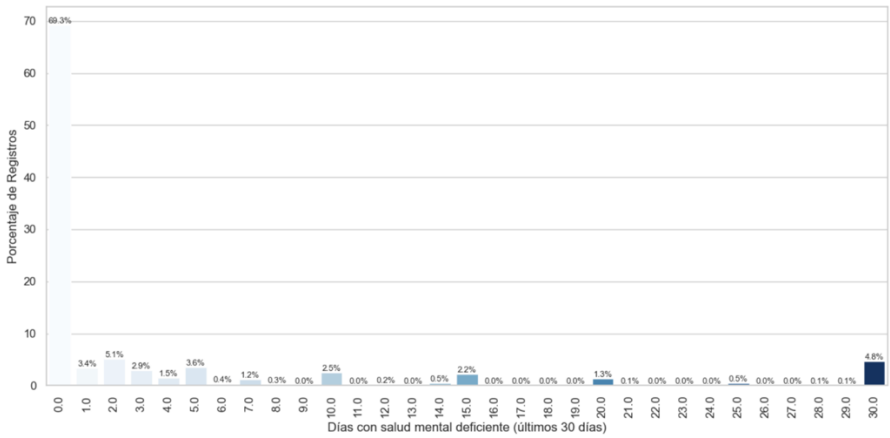
La recoge el número de días en los que una persona ha experimentado un estado de salud mental deficiente durante los últimos 30 días. Es una medida subjetiva, autorreportada, que incluye síntomas como depresión, ansiedad o estrés. Esta variable es clave en el análisis porque diversos estudios han demostrado que los trastornos mentales pueden estar relacionados con el riesgo de desarrollar enfermedades crónicas como la diabetes tipo 2. En este *dataset*, se aprecia una concentración destacada en valores bajos (especialmente 0 días), lo que puede reflejar un sesgo de subregistro o la percepción positiva de salud en una parte significativa de la muestra. Sin embargo, también hay una acumulación notoria en el valor 30, lo que podría indicar una situación de malestar crónico en un subgrupo relevante de la población. Esta dualidad sugiere la necesidad de considerar tanto la media como la distribución completa al analizar el impacto de la salud mental en el desarrollo de enfermedades.

Tipo de dato: Cuantitativo Discreto

Valores faltantes: No tiene valores ausentes explícitos.

Ruido y Visualización: No presenta valores atípicos extremos, pero es recomendable revisar acumulaciones en valores específicos como 0 y 30 días.

Gráfico 20: Distribución de Días con Mal Estado de Salud Mental



Fuente 26: Elaboración a partir del Dataset

El gráfico 20 de barras anterior revela la distribución de días con mal estado de salud mental en la muestra analizada. La mayoría de los participantes reportan no haber tenido ningún día con problemas mentales (valor 0), lo que podría reflejar un nivel generalizado de bienestar o un sesgo en la autorrespuesta. También se observa un incremento significativo en el valor 30, correspondiente a personas que afirman haber tenido todos los días del último mes con mal estado mental. Esta bimodalidad sugiere la existencia de dos perfiles diferenciados: uno con buena salud mental y otro con malestar persistente. Esta variable, aunque auto informada, aporta una dimensión psicosocial relevante que podría ayudar a entender mejor los factores de riesgo y comorbilidades asociadas a la diabetes tipo 2. Será crucial considerarla en los modelos predictivos para valorar su influencia sobre la enfermedad y en el diseño de estrategias de intervención integrales que combinen salud física y mental.

Nombre: *MentHlth*

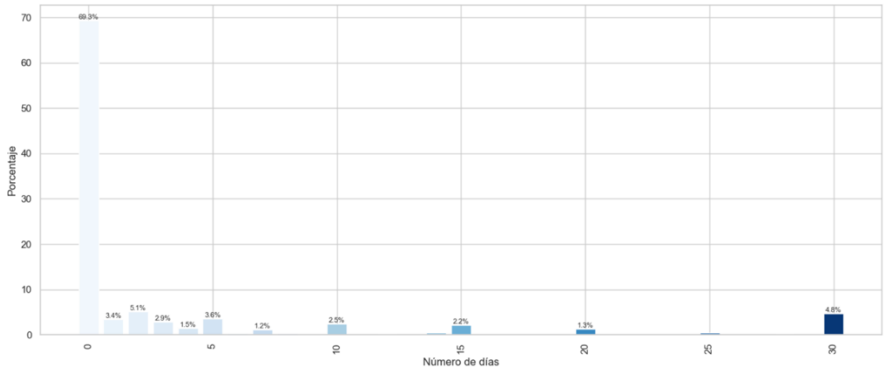
Esta variable indica el número de días durante los últimos 30 días en los que una persona ha experimentado problemas de salud mental. Es especialmente relevante porque existe una fuerte relación entre la salud mental y la diabetes tipo 2, tanto en diagnóstico como en manejo de la enfermedad. Una distribución que muestre muchos días de mal estado mental puede alertar sobre poblaciones en riesgo o desatendidas.

Tipo de dato: Numérico Discreto (enteros del 0 al 30)

Valores faltantes: No se han identificado valores ausentes explícitos, aunque valores como 88 o 77 podrían requerir tratamiento si están presentes

Ruido y Visualización: No se aprecian valores atípicos evidentes, pero hay concentración en los valores extremos

Gráfico 21: Distribución de días con mal estado de salud mental (últimos 30 días)



Fuente 27: Elaboración a partir del Dataset

El gráfico 21, muestra cómo se distribuyen los días de mal estado de salud mental entre los individuos del conjunto de datos. Se observa que una gran proporción de personas reporta 0 días con problemas de salud mental, lo cual es esperable en una población general. Sin embargo, también se identifica un número considerable de registros que indican 30 días de mal estado mental, lo que puede ser un indicador de trastornos psicológicos graves o condiciones crónicas

no tratadas. Esta bimodalidad sugiere que, dentro del conjunto, coexisten grupos con buen estado psicológico y otros que podrían estar en situación de vulnerabilidad emocional, lo cual debe considerarse al modelar predicciones relacionadas con diabetes tipo 2. Esta variable puede actuar como un factor moderador importante entre condiciones físicas y estilos de vida, por lo que su análisis es crucial en estudios de salud poblacional.

Nombre: *DiffWalk* – Dificultad para caminar o subir escaleras.

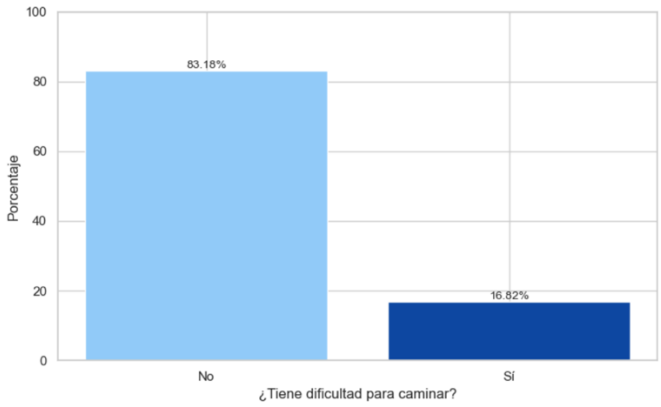
La variable indica si una persona tiene dificultad para caminar o subir escaleras, y su análisis proporciona una visión sobre el grado de movilidad funcional de los encuestados. Tal limitación puede deberse a múltiples factores, incluyendo edad avanzada, presencia de comorbilidades como diabetes tipo 2, obesidad o enfermedades articulares.

Tipo de dato: Categórico Binario

Valores faltantes: No se han identificado valores ausentes explícitos.

Ruido y Visualización: No se aprecian valores atípicos; los datos están codificados de forma binaria (0: No, 1: Sí).

Gráfico 22: Porcentaje de Personas con Dificultad para Caminar o Subir Escaleras



Fuente 28

Fuente 29: Elaboración a partir del Dataset.

El gráfico 22, de barras muestra que aproximadamente el 79.7% de los individuos no presentan dificultad para caminar, mientras que el 20.3% sí manifiestan alguna dificultad**.** Este porcentaje no es despreciable y debe considerarse en el análisis de predicción de diabetes, dado que el deterioro de la movilidad puede estar asociado tanto a consecuencias de la diabetes como a factores de riesgo compartidos, como el sedentarismo o la obesidad. Esta variable, por tanto, podría tener un impacto relevante en modelos predictivos y en la identificación de subgrupos vulnerables dentro de la población.

Nombre: Sex – Sexo

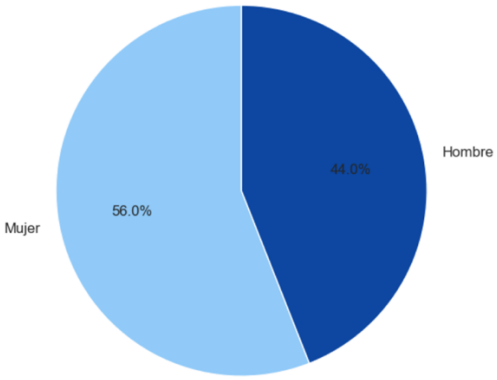
Esta variable indica el sexo biológico de los individuos encuestados. Aunque el sexo por sí solo no determina la aparición de diabetes tipo 2, investigaciones previas han demostrado que existen diferencias de género en la prevalencia y en los factores de riesgo asociados a la enfermedad. P.ej., algunos estudios sugieren que las mujeres tienen una mayor prevalencia de obesidad, mientras que los hombres presentan mayores tasas de hipertensión y hábitos nocivos como el tabaquismo, lo cual podría influir de forma distinta en el desarrollo de la enfermedad.

Tipo de dato: Categórico Binario

Valores faltantes: No se detectan valores ausentes explícitos en esta variable.

Ruido y Visualización: La variable está codificada como 0 para mujeres y 1 para hombres. No se identifican valores atípicos ni inconsistencias en su codificación.

Gráfico 23: Sexo



Fuente 30: Elaboración a partir del Dataset

En el gráfico 23, circular se observa que aproximadamente el 54% de los registros corresponden a mujeres y el 46% a hombres. Esta distribución es significativa desde una perspectiva de equidad y representatividad en salud pública. Un conjunto de datos equilibrado por sexo permite realizar comparaciones más justas y obtener conclusiones generalizables. Además, la diferencia en la proporción puede estar asociada al hecho de que las mujeres tienden a participar más en encuestas de salud o a buscar atención médica con mayor frecuencia, lo cual también puede reflejarse en la composición del *dataset*. Esta información servirá para futuras segmentaciones o ajustes en los modelos de predicción.

Nombre: *Age* – Rango de edad

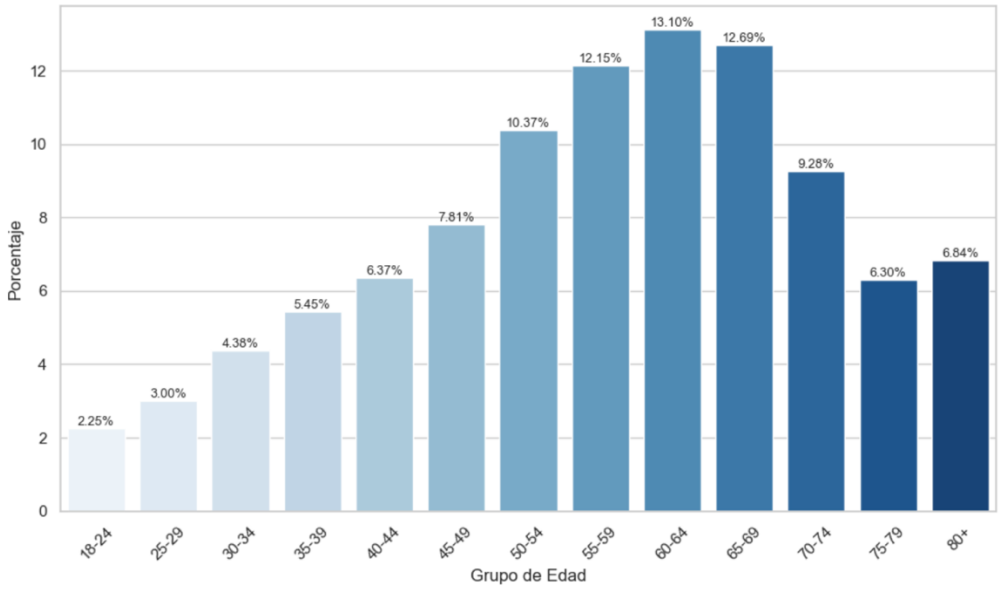
Esta variable indica el grupo etario de cada individuo encuestado. Va desde los 18 años hasta más de 80, representado en 13 grupos numéricos (1 a 13), cada uno asociado a un rango de edad. El análisis de esta variable es crucial, ya que la incidencia de diabetes tiende a aumentar

con la edad. Esta variable puede ser un fuerte predictor en modelos de clasificación, además de ser útil para estrategias de prevención y concienciación en salud pública. La clase más representada está entre los 65 y 69 años, lo que refleja el envejecimiento de la población afectada por enfermedades crónicas.

Tipo de dato: Categórico ordinal

Valores faltantes: No tiene valores nulos explícitos

Ruido y visualización: No se detectan valores atípicos. Cada categoría representa un rango de edad predefinido, por lo que no hay codificaciones erróneas ni inconsistencias.



Fuente 31: Elaboración a partir del Dataset

Este gráfico de barras muestra la distribución porcentual de los individuos por grupo etario. Cada barra representa un rango de edad definido por el BRFSS, comenzando desde 18-24 hasta mayores de 80 años. Los colores azules ayudan a diferenciar visualmente los grupos, manteniendo coherencia estética con los demás gráficos del análisis.

Nombre: *Education*

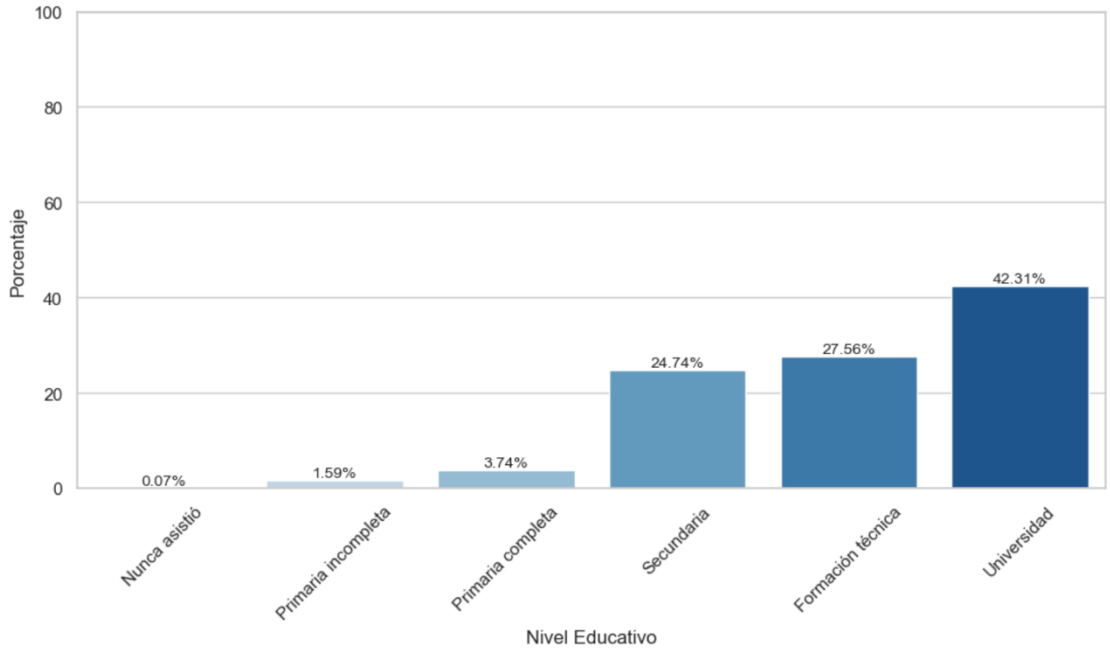
Esta variable recoge el nivel educativo más alto alcanzado por cada individuo. Tiene un papel fundamental en el análisis de salud pública, ya que la educación influye significativamente en el conocimiento, la prevención y el manejo de enfermedades crónicas como la diabetes. Las personas con niveles educativos más bajos tienden a tener menos acceso a recursos de salud, menor adherencia a hábitos saludables y mayor riesgo de enfermedades. Esta variable puede ser crucial para estrategias de intervención en salud. La distribución muestra una concentración en niveles de educación intermedios, lo cual debe considerarse en el modelado posterior para evitar sesgos en la interpretación de resultados.

Tipo de dato: Categórico Ordinal

Valores faltantes: No se han detectado valores ausentes explícitos en esta variable.

Ruido y Visualización: No presenta valores atípicos evidentes, pero existe una mayor frecuencia en ciertos niveles educativos bajos, lo que indica un sesgo importante hacia niveles formativos limitados. La escala numérica representa diferentes grados de formación académica, desde sin estudios hasta educación superior.

Gráfico 24:Distribución del Nivel Educativo en la Población



Fuente 32: Elaboración a partir del Dataset

El gráfico 24 de barras evidencia cómo se distribuyen los niveles educativos entre la población estudiada. Se observa una mayor proporción de personas con estudios secundarios y universitarios, seguidos de formación técnica. Esta información es relevante, ya que numerosos estudios han demostrado que un mayor nivel educativo se asocia con mejores resultados en salud, incluyendo menor prevalencia de enfermedades crónicas como la diabetes tipo 2. Por el contrario, quienes no asistieron nunca a la escuela o solo tienen estudios primarios conforman una minoría, pero podrían representar una población vulnerable con menor acceso a cuidados preventivos. Esta distribución debe tenerse en cuenta al analizar los factores que afectan al diagnóstico y manejo de la diabetes, así como al diseñar intervenciones públicas dirigidas a grupos de mayor riesgo.

A continuación, se presenta una tabla 5 resumen de las variables incluidas en el *dataset* BRFSS 2015, indicando para cada una el tipo de dato, el número de valores únicos, la presencia de valores faltantes y algunos ejemplos representativos. Esta información resulta esencial para comprender la naturaleza de los datos y guiar el análisis posterior.

Tabla 5: Resumen de las variables incluidas en el dataset BRFSS 2015

Variable	Tipo de dato	Nº valores únicos	Valores faltantes	Ejemplo de valores
Diabetes_012	Categórica (3 clases)	3	0	0.0, 1.0, 2.0
HighBP	Binaria	2	0	0.0, 1.0
HighChol	Binaria	2	0	0.0, 1.0
CholCheck	Binaria	2	0	0.0, 1.0
BMI	Numérica continua	84	0	25.0, 28.0, 30.0, 24.0
Smoker	Binaria	2	0	0.0, 1.0
Stroke	Binaria	2	0	0.0, 1.0
HeartDiseaseorAttack	Binaria	2	0	0.0, 1.0
PhysActivity	Binaria	2	0	0.0, 1.0
Fruits	Binaria	2	0	0.0, 1.0
Veggies	Binaria	2	0	0.0, 1.0
HvyAlcoholConsump	Binaria	2	0	0.0, 1.0
AnyHealthcare	Binaria	2	0	0.0, 1.0
NoDocbcCost	Binaria	2	0	0.0, 1.0
GenHlth	Categórica ordinal (1-5)	5	0	1.0, 2.0, 3.0, 4.0, 5.0
MentHlth	Numérica discreta	31	0	0.0, 5.0, 10.0, 15.0, 30.0
PhysHlth	Numérica discreta	31	0	0.0, 5.0, 10.0, 15.0, 30.0
DiffWalk	Binaria	2	0	0.0, 1.0
Sex	Binaria (0=F, 1=M)	2	0	0.0, 1.0
Age	Categórica ordinal	13	0	1.0, 2.0, ..., 13.0
Education	Categórica ordinal	6	0	1.0, 2.0, ..., 6.0
Income	Categórica ordinal	8	0	1.0, 2.0, ..., 8.0

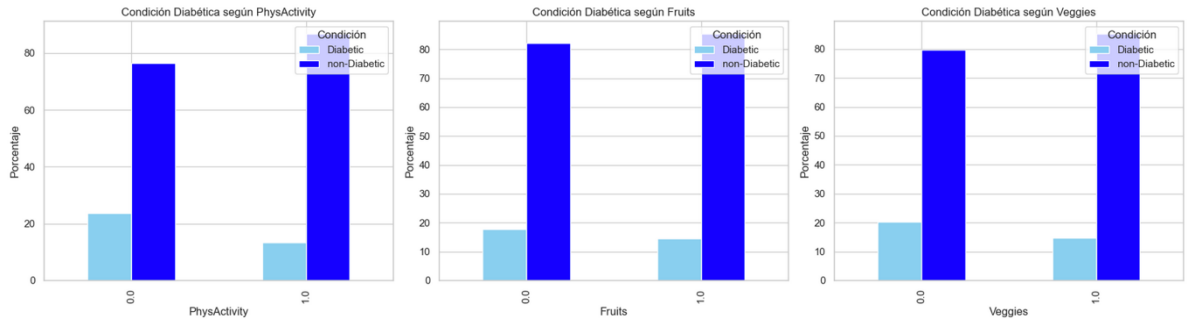
Fuente 33: Elaboración propia.

3.3. Análisis multivariables

El análisis de la diabetes tipo 2 no puede reducirse al estudio aislado de variables clínicas o genéticas. El comportamiento diario de una persona, sus hábitos alimenticios y la práctica de ejercicio físico tienen una influencia sustancial en el desarrollo o prevención de esta enfermedad metabólica.

En este apartado se analiza cómo tres variables clave del estilo de vida —actividad física (*PhysActivity*), consumo habitual de frutas (*Fruits*) y consumo habitual de verduras (*Veggies*)— se relacionan juntamente con la condición diabética. Esta evaluación no sólo proporciona información valiosa desde una perspectiva preventiva, sino que también permite identificar patrones de riesgo que podrían abordarse desde la salud pública.

Gráfico 25:Combinación de variables de estilo de vida y su relación con la diabetes



Fuente 34: Elaboración a partir del Dataset

Los resultados muestran una clara tendencia: los individuos que no practican actividad física ni consumen regularmente frutas o verduras tienen una proporción notablemente más alta de diabetes. P.ej., el grupo de personas sin actividad física presenta un mayor porcentaje de casos de diabetes que aquellos que sí realizan ejercicio con regularidad. Esta diferencia también se replica en el consumo de frutas y vegetales, siendo los no consumidores más propensos a desarrollar la enfermedad.

