MODELO DE *MACHINE LEARNING* PARA LA PREDICCIÓN DE DIABETES MELLITUS TIPO 2, BASADO EN DATOS DE EXÁMENES DE LABORATORIO RUTINARIOS

Natalia Alejandra Macana Díaz (Código 9026993)
Iván René Valderrama Corredor (Código 9027505)

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Henry Fabian Tobar Tosse

Codirector
José Palencia Téllez

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO DE 2026

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

La Diabetes mellitus tipo 2 (DM2) es una enfermedad crónica que representa una de las principales causas de morbilidad y mortalidad a nivel mundial. En Colombia, la DM2 representa una de las enfermedades con mayores casos reportados, afectando la calidad de vida de los pacientes y los sobre costos en la atención médica. Ante este hecho, es indispensable fortalecer las estrategias de promoción y prevención mediante la detección temprana y oportuna del riesgo de padecer DM2. Por lo anterior, los exámenes de laboratorio rutinarios de bajo costo constituyen una fuente confiable de datos que al ser correctamente utilizados y analizados podrían revelar patrones importantes que permitan antelar el desarrollo de la enfermedad.

En la actualidad, los modelos de *machine learning* han demostrado tener un alto potencial de usabilidad en datos clínicos debido a su capacidad de estructurar relaciones complejas entre los datos. El presente proyecto desarrolla una herramienta basada en técnicas de *machine learning* utilizando datos de exámenes de laboratorio rutinarios que permita estimar el riesgo de desarrollar DM2. Este prototipo es capaz de presentar predicciones de riesgo de manera clara y con explicaciones comprensibles para el personal médico.

1. DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

En la actualidad, la diabetes mellitus tipo 2 (DM2) representa un problema de salud pública a nivel mundial y en Colombia este tipo de diabetes representa la mayoría de los casos actuales en el país. La DM2 comprende un grupo de enfermedades metabólicas que se caracteriza por una serie de síntomas como consecuencia de la deficiencia en la secreción o acción de la insulina [1]. Por otro lado, se entiende que la etiología de la DM2 es multifactorial y compleja, englobando tanto factores de riesgo genéticos como ambientales [2].

En el contexto del sistema de salud, el incremento de pacientes con diabetes mellitus tipo 2 (DM2) representa una carga significativa en términos asistenciales, de morbilidad y de deterioro de la calidad de vida de los pacientes. Frente a ello, los exámenes de laboratorio rutinarios constituyen una fuente confiable de datos que, al integrarlos en procesos analíticos avanzados, pueden ofrecer un valor predictivo más allá del diagnóstico convencional. Aprovechar esta información mediante técnicas de ciencia de datos permitiría anticipar el riesgo de DM2 en etapas preclínicas, facilitando intervenciones oportunas y personalizadas por parte del equipo médico.

En consecuencia, se planteó la necesidad de desarrollar modelos predictivos basados en algoritmos de aprendizaje automático (*machine learning*), que utilicen datos provenientes de exámenes de laboratorio rutinarios para proporcionar herramientas de apoyo diagnóstico. Dichos modelos permitieron realizar estimaciones tempranas del riesgo de desarrollar DM2, optimizando la toma de decisiones clínicas y promoviendo estrategias de prevención más eficaces y sostenibles en la práctica médica.

1.2 FORMULACIÓN DEL PROBLEMA

¿Cómo podemos desarrollar un modelo de *machine learning*, utilizando datos de exámenes de laboratorio rutinarios, que estime el riesgo asociado a Diabetes Mellitus Tipo 2 y presente esta estimación de forma clara y comprensible para el personal médico como apoyo al tamizaje clínico?

1.3 SISTEMATIZACIÓN

- ¿Cuáles son las variables clínicas más significativas extraídas de exámenes de laboratorio rutinarios, que permiten estimar con mayor precisión el riesgo de desarrollar diabetes mellitus tipo 2?
- ¿Qué modelos de aprendizaje automático presentan un mejor desempeño en la predicción del riesgo de diabetes mellitus tipo 2 a partir de datos clínicos estructurados?

- ¿Cuáles son las métricas más adecuadas para evaluar la validez y la capacidad predictiva de un modelo de *machine learning* en contextos clínicos?
- ¿Qué estrategias de visualización de resultados pueden facilitar la interpretación clínica de las predicciones por parte del personal médico?

2. OBJETIVOS

2.1 OBJETIVO GENERAL

Desarrollar un modelo de *machine learning* basado en datos de exámenes de laboratorio de bajo costo, que estime el riesgo asociado a Diabetes Mellitus Tipo 2 y apoye el tamizaje clínico oportuno por parte del personal médico.

2.2 OBJETIVOS ESPECÍFICOS

- Identificar y seleccionar las variables clínicas más relevantes, provenientes de exámenes de laboratorio rutinarios, que permitan construir una base de datos adecuada para la predicción del riesgo de desarrollar diabetes mellitus tipo 2.
- Implementar modelos de aprendizaje automático capaces de estimar la probabilidad de pertenencia a la clase de riesgo asociada a DM2
- Evaluar el desempeño del modelo de *machine learning* utilizando métricas estadísticas pertinentes y validarlo con el apoyo de profesionales del área médica para garantizar su precisión, confiabilidad e interpretabilidad clínica.
- Presentar una herramienta que permita representar gráficamente los resultados del modelo predictivo.