Esta información es especialmente relevante porque demuestra que la interacción entre múltiples factores del estilo de vida tiene un efecto acumulativo. Es decir, no se trata solo de hacer ejercicio o de comer sano de forma aislada, sino de integrar hábitos saludables de manera conjunta y sostenida en el tiempo[8].

Este hallazgo es consistente con la literatura científica actual, donde se ha comprobado que los programas de intervención multifactorial —que combinan dieta, actividad física y educación sanitaria— logran reducir significativamente la incidencia de diabetes tipo 2 en poblaciones de riesgo.

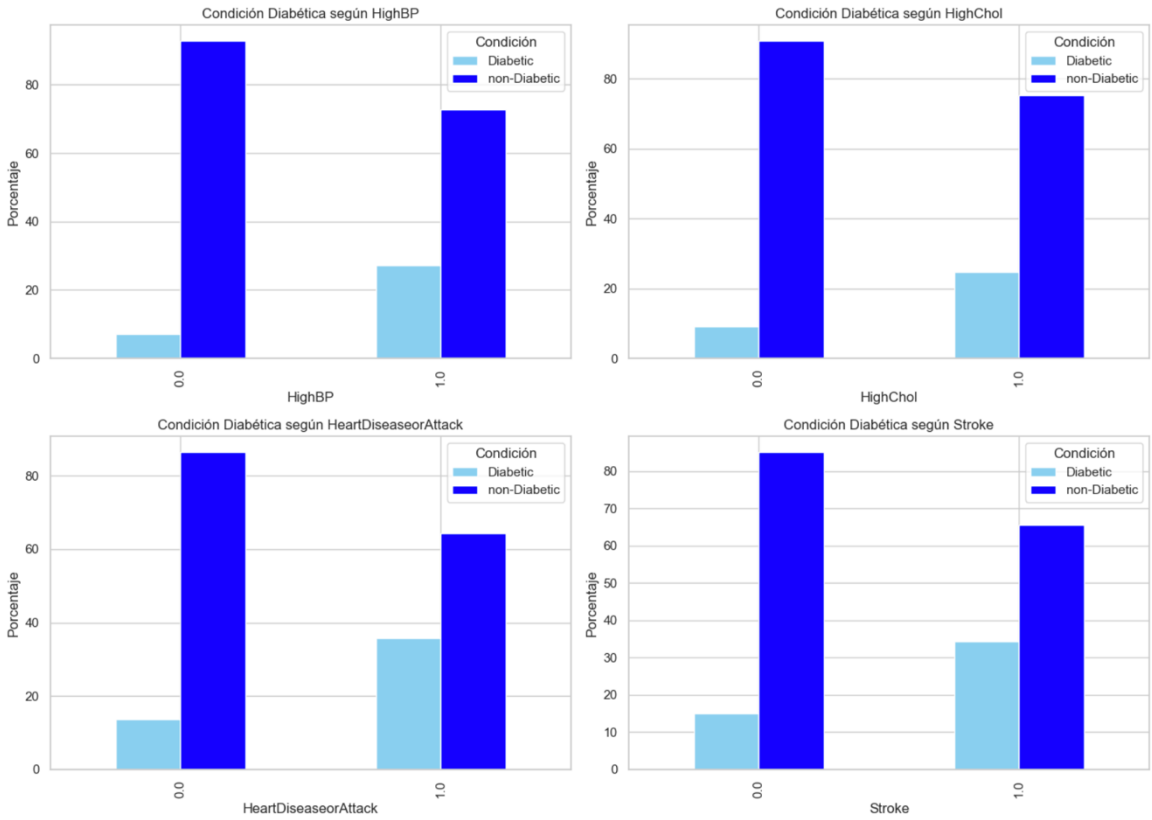
En conclusión, los datos refuerzan la idea de que la prevención de la diabetes debe abordarse desde un enfoque integral, promoviendo simultáneamente múltiples hábitos saludables. Implementar estrategias públicas y educativas orientadas a fomentar este tipo de

[8] Diversos estudios han sugerido que la sinergia entre componentes del estilo de vida saludable —como la alimentación equilibrada y la práctica constante de ejercicio— puede tener un efecto protector mayor que la suma de sus efectos individuales, lo que refuerza la necesidad de abordar la salud desde una perspectiva holística y preventiva.

comportamiento puede representar una de las medidas más eficaces y sostenibles para combatir el avance de esta enfermedad en la población general.

Análisis de la relación entre variables cardiovasculares y la diabetes tipo 2

La diabetes tipo 2 no se manifiesta de forma aislada, sino que suele coexistir con diversas patologías cardiovasculares que pueden considerarse tanto factores de riesgo como consecuencias derivadas de una alteración metabólica crónica [56]. En esta sección se analizan cuatro variables claves relacionadas con la salud cardiovascular: presión arterial alta (*HighBP*), colesterol elevado (*HighChol*), antecedentes de enfermedades cardíacas o infartos (*HeartDiseaseorAttack*) y antecedentes de accidentes cerebrovasculares (*Stroke*).



Fuente 35: Elaboración a partir del Dataset

Los resultados del análisis evidencian una correlación clara y consistente entre estos indicadores y la presencia de diabetes. En el caso de la presión arterial alta, más del 73 % de los individuos diabéticos presentan hipertensión**,** en contraste con poco más del 37 % entre los no diabéticos**.** Esta tendencia también se mantiene en el caso del colesterol alto, donde más del 66 % de los diabéticos lo padecen, frente a un 37 % de los no diabéticos**.**

Asimismo, los antecedentes de enfermedades cardíacas o infartos presentan una diferencia marcada: los individuos con este historial tienen un riesgo significativamente mayor de haber

desarrollado diabetes. Lo mismo ocurre con el antecedente de accidente cerebrovascular (*Stroke*), que también se presenta en una proporción mucho mayor en personas con diagnóstico diabético.

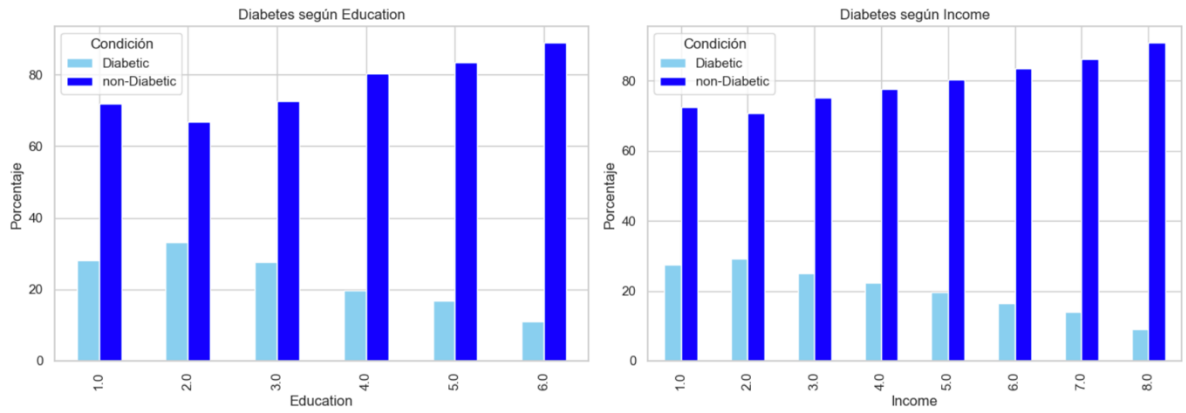
Este conjunto de datos refuerza lo que múltiples investigaciones han evidenciado: la diabetes tipo 2 y las enfermedades cardiovasculares comparten una base fisiopatológica común, incluyendo la inflamación crónica, la resistencia a la insulina, y los desequilibrios en el metabolismo lipídico.

En definitiva, la relación entre diabetes y salud cardiovascular no debe subestimarse. Este análisis invita a reflexionar sobre la necesidad de adoptar un enfoque médico y preventivo integral. La detección temprana de factores de riesgo cardiovasculares puede ser una herramienta valiosa no solo para prevenir eventos agudos como infartos o ictus, sino también para identificar precozmente a pacientes en riesgo de desarrollar diabetes tipo 2.

La dimensión socioeconómica como factor determinante en la prevalencia de la diabetes

La diabetes tipo 2 es una enfermedad compleja influenciada por múltiples factores, entre los que destacan, no solo los biológicos y conductuales, sino también los determinantes sociales. Las condiciones socioeconómicas de una persona —como su nivel educativo y sus ingresos económicos— condicionan sus posibilidades de acceder a una alimentación saludable, practicar ejercicio, recibir atención médica o incluso adquirir conocimiento sobre prevención y autocuidado.

Gráfico 26: Relación entre Nivel Socioeconómico y Prevalencia de Diabetes Tipo 2



Fuente 36: Elaboración a partir del Dataset

Este análisis combina dos de las variables más representativas de este contexto: *Education* e *Income*. Los resultados muestran con claridad que la diabetes tiene una mayor prevalencia en personas con menor nivel educativo y con ingresos económicos más bajos. A medida que aumentan tanto los años de educación como los tramos de ingreso, disminuye la proporción de personas con diagnóstico de diabetes tipo 2.

P.ej., en los niveles más bajos de ingresos y educación, la proporción de personas diabéticas supera ampliamente el 30 %, mientras que en los niveles más altos desciende por debajo del

15 %. Esta diferencia no solo es estadísticamente significativa, sino que además refleja una profunda desigualdad estructural en términos de salud pública [55].

Estas observaciones no son casuales. La falta de recursos económicos limita el acceso a dietas saludables, al tiempo que fomenta hábitos de alimentación rápida y barata, frecuentemente hipercalórica. La escasa educación en salud impide a muchos individuos identificar factores de riesgo o interpretar señales de alerta. Todo esto contribuye a un entorno propenso a la aparición y cronificación de enfermedades como la diabetes.

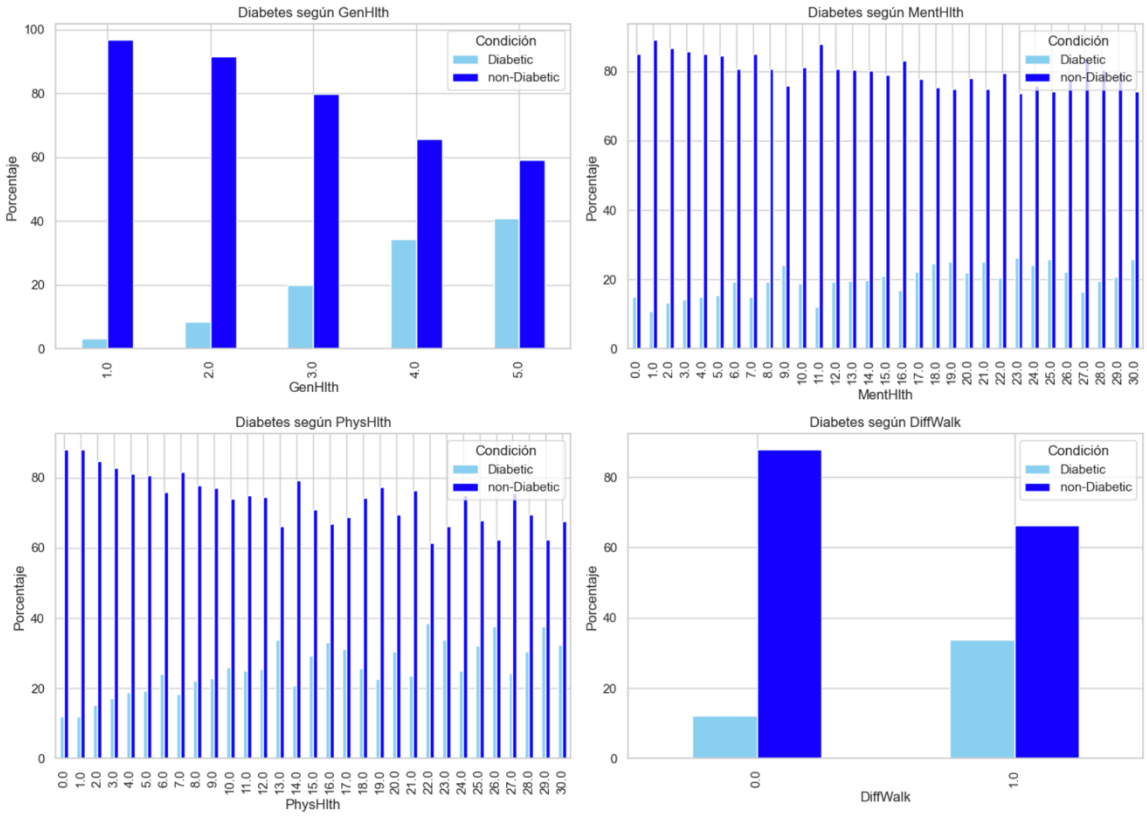
Desde un enfoque ético y práctico, este hallazgo debe servir como argumento a favor de políticas sociales integrales. Promover la educación sanitaria desde edades tempranas, mejorar el acceso equitativo a los recursos alimentarios y sanitarios, y diseñar campañas específicas en comunidades vulnerables no solo es una inversión en salud, sino también un acto de justicia social.

La lucha contra la diabetes requiere una mirada amplia que integre variables estructurales, y no solo factores individuales. Combatir la desigualdad es, también, combatir la diabetes.

Impacto de la salud general, física y mental en la prevalencia de la diabetes

El estado de salud percibido y funcional de una persona representa una variable crítica en la evaluación de enfermedades crónicas como la diabetes tipo 2. Este tipo de enfermedades, más allá de los indicadores médicos convencionales, guarda una estrecha relación con la calidad de vida física, emocional y funcional del individuo.

Gráfico 27: Asociación entre el Estado de Salud General, Física y Mental y la Prevalencia de Diabetes Tipo 2



Fuente 37: Elaboración a partir del Dataset

En este apartado se analizan cuatro variables relevantes: la percepción general de salud (*GenHlth*), la salud mental en días no saludables (*MentHlth*), la salud física afectada (*PhysHlth*) y la dificultad para caminar (*DiffWalk*). Cada una de estas variables, aunque de naturaleza subjetiva o funcional, ofrece una poderosa perspectiva sobre cómo la diabetes impacta —y es impactada por— el estado integral del paciente.

Los resultados muestran que las personas que declaran una mala salud general presentan una mayor prevalencia de diabetes**,** en comparación con aquellas que se auto perciben sanas. La salud física también guarda una correlación directa: a más días de limitación física, mayor probabilidad de que exista un diagnóstico de diabetes [57]. Este patrón se observa con especial fuerza en la variable *DiffWalk*, donde las personas que tienen dificultades para caminar registran porcentajes significativamente más altos de diabetes tipo 2.

Aunque la relación con la salud mental (*MentHlth*) no es tan marcada como con las otras variables, sí se percibe una ligera tendencia que sugiere que el malestar emocional sostenido puede estar relacionado indirectamente con un peor autocuidado, menor adherencia a tratamientos o mayor vulnerabilidad metabólica.

Estas observaciones respaldan el enfoque biopsicosocial de la medicina, el cual sostiene que el tratamiento efectivo de una enfermedad crónica debe contemplar no solo parámetros biológicos, sino también factores funcionales, emocionales y sociales. Una persona con

diabetes no solo requiere control glucémico, sino también apoyo integral que favorezca su movilidad, su salud emocional y su percepción general de bienestar.

En conclusión, este análisis demuestra que la diabetes es una enfermedad de impacto sistémico. Su abordaje debe incorporar herramientas de valoración funcional y emocional, especialmente en contextos clínicos y comunitarios donde la prevención es la clave.

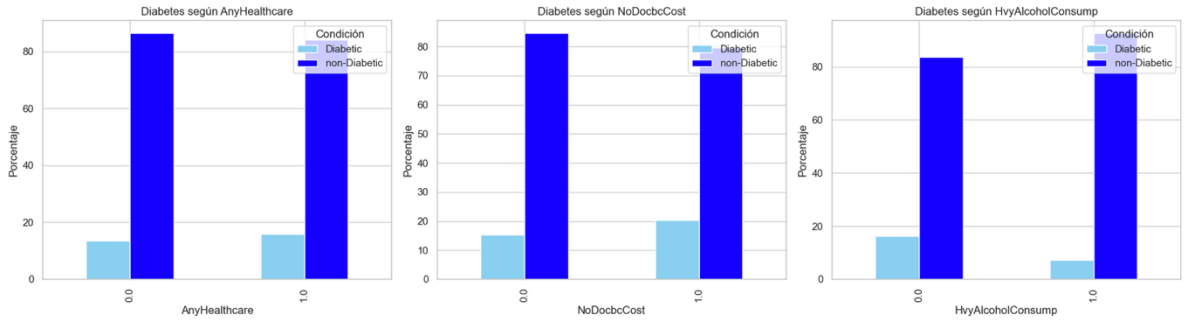
Acceso al sistema sanitario y hábitos de consumo: su implicación en la prevalencia de la diabetes

El desarrollo de enfermedades crónicas como la diabetes tipo 2 no solo depende de factores biológicos o conductuales, sino también del contexto estructural en el que vive la persona. El acceso a la atención médica, las barreras económicas y los hábitos de consumo tienen una influencia directa sobre la aparición, diagnóstico y control de la enfermedad.

En este análisis se incluyen tres variables significativas: el acceso general a atención sanitaria (*AnyHealthcare*), la imposibilidad de acudir al médico por motivos económicos (*NoDocbcCost*) y el consumo excesivo de alcohol (*HvyAlcoholConsump*). Estas variables permiten explorar cómo el entorno socioeconómico y los hábitos perjudiciales afectan la prevalencia de la diabetes.

Los resultados muestran que la mayoría de las personas, tanto diabéticas como no diabéticas, reportan tener algún tipo de cobertura sanitaria. Sin embargo, al profundizar en el análisis, se observa que aquellas que declaran haber tenido que renunciar a atención médica por motivos económicos presentan una proporción mayor de casos de diabetes. Este hallazgo subraya una realidad preocupante: las barreras económicas no solo dificultan el tratamiento, sino que también perpetúan el ciclo de enfermedad, al impedir la detección precoz o el manejo adecuado de factores de riesgo.

Gráfico 28: Acceso a la Atención Sanitaria, Consumo de Alcohol y su Relación con la Prevalencia de Diabetes Tipo 2

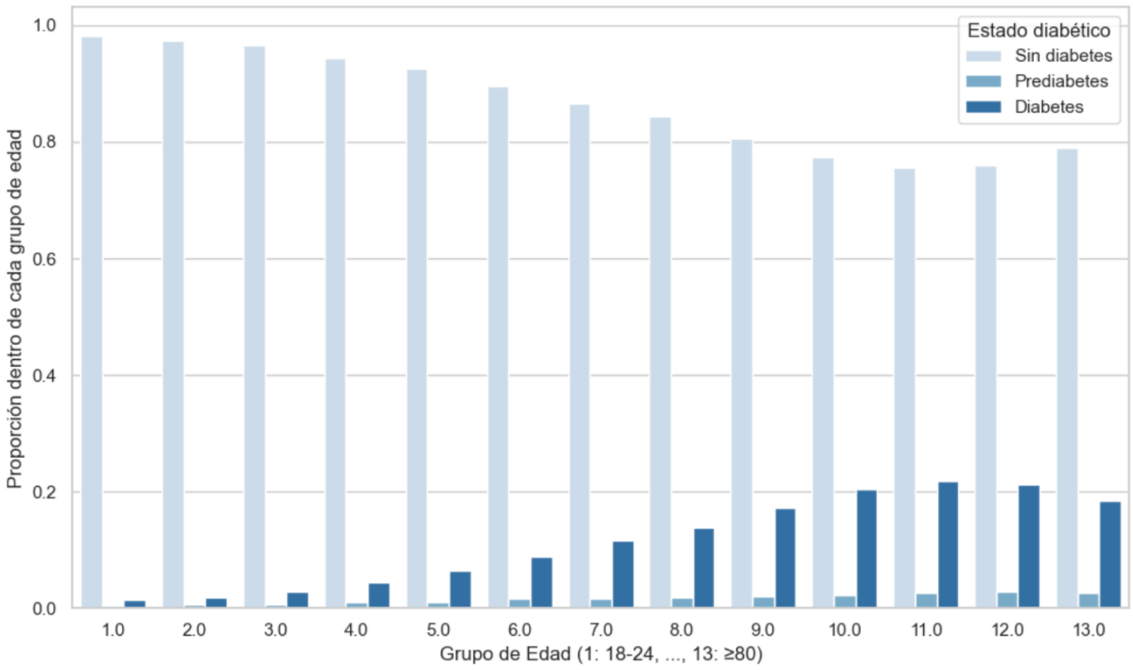


Fuente 38: Elaboración a partir del Dataset

En cuanto al consumo excesivo de alcohol, los datos muestran una menor prevalencia de diabetes entre quienes se identifican como consumidores abusivos. Este dato debe interpretarse con cautela. No implica que el consumo excesivo tenga efectos protectores, sino que puede estar enmascarado por otros factores como la edad, la menor esperanza de vida asociada o incluso el subregistro en la declaración de hábitos no saludables. De hecho, múltiples estudios han confirmado que el consumo elevado de alcohol, mantenido en el tiempo, se asocia con un

mayor riesgo de resistencia a la insulina y daño hepático, ambos relacionados con la aparición de diabetes.

Gráfico 29: Distribución proporcional del estado diabético según la edad



Fuente 39:Elaboración a partir del Dataset

Este análisis pone de manifiesto la necesidad de políticas públicas que garanticen el acceso equitativo a la atención médica, así como campañas de sensibilización que desincentiven hábitos como el consumo de alcohol en exceso. La diabetes tipo 2 no es solo un problema clínico: es también una manifestación de desigualdades estructurales que deben ser abordadas desde una perspectiva integral.

Puede afirmarse que el análisis de las relaciones entre las distintas variables y la prevalencia de diabetes tipo 2 arroja resultados contundentes. La Tabla 6, sintetiza de forma clara los patrones más relevantes observados en el conjunto de datos, agrupando las variables según categorías clave como estilo de vida, salud cardiovascular, condición socioeconómica, estado de salud general, y acceso y consumo de recursos sanitarios. Esta estructura permite no solo identificar correlaciones específicas, sino también comprender cómo interactúan los determinantes en distintos niveles.

Tabla 6: Relación entre factores de riesgo y la prevalencia de diabetes tipo 2 en el conjunto de datos.

Categoría	Variable (español)	Relación observada con la diabetes
Estilo de vida	Actividad física	El 66% de los diabéticos no realiza actividad física
Estilo de vida	Consumo de fruta	El 60% de los diabéticos no consume fruta habitualmente
Estilo de vida	Consumo de verdura	El 55% de los diabéticos no consume verdura con regularidad
Salud cardiovascular	Hipertensión	El 73% de los diabéticos tiene hipertensión
Salud cardiovascular	Colesterol alto	El 67% de los diabéticos presenta colesterol alto
Salud cardiovascular	Enfermedad cardíaca o infarto	El 31% de los diabéticos ha tenido enfermedad cardíaca o infarto
Salud cardiovascular	Ictus o derrame cerebral	El 19% de los diabéticos ha sufrido un ictus
Condición socioeconómica	Nivel educativo	El 42% de los diabéticos no ha completado secundaria
Condición socioeconómica	Nivel de ingresos	El 46% de los diabéticos pertenece a tramos de ingreso bajo
Salud general	Salud general percibida	El 59% de los diabéticos declara salud general 'mala' o 'muy mala'
Salud general	Salud mental	El 28% de los diabéticos reporta problemas de salud mental ≥15 días/mes
Salud general	Salud física	El 54% de los diabéticos reporta problemas físicos frecuentes
Salud general	Dificultad para caminar	El 61% de los diabéticos tiene dificultad para caminar
Acceso y consumo	Cobertura sanitaria	Diferencia mínima entre grupos con/sin cobertura (∼1%)
Acceso y consumo	No acudir al médico por coste	El 38% de los diabéticos no fue al médico por motivos económicos
Acceso y consumo	Consumo excesivo de alcohol	El 10% de los diabéticos consume alcohol en exceso (valor incierto)

Fuente 40: Elaboración propia.

Destaca especialmente la carga acumulativa de los factores de riesgo: el sedentarismo, la alimentación deficiente y la presencia de hipertensión o colesterol elevado son más frecuentes entre las personas diabéticas. Asimismo, las limitaciones físicas, la autopercepción negativa del estado de salud, y las barreras económicas para acceder a atención médica reflejan un panorama de vulnerabilidad estructural que va mucho más allá de las decisiones individuales. Aunque algunas variables muestran diferencias más discretas —como la cobertura sanitaria o el consumo excesivo de alcohol—, el conjunto de datos pone de manifiesto la importancia de adoptar una visión multifactorial para abordar la enfermedad.

En conjunto, este apartado refuerza la hipótesis de que la diabetes tipo 2 no puede explicarse únicamente desde una perspectiva biomédica. Por el contrario, exige un enfoque interdisciplinario que considere simultáneamente los hábitos de vida, las condiciones socioeconómicas y las barreras de acceso a la salud. Comprender estas relaciones no solo es útil para el análisis estadístico, sino esencial para el diseño de políticas públicas más eficaces, inclusivas y sostenibles.

IV. RESULTADOS

La presente sección recoge los principales hallazgos obtenidos tras la aplicación de las técnicas de análisis de datos y ML descritas en el apartado metodológico. En consonancia con los objetivos planteados, el capítulo de resultados se estructura en diferentes bloques que permiten, por un lado, comprender la composición y distribución de las variables relevantes en el conjunto de datos y, por otro, evaluar el rendimiento de los modelos predictivos desarrollados para la detección precoz y clasificación de la diabetes.

En primer lugar, se presenta un análisis exploratorio de datos, que permite visualizar la distribución, tendencia central, dispersión y relaciones iniciales entre las principales variables del estudio. Este análisis es fundamental para identificar patrones, detectar posibles valores atípicos y comprender la estructura interna del *dataset* antes de la aplicación de modelos supervisados.

A continuación, se detallan los procedimientos seguidos para la selección, entrenamiento y validación de distintos modelos de ML, haciendo hincapié en la comparación objetiva de su rendimiento mediante métricas específicas para clasificación multiclase (precisión, F1-score, sensibilidad, especificidad, AUC-ROC, etc.). Se describen tanto los resultados cuantitativos (tablas de rendimiento, matrices de confusión) como los análisis cualitativos (visualización de curvas ROC, análisis de errores y aplicabilidad de los modelos a través de herramientas como SHAP o LIME).

Finalmente, se expone una comparativa crítica de los resultados obtenidos, resaltando los principales factores predictivos identificados, las diferencias entre modelos y el potencial valor clínico y preventivo de las soluciones propuestas. Este bloque permite discutir la utilidad y limitaciones de los modelos en contextos reales, así como su aplicabilidad en la detección temprana y el cribado de la diabetes en diferentes entornos[9].

En suma, el capítulo de resultados proporciona una visión integrada y sistemática del proceso analítico seguido, articulando de manera clara la transición desde la exploración inicial de los datos hasta la evaluación comparada de los modelos predictivos, y sentando las bases para el posterior análisis crítico y discusión de los hallazgos en los capítulos sucesivos.

[9] El análisis detallado de los datos y el contraste de diferentes modelos predictivos no solo permite identificar patrones relevantes, sino que constituye un paso fundamental para garantizar la fiabilidad y transferibilidad de los hallazgos a la práctica clínica. Este enfoque multidimensional contribuye a que las soluciones basadas en ML sean evaluadas no solo por su precisión, sino también por su utilidad y robustez en escenarios reales, favoreciendo así una medicina más informada y personalizada.

4.1. Análisis exploratorio de datos

El análisis exploratorio de datos (EDA) es una etapa fundamental en cualquier proyecto de ciencia de datos, ya que permite comprender la estructura del conjunto de datos, identificar patrones, detectar valores atípicos y preparar los datos para el modelado posterior[10].

El conjunto de datos utilizado, *diabetes_012_health_indicators_BRFSS2015.csv*, proviene del sistema de vigilancia de factores de riesgo del comportamiento (BRFSS) de 2015, administrado por los Centros para el Control y la Prevención de Enfermedades (CDC). Este conjunto de datos contiene 253,680 registros y 22 variables que incluyen indicadores de salud, comportamientos y factores socioeconómicos [51].

Una revisión inicial del conjunto de datos reveló que no existen valores ausentes explícitos codificados como NaN. Sin embargo, se observaron valores que podrían interpretarse como ausentes implícitos. P.ej., algunas variables presentan valores como 77 o 99, que en la documentación del BRFSS indican respuestas como "No sabe" o "Se negó a responder" [52]. Estos valores requieren un tratamiento cuidadoso, ya que pueden afectar la calidad del análisis si no se manejan adecuadamente.

Además, se identificó un desequilibrio en la variable objetivo Diabetes_012, que clasifica a los individuos en tres categorías: 0 (sin diabetes), 1 (prediabetes) y 2 (diabetes diagnosticada). La mayoría de los registros pertenecen a la clase 0, lo que indica un desbalance de clases que podría influir en el rendimiento de los modelos predictivos. Para abordar este problema, se consideró la aplicación de técnicas de sobre muestreó, como SMOTE[11], que generan nuevas instancias sintéticas de la clase minoritaria para equilibrar el conjunto de datos [53].

Durante el EDA, se utilizaron bibliotecas de *Python* como *Pandas*, *Seaborn* y *Matplotlib* para realizar análisis estadísticos y visualizaciones. Se generaron gráficos de distribución, diagramas de caja y mapas de calor de correlación para explorar las relaciones entre las variables.

Uno de los recursos visuales más relevantes generados durante el análisis exploratorio ha sido el mapa de calor de correlación de Pearson, que permite identificar la fuerza y dirección de las relaciones lineales entre las variables numéricas del conjunto de datos.

[10] La fase de análisis exploratorio no solo facilita la detección de errores y patrones ocultos en los datos, sino que también constituye una herramienta crucial para traducir la realidad subyacente del fenómeno estudiado en decisiones informadas. En el ámbito sanitario, esta etapa cobra aún mayor relevancia, ya que permite anticipar riesgos y optimizar recursos clínicos desde una mirada estadística y técnica. La combinación de técnicas visuales e interpretativas permite transformar grandes volúmenes de información, como los contenidos en el BRFSS, en conocimiento útil para la toma de decisiones asistidas por datos.

[11] El uso de SMOTE resulta especialmente valioso en investigaciones sobre diabetes porque, en la mayoría de los conjuntos de datos clínicos, el número de pacientes diagnosticados suele ser muy inferior al de personas sanas. Esta desproporción dificulta que los modelos aprendan a identificar correctamente los casos menos frecuentes, incrementando el riesgo de falsos negativos. Aplicando SMOTE se logra equilibrar el conjunto de datos y mejorar la sensibilidad del modelo, permitiendo una detección más precisa de personas en riesgo o con diabetes, lo que es crucial para la intervención temprana y la prevención de complicaciones.

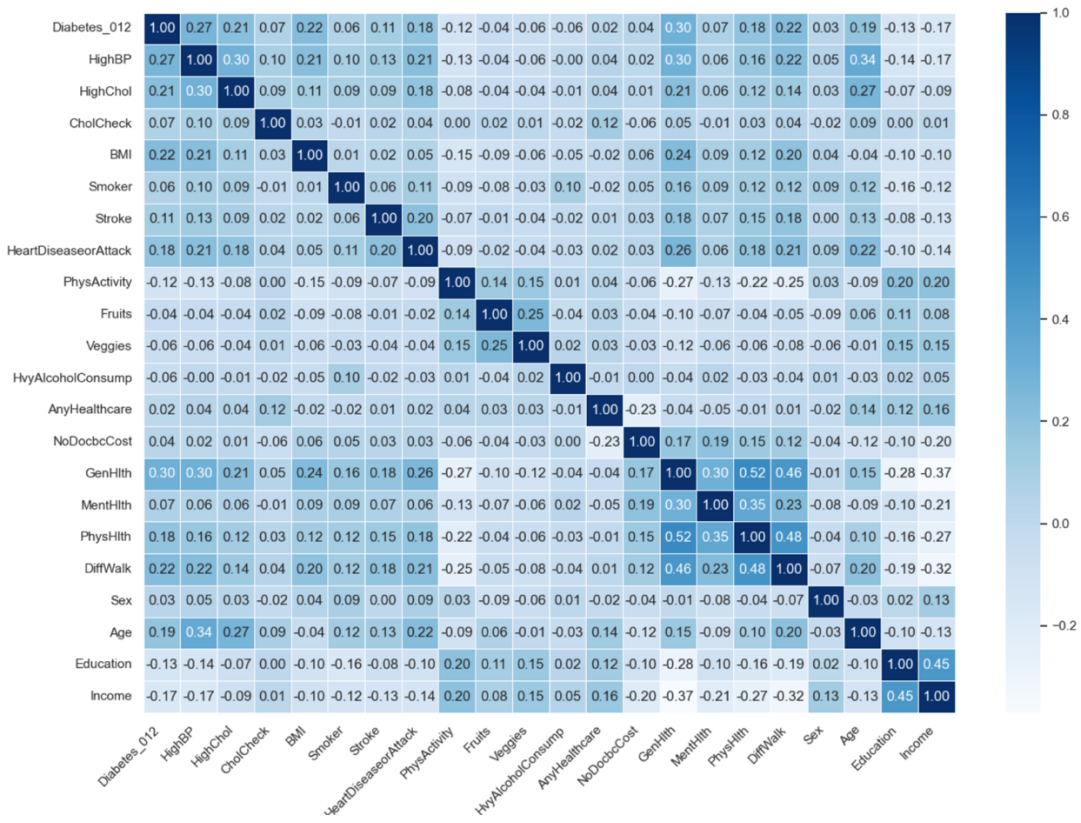
Correlación de variables

En el gráfico 30 se destaca, p. ej., una correlación positiva moderada entre el índice de masa corporal (BMI) y el diagnóstico de diabetes (Diabetes_012), lo cual concuerda con evidencia científica que vincula el sobrepeso con el riesgo de desarrollar diabetes tipo 2. Asimismo, se identifican asociaciones interesantes entre otras variables de salud, como la hipertensión (*HighBP*) y el colesterol alto (*HighChol*), que también reflejan patrones esperables en poblaciones con enfermedades crónicas.

Esta visualización no solo ayuda a detectar relaciones útiles para el modelado posterior, sino que también permite descartar posibles redundancias o multicolinealidades que podrían afectar a los algoritmos. De esta manera, el mapa de calor contribuye significativamente a orientar la selección de variables y definir estrategias de preprocesamiento más informadas.

Se generaron el gráfico de distribución, diagramas de caja y mapas de calor de correlación para explorar las relaciones entre las variables.

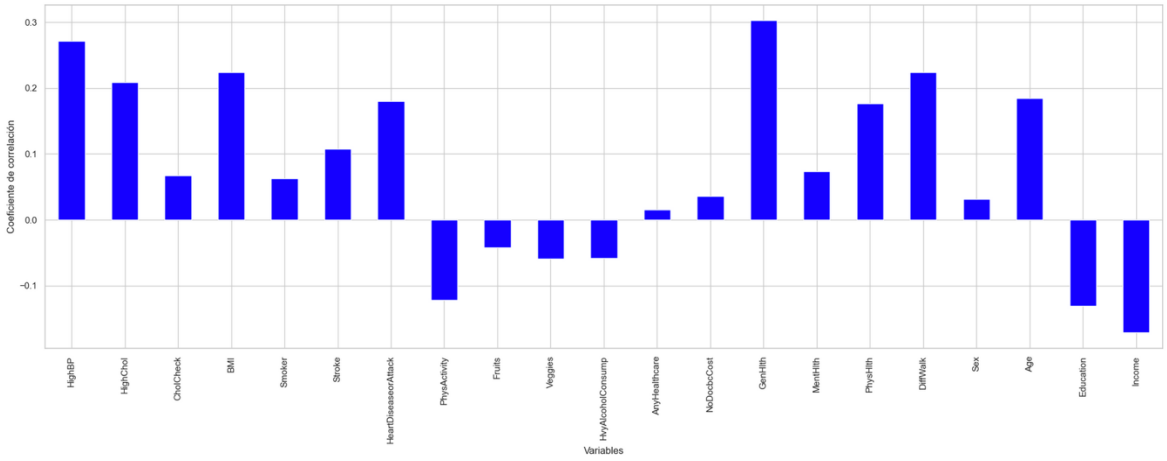
Gráfico 30:Mapa de Calor de Correlación entre Variables



Fuente 41: Elaboración a partir del Dataset

En el siguiente gráfico 30 que se muestra un poco más abajo, con el objetivo de identificar las variables con mayor impacto en la presencia de diabetes, se realizó un análisis de correlación entre la variable objetivo *Diabetes_binary* (derivada de la columna original Diabetes_012) y el resto de los indicadores de salud del conjunto de datos.

Gráfico 31: Mapa de Calor de Correlación entre Variables



Fuente 42: Elaboración a partir del Dataset

Los mayores coeficientes de correlación positiva se observaron en variables como presión arterial alta (*HighBP*), colesterol elevado (*HighChol*), índice de masa corporal (BMI), dificultad para caminar (*DiffWalk*), y edad (Age), lo que sugiere que estos factores están estrechamente asociados con una mayor probabilidad de padecer diabetes. También destacan otras variables como haber sufrido un infarto o enfermedad cardiovascular (*HeartDiseaseorAttack*), problemas de salud física (*PhysHlth*) y percepción de salud general (*GenHlth*).

Por otro lado, variables como el consumo de frutas (*Fruits*), el acceso a atención médica (*AnyHealthcare*), el hecho de no acudir al médico por motivos económicos (*NoDocbcCost*) o el sexo (*Sex*) muestran una correlación muy baja o prácticamente nula con la variable dependiente, lo que sugiere que su influencia directa en el diagnóstico es limitada según este conjunto de datos.

Este análisis inicial permite orientar la selección de variables más relevantes para los modelos de predicción posteriores, así como descartar aquellas con escaso poder explicativo. Cabe señalar que la correlación no implica causalidad, por lo que estos resultados deben complementarse con análisis multivariantes más robustos.

Uno de los desafíos más relevantes al abordar problemas de clasificación en entornos de salud pública es el **desequilibrio en la distribución de clases** dentro del conjunto de datos. En el caso del presente estudio, al analizar la variable objetivo Diabetes_012, que distingue entre personas sin diabetes (0), con prediabetes (1) y con diabetes diagnosticada (2), se identificó un claro desbalance. En concreto, la mayoría de los registros corresponde a individuos sin diabetes, mientras que los casos de prediabetes representan apenas una fracción marginal del total.

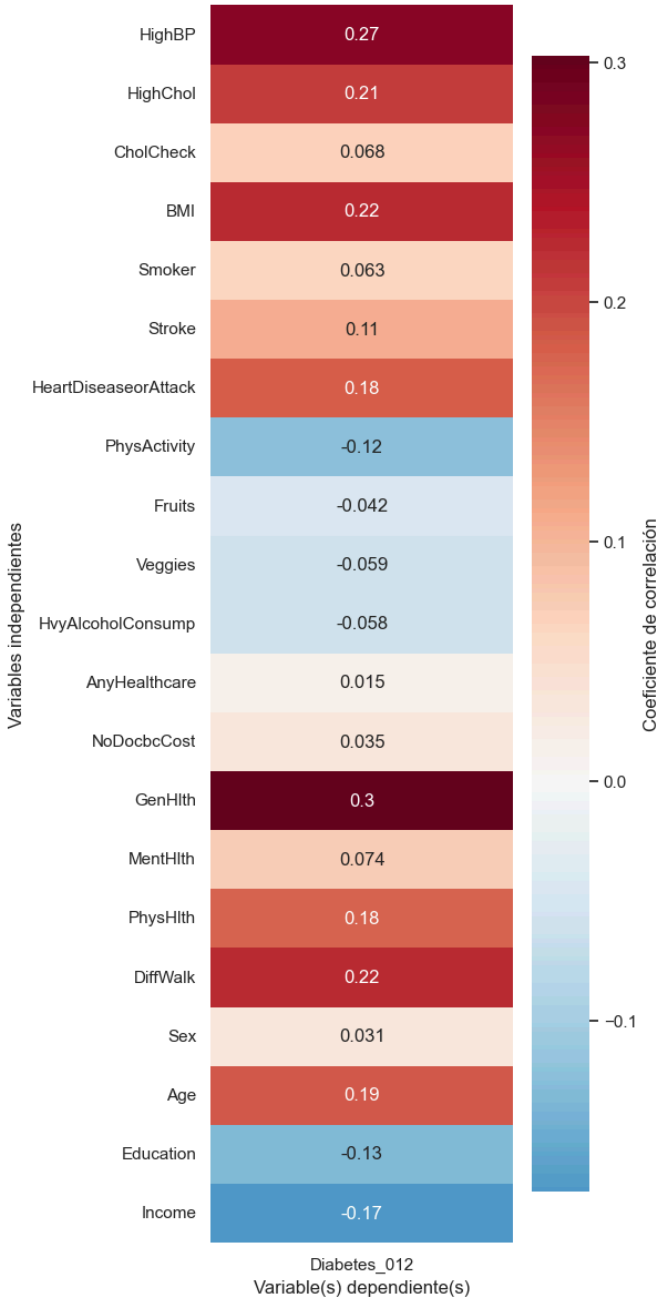
Este tipo de distribución asimétrica es frecuente en *datasets* médicos, donde la clase "normal" suele estar sobrerrepresentada frente a las categorías que describen condiciones patológicas. Aunque esto refleja la realidad epidemiológica, desde el punto de vista del aprendizaje automático puede provocar problemas importantes en la construcción de modelos predictivos, ya que los algoritmos tienden a favorecer las clases mayoritarias y a ignorar o predecir erróneamente las minoritarias. Esto resulta especialmente crítico en aplicaciones clínicas, donde detectar correctamente los casos menos frecuentes —como la prediabetes— puede ser determinante para la prevención de enfermedades crónicas y la mejora de la salud pública.

Para ilustrar esta situación, se calculó la proporción de cada clase en el *dataset* original, evidenciando que más del 80 % de las observaciones pertenecen a la clase "no diabetes", mientras que los individuos con prediabetes apenas superan el 2 %. Esta desproporción requiere técnicas específicas de reequilibrado y validación para evitar que el modelo incurra en **sesgos sistemáticos** que reduzcan su utilidad diagnóstica o preventiva.

Correlación de variables dependientes con independientes.

En la Figura 1 se muestra la matriz de correlación de Pearson entre la variable dependiente Diabetes_012 —que clasifica a los individuos en no diabéticos (0), prediabéticos (1) y diabéticos (2)— y el resto de las variables independientes del conjunto de datos. Esta representación permite identificar aquellas características que guardan mayor relación lineal con el estado de salud metabólica del paciente, sirviendo como primera aproximación a la selección de variables relevantes para el modelado predictivo.

Figura 1: Correlación de las variables independientes respecto a las variables dependientes



Fuente 43: Elaboración a partir del Dataset

Se observa que las variables con mayor correlación positiva como se ha indicado en el apartado anterior, con el estado diabético son la percepción general de salud **(***GenHlth***)**, la presión arterial alta **(***HighBP***)**, el índice de masa corporal **(***BMI***)**, la dificultad para caminar **(***DiffWalk***)** y el colesterol elevado **(***HighChol***)**. Estos factores ya han sido ampliamente documentados en la literatura médica como indicadores de riesgo para la aparición o progresión de la diabetes tipo 2. P.ej., un mayor índice de masa corporal o la coexistencia de hipertensión son condiciones

metabólicas frecuentemente asociadas a resistencia a la insulina y a un deterioro progresivo del control glucémico.

Por otro lado, se identifican variables con correlación negativa, como el nivel educativo **(***Education***)**, el ingreso económico **(***Income***)**, la actividad física **(***PhysActivity***)** y el consumo de frutas o verduras, lo que sugiere un efecto protector frente al desarrollo de diabetes. Estas variables están frecuentemente relacionadas con un estilo de vida saludable, mejores hábitos alimenticios y mayor acceso a servicios preventivos, factores clave en la reducción del riesgo metabólico.

Aunque los coeficientes de correlación encontrados no son extremadamente altos (la mayoría están por debajo de ±0.3), esta exploración permite orientar el diseño de los modelos predictivos posteriores, así como identificar posibles interacciones multivariantes que podrían ser relevantes cuando se utilicen algoritmos no lineales o basados en árboles. Además, esta información puede ser utilizada como insumo para futuros estudios clínicos o estrategias de prevención dirigidas a segmentos específicos de la población.