3. MARCO TEORICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

A continuación, se exponen los fundamentos conceptuales y científicos que respaldan el desarrollo del proyecto. Se abordan los principales aspectos relacionados con la Diabetes Mellitus tipo 2; sus factores de riesgo, criterios diagnósticos y complicaciones, así como el uso de técnicas de *machine learning* aplicadas al análisis de datos clínicos, incluyendo las métricas de evaluación de modelos y los aspectos normativos y éticos relevantes para su implementación en contextos de salud.

3.1.1 DIABETES MELLITUS TIPO 2

La Diabetes mellitus tipo 2 (DM2) es una enfermedad metabólica crónica de origen complejo y multifactorial que se desarrolla a través de dos mecanismos fisiopatológicos principales:

- El deterioro de la función celular de los islotes pancreáticos provocando una disminución de la síntesis de la insulina
- La resistencia de los tejidos periféricos a la insulina dando como resultado una disminución de la respuesta metabólica a la insulina.

La interacción correcta de ambos mecanismos es importante para el mantenimiento de la homeostasis de la glucosa cuando uno o ambos fallan de forma sostenida, se establece el estado de hiperglucemia crónica que define clínicamente la enfermedad [3].

En términos fisiopatológicos, la DM2 es precedida por un periodo asintomático prolongado, que puede abarcar de 5 a 12 años, conocido como estado prediabético. Este estado se caracteriza por una hiperglucemia lo suficientemente elevada para aumentar la incidencia de microangiopatías, e incluye condiciones como la glucosa plasmática en ayunas (GPA) y la intolerancia a la glucosa (IGT), las cuales ya presentan niveles elevados de insulina circulante como intento compensatorio. Si no se interviene, entre el 30% y el 70% de las personas con prediabetes desarrollarán DM2 en los siguientes 5-10 años [4].

La clasificación diagnóstica de la DM2 se basa en los siguientes criterios definidos por la *American Diabetes Association* (por sus siglas en inglés, ADA) [5]:

- Glucosa plasmática en ayunas ≥ 126 mg/dL
- Glucosa plasmática a las 2 horas durante la prueba de tolerancia la glucosa oral ≥ 200 mg/dL
- HbA1c ≥ 6.5%

- Glucosa plasmática casual ≥ 200 mg/dL en presencia de síntomas.

3.1.1.1 FACTORES DE RIESGO BIOQUÍMICOS ASOCIADOS A LA DIABETES MELLITUS TIPO 2

La glucosa plasmática en ayunas (FPG, por sus siglas en inglés) y la hemoglobina glicosilada (HbA1c) constituyen los marcadores diagnósticos por excelencia de la DM2, según los criterios establecidos por la American Diabetes Association (ADA) [5].

- **GLUCOSA PLASMÁTICA EN AYUNAS (GPA)**

La glucosa en ayunas refleja el estado glicémico inmediato tras un ayuno de al menos 8 horas. Su elevación indica disfunción del eje insulina-glucagón en el ayuno, donde el hígado libera glucosa de forma incontrolada por resistencia a la acción supresora de la insulina [6]. Los puntos de corte diagnósticos establecidos por ADA son:

  - Normal: FPG < 100 mg/dL
  - Prediabetes (glucosa alterada en ayunas): FPG entre 100-125 mg/dL
  - Diabetes: FPG ≥ 126 mg/dL en dos mediciones separadas

- **HEMOGLOBINA GLICOSILADA (HbA1c)**

La HbA1c representa el promedio ponderado de la glucemia durante los últimos 2-3 meses, gracias a que la glucosa se une de forma irreversible y no enzimática a la hemoglobina de los eritrocitos circulantes [7]. A diferencia de la glucosa puntual, la HbA1c no se afecta por el ayuno reciente o el estrés agudo, lo que la hace especialmente útil para el diagnóstico y seguimiento de la DM2. Los puntos de corte son:

  - Normal: HbA1c < 5.7%
  - Prediabetes: HbA1c entre 5.7% y 6.4%
  - Diabetes: HbA1c ≥ 6.5%

La asociación entre dislipidemia y diabetes tipo 2 está ampliamente documentada y constituye uno de los pilares fisiopatológicos del síndrome metabólico [8]. Los pacientes con DM2 presentan típicamente un patrón lipídico característico denominado "dislipidemia diabética", que incluye Triglicéridos elevados, HDL disminuido y partículas LDL pequeñas y densas, todas consecuencias directas o indirectas de la resistencia a la insulina.

- **TRIGLICÉRIDOS (TG)**

Los Triglicéridos son los principales lípidos circulantes con función de almacenamiento energético. En la DM2, la resistencia hepática a la insulina aumenta la producción de lipoproteínas VLDL ricas en Triglicéridos, y simultáneamente disminuye la actividad de la lipoproteína lipasa,

reduciendo su aclaramiento [9]. Como resultado, la hipertrigliceridemia es uno de los hallazgos lipídicos más tempranos y consistentes en pacientes con resistencia a la insulina, incluso antes del diagnóstico de DM2.