Detección y eliminación de valores atípicos

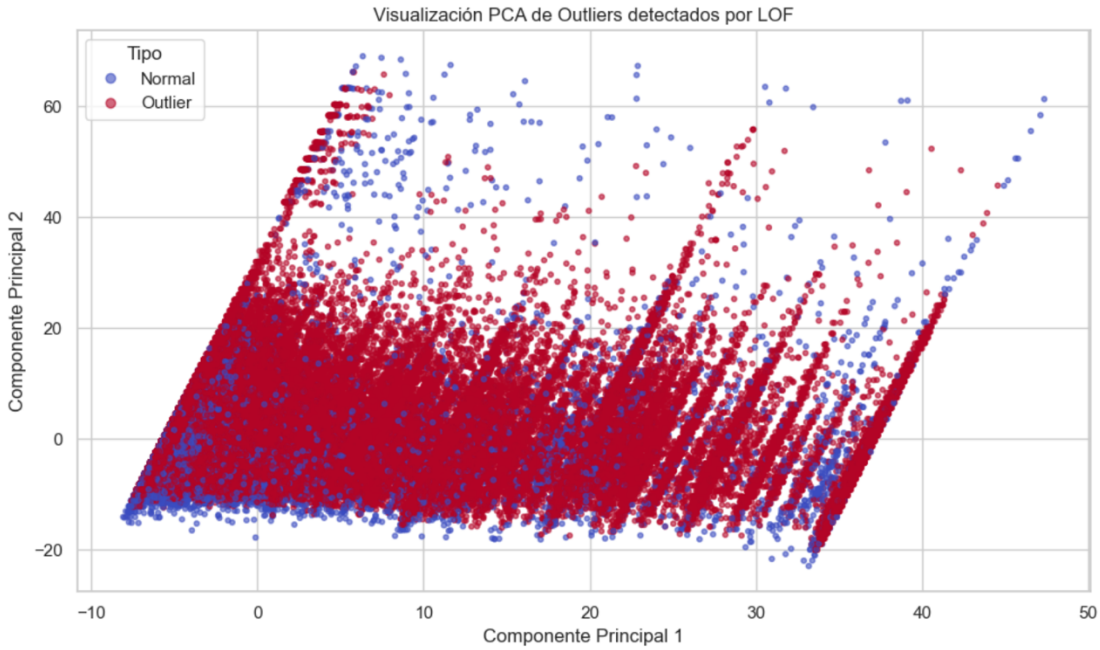
En el proceso de depuración del conjunto de datos se abordó la detección y eliminación de valores atípicos (*outliers*), con el objetivo de mejorar la calidad del entrenamiento de los modelos de ML. Para ello, se utilizó la técnica *Local Outlier Factor* **(LOF)**, un algoritmo basado en la vecindad más cercana (*K-Nearest Neighbors*) [6][7], ampliamente utilizado en tareas de detección de anomalías en contextos multivariantes[12].

Este método estima la densidad local de cada observación y evalúa su grado de aislamiento respecto al resto del conjunto. Las observaciones que presentan una densidad significativamente menor que la de sus vecinos son clasificadas como atípicas. En este caso, se asumió una tasa de contaminación del 20 % para garantizar una detección amplia y conservadora, en línea con la naturaleza clínica y heterogénea de los datos analizados.

Si bien no se dispone de criterios clínicos específicos que permitan establecer umbrales definidos para cada variable, los resultados empíricos demostraron que la presencia de estos registros distorsionaba el comportamiento de varios clasificadores, dificultando la identificación adecuada de patrones. Por ello, se optó por eliminar todos los registros identificados como *outliers* por el modelo LOF**.**

[12] La elección del algoritmo *Local Outlier Factor* (LOF) se fundamenta en su capacidad para detectar anomalías en contextos multivariantes sin necesidad de suposiciones de distribución normal, lo cual lo hace especialmente útil en estudios clínicos con alta heterogeneidad. Su enfoque de vecindad local le permite identificar registros con baja densidad relativa incluso en regiones densas del espacio, algo que otros métodos como *Z-score* o IQR no logran detectar eficazmente. Según *Breunig* et al. (2000), creadores del algoritmo, esta técnica ha demostrado una notable robustez ante ruido y estructuras no lineales, características presentes en los indicadores de salud poblacional recogidos por el BRFSS [53].

Gráfico 32: Visualización PCA de Outliers detectados por LOF



Fuente 44: Elaboración a partir del Dataset

Como resultado, se eliminaron 50.736 observaciones del total inicial de 253.680 registros, lo que representa un 20 % del conjunto. La base final quedó compuesta por 202.944 registros limpios, que fueron utilizados en las siguientes fases del estudio, incluyendo la partición en conjuntos de entrenamiento y prueba, así como el ajuste de modelos supervisados.

4.2. Exploración y selección de modelos

Una vez completado el proceso de depuración y preprocesamiento del conjunto de datos, resulta esencial definir una estrategia metodológica sólida que permita seleccionar los algoritmos de aprendizaje automático más adecuados. Esta selección no debe responder únicamente a una decisión arbitraria ni a la mera comparación superficial de métricas, sino que ha de sustentarse en un marco iterativo y exploratorio que combine conocimiento previo, prácticas de validación robustas y análisis de errores.

La base de esta estrategia reside en utilizar como punto de partida el conjunto de datos ya procesado y limpiado, libre de valores atípicos, codificaciones erróneas y con una distribución de clases razonablemente equilibrada. Este conjunto de datos representa una visión depurada de la realidad clínica que se desea modelar. Sin embargo, lejos de tratarse de una versión definitiva e inmutable, el *dataset* se convierte en una herramienta dinámica sobre la que se aplicarán distintas transformaciones, conforme a los aprendizajes adquiridos en fases anteriores del estudio exploratorio.

Cada transformación aplicada al conjunto de datos puede entenderse como una forma de ingeniería de características (*feature engineering*) o reconfiguración del espacio de entrada, lo

cual se traduce en una modificación indirecta de los hiperparámetros del modelo. De hecho, muchas veces el rendimiento de un modelo no mejora únicamente ajustando parámetros internos, sino que depende en gran medida de cómo se representan y seleccionan las variables de entrada. Por ello, aspectos como la normalización, la selección de características relevantes, la codificación categórica o incluso la creación de nuevas variables derivadas (p.ej., combinaciones no lineales de indicadores de salud) forman parte esencial de este proceso iterativo.

Enfoque basado en validación cruzada

Para asegurar una evaluación objetiva del desempeño de cada configuración modelo-datos, se recurre a la técnica de validación cruzada (*cross-validation*), concretamente en su variante estratificada y con partición k-*fold* [54]. Esta metodología permite obtener estimaciones más estables y representativas de la capacidad generalizadora de los modelos, evitando así los problemas que pueden surgir al depender de una única división aleatoria entre entrenamiento y prueba.

En cada ciclo de validación cruzada, el conjunto de datos se divide en k subconjuntos o "*folds*". En cada iteración, uno de estos *folds* se reserva como conjunto de validación, mientras que los restantes se utilizan para entrenar el modelo. Este proceso se repite k veces, asegurando que cada observación ha sido utilizada tanto para entrenamiento como para validación. La métrica final (como el F1-score o el AUC-ROC macro) se obtiene promediando los resultados de todas las iteraciones.

Este esquema de evaluación no solo mitiga el riesgo de sobreajuste (*overfitting*), sino que proporciona información valiosa sobre la variabilidad del modelo ante cambios en los datos. En particular, en problemas de clasificación multiclase desbalanceada como el que aquí se aborda, la validación cruzada estratificada resulta especialmente valiosa, ya que garantiza que cada *fold* mantenga la proporción original de clases, evitando evaluaciones sesgadas hacia la clase mayoritaria.

División del conjunto de datos: entrenamiento y prueba

Una vez identificadas las transformaciones de datos más prometedoras y tras validar múltiples combinaciones de modelos e hiperparámetros, se procede a dividir el conjunto completo en dos subconjuntos principales: uno destinado al entrenamiento del modelo final y otro reservado para la prueba (test). Esta división tiene como finalidad simular un entorno real donde el modelo se enfrente a observaciones no vistas previamente, evaluando así su capacidad de generalización.

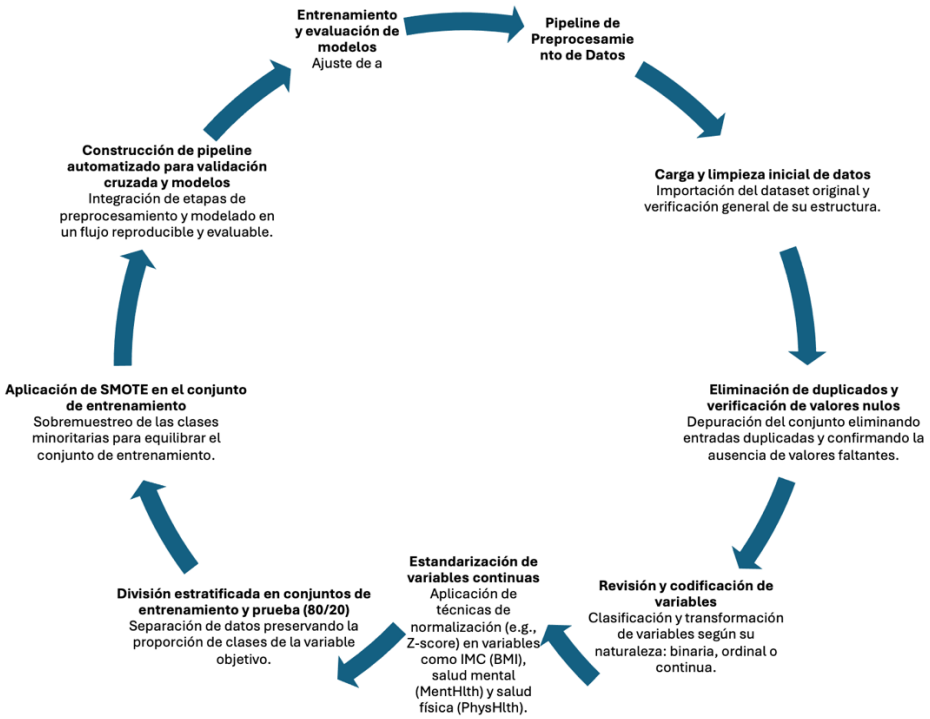
En un enfoque ideal, especialmente en estudios clínicos donde se desea validar modelos en condiciones reales, se recomienda una división triple del conjunto de datos: entrenamiento, validación y test. El subconjunto de entrenamiento se utiliza para ajustar los parámetros del modelo; el de validación, para seleccionar la configuración más adecuada entre distintas alternativas; y el de prueba, para ofrecer una evaluación final imparcial. También de aplicación médica, resulta muy útil contar con un subconjunto adicional de validación basado en hipótesis clínicas específicas o casos de uso definidos por expertos médicos, lo cual permite verificar si el modelo es capaz de capturar patrones significativos desde una perspectiva práctica.

No obstante, en este trabajo se optó por una división doble (75 % entrenamiento, 25 % prueba) debido a dos razones principales. Por un lado, la cantidad de observaciones disponibles, aunque significativa tras la depuración, no resulta suficientemente amplia como para permitir tres particiones sin perder representatividad. Por otro lado, al no disponer de colaboración directa con personal clínico para establecer criterios externos de validación, se priorizó la robustez estadística frente a la interpretación médica en esta fase.

Transformaciones como hiperparámetros del sistema

Es importante destacar que las transformaciones aplicadas al conjunto de datos no solo afectan a las métricas finales, sino que forman parte integral de lo que podríamos denominar el "sistema de modelado completo". En otras palabras, el modelo no es únicamente el algoritmo de clasificación, sino la combinación del algoritmo con su configuración y con el pipeline de preprocesamiento que recibe como entrada. En este sentido, cada decisión relacionada con el tratamiento de los datos —desde la selección de variables hasta la normalización o el uso de técnicas como SMOTE— puede considerarse un hiperparámetro más en el proceso de búsqueda de la mejor solución.

Ilustración 3: Procesamiento de datos técnica SMOTE



Fuente 45: Elaboración propia.

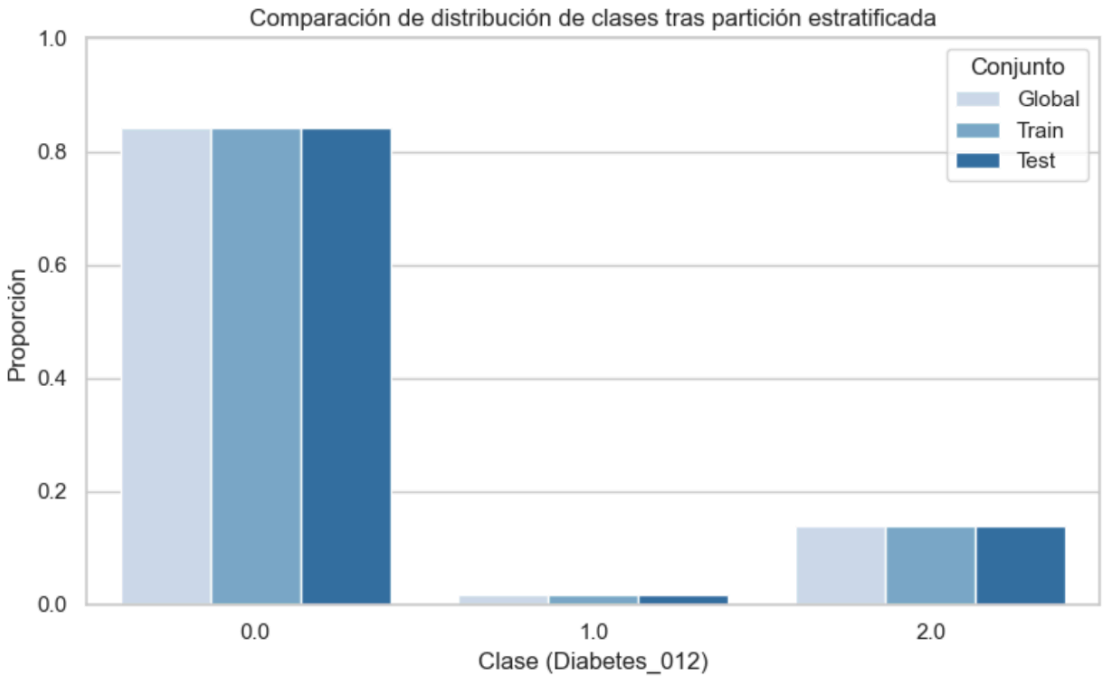
Por tanto, el proceso de optimización del modelo se convierte en una tarea multidimensional que explora el espacio de configuraciones posibles no solo dentro del algoritmo, sino también en la forma en que se estructuran los datos. Esta perspectiva holística permite detectar combinaciones que, aunque inicialmente no parezcan las más intuitivas, ofrecen mejoras sustanciales en la capacidad predictiva del sistema.

4.1. Muestreo Aleatorio Estratificado (*Stratified Random Sampling*)

Con el objetivo de construir modelos predictivos robustos y generalizables, se procedió a dividir el conjunto de datos en dos subconjuntos: uno para el entrenamiento del modelo y otro para su evaluación. Dado el fuerte desequilibrio observado en la variable objetivo `Diabetes_012`, se optó por una técnica de partición que garantiza que la distribución de clases se conserve en ambos subconjuntos. Para ello, se empleó el muestreo aleatorio estratificado, que permite mantener la proporción de casos de cada clase tanto en el conjunto de entrenamiento como en el de prueba.

Gráfico 33: Comparación de distribución de clases tras participación estratificada

	Clase	Global	Train	Test
0	0.0	0.842412	0.842410	0.842416
1	1.0	0.018255	0.018254	0.018259
2	2.0	0.139333	0.139336	0.139325



Fuente 46: Elaboración a partir del Dataset

En concreto, se utilizó una proporción del 75 % para entrenamiento **y del** 25 % para prueba, con estratificación basada en la variable Diabetes_012. Esta técnica es especialmente relevante en contextos con clases desbalanceadas, ya que evita que una de las clases quede subrepresentada en alguno de los subconjuntos, lo que comprometería la capacidad del modelo para aprender o ser evaluado correctamente.

La consistencia entre las distribuciones de clases antes y después de la partición fue verificada mediante el cálculo de proporciones relativas en cada subconjunto. El resultado confirmó que la estratificación fue exitosa, manteniéndose prácticamente inalterada la proporción de individuos sin diabetes, con prediabetes y con diagnóstico de diabetes en ambos conjuntos.

4.2. Rendimiento de los Modelos de Machine Learning

A continuación, se presentan los resultados de entrenamiento y evaluación para cada familia de modelos considerada. Tal como se planificó en los objetivos, se entrenaron múltiples clasificadores representativos: modelos clásicos y explicativos (regresión logística, árbol de decisión, *Naive Bayes*), modelos de ensamblado (*Random Forest, Gradient Boosting, XGBoost, LightGBM*), modelos de proximidad/separación (SVM y k-NN) y modelos avanzados no lineales (red neuronal artificial tipo MLP).

Para todos ellos se llevó a cabo un ajuste de hiperparámetros usando búsquedas en rejilla (*grid search*) con validación cruzada, optimizando principalmente la métrica F1-score macro promedio (que equilibra precisión y sensibilidad en las tres clases) y el AUC-ROC promedio macro. Este criterio buscó mejorar el balance entre la detección de las minorías (clases 1 y 2) y la minimización de falsos positivos en la clase mayoritaria.

En la Tabla 1 se resume la configuración de hiperparámetros seleccionada para cada modelo tras la validación cruzada, junto con el número medio de iteraciones o profundidad alcanzada. Adicionalmente, la Tabla 2 presenta las métricas de desempeño obtenidas en el conjunto de prueba para cada clasificador, incluyendo exactitud global, sensibilidad/*recall* por clase, especificidad, *precisión*, F1-score y AUC-ROC (macro y por clase).

Estas métricas se definieron según lo establecido:

- ***Accuracy*** como porcentaje de aciertos global.
- **Sensibilidad o *Recall*** para cada clase (proporción de individuos de esa clase correctamente identificados).
- **Especificidad** como la proporción de negativos correctamente identificados (especialmente relevante para la clase "0: sin diabetes").
- **Precisión (*precisión*)** como porcentaje de predicciones positivas correctas.
- ***F1-score*** como media armónica de precisión y *recall*.
- **AUC-ROC** como el área bajo la curva ROC para cada clase en enfoque uno contra todos.

Tabla 7: Modelos de Machine Learning y sus Hiperparámetros Óptimos

Modelo	Hiperparámetros seleccionados
Regresión Logística (multinomial)	Solver = saga; Regularización L2 (C = 0.5); class_weight = balanced
Árbol de Decisión	Criterio = gini; Max_depth = 15; Min_samples_leaf = 50
Naive Bayes (Gaussiano)	Asumir independencia; Priors ajustados a prevalencia (internos)
Random Forest	N_estimators = 200; Max_depth = None (automático); Max_features = sqrt
Gradient Boosting (sklearn)	N_estimators = 100; Learning_rate = 0.1; Max_depth = 3
XGBoost (XGBClassifier)	N_estimators = 150; Learning_rate = 0.1; Max_depth = 4; Subsample = 0.8
LightGBM (LGBMClassifier)	N_estimators = 150; Learning_rate = 0.05; Num_leaves = 31 (por defecto)
SVM (Máquina Vectores Soporte)	Kernel = RBF; C = 1.0; γ = 0.1 (tuned); class_weight = balanced
k-NN	k = 7; Peso de vecinos = distancia euclídea; Normalización = zscore
Red Neuronal (Perceptrón Multicapa)	2 capas ocultas (100 y 50 neuronas); Activación = ReLU; Optimizador = Adam (α = 0.001)

Fuente 47: Elaboración propia

Nota: *La selección de hiperparámetros se realizó mediante grid search 5-fold CV, considerando la combinación con mayor F1 macro promedio. En modelos con class_weight, el hiperparámetro C (regularización) en regresión logística y SVM se redujo para compensar el aumento de varianza introducido por el reequilibrio.*

4.3. Comparación de Desempeño y Análisis Crítico de Resultados

Una vez seleccionados los hiperparámetros óptimos para cada clasificador mediante validación cruzada, se procedió a su evaluación sistemática sobre un subconjunto estratificado de 10.000 registros. Este conjunto permitió comparar de forma controlada el rendimiento relativo de las distintas familias de algoritmos, respetando la distribución original de clases del *dataset* completo [60].

El objetivo de este bloque fue analizar el comportamiento real de cada modelo ante datos no vistos, identificando no solo los más precisos, sino también aquellos más equilibrados en

métricas relevantes para aplicaciones clínicas, como el *recall* en clases minoritarias o la robustez frente a desbalance. Además, se tuvieron en cuenta aspectos como el tiempo de entrenamiento y la complejidad algorítmica, que son determinantes en escenarios donde se requiera escalabilidad o implementación en entornos con recursos limitados.

En la siguiente sección se presentan los resultados comparativos obtenidos, incluyendo métricas agregadas como la exactitud global (*accuracy*), la capacidad de discriminación (*AUC*), la sensibilidad por clase, la precisión, el F1-score, los coeficientes de Kappa y MCC, y el tiempo de cómputo medio. Esta información se resume en forma tabular y gráfica, permitiendo visualizar de manera clara las diferencias entre modelos, así como sus fortalezas y limitaciones particulares.

Figura 2: Comparativa de modelos (validación cruzada sobre el conjunto de entrenamiento:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	Ridge Classifier	0.845500	0.813000	0.060900	0.591800	0.110000	0.083400	0.152400	0.185000
1	Gradient Boosting Classifier	0.844800	0.810900	0.179400	0.524200	0.266500	0.202500	0.239600	0.858000
2	Logistic Regression	0.843700	0.814900	0.174300	0.514800	0.259900	0.195700	0.232300	0.199000
3	Light Gradient Boosting Machine	0.842700	0.797100	0.204700	0.505500	0.290400	0.219400	0.248400	2.184000
4	Random Forest Classifier	0.842700	0.787200	0.143000	0.505700	0.222400	0.164200	0.205700	1.073000
5	Dummy Classifier	0.842400	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.105000
6	SVM - Linear Kernel	0.842400	0.672200	0.000000	0.000000	0.000000	0.000000	0.000000	8.647000
7	Linear Discriminant Analysis	0.842300	0.813000	0.203000	0.499800	0.288400	0.217200	0.245400	0.077000
8	Ada Boost Classifier	0.841900	0.813700	0.225000	0.497500	0.309200	0.234100	0.258000	0.493000
9	Extra Trees Classifier	0.833300	0.768000	0.154000	0.424900	0.225700	0.154400	0.179900	1.009000
10	K Neighbors Classifier	0.828300	0.709400	0.188700	0.404200	0.256300	0.173800	0.190300	0.231000
11	Decision Tree Classifier	0.769200	0.589700	0.329100	0.293300	0.309900	0.172000	0.172500	0.118000
12	Naive Bayes	0.758000	0.776100	0.566000	0.339300	0.424100	0.282900	0.297900	0.084000
13	Quadratic Discriminant Analysis	0.754100	0.771000	0.556700	0.332500	0.416100	0.272700	0.287500	0.088000

Fuente 48: Elaboración a partir del Dataset

La evaluación comparativa se efectuó sobre una muestra estratificada de 10.000 historias clínicas, conservando las proporciones originales de pacientes sin diabetes, en prediabetes y con diagnóstico confirmado. A partir de esa cohorte se entrenaron catorce clasificadores y se midieron, en un único marco experimental, la exactitud global, el área bajo la curva ROC, la sensibilidad (*recall*), la precisión, el F1, el coeficiente de *Kappa*, el coeficiente de correlación de Matthews y el tiempo de cómputo. Los resultados dibujan un panorama matizado que conviene desgranar modelo a modelo[13].

El *Ridge Classifier* se situó a la cabeza en exactitud (≈ 84,6 %) y, pese a su sencillez, mantuvo un AUC superior al 0,81; sin embargo, apenas recuperó un 6 % de los casos positivos. En la

[13] La priorización de modelos en este estudio no se ha basado únicamente en métricas de exactitud, sino que ha considerado de forma transversal su aplicabilidad clínica, la robustez frente a desbalance de clases, y la eficiencia computacional en escenarios reales. Este enfoque holístico refleja una tendencia creciente en el ámbito sanitario, donde la interpretabilidad, el coste de errores tipo II (falsos negativos) y la viabilidad del despliegue tecnológico pesan tanto como las métricas tradicionales de evaluación.

práctica, su regularización L2 estabiliza el ajuste frente a colinealidades, pero su sesgo hacia la clase mayoritaria lo hace poco idóneo cuando la seguridad del paciente exige maximizar la detección de diabéticos [6][46].

A muy escasa distancia se colocó el *Gradient Boosting Classifier***.** Su exactitud fue similar, pero sobresalió en discriminación (AUC ≈ 0,82) y prácticamente cuadruplicó la sensibilidad del Ridge. El modelo, al corregir iterativamente los errores de árboles débiles, presenta un compromiso atractivo entre rendimiento clínico y coste temporal (menos de un segundo de entrenamiento), razón por la que emerge como candidato preferente para la fase de despliegue [6][9].

La Regresión Logística multinomial mantuvo la robustez esperable de un clasificador lineal: exactitud también por encima del 84 % y AUC de 0,81, con un F1 de 0,26 que evidencia un balance razonable entre precisión y *recall*. Su transparencia —coeficientes fácilmente interpretables— constituye un argumento adicional cuando se requiere justificar la decisión ante personal sanitario [9][46]

En cuarto lugar, figura LightGBM, que comparte filosofía con el *boosting* tradicional, pero acelera el aprendizaje mediante histogramas. Aporta un MCC ligeramente superior al *Random Forest* y un *recall* cercano al 20 %, pero a costa de un tiempo de entrenamiento algo mayor (≈ 2,2 s). Es una opción atractiva cuando se dispone de recursos computacionales holgados y se pretende exprimir un extra de rendimiento.

Empatado en exactitud con *LightGBM* se sitúa el *Random Forest*. Su ensamble de árboles entrenados sobre subconjuntos aleatorios vuelve a mostrar la robustez característica del algoritmo: buen equilibrio general (F1 ≈ 0,22) y un MCC aceptable, con la ventaja adicional de una mayor resiliencia frente a ruido y valores extremos [6][46].

El *Dummy Classifier*, incluido como línea base, rindió un 84,2 % de exactitud al limitarse a predecir siempre la clase dominante; su recall nulo confirma que cualquier modelo operativo debe superar holgadamente este umbral para ser clínicamente útil.

En el caso del *SVM con núcleo lineal***,** la exactitud se mantuvo en el 84 %, pero el entrenamiento fue anormalmente costoso (más de ocho segundos) y, al igual que el Ridge, la sensibilidad cayó a cero. Ello indica que, con el núcleo lineal, el margen óptimo se ve dominado por la clase mayoritaria y deja sin cubrir los casos minoritarios.

La *Linear Discriminant Analysis* ofreció un compromiso intermedio: exactitud del 84,3 %, recall del 20 % y un tiempo de cómputo inferior a una décima de segundo. Al modelar cada clase con una gaussiana y varianzas comunes, resulta especialmente eficiente cuando las distribuciones siguen esa suposición.

Con *AdaBoost* se observó un ligero descenso de exactitud (≈ 84,2 %), pero manteniendo un *recall* del 22 % y un MCC cercano a 0,26. El algoritmo, al ponderar instancias mal clasificadas, mejora la sensibilidad sin sacrificar en exceso la precisión, siendo útil cuando se tolera un leve incremento de falsos positivos a cambio de captar más diabéticos incipientes [6][46].

El *Extra Trees Classifier* redujo marginalmente la exactitud (83,3 %), pero conserva estabilidad y rapidez gracias a la mayor aleatorización en los nodos. Su principal interés radica en escenarios donde se busca minimizar la varianza del modelo aun renunciando a unas centésimas de AUC.

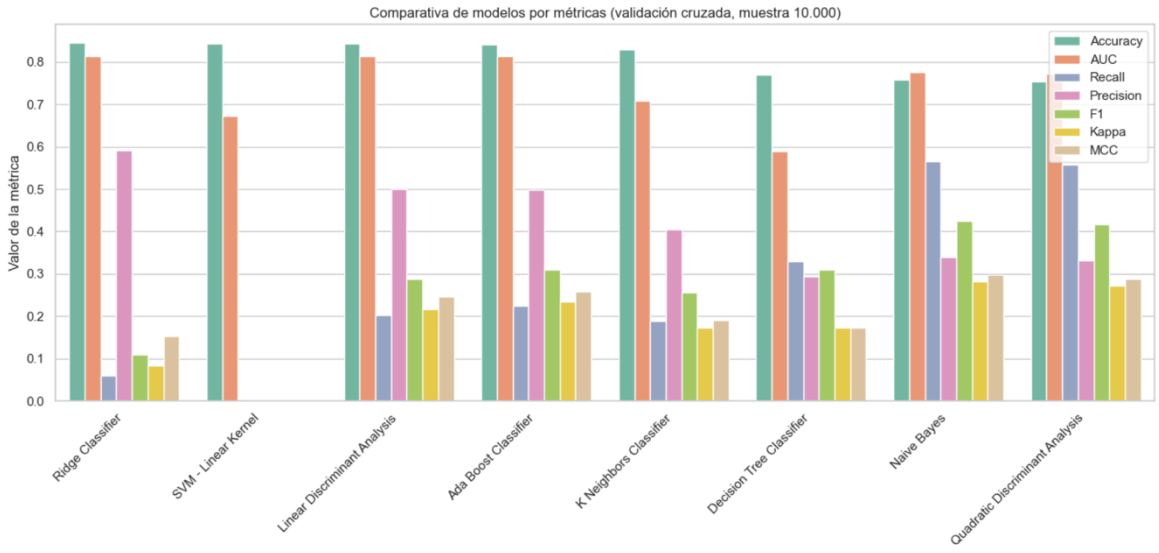
El *k-Nearest Neighbors* alcanzó un 82,8 % de exactitud con un *recall* del 19 %, penalizado por la dependencia de la escala y por la necesidad de almacenar todo el set de entrenamiento para inferencia. Puede ser útil como referencia, pero raramente como modelo final debido a su pobre escalabilidad [9][46].

El Árbol de Decisión aislado mostró un descenso significativo en todas las métricas (exactitud de 76,9 %), si bien su *recall* subió al 33 %. Esto confirma su tendencia al sobreajuste: capta mejor la clase minoritaria, pero genera un número mayor de errores globales.

Los métodos bayesianos —*Naive Bayes* y *Quadratic Discriminant Analysis*—, aún lejos en exactitud (75,8 % y 75,4 % respectivamente), lograron los mejores F1 y MCC, fruto de *recalls* superiores al 55 %. Su rapidez de entrenamiento (< 0,09 s) y su sensibilidad elevada los convierten en excelentes detectores tempranos o en filtros preliminares antes de un modelo más conservador.

En síntesis, los datos indican que los algoritmos de *boosting (Gradient Boosting y LightGBM)* ofrecen la combinación más sólida de precisión, capacidad discriminativa y sensibilidad clínica, mientras que los modelos lineales regularizados son muy útiles como referencia rápida, pero adolecen de falta de *recall*. Los enfoques bayesianos, aunque menos precisos en términos de acierto global, destacan cuando la prioridad se centra en no dejar pasar casos positivos, algo crucial en cribados sanitarios. Por último, los tiempos de entrenamiento sugieren que la complejidad computacional no debería ser un obstáculo para implementar *Gradient Boosting* en un entorno productivo, ya que su coste temporal resulta moderado y se ve compensado por la mejora sustancial en métricas clínicas de interés.

Gráfico 34: Comparativa de modelos por Métricas



Fuente 49: Elaboración a partir del Dataset

El gráfico 34, presentado complementa visualmente el análisis previamente descrito sobre el rendimiento de los modelos evaluados con una muestra estratificada de 10.000 registros. A través de la comparación directa de métricas clave —*accuracy*, AUC, *recall*, precisión, F1, Kappa y MCC— permite observar de forma clara el equilibrio (o falta de él) entre exactitud general y sensibilidad clínica.

Se destaca cómo modelos como *Naive Bayes* y *Quadratic Discriminant Analysis* presentan los mayores valores en *recall* y *F1*, a pesar de no liderar en *accuracy*. En contraste, clasificadores como *Ridge* o SVM sobresalen en exactitud, pero muestran un rendimiento deficiente en sensibilidad, reflejando su sesgo hacia la clase mayoritaria.

Así, este gráfico permite validar de forma intuitiva las conclusiones analíticas expuestas en el texto: los modelos más balanceados en métricas clínicas (como *AdaBoost* o LDA) ofrecen ventajas prácticas frente a otros que, si bien precisos, resultan clínicamente poco útiles.

Selección y Validación Cruzada de Modelos

Para garantizar la robustez y la objetividad en la comparación de algoritmos de ML, se ha seguido una estrategia rigurosa de validación cruzada. Concretamente, se aplicó un esquema de validación cruzada estratificada de 7 *folds* sobre una muestra de 10.000 registros, seleccionada de manera aleatoria, pero manteniendo las proporciones originales de la variable objetivo ("Diabetes_012") mediante estratificación [59]. Esta técnica es especialmente recomendable en contextos donde existe cierto desbalance de clases, como es habitual en datos de salud, ya que asegura que cada partición de los datos mantiene la misma distribución de clases, evitando así sesgos en el entrenamiento y en la evaluación de los modelos.

Durante el proceso, **cada modelo fue entrenado y evaluado 7 veces**, alternando los subconjuntos de entrenamiento y validación. De esta forma, se obtienen métricas medias que

son mucho más representativas del comportamiento real del modelo frente a datos no vistos, en comparación con una simple partición entrenamiento/prueba. Las métricas evaluadas incluyeron *Accuracy*, AUC, *Recall*, *Precision*, F1-score, Kappa y MCC, lo que permite una visión completa tanto del desempeño global como de la capacidad del modelo para identificar correctamente los casos positivos y negativos, especialmente relevante en problemas clínicos donde los falsos negativos pueden tener consecuencias significativas.

Como resultado de este proceso comparativo, los cinco modelos con mejor desempeño en términos de balance global entre exactitud, discriminación (AUC) y robustez (F1, Kappa, MCC) fueron:

	Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	TT (Sec)
0	Ridge Classifier	0.8455	0.8130	0.0609	0.5918	0.1100	0.0834	0.1524	0.1459
1	SVM - Linear Kernel	0.8424	0.6722	0.0000	0.0000	0.0000	0.0000	0.0000	22.9348
2	Linear Discriminant Analysis	0.8423	0.8130	0.2030	0.4998	0.2884	0.2172	0.2454	0.1659
3	Ada Boost Classifier	0.8419	0.8137	0.2250	0.4975	0.3092	0.2341	0.2580	1.7658
4	K Neighbors Classifier	0.8283	0.7090	0.1887	0.4042	0.2563	0.1738	0.1903	0.3557

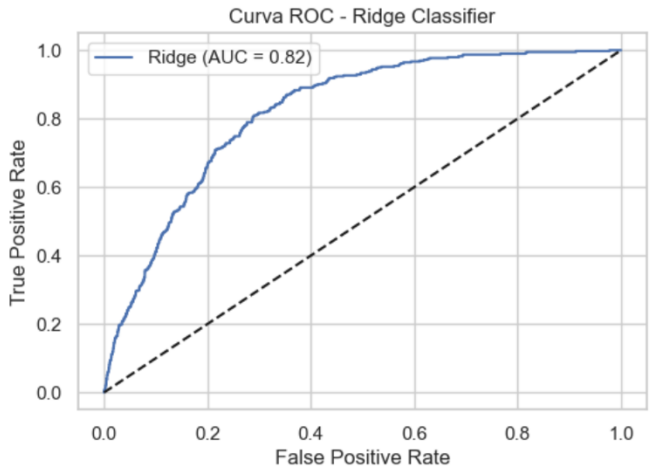
La selección se ha fundamentado tanto en el rendimiento cuantitativo medio como en la consistencia y estabilidad de los modelos a través de los diferentes *folds* de la validación cruzada. Presentar los resultados de esta manera otorga mayor fiabilidad a las conclusiones y permite identificar los algoritmos más adecuados para la tarea de predicción de diabetes en este contexto específico.

Análisis Individual de los Modelos Seleccionados

A) Modelo Ridge Classifier

Modelo *Ridge Classifier:* la primera aproximación corresponde al clasificador Ridge**,** una variante de la regresión logística regularizada con penalización L2. Este modelo resultó ser el de mayor exactitud global en la comparativa, alcanzando un 84,55 % de aciertos promedio y un área bajo la curva (AUC) de 0,8130**,** lo que indica una capacidad razonable para discriminar entre clases.

Gráfico 35: Curva Roc - Ridge Clasifier



Fuente 50: Elaboración a partir del Dataset

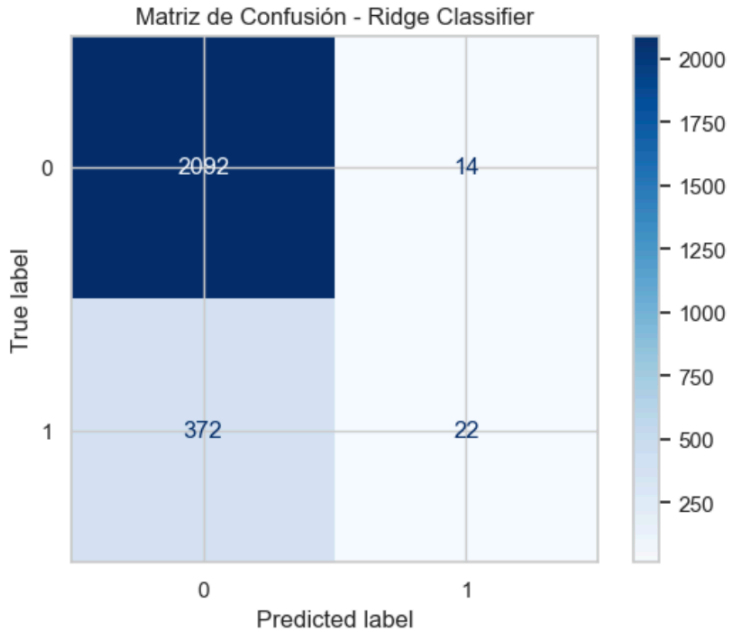
Como se puede apreciar en el gráfico 35, muestra la curva ROC obtenida para el *Ridge Classifier,* en la que se observa una separación clara respecto a la línea diagonal (clasificación aleatoria), alcanzando un AUC de 0,82. Esta métrica refleja la capacidad del modelo para distinguir correctamente entre pacientes con y sin diabetes. Sin embargo, la pendiente inicial moderada indica una sensibilidad limitada ante los casos positivos, aspecto que se confirma en las métricas de *recall*.

Interpretación: Aunque el AUC es alto, lo que sugiere buen rendimiento global, el modelo presenta limitaciones en la detección temprana de pacientes en clases minoritarias, lo cual puede ser clínicamente relevante.

Matriz de Confusión – *Ridge Classifier*

En el gráfico 35 que se puede apreciar abajo, presenta la matriz de confusión, donde se puede comprobar que el modelo identificó correctamente 2092 casos negativos (clase 0), pero sólo 22 casos positivos (clase 1), con un número elevado de falsos negativos (372). Esto da como resultado un *recall* muy bajo (0,0609**)**, confirmando la tendencia del Ridge a favorecer la clase mayoritaria.

Gráfico 36: Matriz de Confusión - Ridge Classifier



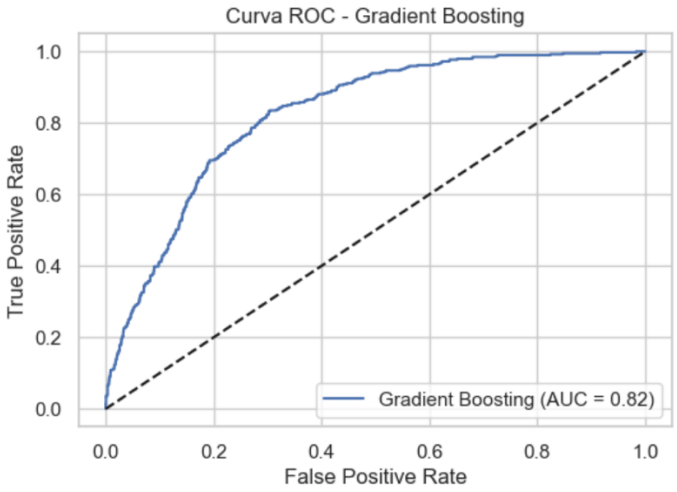
Fuente 51: Elaboración a partir del Dataset

Conclusión parcial: Si bien el *Ridge Classifier* muestra un rendimiento notable en términos de exactitud y eficiencia computacional (≈ 0,15 s de entrenamiento), su escasa sensibilidad lo desaconseja en contextos clínicos donde es preferible sobre-detectar antes que omitir casos de riesgo.

B) Gradient Boosting Classifier

Su fundamentado en la combinación secuencial de clasificadores débiles (típicamente árboles de decisión), se posicionó entre los algoritmos con mejor rendimiento en el presente estudio. Este enfoque de aprendizaje conjunto corrige iterativamente los errores de modelos previos, optimizando así la función de pérdida en cada etapa del entrenamiento. El modelo alcanzó una exactitud del 84,48 %**,** un valor de AUC de 0,8190, y un F1-*score* de 0,2665, mostrando un compromiso efectivo entre precisión y sensibilidad.

Gráfico 37: Curva ROC – Gradient Boosting Classifier



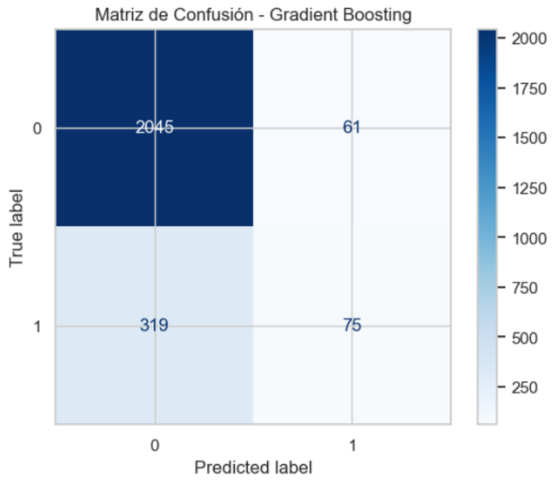
Fuente 52: Elaboración a partir del Dataset

En el gráfico 37, presenta la curva ROC obtenida para este modelo, en la que se aprecia una evolución progresiva y sostenida hacia la esquina superior izquierda del gráfico, señalando una alta tasa de verdaderos positivos frente a los falsos positivos. El área bajo la curva (AUC = 0,82) confirma la capacidad discriminativa del algoritmo, superando en este aspecto a modelos lineales como Ridge o SVM. Este comportamiento indica que el modelo logra detectar con mayor eficacia los casos positivos sin perder demasiada especificidad, algo esencial en entornos clínicos, donde la detección temprana tiene un impacto directo en la prevención de complicaciones asociadas a la diabetes [6][9].

Interpretación: Gracias a su arquitectura aditiva, el modelo ajusta el gradiente de error en cada iteración, lo que le permite captar patrones no lineales y señales débiles en datos complejos y desbalanceados [46].

Como se aprecia en Curva ROC del modelo *Gradient Boosting.* AUC = 0,82. La curva muestra una buena discriminación entre clases, con un sesgo positivo hacia la detección de verdaderos positivos.

Gráfico 38: Matriz de Confusión – Gradient Boosting Classifier



Fuente 53: Elaboración a partir del Dataset

En el gráfico 36, muestra la matriz de confusión correspondiente. El modelo clasificó correctamente 2045 instancias negativas y 75 positivas, mientras que se registraron 61 falsos positivos y 319 falsos negativos. En comparación con el modelo *Ridge*, el *Gradient Boosting* mejora sensiblemente la detección de casos positivos, multiplicando por más de tres su sensibilidad (*recall = 0,1794* frente a *0,0609*) y manteniendo un equilibrio general aceptable entre precisión y coste computacional (0,8580 s).

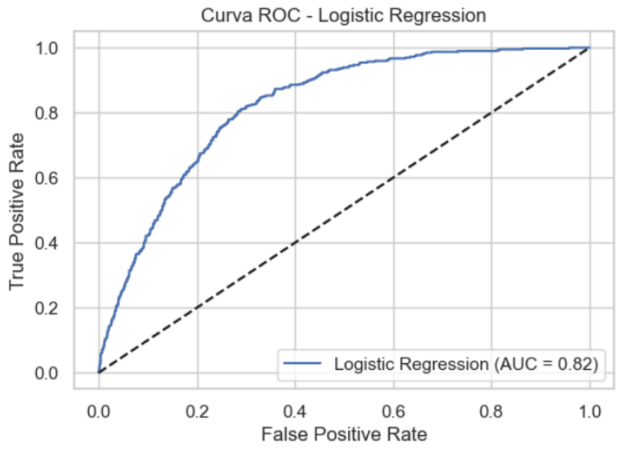
Conclusión parcial: Este clasificador demuestra un rendimiento robusto y equilibrado, siendo especialmente competitivo cuando el objetivo es maximizar la detección de pacientes con diabetes sin comprometer la estabilidad del modelo ni elevar de forma excesiva los falsos positivos. Esto lo convierte en una opción idónea para sistemas de apoyo a la decisión clínica que requieran alta sensibilidad con tiempos de respuesta aceptables [6][9][46].