Los puntos de corte clínicos según ATP III [10] son:

- Normal: TG < 150 mg/dL
- Limítrofe alto: TG entre 150-199 mg/dL
- Alto: TG entre 200-499 mg/dL
- Muy alto: TG ≥ 500 mg/dL

- **COLESTEROL HDL**

El Colesterol HDL (lipoproteínas de alta densidad) cumple una función protectora cardiovascular mediante el transporte reverso de colesterol desde los tejidos periféricos al hígado. En estados de resistencia a la insulina, los niveles de HDL disminuyen debido al aumento del intercambio de ésteres de colesterol con VLDL, aumento de la actividad de la lipasa hepática y disminución de la apolipoproteína A-I [11]. La disminución del HDL es uno de los criterios definitorios del síndrome metabólico según ATP III:

- HDL bajo en hombres: < 40 mg/dL
- HDL bajo en mujeres: < 50 mg/dL

- **COLESTEROL LDL**

El Colesterol LDL (lipoproteínas de baja densidad), conocido como "colesterol malo", aunque no muestra elevaciones tan marcadas como TG y HDL en DM2, presenta cambios cualitativos importantes: predominio de partículas pequeñas y densas que son más aterogénicas por mayor susceptibilidad a la oxidación y mayor afinidad por la pared arterial [12]. Los puntos de corte son:

- Óptimo: LDL < 100 mg/dL
- Casi óptimo: LDL entre 100-129 mg/dL
- Limítrofe alto: LDL entre 130-159 mg/dL
- Alto: LDL entre 160-189 mg/dL
- Muy alto: LDL ≥ 190 mg/dL

- **COLESTEROL TOTAL**

El Colesterol Total (CT) es la suma de todas las fracciones de colesterol circulante (HDL, LDL, VLDL e IDL). Aunque por sí solo es menos específico que sus componentes, su interpretación

combinada con HDL mediante el Índice Aterogénico ofrece información valiosa sobre el riesgo cardiovascular global [13]. Los puntos de corte ATP III son:

- Deseable: CT < 200 mg/dL
- Limítrofe alto: CT entre 200-239 mg/dL
- Alto: CT ≥ 240 mg/dL

Más allá de los valores absolutos de cada lípido, las relaciones (ratios) entre ellos han demostrado ser predictores más robustos que las medidas individuales para evaluar el riesgo metabólico y cardiovascular [14]. Estos ratios capturan información sobre el balance entre fracciones aterogénicas y protectoras del perfil lipídico, siendo particularmente útiles en modelos predictivos.

- **ÍNDICE ATEROGÉNICO (COLESTEROL TOTAL / HDL)**

El Índice Aterogénico (IA), también conocido como índice de Castelli I, fue propuesto por Castelli y colaboradores [13] tras el estudio Framingham. Calculado como CT/HDL, el IA proporciona una medida integradora del riesgo cardiovascular global, capturando simultáneamente el exceso de fracciones aterogénicas y el déficit de protección por HDL. Sus puntos de corte clínicos son:

- Riesgo bajo: IA < 3.5
- Riesgo medio: IA entre 3.5-4.5
- Riesgo alto: IA > 4.5

En el contexto de DM2, valores elevados del IA son frecuentes y reflejan la dislipidemia diabética característica.

- RATIO TRIGLICÉRIDOS / HDL

El ratio TG/HDL ha sido propuesto como un marcador específico de resistencia a la insulina, particularmente útil en poblaciones donde no se cuenta con técnicas más sofisticadas como el *clamp* euglicémico o el HOMA-IR [15]. Valores superiores a 3.5 sugieren resistencia a la insulina y correlacionan con la presencia de partículas LDL pequeñas y densas. En el contexto del tamizaje de DM2, este ratio ofrece información complementaria sobre el estado metabólico del paciente.

- **RATIO LDL / HDL**

El ratio LDL/HDL, también conocido como índice de Castelli II, complementa al IA mostrando específicamente el balance entre las dos fracciones más relevantes para el riesgo cardiovascular. Valores superiores a 3.0 se asocian con riesgo cardiovascular elevado [14].

La nefropatía diabética constituye una de las complicaciones microvasculares más frecuentes y severas de la DM2, afectando aproximadamente al 30-40% de los pacientes diabéticos a lo largo de su evolución [16]. La estrecha asociación entre función renal y diabetes hace de los marcadores renales predictores valiosos en modelos de riesgo metabólico.

- **CREATININA SÉRICA**

La Creatinina es un producto de degradación del fosfato de creatina presente en el músculo esquelético, eliminada principalmente por filtración glomerular. Sus niveles séricos son inversamente proporcionales a la tasa de filtración glomerular (TFG), constituyendo el biomarcador más utilizado para evaluar la función renal en la práctica clínica rutinaria [17].

Los puntos de corte de referencia son:

- Normal en hombres: 0.7-1.3 mg/dL
- Normal en mujeres: 0.6-1.1 mg/dL
- Elevación leve: 1.4-2.0 mg/dL
- Elevación moderada-severa: > 2.0 mg/dL

La relación entre Creatinina y DM2 es bidireccional:

- Creatinina elevada en pacientes diabéticos refleja deterioro renal secundario a la nefropatía diabética, complicación crónica establecida.
- Creatinina elevada en pacientes sin diabetes diagnosticada puede preceder al diagnóstico, ya que la disfunción renal subclínica se asocia con resistencia a la insulina y predispone al desarrollo de DM2[16].

Desde el punto de vista epidemiológico, la DM2 constituye una de las mayores amenazas sanitarias del siglo XXI. Según la *International Diabetes Federation* (por sus siglas en inglés, IDF), para el año 2024, 589 millones de adultos entre 20 y 79 años viven con diabetes en el mundo, el 81% de personas con diabetes viven en países de ingresos bajos y medios, siendo América latina una de las regiones más afectadas. Además, se prevé que el número total de personal con diabetes aumente hasta 853 millones en 2050, indicando que 1 de cada 8 adultos vivirá con diabetes, lo que supone un aumento del 46% [18].

Según cifras previas del Sistema Integrado de Información de la Protección Social (SISPRO) y la Organización Panamericana de la Salud (OPS), para el año 2022, la diabetes fue la octava causa de mortalidad general en Colombia, presentándose aproximadamente 13,3 muertes por cada 100.000 habitantes con Edades entre 30 a 70 años [19].

- Los factores de riesgo más relevantes incluyen:

- Edad avanzada (≥45 años)
- Obesidad (aumento de la circunferencia abdominal)
- Sedentarismo
- Dislipidemias
- Antecedentes familiares de primer grado con DM2
- Síndrome de ovario poliquístico en mujeres [20].

Además de su impacto individual en la calidad de vida, la DM2 representa según cifras reportadas por la IDF un aumento en el gasto global en salud creciendo de 232 mil millones de dólares en 2007 a más de un billón de dólares (1.015 billones) en 2024 para adultos entre 20 y 79 años, lo que equivale a un incremento acumulado del 338% en 17 años [5].

Asimismo, las personas con DM2 tienen un mayor riesgo de padecer enfermedades cardiovasculares, incluyendo un 72% más de riesgo de infarto, un 52% más de riesgo de accidente cerebrovascular y un 84% más de riesgo de insuficiencia cardíaca. Además, las personas con diabetes tienen un 56% más de riesgo de desarrollar demencia en comparación con aquellas que no tienen diabetes. Casi uno de cada cuatro adultos con diabetes presenta algún tipo de retinopatía diabética. Estas complicaciones generan altos niveles de hospitalización, discapacidad y mortalidad prematura, y son la causa de más del 11% de las muertes globales por todas las causas en adultos [5].

Este panorama epidemiológico, clínico y económico resalta la necesidad de estrategias predictivas y preventivas apoyadas en ciencia de datos, que permitan anticipar el riesgo de DM2 en etapas tempranas utilizando datos clínicos disponibles, como los exámenes de laboratorio rutinarios.

3.1.1.2 IMPORTANCIA DEL DIAGNÓSTICO TEMPRANO

El diagnóstico precoz de la Diabetes mellitus tipo 2 (DM2) constituye una prioridad tanto clínica como en salud pública, debido a su capacidad para reducir significativamente el riesgo de complicaciones crónicas y optimizar el uso de los recursos sanitarios. Detectar la enfermedad en etapas tempranas, o incluso en su fase prediabética, permite implementar intervenciones costo-efectivas que retrasan o previenen su progresión. El programa de prevención de diabetes (por sus siglas en inglés, DPP), demostraron que los cambios en el estilo de vida como; dieta balanceada, incremento de la actividad física y reducción del peso corporal, podrían disminuir considerablemente el riesgo de conversión de prediabetes a DM2 de 42 al 58% [21].

Sin embargo, en la práctica clínica, el diagnóstico de DM2 suele realizarse en fases avanzadas, ya que la enfermedad puede permanecer asintomática durante largos periodos, inclusive años. Como consecuencia, muchos pacientes debutan con complicaciones microvasculares

(retinopatía, nefropatía, neuropatía) o macrovasculares (enfermedad cardiovascular o accidente cerebrovascular) [18]. Esta situación representa una brecha crítica en los sistemas de salud, particularmente en regiones de bajos y medianos ingresos donde el tamizaje oportuno es escaso.

En este contexto, se hace necesario adoptar una estrategia de detección proactiva basada en datos clínicos, como los resultados de exámenes rutinarios (glucosa, colesterol, Triglicéridos, hemoglobina glicosilada, Creatinina, entre otros), que suelen estar disponibles el historial de sistemas de laboratorio. El análisis automatizado de estos datos mediante técnicas de ciencia de datos y aprendizaje automático podrían identificar patrones asociados con el riesgo de DM2 antes de que se alcancen los umbrales diagnósticos convencionales (por ejemplo, glucosa en ayunas ≥126 mg/dL o HbA1c ≥6.5%) [5].

3.1.2 MARCO TEÓRICO *MACHINE LEARNING*

El presente estudio se fundamenta en técnicas de aprendizaje automático supervisado, una rama de la inteligencia artificial cuya ejecución en distintos sectores ha mostrado un crecimiento importante en la última década [22]. A continuación, se presentan los fundamentos teóricos necesarios para comprender los algoritmos, técnicas de validación, métricas de evaluación utilizados en el desarrollo del modelo predictivo de este trabajo.

3.1.2.1 INTRODUCCIÓN AL *MACHINE LEARNING* (ML)

Según Arthur Samuel [23], el *machine learning* es el campo de estudio que da a los ordenadores la capacidad de aprender sin ser programados de manera explícita. Oracle define el aprendizaje automático (ML) como una subcategoría de la inteligencia artificial enfocada a construir sistemas que aprenden y mejoran a medida que consumen más información, esto con el fin de desarrollar sistemas o máquinas que imiten la inteligencia humana. El aprendizaje automático es una técnica que descubre patrones y tendencias desconocidas en grandes conjuntos de datos al explorarlos y que van más allá de un análisis estadístico. El aprendizaje automático usa algoritmos altamente sofisticados y entrenados para identificar relaciones relevantes entre los datos, creando modelos. Estos modelos se usan para realizar predicciones, recomendaciones y clasificar datos.