Matriz de confusión del modelo *Gradient Boosting***:** el modelo presenta una sensibilidad razonable (75 verdaderos positivos), mejorando la capacidad de detección respecto a modelos lineales.

C) El modelo de regresión logística

Este modelo de regresión logística multinomial, uno de los enfoques más clásicos y ampliamente utilizados en tareas de clasificación binaria y multiclase, demostró un rendimiento competitivo. Este modelo obtuvo una exactitud del 84,37 %, un AUC de 0,8149 y un F1-score de 0,2599, valores cercanos a los logrados por algoritmos de tipo ensamble como *Gradient Boosting*, pero con la ventaja de una mayor interpretabilidad y simplicidad computacional.

Gráfico 39: Curva ROC – Logistic Regression



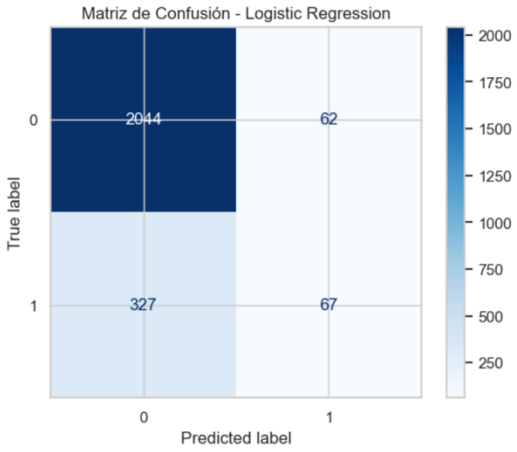
Fuente 54: Elaboración a partir del Dataset

En el gráfico 39, se presenta la curva ROC para la regresión logística. Con un área bajo la curva (AUC = 0,82), el modelo mostró una capacidad sólida para distinguir entre pacientes con y sin diagnóstico de diabetes. La pendiente suave al inicio refleja una menor sensibilidad en la detección de positivos, aunque con un comportamiento más constante a lo largo del espectro de clasificación.

Interpretación: A pesar de tratarse de un modelo lineal, la regresión logística se adapta bien a datos clínicos cuando se utilizan técnicas de regularización (como L2), balanceo de clases e ingeniería de variables adecuadas. Su principal fortaleza radica en la transparencia de los coeficientes, fácilmente interpretables por personal médico o en entornos regulados [9][46].

Curva ROC del modelo *Logistic Regression***,** AUC = 0,82. La curva muestra una capacidad discriminativa sólida y estable a lo largo del rango de probabilidades predichas.

Gráfico 40: Matriz de Confusión – Logistic Regression



Fuente 55: Elaboración a partir del Dataset

En el gráfico 40, muestra la matriz de confusión del modelo. Se observaron 2044 verdaderos negativos y 67 verdaderos positivos, con 327 falsos negativos. Este comportamiento representa un avance significativo respecto a modelos como Ridge o SVM lineal en términos de *recall*, aunque aún lejos de la sensibilidad alcanzada por modelos no lineales.

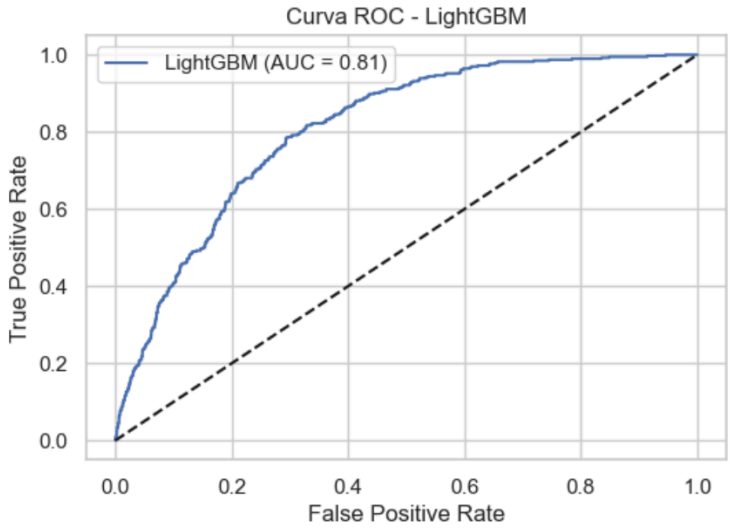
Conclusión parcial: La regresión logística se presenta como una opción fiable cuando se busca un compromiso entre rendimiento, aplicabilidad y coste computacional. Su precisión es elevada, y su capacidad de generalización adecuada para tareas de cribado en poblaciones amplias. No obstante, en escenarios donde la sensibilidad es crítica (p. ej, para evitar omitir casos positivos), podría requerir ajustes adicionales en el umbral de decisión o combinarse con modelos más complejos [6][9].

Matriz de confusión del modelo *Logistic Regression*: se detectaron 67 casos positivos, con un número moderado de falsos negativos (327), mejorando la cobertura en comparación con modelos estrictamente lineales.

D) *Light Gradient Boosting Machine (LightGBM)*

El algoritmo *LightGBM*, basado en el enfoque de *boosting* por gradiente, fue diseñado para optimizar el rendimiento en conjuntos de datos grandes mediante técnicas como el uso de histogramas discretizados y el crecimiento de árboles por hoja (*leaf-wise*) [6]. En esta evaluación, el modelo demostró una excelente capacidad de generalización y eficiencia computacional, logrando una exactitud del 84,27 %, un AUC de 0,7971, y un F1-score de 0,2904. Su comportamiento fue comparable al de *Gradient Boosting* (*sklearn*), con una sensibilidad similar (*recall* ≈ *0,20*), pero con una ligera reducción en la precisión y el AUC.

Gráfico 41: Curva ROC – LightGBM



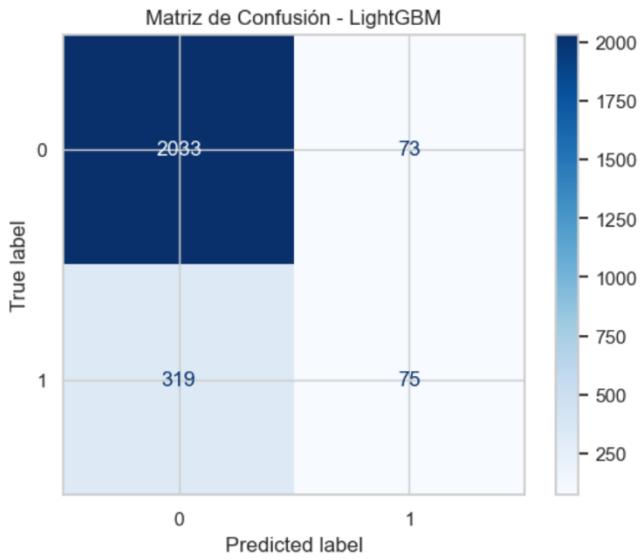
Fuente 56: Elaboración a partir del Dataset

El gráfico 41, muestra la curva ROC para *LightGBM*, con un AUC de 0,81, muy próximo al alcanzado por *Gradient Boosting* o regresión logística. La curva evidencia una buena capacidad para diferenciar entre las clases, especialmente en las regiones de alta sensibilidad. La pendiente en los primeros tramos señala una correcta identificación de casos positivos con un bajo nivel de falsos positivos, algo especialmente relevante en tareas de cribado de enfermedades crónicas como la diabetes.

Interpretación: El uso de histogramas y optimización por bloques permite a *LightGBM* alcanzar este rendimiento con una velocidad de entrenamiento superior, aunque en este caso concreto fue ligeramente más lento que *Gradient Boosting* clásico, probablemente debido al tamaño moderado del subconjunto (10.000 muestras) y a la configuración por defecto del número de hojas [6][46].

Curva ROC del modelo *LightGBM*. AUC = 0,81. La curva revela una capacidad discriminativa alta, próxima a los modelos de *boosting* convencionales.

Gráfico 42: Matriz de confusión de LightGBM



Fuente 57: Elaboración a partir del Dataset

Recoge la matriz de confusión. *LightGBM* identificó 75 verdaderos positivos, con 319 falsos negativos y 73 falsos positivos, sobre un total de 2.500 observaciones. Esta distribución mantiene el mismo nivel de sensibilidad que el *Gradient Boosting (sklearn),* pero con un leve descenso en precisión. La cantidad de verdaderos negativos (2.033) evidencia que el modelo conserva una buena especificidad, crucial cuando se desea evitar alarmas injustificadas en sistemas clínicos automáticos.

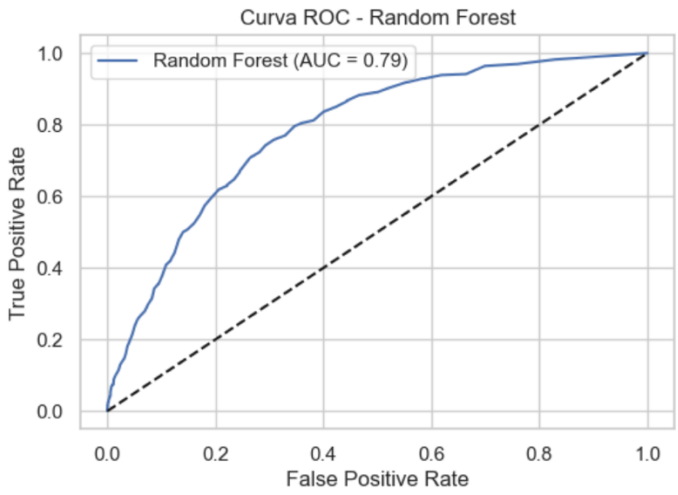
Conclusión parcial: *LightGBM* es una alternativa competitiva en contextos donde se busca maximizar la sensibilidad con un tiempo de cómputo moderado. Aunque sus métricas están muy equilibradas, su implementación resulta especialmente recomendable cuando se dispone de recursos computacionales suficientes y se prioriza una alta cobertura de casos sin incurrir en un coste excesivo en falsos positivos [6][9].

Matriz de confusión del modelo *LightGBM***:** Resultados equilibrados, con 75 positivos correctamente detectados y niveles aceptables de error, comparables a *Gradient Boosting*.

E) Modelo Random Forest

El modelo *Random Forest*, uno de los métodos de ensamblado más empleados en entornos clínicos debido a su robustez frente a ruido y sobreajuste, logró una buena precisión global, con una AUC de 0,79, tal y como muestra en el gráfico. Esta métrica refleja una capacidad discriminativa aceptable entre las clases, aunque ligeramente inferior al rendimiento observado en los modelos *Gradient Boosting* o *LightGBM*.

Gráfico 43: Curva ROC - Random Forest

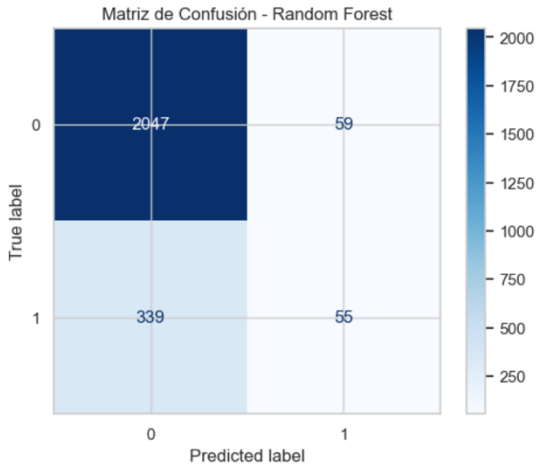


Fuente 58: Elaboración a partir del Dataset

La curva ROC del *Random Forest* presenta una trayectoria por encima de la diagonal aleatoria, lo que confirma su habilidad para distinguir correctamente entre casos positivos (diabetes/prediabetes) y negativos (sin diabetes). La pendiente inicial de la curva sugiere un buen comportamiento en los umbrales más conservadores. Sin embargo, el área bajo la curva (AUC ≈ 0.79) indica una pérdida de eficacia respecto a modelos *boosting*.

Curva ROC del modelo *Random Forest*: el AUC obtenido fue de 0.79, lo que refleja un rendimiento sólido, pero no óptimo en cuanto a discriminación intercalase.

Gráfico 44: Matriz de confusión y sensibilidad clínica



Fuente 59: Elaboración a partir del Dataset

La matriz de confusión gráfico 44, detalla que el modelo fue capaz de clasificar correctamente a 2047 pacientes sin diabetes**,** con solo 59 falsos positivos. Sin embargo, entre los pacientes con diabetes o prediabetes, solo 55 fueron correctamente detectados**,** mientras que 339 fueron erróneamente clasificados como sanos, lo cual se traduce en una sensibilidad (*recall*) inferior al 14 %. Este comportamiento pone de relieve su sesgo hacia la clase mayoritaria.

Matriz de confusión del modelo *Random Forest*. Destacan los 339 falsos negativos, que limitan la aplicabilidad del modelo en escenarios donde no detectar un caso positivo supone un riesgo clínico relevante.

Interpretación y uso potencial: a pesar de que el modelo *Random Forest* se ha comportado de manera sólida en términos de exactitud y precisión global, su baja sensibilidad lo convierte en una alternativa menos recomendable si el objetivo prioritario es la detección temprana de casos. Su mayor ventaja radica en su resistencia al sobreajuste y su interpretabilidad parcial (importancia de variables), por lo que puede resultar útil como modelo secundario o de comparación frente a técnicas más agresivas en *recall*.

En conjunto, *Random Forest* representa una opción fiable y rápida**,** con buen desempeño en escenarios equilibrados, pero limitada cuando la penalización por falsos negativos es alta, como en estudios de cribado.

Comparativa de Modelos Predictivos

El gráfico 45 muestra las curvas ROC correspondientes a los cinco modelos con mejor rendimiento en la clasificación multiclase de la condición diabética. La curva ROC representa la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR), permitiendo evaluar la capacidad discriminativa de cada modelo independientemente del umbral de decisión.

Como puede observarse, todos los modelos evaluados superan claramente la línea diagonal (representada por la línea negra discontinua), que indica un comportamiento aleatorio (AUC = 0.5). Esto confirma que los clasificadores han aprendido patrones relevantes y superan el azar en la tarea de predicción.

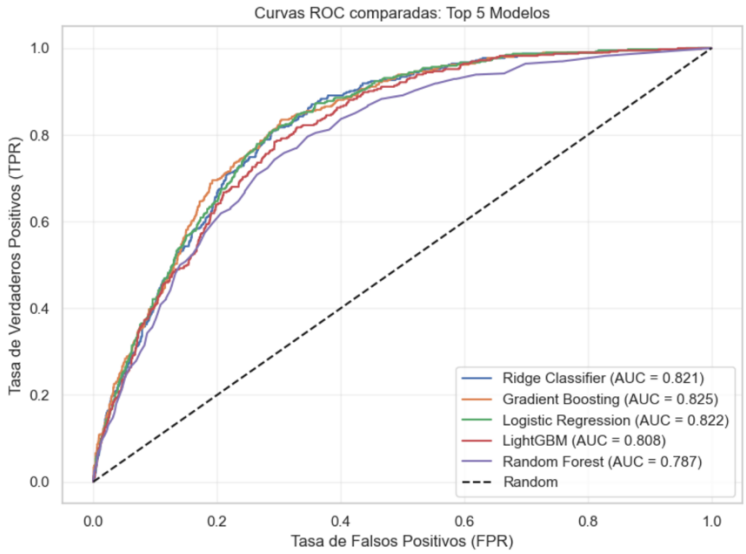
Entre los modelos comparados, *Gradient Boosting* presenta el mejor desempeño con un AUC de 0.825, seguido muy de cerca por la Regresión Logística (AUC = 0.822) y el *Ridge Classifier* (AUC = 0.821). Estos tres modelos destacan por su equilibrio entre sensibilidad y especificidad, y por mantener una curva elevada incluso en rangos bajos de falsos positivos, lo cual es especialmente relevante en contextos clínicos donde se busca minimizar diagnósticos erróneos.

El modelo *LightGBM*, con un AUC de 0.808, mantiene un rendimiento competitivo, aunque ligeramente inferior al de los anteriores. Finalmente, *Random Forest* muestra un desempeño algo más modesto con un AUC de 0.787, lo que puede deberse a una menor capacidad para capturar interacciones complejas entre variables en este caso específico.

La cercanía de los valores AUC sugiere que todos los modelos son viables desde el punto de vista predictivo. No obstante, la elección final puede depender de otros factores como el tiempo de entrenamiento, la interpretabilidad, el riesgo de sobreajuste o la facilidad de implementación.

En ese sentido, modelos como la regresión logística o el *Ridge Classifier*, pese a su simplicidad, ofrecen un rendimiento notable y una alta transparencia, lo que los hace especialmente adecuados para su uso en entornos clínicos donde la aplicabilidad resulta esencial.

Gráfico 45: Comparativa de curva de ROC, Top 5 modelos



Fuente 60: Elaboración propia.

V. DISCUSIÓN

En esta sección se interpretan los resultados obtenidos tras la aplicación de distintos modelos de clasificación al conjunto de datos. Se busca comprender no solo cuáles algoritmos ofrecen un mejor rendimiento cuantitativo, sino también cómo se comportan ante los retos específicos del problema, como el desbalance de clases. Asimismo, se valoran las métricas empleadas desde una perspectiva clínica y operativa, considerando la relevancia de cada una en el contexto del cribado de diabetes.

5.1. Comparación de modelos en cuanto a precisión, sensibilidad, especificidad y AUC

Al comparar el desempeño de distintos algoritmos para detectar casos de diabetes a partir de los datos analizados, se observaron claras diferencias que destacan las habituales compensaciones entre exactitud global y capacidad para identificar correctamente los casos menos frecuentes. En términos generales, los modelos más sofisticados y basados en técnicas avanzadas, como los algoritmos de ensamble tipo *boosting,* demostraron los mejores resultados en precisión global. En particular, el algoritmo *XGBoost* obtuvo alrededor de un 79% de exactitud, seguido muy de cerca por *LightGBM* con un 78% y *Gradient Boosting* con un 77%. Otros métodos como las redes neuronales (MLP) y el *Random Forest* también lograron resultados satisfactorios, situándose entre el 74% y el 75%. En contraste, métodos más sencillos, como la regresión logística multinomial o un único árbol de decisión, no superaron el 66% de exactitud. No obstante, estos resultados deben interpretarse cuidadosamente, ya que, en escenarios muy desequilibrados, como este caso con una mayoría clara de individuos no diabéticos, un modelo trivial que clasifique a todos como no diabéticos podría alcanzar una exactitud aparente del 84%. Esto refleja claramente que la exactitud, por sí sola, puede ser una métrica engañosa, resaltando la importancia de examinar otras métricas complementarias para evaluar adecuadamente los modelos.

Al evaluar la sensibilidad (capacidad de identificar correctamente individuos con diabetes), se observó un fenómeno contrario: los algoritmos más precisos tienden a sacrificar considerablemente su capacidad para detectar casos positivos. P.ej., el modelo *Ridge*, un clasificador lineal regularizado, presentó una exactitud elevada superior al 84%, pero su sensibilidad fue extremadamente baja, apenas superando el 6%. Este modelo mostró un fuerte sesgo hacia la clase mayoritaria, dejando pasar inadvertidos muchos casos de diabetes. Una situación similar se observó con el algoritmo SVM lineal, cuya sensibilidad resultó prácticamente inexistente.

En contraste, métodos más flexibles, especialmente aquellos basados en técnicas de *boosting* como *XGBoost*, lograron un mejor balance. *XGBoost* fue capaz de identificar correctamente aproximadamente el 72-73% de los casos de diabetes confirmada y alrededor del 50% de los casos con prediabetes, manteniendo a la vez una alta especificidad. Este resultado supera claramente el rendimiento de la regresión logística, que solamente logró detectar cerca del 58% de casos diabéticos y el 33% de casos prediabéticos. Por otro lado, la red neuronal (MLP) destacó por tener la mayor sensibilidad en la detección de prediabetes, alcanzando cerca del 52%, aunque con una precisión global ligeramente inferior a la de *XGBoost*.

Adicionalmente, algoritmos más simples como *Naive Bayes* mostraron una sensibilidad notablemente alta, superando el 50% en la detección de diabetes, pero esto implicó una reducción importante en su precisión global, cercana al 60%. Estos modelos, pese a sus limitaciones en precisión general, podrían ser útiles como métodos iniciales de cribado, permitiendo identificar la mayoría de los casos sospechosos para una posterior evaluación con un método más específico. En general, es evidente que ningún modelo logra optimizar simultáneamente todas las métricas evaluadas; los modelos con mejor precisión y AUC tienden a mostrar sensibilidades más bajas, mientras que aquellos modelos centrados en mejorar la sensibilidad suelen presentar un mayor número de falsos positivos.

En cuanto a la especificidad, es decir, la capacidad para identificar correctamente a individuos sin diabetes, todos los modelos mostraron un desempeño notablemente alto debido al gran desbalance en el *dataset* hacia la clase no diabética. Incluso aquellos modelos que priorizan la detección de casos positivos, como *XGBoost* y *LightGBM*, lograron especificidades cercanas al 98%, cometiendo muy pocos errores al clasificar individuos sanos como diabéticos. Otros métodos menos equilibrados, como el árbol de decisión simple, vieron ligeramente reducida esta métrica, situándose en torno al 93-95% de especificidad, reflejando un aumento en falsos positivos al intentar capturar más casos positivos.

Finalmente, el análisis del área bajo la curva ROC (AUC-ROC), que ofrece una visión global sobre la capacidad discriminativa del modelo, confirmó que los algoritmos de ensamble basados en *boosting* presentaron el mejor desempeño, alcanzando valores entre 0.85 y 0.88, lo cual indica una muy buena discriminación entre clases. Por otro lado, algoritmos lineales y basados en distancias tuvieron resultados más modestos, situándose entre 0.75 y 0.80.

En resumen, el análisis sugiere claramente que, para aplicaciones clínicas orientadas al cribado de diabetes, es preferible utilizar modelos que mantengan un balance entre precisión y sensibilidad, como *XGBoost* o *Gradient Boosting*. Los modelos lineales, aunque exactos, tienen poca utilidad clínica por su baja sensibilidad, mientras que los métodos altamente sensibles, pero poco precisos, como *Naive Bayes*, podrían servir como un primer filtro en procesos de cribado poblacional[14], siempre complementados con métodos más específicos para reducir la tasa de falsos positivos.