En el área médica, se ha demostrado la capacidad para detectar patrones complejos en grandes volúmenes clínicos que escapan el análisis estadístico tradicional, permitiendo aplicaciones como diagnóstico asistido por imagen, estratificar riesgo, tamizaje de enfermedades crónicas [24].

Entre las diferentes técnicas de ML se clasifican en 4 categorías:

3.1.2.1.1 APRENDIZAJE SUPERVISADO

Los modelos de aprendizaje automático supervisado se entrenan con conjuntos de datos etiquetados, lo que permite aprender y aumentar su precisión con el tiempo. Por ejemplo, un algoritmo se entrena con imágenes de algún objeto, todos etiquetados previamente por humanos [23]. Este tipo de aprendizaje supervisado requiere que un experto humano proporcione las respuestas correctas etiquetando los datos para que el algoritmo pueda aprender y hacer predicciones precisas en el futuro.

- **APRENDIZAJE SUPERVISADO PARA CLASIFICACIÓN BINARIA**

La clasificación binaria es una tarea de aprendizaje supervisado donde la variable objetivo toma solo dos valores posibles (típicamente 0 y 1). Formalmente, dado un conjunto de entrenamiento

$$D = (x^1, y^1), (x^2, y^2), \dots, (x_n, y_n) \quad (1)$$

donde $x_i \in \mathbb{R}^d$ es un vector de características y $y_i \in \{0, 1\}$ es la etiqueta, el objetivo es aprender una función $f: \mathbb{R}^d \rightarrow \{0, 1\}$ que minimice el error de predicción en datos no observados [26].
En aplicaciones clínicas, los modelos de clasificación binaria suelen producir una probabilidad continua $P(y=1|x) \in [0,1]$ que representa la confianza del modelo en que la observación pertenezca a la clase positiva. Esta probabilidad se convierte en una etiqueta binaria mediante la aplicación de un umbral de decisión (típicamente 0.5), aunque el umbral óptimo puede ajustarse según el balance deseado entre sensibilidad y especificidad.

- **ALGORITMOS SELECCIONADOS PARA CLASIFICACIÓN CLÍNICA**

La elección de algoritmos se fundamentó en tres criterios:

a. Capacidad para manejar datos tabulares heterogéneos.
b. Interpretabilidad de los resultados en contexto clínico.
c. Evidencia previa de buen desempeño en problemas de tamizaje [27].

- **REGRESIÓN LOGÍSTICA**

La regresión logística es un modelo lineal probabilístico ampliamente utilizado en epidemiología y medicina basada en evidencia [28]. Modela la probabilidad de pertenencia a la clase positiva mediante la función sigmoidea:

$$P(y = 1|x) = 1/(1 + e\text{^}(-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p))) \quad (2)$$

donde β_0 es el intercepto y β_i son los coeficientes asociados a cada variable predictora. La ventaja principal de la regresión logística radica en la interpretabilidad de sus coeficientes en términos de odds ratios, lo que permite cuantificar el riesgo relativo asociado a cada predictor. En el presente proyecto, la regresión logística se implementó como modelo de referencia (*baseline* interpretable) que permite comparar el desempeño de algoritmos más complejos [28].

- ***RANDOM FOREST***

Random Forest es un algoritmo de ensamble basado en árboles de decisión, propuesto por Breiman. Su funcionamiento se basa en el principio de *bagging* (*Bootstrap Aggregating*): se entrenan múltiples árboles de decisión sobre muestras *bootstrap* del conjunto de entrenamiento, y la predicción final se obtiene mediante votación mayoritaria (clasificación) o promedio (regresión).

Cada árbol del bosque aleatorio se construye considerando solo un subconjunto aleatorio de las variables en cada división de nodo, lo que reduce la correlación entre árboles y mejora la generalización del modelo. Las ventajas principales de Random Forest en el contexto biomédico son: capacidad de capturar interacciones no lineales entre variables, robustez frente a *outliers* y datos faltantes, manejo nativo de variables continuas y categóricas, y provisión de una medida de importancia de variables que facilita la interpretabilidad [29].

- **XGBOOST (*EXTREME GRADIENT BOOSTING)***

XGBoost es una implementación optimizada del algoritmo de *gradient boosting* propuesta por Chen y Guestrin [30]. A diferencia del *bagging* utilizado por Random Forest, el *boosting* construye los árboles de forma secuencial: cada nuevo árbol se entrena para corregir los errores de los anteriores, ponderando más las observaciones que fueron clasificadas incorrectamente.

Las características que distinguen a XGBoost de implementaciones tradicionales de *gradient boosting* son: regularización L1 y L2 integrada para prevenir el sobreajuste, poda eficiente de árboles mediante el parámetro max_depth, paralelización del cómputo, manejo nativo de valores faltantes, y soporte para validación cruzada interna. Estas optimizaciones han posicionado a XGBoost como uno de los algoritmos más utilizados en problemas de clasificación con datos tabulares, con desempeño consistentemente superior en múltiples estudios médicos [27,30].

3.1.2.1.2 APRENDIZAJE NO SUPERVISADO

En el aprendizaje automático no supervisado, un modelo se encarga de buscar patrones en datos sin etiquetar. Este aprendizaje puede encontrar patrones o tendencias que los humanos no buscan específicamente. Este aprendizaje tiende a detectar agrupaciones de datos similares, creando clústeres. Una vez entrenado, el modelo puede identificar patrones similares y agrupar

esos datos en su categoría correspondiente [23]. Por ejemplo, un modelo de aprendizaje automático no supervisado puede analizar datos de una plataforma de ventas en línea e identificar diferentes tipos y patrones en los clientes y las compras que realizan [25].

3.1.2.1.3 APRENDIZAJE SEMISUPERVISADO

El aprendizaje semisupervisado ofrece un equilibrio ideal entre el aprendizaje supervisado y el no supervisado. Durante el entrenamiento, utiliza un conjunto de datos etiquetados más pequeño para guiar la clasificación y la extracción de características de un conjunto de datos sin etiquetar más grande. El aprendizaje semisupervisado puede resolver el problema de no tener suficientes datos etiquetados para un algoritmo de aprendizaje supervisado. También es útil si etiquetar suficientes datos resulta demasiado costoso [31].

3.1.2.1.4 APRENDIZAJE POR REFUERZO

El aprendizaje automático por refuerzo entrena los modelos mediante ensayo y error para que tomen acción mediante el establecimiento de un sistema de recompensas. Este aprendizaje puede entrenar modelos para jugar o tomar decisiones autónomas, indicando a la máquina cuándo tomó las decisiones correctas, lo que le ayuda a aprender con el tiempo qué acciones debe tomar [25].

3.1.2.2 METRICAS PARA EVALUAR EL DESEMPEÑO DE LOS MODELOS

La selección de métricas de evaluación apropiadas es uno de los aspectos más críticos en aplicaciones clínicas de *machine learning*. La métrica *accuracy* (exactitud), aunque ampliamente utilizada, es insuficiente cuando las consecuencias de los errores tipo I (falsos positivos) y tipo II (falsos negativos) son asimétricas, como ocurre en tamizaje de enfermedades crónicas [32].

3.1.2.2.1 MÉTRICAS DERIVADAS DE LA MATRIZ DE CONFUSIÓN

La matriz de confusión es la base sobre la cual se construyen la mayoría de métricas de clasificación binaria. Sus cuatro componentes son: Verdaderos Positivos (VP), Falsos Positivos (FP), Verdaderos Negativos (VN) y Falsos Negativos (FN), como se muestra en la Figura 1. De ellos se derivan:

- Sensibilidad (*Recall, True Positive Rate*):

$$\frac{VP}{VP + FN} \quad (3)$$

  Indica la proporción de casos positivos correctamente identificados. Crítica en tamizaje, donde un FN representa un paciente diabético no detectado.

- Especificidad *(True Negative Rate*):

$$\frac{VN}{VN + FP} \quad (4)$$

Indica la proporción de casos negativos correctamente identificados. Importante para evitar derivaciones innecesarias y costos asociados.

- Precisión (Valor Predictivo Positivo):

$$\frac{VP}{VP + FP} \quad (5)$$

Indica la confiabilidad del diagnóstico positivo.

- F1-score: media armónica de precisión y sensibilidad. Útil como métrica balanceada cuando ambas son relevantes.

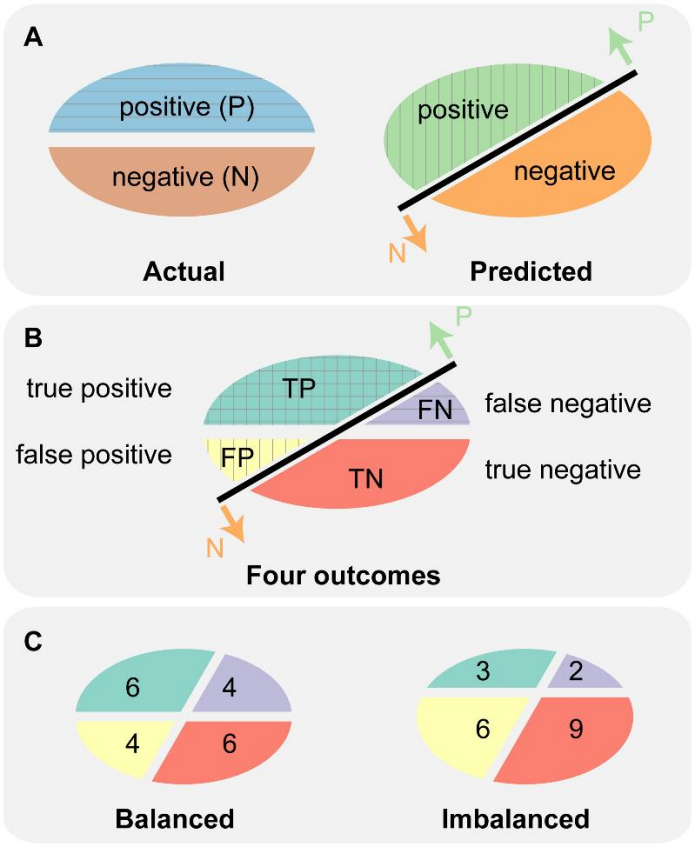


Figura 1. Matriz de confusión

Nota. (A) El óvalo izquierdo muestra dos etiquetas reales: positivas (P; azul; mitad superior) y negativas (N; rojo; mitad inferior). El óvalo derecho muestra dos etiquetas predichas: "predicho como positivo" (verde claro; mitad superior izquierda) y "predicho como negativo" (naranja; mitad inferior derecha). Una línea negra representa un clasificador que separa los datos en "predicho como positivo", indicado por la flecha hacia arriba "P", y "predicho como negativo", indicado por la flecha hacia abajo "N". (B) La combinación de dos etiquetas reales y dos predichas produce cuatro resultados: verdadero positivo (VP; verde), falso negativo (FN; morado), falso