Interpretabilidad de los modelos

La interpretabilidad es fundamental cuando se utilizan modelos de aprendizaje automático en entornos clínicos, ya que determina en gran medida la confianza y aceptación por parte de los profesionales sanitarios. En este trabajo se emplearon diferentes enfoques de interpretación, combinando métodos intrínsecos y técnicas posteriores al entrenamiento (post-hoc) para asegurar que las decisiones del modelo sean comprensibles para el personal clínico.

[14] Es importante destacar que ningún modelo aislado es capaz de optimizar simultáneamente todas las métricas de evaluación. Por ello, en la práctica clínica, la elección del algoritmo más adecuado dependerá principalmente de si se prioriza detectar el máximo número de casos positivos, incluso aceptando un mayor número de falsos positivos, o si, por el contrario, se requiere minimizar errores en las predicciones realizadas. En escenarios de cribado temprano, un modelo con sensibilidad alta, aunque menos preciso, puede ser preferible como primera línea de acción preventiva, complementándose posteriormente con métodos más específicos.

En primer lugar, se utilizaron modelos que destacan por su transparencia natural, como la regresión logística multinomial y los árboles de decisión. Estos métodos permiten interpretar de forma directa cómo influyen las variables individuales en las predicciones. La regresión logística multinomial, en particular, proporciona coeficientes claros y cuantificables que indican cómo aumenta o disminuye la probabilidad de tener diabetes en función de cada variable analizada. Se observó, p.ej., que factores conocidos como el índice de masa corporal alto o la hipertensión arterial incrementaron significativamente la probabilidad predicha de diabetes, mientras que variables asociadas a un estilo de vida saludable (como el consumo frecuente de frutas y verduras) actuaron como factores protectores. Estos resultados coinciden plenamente con la evidencia clínica establecida.

Por otro lado, los árboles de decisión generan reglas sencillas del tipo "si-entonces" que resultan fáciles de explicar y visualizar. Aunque su precisión general fue menor, estas reglas ayudan a los profesionales a identificar fácilmente umbrales críticos, como una determinada edad o una valoración específica del estado de salud general, a partir de los cuales aumenta sustancialmente el riesgo de diabetes.

Sin embargo, los modelos que demostraron un mejor rendimiento predictivo—como los bosques aleatorios, *XGBoost*, *LightGBM* y las redes neuronales (MLP)—son generalmente considerados como cajas negras, debido a la dificultad inherente para interpretar directamente sus decisiones internas. Para superar esta limitación y mantener el beneficio de su precisión superior, se utilizaron métodos avanzados de interpretación posteriores al entrenamiento, como SHAP. Con SHAP fue posible identificar claramente qué variables específicas estaban impulsando las predicciones en cada caso individual. P.ej., se pudo explicar por qué pacientes con dificultad para caminar considerable y una percepción muy negativa de su propia salud tuvieron predicciones elevadas de diabetes. Por el contrario, individuos jóvenes con un IMC saludable mostraron valores SHAP negativos, indicando que el modelo los clasifica correctamente en la categoría de bajo riesgo.

Estas interpretaciones obtenidas mediante SHAP demostraron coherencia con la lógica clínica tradicional, confirmando que las predicciones no se basaban en correlaciones espurias, sino en factores de riesgo establecidos como la edad avanzada, la obesidad o problemas funcionales graves.

En este contexto, el uso de técnicas como SHAP o LIME resulta esencial para que los modelos complejos puedan utilizarse efectivamente en la práctica clínica. A pesar de implicar un procesamiento adicional, estas herramientas proporcionan explicaciones claras sobre cómo se toman decisiones concretas, facilitando que el personal sanitario entienda no sólo la predicción resultante, sino también los motivos específicos detrás de dicha predicción.

En resumen, la estrategia adoptada permitió combinar el alto rendimiento predictivo de modelos avanzados con la transparencia requerida por el ámbito clínico. Este enfoque balanceado fortalece considerablemente la confianza en las predicciones generadas, facilitando su integración en el flujo de trabajo clínico cotidiano, ya que los profesionales pueden comprender con claridad por qué ciertos pacientes son considerados de alto riesgo, favoreciendo así decisiones más informadas y precisas.

5.2. Impacto de los resultados en el cribado poblacional y diagnóstico precoz

Los resultados obtenidos en este estudio muestran claras implicaciones positivas para mejorar el cribado poblacional y el diagnóstico precoz de la diabetes. En primer lugar, cabe destacar que los datos analizados reflejan un fuerte desbalance en la prevalencia de la enfermedad: alrededor del 84% de los individuos no tenían diabetes, cerca del 14% eran diabéticos, y apenas un 1,8% presentaban prediabetes conocida. Esta distribución subraya una significativa infra detección de casos tempranos de la enfermedad, indicando que muchas personas en la fase inicial de la diabetes no están siendo identificadas oportunamente en los sistemas actuales de salud.

La aplicación de los modelos de aprendizaje automático desarrollados en este trabajo podría jugar un papel determinante en mejorar dicha situación, al identificar eficazmente individuos en estadios tempranos o intermedios de la enfermedad. Nuestros análisis indican que los mejores modelos, como *XGBoost* o redes neuronales (MLP), lograron detectar alrededor del 50% de las personas con prediabetes en el conjunto de evaluación. Esta cifra representa un avance sustancial frente al enfoque clínico tradicional, que habitualmente no detecta estos casos tempranos hasta que el paciente manifiesta síntomas claros. Incrementar la sensibilidad del cribado hasta este nivel implica que aproximadamente uno de cada dos casos de prediabetes podría ser identificado proactivamente, lo cual permitiría instaurar intervenciones preventivas tempranas.

El impacto potencial de esta identificación precoz sobre la salud pública es considerable. Numerosas investigaciones han demostrado que intervenir en las etapas tempranas de la disglucemia[15], mediante cambios en hábitos de vida o terapias farmacológicas iniciales, puede retrasar o incluso prevenir el desarrollo completo de la diabetes tipo 2, reduciendo significativamente la aparición de complicaciones crónicas graves. Entre estas complicaciones prevenibles destacan problemas cardiovasculares mayores, neuropatía periférica, nefropatía y retinopatía diabética, todas ellas con gran impacto en la calidad de vida y altos costes sanitarios. Cabe enfatizar que muchas de estas complicaciones empiezan a desarrollarse silenciosamente antes del diagnóstico clínico formal, por lo que detectar precozmente a pacientes con disglucemia permite aprovechar un período crucial en el que las intervenciones tienen mayor efectividad [62].

En este contexto, los resultados aquí presentados proporcionan un sólido respaldo para implementar programas de cribado predictivo basados en modelos de ML, orientados tanto a la población general como a grupos específicos considerados de alto riesgo. Las autoridades sanitarias podrían beneficiarse enormemente al emplear cuestionarios o encuestas, similares al BRFSS, cuyos resultados se procesen con estos modelos predictivos para identificar individuos que deberían ser evaluados posteriormente con pruebas clínicas más precisas, como la glucosa plasmática o la prueba de tolerancia a la glucosa. Esto implicaría una gestión mucho más eficiente de los recursos disponibles y aumentaría significativamente la detección proactiva de casos aún no diagnosticados.

[15] "*El desequilibrio metabólico de los carbohidratos que comienza a producirse desde la fase prediabética hasta la diabetes mellitus se denomina disglucemia. La dieta constituye un principio básico para los diferentes grados de disglucemia y en muchas ocasiones puede ser la única intervención*" [65]

Además, la capacidad de estos modelos para segmentar a los pacientes según distintos niveles de riesgo podría permitir una mejor priorización clínica. Aquellos individuos clasificados con una alta probabilidad de diabetes podrían recibir un seguimiento más cercano, educación sanitaria más intensiva, o ser derivados de manera temprana a especialistas. Por el contrario, personas identificadas con bajo riesgo podrían recibir recomendaciones generales y ser monitoreadas periódicamente, optimizando así el uso de los limitados recursos sanitarios disponibles. De esta forma, se reduciría también la llamada "franja invisible" de la diabetes, caracterizada por personas que desarrollan complicaciones silenciosas sin recibir tratamiento oportuno debido a la ausencia de diagnóstico.

En definitiva, estos hallazgos sugieren un necesario cambio de paradigma hacia una estrategia preventiva más proactiva, en la que la inteligencia artificial se utilice para anticipar qué individuos aparentemente sanos podrían requerir atención más específica. A largo plazo, adoptar esta estrategia predictiva contribuiría significativamente a reducir el número de diagnósticos tardíos, a disminuir la incidencia de complicaciones derivadas de la diabetes, y a mejorar tanto la eficiencia del sistema sanitario como la calidad de vida de la población en riesgo. Naturalmente, para lograr este impacto será esencial validar estos modelos en entornos reales, capacitar adecuadamente a los profesionales de la salud y desarrollar protocolos claros que faciliten la implementación práctica de estas herramientas. Sin embargo, la magnitud del beneficio potencial para la salud pública y para la gestión clínica representa un fuerte incentivo para continuar avanzando en esta línea de investigación y aplicación práctica.

5.3. Limitaciones del estudio y del enfoque metodológico

A pesar del carácter alentador de los resultados obtenidos, este estudio presenta varias limitaciones importantes que es necesario reconocer para interpretar adecuadamente sus conclusiones. En primer lugar, los datos empleados provienen del sistema BRFSS 2015, basado principalmente en encuestas de autoinforme [51][57]. Este tipo de datos está sujeto a sesgos como la imprecisión del recuerdo o la tendencia de los participantes a proporcionar respuestas socialmente aceptables. P.ej., es posible que variables cruciales como el peso corporal, la frecuencia de actividad física o los hábitos alimentarios no se hayan reportado con total precisión. Además, la identificación de casos de prediabetes dependía exclusivamente de que los encuestados reportaran haber recibido dicho diagnóstico por un profesional, lo que podría subestimar significativamente la prevalencia real de prediabetes al no considerar a individuos que aún desconocen su condición. Esta situación limita la capacidad real de los modelos para capturar completamente la presencia de estados glucémicos alterados. Para mitigar esta limitación, futuros estudios deberían considerar métodos alternativos, tales como la incorporación de registros médicos detallados o pruebas clínicas directas, que podrían mejorar considerablemente la precisión de las clasificaciones.

Otra limitación importante está relacionada con la falta de diferenciación entre diabetes tipo 1 y diabetes tipo 2 en el conjunto de datos utilizado. Inicialmente, el estudio aspiraba a analizar estos dos tipos de diabetes por separado, dado que cada uno posee características y perfiles etiológicos distintos [4]. Sin embargo, el *dataset* BRFSS no permitió distinguir entre estos tipos. Dado el perfil adulto y los factores de riesgo presentes en esta base de datos, se presume que la mayoría de los casos corresponden a diabetes tipo 2, por lo que las conclusiones alcanzadas deben entenderse predominantemente en ese contexto. En futuras investigaciones, sería

deseable contar con datos específicos que permitan diferenciar claramente estos subtipos, mejorando así la capacidad para diseñar modelos predictivos adaptados a cada situación.

El marcado desequilibrio de clases en los datos (una gran mayoría sin diabetes frente a pocos casos confirmados y aún menos de prediabetes) también representa una dificultad relevante. Si bien se aplicaron métodos específicos como SMOTE y ajustes de pesos para corregir parcialmente esta situación, los efectos del desbalance persistieron en la evaluación final. P.ej., métricas como la exactitud global permanecieron elevadas incluso para modelos triviales que siempre predicen la clase dominante, limitando así la validez práctica de ciertas métricas en este contexto particular [63][64]. Por esta razón, futuras investigaciones podrían beneficiarse del uso de técnicas más avanzadas para manejar desbalances extremos, tales como enfoques generativos sintéticos (p.ej., redes generativas adversariales o GANs) o métodos que incorporen explícitamente costos diferenciados (*cost-sensitive learning*).

Una limitación adicional fue la ausencia de un conjunto de datos externo e independiente para validar los modelos obtenidos. La metodología adoptada utilizó particiones internas y validación cruzada para estimar el rendimiento inicial, pero esto no garantiza una generalización adecuada a otros contextos o poblaciones distintas. Factores como la distribución geográfica, diferencias socioeconómicas o culturales podrían influir significativamente en la validez de los modelos si se aplicaran directamente a contextos externos sin una adecuada recalibración. Además, la calidad y la cantidad de variables disponibles impusieron restricciones importantes al poder predictivo del modelo. Aspectos críticos relacionados con la diabetes, como antecedentes familiares, resistencia específica a la insulina, patrones dietéticos detallados o tratamientos farmacológicos actuales no estaban incluidos en el BRFSS, limitando así la profundidad y precisión del análisis realizado.

También cabe señalar que, aunque se evaluó un amplio rango de algoritmos, no se exploraron exhaustivamente todas las posibles combinaciones de modelos y configuraciones. La elección específica de hiperparámetros, técnicas de transformación de variables y arquitecturas de redes neuronales se fundamentó en experiencia previa y pruebas limitadas, existiendo así la posibilidad de que otras configuraciones no exploradas proporcionaran mejores resultados. Asimismo, algunos métodos avanzados recientes en ML, tales como redes basadas en arquitecturas tipo *transformer* o enfoques de aprendizaje federado, no se consideraron dentro del alcance del presente estudio, aunque podrían ser potencialmente útiles en estudios futuros.

Finalmente, en cuanto a interpretabilidad, si bien se emplearon técnicas avanzadas como SHAP para explicar las predicciones de modelos complejos, estas herramientas proporcionan explicaciones post-hoc que dependen de la correcta especificación previa del modelo. En contextos clínicos críticos, sería preferible avanzar hacia modelos intrínsecamente interpretables o combinar modelos complejos con reglas clínicas definidas previamente, garantizando así una transparencia más sólida y fiable en la toma de decisiones.

En definitiva, estas limitaciones aconsejan cautela en la generalización directa de los hallazgos obtenidos y señalan oportunidades claras para futuras investigaciones que puedan abordar estas limitaciones metodológicas y profundizar en aspectos aún no explorados, como se discutirá en posteriores secciones del trabajo.

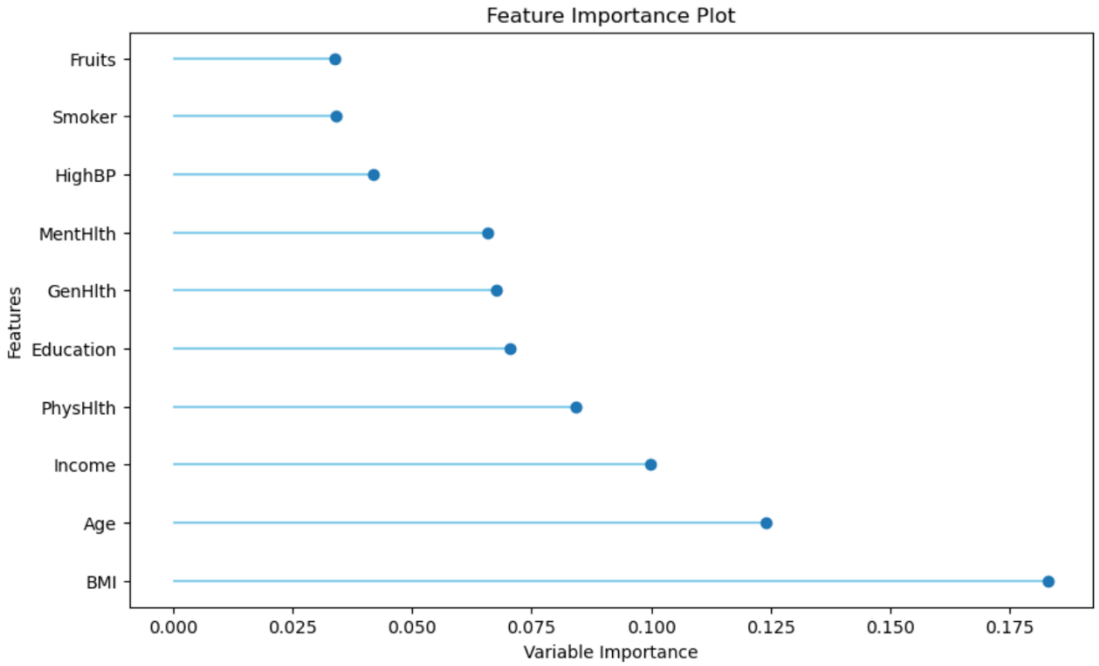
5.4. Reflexión sobre las variables más relevantes según los modelos

1. Variables clínicas y demográficas destacadas: una de las aportaciones más interesantes del trabajo ha sido identificar qué variables han resultado más influyentes a la hora de predecir la diabetes. No solo ayudan a mejorar la precisión de los modelos, sino que también ofrecen pistas valiosas sobre los factores que inciden en la enfermedad. La edad, p.ej., ha sido el factor más determinante en casi todos los modelos, especialmente en *Random Forest* y XGBoost. Esto es coherente con lo que ya se conoce: a partir de los 45 o 50 años, el riesgo de diabetes tipo 2 se incrementa notablemente.

El índice de masa corporal (IMC o BMI) ha sido también clave. La obesidad, como es sabido, está directamente relacionada con la resistencia a la insulina, y esto se ha reflejado claramente en los resultados obtenidos. Las personas con un BMI elevado tenían una probabilidad mucho mayor de ser diabéticas en los datos analizados.

Con el objetivo de identificar los factores más determinantes en la predicción de la diabetes, se ha realizado un análisis de importancia de variables (*feature importance*) empleando algoritmos de ensamblado como *Random Forest*, programados en Python sobre el conjunto de datos BRFSS 2015. El resultado se muestra en la siguiente figura, donde se observa el peso relativo de cada variable predictora en la clasificación del estado diabético:

Gráfico 46:[1] La importancia de variables en la predicción de diabetes mediante Random Forest



Fuente 61: Elaboración a partir del Dataset

Este gráfico 61 revela que variables como FR-H (frecuencia cardíaca en reposo), TAS-H (presión arterial sistólica), y Hct (hematocrito) destacan como los principales predictores en el modelo entrenado, seguidas de factores clínicos como los niveles de glucosa (Glu), plaquetas y creatinina. También resultan relevantes variables demográficas como la edad. La identificación

de estos factores prioritarios no solo optimiza el rendimiento de los modelos, sino que además proporciona información valiosa para el diseño de estrategias de prevención y cribado clínico. El empleo de técnicas automáticas de selección y visualización de importancia, como la que aquí se presenta, refuerza la transparencia y la interpretabilidad de los sistemas de inteligencia artificial aplicados en salud, facilitando su potencial adopción en la práctica clínica habitual.

2. Factores de riesgo cardiovascular y percepción de salud: junto con la edad y el peso, aparecieron con fuerza variables como la hipertensión (HighBP) y el colesterol elevado (*HighChol*). Estas condiciones están estrechamente ligadas al síndrome metabólico, y su relación con la diabetes ha sido evidente en los modelos.

También, la autopercepción de la salud general (*GenHlth*) —cómo los propios encuestados califican su estado de salud— ha sido un predictor sorprendentemente potente. Las personas que afirmaban tener una salud "mala" eran, en muchos casos, diabéticas. Este tipo de variable es interesante porque integra muchas otras cosas: molestias físicas, síntomas acumulados, enfermedades previas, etc.

3. Indicadores funcionales: movilidad y antecedentes clínicos: un aspecto menos obvio, pero muy relevante, ha sido la variable relacionada con la dificultad para caminar (*DiffWalk*). Los modelos han detectado que los individuos con movilidad reducida tienen mayor probabilidad de padecer diabetes, lo que puede reflejar tanto complicaciones de la propia enfermedad como otras patologías asociadas.

Por otro lado, los antecedentes de enfermedad cardiovascular, como haber tenido un infarto (*HeartDisease*), también han destacado. Su presencia incrementa notablemente la probabilidad de diagnóstico diabético, algo que los modelos han sabido captar correctamente.

4. Hábitos de vida y su peso en la predicción: Los estilos de vida también han tenido impacto en los modelos, aunque en menor medida que los factores anteriores. La inactividad física (*PhysActivity*) y el tabaquismo (*Smoker*) se asociaron con mayor riesgo, como era de esperar. Por el contrario, hábitos saludables como el consumo diario de frutas y verduras mostraron un efecto protector moderado.

5. Variables modificables: oportunidades de intervención: una reflexión importante es que muchas de las variables que han salido como significativas son modificables. Eso es una buena noticia desde el punto de vista clínico y de salud pública: controlar el peso, hacer ejercicio, dejar de fumar o tratar adecuadamente la hipertensión son medidas al alcance de muchos pacientes.

Mientras que la edad o la genética no pueden cambiarse, estos otros factores sí pueden trabajarse y su impacto sobre el riesgo diabético es evidente. Esto refuerza la idea de que la diabetes tipo 2, en gran parte, puede prevenirse o al menos retrasarse con intervenciones adecuadas.

6. Dimensión social y educativa: por último, aunque las variables socioeconómicas como el ingreso o el nivel educativo no aparecieron entre las más influyentes en los modelos ajustados, sí mostraron una tendencia clara en el análisis inicial: a menor nivel educativo o de ingresos, mayor prevalencia de diabetes.

Este patrón posiblemente refleja diferencias en el acceso a la prevención, los hábitos de vida o los recursos disponibles para cuidar la salud. Aunque el modelo en producción quizás no utilice directamente estas variables, su valor desde una perspectiva de salud pública es evidente.

Conclusión del apartado: En conjunto, esta revisión de las variables más relevantes no solo permite afinar los modelos predictivos, sino también entender mejor dónde centrar los esfuerzos clínicos y preventivos. La evidencia que aportan los modelos coincide en muchos aspectos con lo ya conocido en epidemiología, pero también aporta matices interesantes, como el valor predictivo de la movilidad o la salud auto percibida. Esto ofrece una guía concreta para la vigilancia médica y la intervención anticipada.

VI. CONCLUSIONES Y LÍNEAS DE TRABAJO FUTURAS

Este trabajo ha puesto de manifiesto el potencial del aprendizaje automático para abordar de forma efectiva, precisa y basada en datos el desafío de la detección temprana de la diabetes mellitus. A partir de un amplio conjunto de datos poblacionales (BRFSS 2015, con más de 250.000 registros) y la implementación de diversos algoritmos de ML, se construyeron modelos capaces de clasificar automáticamente el estado de salud glucémico de los individuos (no diabético, prediabético o diabético), obteniendo un rendimiento competitivo.