positivo (FP; amarillo) y verdadero negativo (VN; rojo). (C) Dos óvalos muestran ejemplos de VP, FP, VN y FN para datos equilibrados (izquierda) y desequilibrados (derecha). Tomado de [33].

3.1.2.2.2 AUC-ROC Y AUC-PR

El Área Bajo la Curva ROC (AUC-ROC) es una métrica de discriminación independiente del umbral de decisión, que evalúa la capacidad del modelo para ordenar correctamente las observaciones positivas y negativas [33]. Un AUC-ROC de 1.0 indica discriminación perfecta, mientras que 0.5 corresponde a una clasificación aleatoria. En contextos clínicos, valores entre 0.7 y 0.8 se consideran aceptables, entre 0.8 y 0.9 buenos, y superiores a 0.9 excelentes.

Sin embargo, en escenarios con desbalance significativo, el AUC-ROC puede ser engañoso al sobreestimar el desempeño. En estos casos se recomienda complementar con el Área Bajo la Curva *Precision-Recall* (AUC-PR), que es más sensible a la clase minoritaria al no incluir los Verdaderos Negativos en su cálculo [32].

3.1.2.2.3 CALIBRACIÓN DEL MODELO (*BRIER SCORE*)

La calibración evalúa si las probabilidades predichas por el modelo reflejan las probabilidades reales de pertenencia a la clase positiva. Un modelo bien calibrado debe satisfacer entre todos los pacientes con probabilidad predicha de 0.7, aproximadamente el 70% sean realmente diabéticos [34].

El *Brier Score* cuantifica la calibración como el error cuadrático medio entre las probabilidades predichas y los resultados reales:

$$Brier = (1/n) \times \Sigma(\hat{p_i} - y_i)^2 \quad (6)$$

donde $\hat{p}_i$ es la probabilidad predicha y y_i es la etiqueta real (0 o 1). Valores cercanos a 0 indican mejor calibración. La calibración es particularmente importante en aplicaciones clínicas porque las decisiones médicas suelen tomarse con base en umbrales de probabilidad específicos. Un modelo con buena discriminación pero pobre calibración puede generar interpretaciones erróneas y decisiones clínicas subóptimas.

Cuando se identifican problemas de calibración, pueden aplicarse técnicas de recalibración como *Platt Scaling* (regresión sigmoidea) o *Isotonic Regression* [35], las cuales ajustan las probabilidades predichas mediante un modelo posterior entrenado sobre datos de validación.

3.1.2.2.4 DECISION CURVE ANALYSIS (DCA)

El Análisis de Curva de Decisión, propuesto por Vickers y Elkin, es una metodología emergente para evaluar la utilidad clínica neta de un modelo predictivo. A diferencia de las métricas tradicionales, DCA incorpora explícitamente las preferencias del clínico sobre el balance entre

falsos positivos y falsos negativos.

El beneficio neto se calcula como:

$$Beneficio\ Neto\ =\ \left(\frac{VP}{n}\right) -\ \left(\frac{FP}{n}\right) \times\ \left(umbral - (1 - umbral)\right) \quad (7)$$

donde el *ratio* (umbral / (1-umbral)) representa el costo relativo de un FP respecto a un FN. La curva DCA grafica el beneficio neto contra distintos umbrales y se compara con dos estrategias de referencia: tratar a todos los pacientes y no tratar a ninguno. Un modelo aporta utilidad clínica si su curva está por encima de ambas estrategias en el rango de umbrales clínicamente relevantes [36].

3.1.3 GENERALIDADES NORMATIVAS Y ÉTICAS

El presente proyecto se enmarca en un riguroso cumplimiento normativo orientado a la protección de los datos clínicos utilizados, articulando para ello el ordenamiento jurídico colombiano y los referentes internacionales aplicables. En el plano nacional, se atiende el derecho fundamental al Habeas Data consagrado en el artículo 15 de la Constitución Política [37], así como las disposiciones de la Ley Estatutaria 1581 de 2012 [38] y su Decreto reglamentario 1377 de 2013 [3], que regulan el tratamiento de datos personales y conceden una protección reforzada a los datos sensibles de salud, en línea con los pronunciamientos jurisprudenciales que reconocen el carácter fundamental y autónomo de este derecho [39]. La investigación se acoge a la excepción prevista en el artículo 6, literal d) de dicha ley, relativa al tratamiento con finalidad científica, condicionada a la disociación irreversible de la identidad del titular.