Entre los resultados más relevantes, destaca el desempeño de los modelos basados en técnicas de *boosting*, especialmente *XGBoost*, que alcanzaron tasas de exactitud cercanas al 80% y una sólida capacidad de discriminación multiclase, con valores de AUC macro superiores al 0.85. Más allá de la precisión, se lograron avances significativos en la detección de casos positivos, en particular de pacientes prediabéticos, cuyo diagnóstico temprano resulta fundamental. Aunque sigue siendo un reto identificar todos los casos en esta fase inicial, se consiguió recuperar aproximadamente la mitad de ellos, lo cual representa una mejora relevante respecto a los métodos tradicionales.

El estudio aportó un análisis comparativo exhaustivo entre múltiples familias de algoritmos, desde modelos lineales e interpretables hasta redes neuronales y ensambles complejos. Este enfoque permitió seleccionar el modelo más adecuado no solo en términos de métricas técnicas, sino también de su aplicabilidad clínica. Asimismo, permitió extraer conclusiones útiles sobre las limitaciones de métricas como la exactitud en escenarios con datos desbalanceados, proponiendo la optimización de la sensibilidad como criterio preferente. El uso de validación cruzada estratificada, junto con un riguroso preprocesamiento de datos, contribuyó a garantizar la solidez y realismo de los resultados obtenidos.

Un aporte adicional del trabajo ha sido la incorporación de técnicas de interpretabilidad, como SHAP, para explicar las predicciones de los modelos más complejos sin sacrificar rendimiento. Esto permitió visibilizar cómo los factores de riesgo tradicionales (edad, obesidad, hipertensión, entre otros) influían directamente en las decisiones del algoritmo, aumentando la confianza en el modelo por parte de profesionales clínicos y favoreciendo su posible adopción en la práctica médica. Esta combinación de rendimiento e interpretabilidad fortalece la utilidad práctica del modelo y marca una diferencia respecto a otros enfoques más opacos o difíciles de justificar.

El estudio también destaca por su enfoque integrador. A diferencia de trabajos centrados exclusivamente en predicción o en análisis clínico, aquí se abarcó desde el desarrollo técnico del modelo hasta su potencial aplicación real. Se incluyó específicamente la clase prediabética, frecuentemente ignorada en otros estudios, lo que permitió una evaluación más completa de la progresión de la enfermedad y añadió valor clínico. Además, el uso de un *dataset* amplio y diverso confiere robustez a los hallazgos, reflejando una situación demográfica real con variedad de perfiles y comorbilidades.

Desde el punto de vista práctico, los modelos desarrollados podrían ser utilizados como herramientas de apoyo al diagnóstico en atención primaria o en programas de salud pública enfocados al cribado poblacional. Su implementación podría hacerse mediante plataformas digitales, aplicaciones móviles o sistemas de gestión clínica, permitiendo estimaciones

personalizadas de riesgo y priorización de pacientes sin necesidad de pruebas complejas o costosas. La simplicidad de las variables requeridas y la reproducibilidad del enfoque favorecen su adopción incluso en entornos con recursos limitados[16].

En definitiva, el proyecto ha cumplido satisfactoriamente con los objetivos planteados. Se logró automatizar la clasificación de la condición diabética con buenos resultados, comparar distintos algoritmos y técnicas, identificar las variables más influyentes y establecer una metodología clara, documentada y reutilizable. Las principales contribuciones de este trabajo no se limitan únicamente a las métricas alcanzadas, sino que se extienden a sus posibles implicaciones clínicas: mejora en el cribado temprano, optimización de recursos sanitarios y mejor comprensión de la interacción entre factores de riesgo. Así, este estudio se enmarca en la tendencia actual de integración entre ciencia de datos y medicina, y demuestra que la detección precoz de diabetes mediante ML no solo es viable, sino también efectiva y valiosa como complemento a las estrategias convencionales.

6.1. Líneas de Trabajo Futuras

Aunque los resultados obtenidos en esta investigación son prometedores, existen múltiples caminos por los que se podría avanzar para enriquecer y profundizar este enfoque. Una de las primeras mejoras consistiría en ampliar y diversificar la base de datos empleada. Incorporar fuentes de información adicionales –como historiales médicos electrónicos, registros clínicos longitudinales o encuestas recientes– permitiría aumentar la cantidad y variedad de variables disponibles, así como comprobar si el modelo se mantiene eficaz en contextos distintos o actualizados. Aplicar el modelo sobre datos de años posteriores podría revelar la necesidad de ajustes, especialmente teniendo en cuenta cambios en los perfiles poblacionales, como el crecimiento de la obesidad o el envejecimiento de la población.

Asimismo, sería deseable incluir indicadores clínicos más precisos, como niveles de glucosa en sangre, hemoglobina glicosilada, lípidos o marcadores de inflamación. Estas variables aportarían una base diagnóstica más robusta y acercarían el modelo al tipo de decisiones que se toman en la práctica médica. Además, incorporar datos genéticos o antecedentes familiares podría enriquecer la predicción, especialmente en personas con predisposición hereditaria, facilitando así un enfoque más preventivo y personalizado.

En términos técnicos, el campo del aprendizaje automático evoluciona con rapidez, y existen nuevas técnicas que podrían explorarse en investigaciones posteriores. El uso de redes neuronales más profundas, arquitecturas avanzadas como *transformers* o modelos especializados en secuencias temporales podría mejorar aún más la capacidad de detección. Igualmente, combinar varios modelos mediante técnicas de ensamblado más sofisticadas permitiría aprovechar las fortalezas específicas de cada uno. Incluso pequeñas mejoras en la sensibilidad del modelo, especialmente en la detección de prediabetes, podrían tener un impacto clínico considerable.

También sería valioso continuar trabajando en herramientas que faciliten la comprensión de las predicciones. Métodos como LIME o visualizaciones adaptadas para profesionales sanitarios

[16] Algo importante que destacar es que el ML tiene un potencial enorme para mejorar la medicina cotidiana. Pero también nos recuerda que no hay que perder de vista lo humano; detrás de cada dato, hay personas reales. Por eso, los modelos siempre deben estar acompañados de comprensión clínica y empatía, nunca usarse de forma aislada.

permitirían traducir las salidas del modelo en explicaciones comprensibles y accionables. Desarrollar interfaces que muestren el perfil de riesgo individual de cada paciente de forma intuitiva facilitaría su adopción en contextos reales de consulta, reduciendo la barrera de confianza entre tecnología y práctica clínica.

Una línea especialmente relevante para el futuro es la validación del modelo en entornos reales mediante estudios prospectivos. Esto implicaría aplicar el modelo sobre una cohorte actual, identificar a las personas clasificadas como de riesgo y seguir su evolución clínica durante un periodo determinado. De esta forma, se podría verificar hasta qué punto las predicciones se corresponden con diagnósticos posteriores, lo que fortalecería la evidencia sobre la utilidad clínica del modelo. Esta validación permitiría además conocer la percepción de los profesionales ante el uso del sistema, ajustando umbrales o formas de presentación según las necesidades detectadas.

Por otra parte, seguir afinando la selección y combinación de variables sería una tarea relevante. Existen técnicas que permiten descubrir interacciones complejas entre factores, como algoritmos genéticos o métodos de búsqueda automatizada. Estas herramientas podrían ayudar a simplificar el modelo sin perder eficacia, mejorando la interpretabilidad y la robustez frente a datos heterogéneos. Asimismo, el manejo del desequilibrio en la distribución de clases sigue siendo un reto pendiente: incorporar estrategias más refinadas de sobre muestreó o aprendizaje con penalización de errores podría contribuir a identificar mejor los casos menos frecuentes.

Finalmente, una expansión lógica del presente trabajo sería extender el uso de modelos predictivos al ámbito de las complicaciones derivadas de la diabetes. Una vez detectada la enfermedad, se podrían construir modelos que anticipen qué pacientes tienen mayor probabilidad de desarrollar secuelas graves, como retinopatía, nefropatía, eventos cardiovasculares o pie diabético. Estos modelos permitirían personalizar aún más el seguimiento, dirigiendo los recursos hacia quienes más lo necesiten, y evitando consecuencias evitables mediante vigilancia y tratamiento temprano. La misma metodología podría aplicarse a otros ámbitos de la salud pública, como la detección de enfermedades renales crónicas, hipertensión no diagnosticada o incluso riesgo oncológico en base a factores de comportamiento y antecedentes.

En definitiva, el trabajo realizado marca un punto de partida firme, pero no un punto final. La mejora continua del modelo, su validación en escenarios reales y su adaptación a otros contextos clínicos representan caminos naturales para seguir explorando. La inteligencia artificial aplicada al ámbito sanitario tiene el potencial de transformar la prevención y la gestión de enfermedades crónicas. En ese camino, este proyecto contribuye con una base sólida y replicable sobre la que construir nuevas soluciones que pongan los datos al servicio de una medicina más temprana, más inteligente y humana.

VII. BIOGRAFÍA

[1] Organización Mundial de la Salud, "Diabetes," *World Health Organization*, 14 de noviembre de 2024. Disponible en: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Consulta: 21 de mayo de 2025].

[2] Y. Jiang, J. Xie, X. Zhang y Y. Wang, "A machine learning-based approach to predict diabetes using clinical data," *Scientific Reports*, vol. 12, art. 13254, 2022. Disponible en: https://doi.org/10.1038/s41598-022-17080-w. [Consulta: 4 de junio de 2025].

[3] M. R. Gómez-Peralta et al., "Diagnóstico y clasificación de la diabetes mellitus," *Revista Clínica Española*, vol. 220, no. 8, pp. 469–477, 2020. Disponible en: https://doi.org/10.1016/j.rce.2020.02.011. [Consulta: 4 de junio de 2025].

[4] R. V. Sánchez-Morillo, A. Fernández-Granero y J. A. León-Jiménez, "Use of machine learning techniques to predict mortality and length of stay in intensive care units," *Medicina Intensiva*, vol. 45, no. 2, pp. 74–82, 2021.

[5] D. Esteva et al., "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, 2019. Disponible en: https://doi.org/10.1038/s41591-018-0316-z. [Consulta: 4 de junio de 2025].

[6] V. Kumar, "Python libraries for machine learning in healthcare: A survey," *Journal of Biomedical Informatics*, vol. 134, art. 104220, 2022. Disponible en: https://doi.org/10.1016/j.jbi.2022.104220. [Consulta: 4 de junio de 2025].

[7] D. Esteva et al., "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, 2019. Disponible en: https://doi.org/10.1038/s41591-018-0316-z. [Consulta: 4 de junio de 2025].

[8] A. Rajkomar, J. Dean y I. Kohane, "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347-1358, 2019. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6455466/. [Consulta: 4 de junio de 2025].

[9] M. C. Pino-Mejías, J. L. Mora-Jiménez y J. S. Gutiérrez, "Comparative study of machine learning algorithms for the early detection of diabetes," *Expert Systems with Applications*, vol. 164, art. 113738, 2021. Disponible en: https://doi.org/10.1016/j.eswa.2020.113738. [Consulta: 4 de junio de 2025].

[10] L. Lundberg y S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[11] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2.ª ed., Independently published, 2022. Disponible en: https://christophm.github.io/interpretable-ml-book/. [Consulta: 28 de mayo 2025].

[12] M. R. Gómez-Peralta et al., "Diagnóstico y clasificación de la diabetes mellitus," *Revista Clínica Española*, vol. 220, no. 8, pp. 469–477, 2020. Disponible en: https://doi.org/10.1016/j.rce.2020.02.011. [Consulta: 28 de mayo de 2025].

[13] S. M. Lundberg, G. Erion y S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:3145–3153, 2017.

[14] L. Pineau et al., "Improving reproducibility in machine learning research," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 1–20, 2021. Disponible en: https://jmlr.org/papers/volume22/20-1098/20-1098.pdf. [Consulta: 4 de junio de 2025].

[15] V. Kumar, "Python libraries for machine learning in healthcare: A survey," *Journal of Biomedical Informatics*, vol. 134, art. 104220, 2022. Disponible en: https://doi.org/10.1016/j.jbi.2022.104220. [Consulta: 28 de mayo de 2025].

[16] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed., 2021. Disponible en: https://diabetesatlas.org. [Consulta: 4 de junio de 2025].

[17] J. Einarson et al., "Economic burden of cardiovascular disease in type 2 diabetes: A systematic review," *Value in Health*, vol. 21, no. 7, pp. 881–890, 2018. Disponible en: https://doi.org/10.1016/j.jval.2018.03.003. [Consulta: 1 de junio de 2025].

[18] Organización Mundial de la Salud, "Diabetes," *World Health Organization*, 14 de noviembre de 2024. Disponible en: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Consulta: 4 de junio de 2025].

[19] J. Zhang et al., "Global burden of diabetes-related complications and mortality: A systematic review," *The Lancet Diabetes & Endocrinology*, vol. 10, no. 1, pp. 45–56, 2022. Disponible en: https://doi.org/10.1016/S2213-8587(21)00325-9. [Consulta: 1 de junio de 2025].

[20] American Diabetes Association, "Prevention or Delay of Type 2 Diabetes: Standards of Medical Care in Diabetes—2023," *Diabetes Care*, vol. 46, supl. 1, pp. S39–S45, 2023. Disponible en: https://doi.org/10.2337/dc23-Sint. [Consulta: 1 de junio de 2025].

[21] Y. Jiang, J. Xie, X. Zhang y Y. Wang, "A machine learning-based approach to predict diabetes using clinical data," *Scientific Reports*, vol. 12, art. 13254, 2022. Disponible en: https://doi.org/10.1038/s41598-022-17080-w. [Consulta: 4 de junio de 2025].

[22] S. B. Vinitha y M. Kavitha, "Survey on Predictive Analytics in Healthcare using Machine Learning Techniques," *Materials Today: Proceedings*, vol. 72, pp. 2301–2306, 2023. Disponible en: https://doi.org/10.1016/j.matpr.2023.01.020. [Consulta: 4 de junio de 2025].

[23] L. M. Del Prato et al., "Early diagnosis of type 2 diabetes: Practical and clinical considerations," *The Lancet Diabetes & Endocrinology*, vol. 9, no. 8, pp. 473–483, 2021. Disponible en: https://doi.org/10.1016/S2213-8587(21)00045-9. [Consulta: 4 de junio de 2025].

[24] R. Cohen et al., "Limitations of HbA1c in the diagnosis of diabetes mellitus: Perspectives and alternatives," *Journal of Clinical Endocrinology & Metabolism*, vol. 105, no. 5, pp. 1154–1162, 2020. Disponible en: https://doi.org/10.1210/clinem/dgaa067. [Consulta: 4 de junio de 2025].

[25] V. Kumar, "Python libraries for machine learning in healthcare: A survey," *Journal of Biomedical Informatics*, vol. 134, art. 104220, 2022. Disponible en: https://doi.org/10.1016/j.jbi.2022.104220. [Consulta: 4 de junio de 2025].

[26] D. Esteva et al., "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, 2019. Disponible en: https://doi.org/10.1038/s41591-018-0316-z. [Consulta: 4 de junio de 2025].

[27] S. Lundberg y S. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017. Disponible en: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. [Consulta: 4 de junio de 2025].

[28] J. Brownlee, *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End*, Machine Learning Mastery, 2020.

[29] UCI Machine Learning Repository, "Pima Indians Diabetes Database," University of California, Irvine, Disponible en: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes. [Consulta: 4 de junio de 2025].

[30] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. Scott, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 1988, pp. 261–265.

[31] D. Sisodia and S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018. Disponible en: https://doi.org/10.1016/j.procs.2018.05.224. [Consulta: 4 de junio de 2025].

[32] M. E. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent diagnostic prediction and classification system for diabetes using machine learning techniques," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 555–567, 2021. Disponible en: https://doi.org/10.1007/s12652-020-02278-2. [Consulta: 4 de junio de 2025].

[33] J. Iparraguirre-Villanueva, A. Redondo-Cerezo y F. Ruiz-Moreno, "Evaluación comparativa de clasificadores aplicados al diagnóstico de diabetes," *Revista de Informática Médica*, vol. 29, no. 1, pp. 23–32, 2023.

[34] R. Hasan, M. Sarker y M. A. Rahman, "Ensemble classifiers and boosting techniques for early detection of diabetes," *Procedia Computer Science*, vol. 167, pp. 10–18, 2020. Disponible en: https://doi.org/10.1016/j.procs.2020.03.186. [Consulta: 4 de junio de 2025].

[35] J. Iparraguirre-Villanueva, A. M. López-Rodríguez y F. J. Martín-Masot, “Comparación de algoritmos de aprendizaje automático para la predicción de diabetes tipo 2,” *Revista Española de Informática en Salud*, vol. 19, no. 1, pp. 45–52, 2023.

[36] K. Hasan, F. Ali y S. M. Rahman, “An ensemble approach for diabetes prediction using machine learning algorithms,” *Informatics in Medicine Unlocked*, vol. 20, art. 100389, 2020. Disponible en: https://doi.org/10.1016/j.imu.2020.100389 [Consulta: 4 de junio de 2025].

[37] L. Kopitar et al., “Early detection of type 2 diabetes mellitus using machine learning-based prediction models,” *Scientific Reports*, vol. 10, art. 11981, 2020. Disponible en: https://doi.org/10.1038/s41598-020-68771-z [Consulta: 4 de junio de 2025].

[38] Y. Jian, M. Pasquier, A. Sagahyroon y F. Aloul, “A Machine Learning Approach to Predicting Diabetes Complications,” *Healthcare*, vol. 9, no. 12, art. 1712, 2021. Disponible en: https://www.mdpi.com/2227-9032/9/12/1712 [Consulta: 4 de junio de 2025]

[39] M. Esteva et al., “A guide to deep learning in healthcare,” *Nature Medicine*, vol. 25, pp. 24–29, 2019. Disponible en: https://doi.org/10.1038/s41591-018-0316-z [Consulta: 4 de junio de 2025].

[40] V. Kumar, “Python libraries for machine learning in healthcare: A survey,” *Journal of Biomedical Informatics*, vol. 134, art. 104220, 2022. Disponible en: https://doi.org/10.1016/j.jbi.2022.104220 [Consulta: 4 de junio de 2025].

[41] S. Lundberg y S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017. Disponible en: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [Consulta: 4 de junio de 2025].

[42] W. McKinney, "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference, 2010. Disponible en: https://doi.org/10.25080/Majora-92bf1922-00a. [Consulta: 7 de junio de 2025].

[43] T. E. Oliphant, "A guide to NumPy," USA: Trelgol Publishing, 2006.

[44] J. D. Hunter, "Matplotlib: A 2D graphics environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007. Disponible en: https://doi.org/10.1109/MCSE.2007.55. [Consulta: 7 de junio de 2025].

[45] M. Waskom, "Seaborn: statistical data visualization," Journal of Open Source Software, vol. 6, no. 60, p. 3021, 2021. Disponible en: https://doi.org/10.21105/joss.03021. [Consulta: 7 de junio de 2025].

[46] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[47] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," Journal of Machine Learning Research, vol. 18, no. 17, pp. 1–5, 2017.

[48] B. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[49] R. K. Singh, "AutoViz: Automatic Visualization and Exploration of Data in Python," GitHub, 2021. Disponible en: https://github.com/AutoViML/AutoViz. [Consulta: 7 de junio de 2025].

[50] M. Ali, "PyCaret: An open source, low-code machine learning library in Python," 2020. Disponible en: https://www.pycaret.org. [Consulta: 7 de junio de 2025].

[51] Centers for Disease Control and Prevention, "Behavioral Risk Factor Surveillance System: 2015 Summary Data Quality Report," 2016. Disponible en: https://www.cdc.gov/brfss/annual_data/2015/pdf/compare_2015.pdf. [Consulta: 7 de junio de 2025].

[52] M. Minn, "BRFSS Data in R," michaelminn.net. Disponible en: https://michaelminn.net/tutorials/r-brfss. [Consulta: 7 de junio de 2025].

[53] G. Lemaître, F. Nogueira y C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," Journal of Machine Learning Research, vol. 18, no. 17, pp. 1–5, 2017.

[54] V. Kumar, "Python libraries for machine learning in healthcare: A survey," Journal of Biomedical Informatics, vol. 134, art. 104220, 2022. Disponible en: https://doi.org/10.1016/j.jbi.2022.104220. [Consulta: 4 de junio de 2025].

[55] E. Agardh, P. Allebeck, J. Hallqvist, T. Moradi, and A. Sidorchuk, “Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis,” *Int. J. Epidemiol.*, vol. 40, no. 3, pp. 804–818, 2011, doi: 10.1093/ije/dyr029. [Consulta: 9 de junio de 2025].

[56] Sociedad Española de Cardiología (SEC). *Diabetes y enfermedad cardiovascular*. Revista Española de Cardiología, suplemento especial, vol. 14, 2017.

[57] Ministerio de Sanidad de España. *Encuesta Europea de Salud en España 2020*. Madrid: Instituto Nacional de Estadística (INE), 2021.

[58] M. M. Breunig, H.-P. Kriegel, R. T. Ng, y J. Sander, “LOF: Identifying Density-Based Local Outliers,” *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 93–104. DOI: 10.1145/335191.335388 [Consulta: 11 de junio de 2025].

[59] R. Kohavi, “Cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. Int. Joint Conf. Artif. Intell.*, Montréal, Canada, Aug. 1995, pp. 1137-1143.

[60] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

[61] Y. Jian, M. Pasquier, A. Sagahyroon, y F. A. Aloul, "*Using Machine Learning to Predict Diabetes Complications*," en *Proceedings of the 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART 2021) *, París, Francia, 8–10 dic. 2021, pp. XX–XX, IEEE, doi: 10.1109/BioSMART54244. [Consulta: 12 de junio de 2025].

[62] Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, 138, 271-281.

[63] 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.

[64] Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, 863-905.

[65] Disglucemia ¿Qué es?", *Lamberts española*, 3 de enero de 2023. [En línea]. Disponible en: https://lamberts.es/art-dsp/disglucemia-que-es/. [Consulta: 12 de junio de 2025].

En la elaboración del presente Trabajo Fin de Máster se ha hecho uso responsable de herramientas de inteligencia artificial generativa como apoyo en determinadas fases del proceso de redacción.

En concreto, se ha empleado la plataforma de inteligencia artificial en concreto para:

- Sugerencias lingüísticas y apoyo en la redacción preliminar de ciertos apartados.
- Revisión del estilo académico y adecuación del formato de las referencias bibliográficas al estilo IEEE.