Desde una perspectiva ética, el proyecto se fundamenta en los cuatro principios bioéticos formulados por Beauchamp y Childress [40] (autonomía, beneficencia, no maleficencia y justicia), los cuales orientaron tanto las decisiones metodológicas como el diseño de la herramienta resultante. La autonomía se preservó mediante el apego estricto a los términos de cesión de los datos y la exclusión deliberada de variables susceptibles de generar discriminación; la beneficencia se materializa en el potencial impacto social del tamizaje temprano de diabetes mellitus tipo 2 en contextos de atención primaria; la no maleficencia se garantiza a través de una anonimización rigurosa, junto con la inclusión de un *disclaimer* explícito y el reconocimiento transparente de las limitaciones del modelo, y la justicia se promueve al sustentar el desarrollo en exámenes de bajo costo y amplia disponibilidad. Adicionalmente, se adoptaron voluntariamente los principios establecidos por la Organización Mundial de la Salud sobre inteligencia artificial en salud [41], lo cual se evidencia en la incorporación de técnicas de explicabilidad como SHAP, en la concepción de la herramienta como apoyo a la decisión clínica, nunca como sustituto del juicio profesional médico.

3.2. ANTECEDENTES

La predicción de Diabetes Mellitus Tipo 2 (DM2) mediante técnicas de aprendizaje automático ha sido abordada en diversos estudios internacionales. A continuación, se presentan los antecedentes más relevantes, organizados según su enfoque metodológico y contribución al estado del arte.

Uno de los estudios más relevantes es el desarrollado por Li y colaboradores [42] utilizaron datos del China Health and Nutrition Survey (CHNS) para construir un modelo predictivo de prediabetes mediante XGBoost, Random Forest, SVM y ANN, alcanzando un AUC de 0.939 con XGBoost e implementando SHAP para cuantificar el impacto de variables como Edad, Triglicéridos, ApoB e IMC. La diferencia con el presente trabajo radica en la adaptación al contexto clínico colombiano y en el uso de una arquitectura de doble modelo que se ajusta a la disponibilidad heterogénea de exámenes de laboratorio en la población local.

Por su parte, un estudio liderado por Lindholm y colaboradores [43] aplicaron Regresión Logística y Random Forest sobre una cohorte europea de más de 70,000 pacientes, incorporando 23 variables clínicas incluyendo HbA1c, demostrando alta precisión predictiva. A diferencia de este estudio, el presente proyecto excluye explícitamente la HbA1c como predictora para evitar fuga de datos, dado que constituye criterio diagnóstico, y opera con un menor número de variables seleccionadas mediante un esquema multinivel (ANOVA F-test, *Mutual Information*, VIF iterativo y RFECV).

En la publicación de Khedkar y colaboradores [44] desarrollaron un modelo predictivo de DM2 sobre datos multivariados con 23 atributos, aplicando técnicas de reducción de dimensionalidad y comparando múltiples algoritmos de aprendizaje automático. Su contribución metodológica destaca la importancia de seleccionar atributos óptimos. El presente proyecto se diferencia por aplicar selección formal de predictores con validación cruzada (RFECV) y análisis de multicolinealidad (VIF iterativo), aproximación más rigurosa que la selección manual.

Así mismo, Cardozo y colaboradores [45] propusieron un modelo basado en datos de laboratorio rutinarios utilizando KNN, SVM, Naive Bayes, Random Forest y ANN, alcanzando una sensibilidad del 78.1%. Aunque el enfoque es similar, el presente proyecto alcanza una sensibilidad superior (87.6% en el Modelo A), excluye HbA1c del conjunto de predictores para evitar fuga de datos y aplica análisis de calibración con recalibración mediante *Isotonic Regression* para garantizar confiabilidad clínica de las probabilidades.

Por otro lado, Kopitar y colaboradores [46] realizaron una detección temprana de DM2 con datos clínicos de laboratorio, comparando algoritmos como Glmnet, Random Forest, XGBoost y LightGBM en una población eslovena. Sus resultados mostraron desempeños similares a la

Regresión Logística sin validación estadística formal de las diferencias. El presente proyecto incorpora el test de DeLong para confirmar estadísticamente la superioridad de XGBoost sobre *Logistic Regression* ($p<0.0001$) y el test de McNemar para comparación de predicciones binarias.

En el año 2023, Zhang y colaboradores [47] aplicaron técnicas de minería de datos para predecir el riesgo de DM2 en una población comunitaria china, integrando variables clínicas, demográficas y de estilo de vida. El presente proyecto se diferencia al utilizar exclusivamente datos clínicos de laboratorio rutinarios disponibles en cualquier consulta médica, lo que aumenta la aplicabilidad operativa de la herramienta en contextos de atención primaria con recursos limitados.

En el estudio realizado por Wang y colaboradores [27] aplicaron XGBoost sobre datos de exámenes de sangre rutinarios (glucosa, HbA1c, perfil lipídico) en población china, demostrando robustez del algoritmo en contextos clínicos reales. El presente proyecto valida estos hallazgos en población colombiana y profundiza la interpretabilidad mediante SHAP global y local, incluyendo análisis por subgrupos etarios que revelan AUC=0.820 en pacientes jóvenes (18-45 años) donde la Edad sola tiene escaso poder discriminativo.

De igual manera, Silva y colaboradores [48] realizaron una revisión sistemática sobre *machine learning en* predicción de DM2, identificando heterogeneidad metodológica entre estudios y recomendando estandarización en transparencia y selección de variables clínicas. El presente proyecto adopta dichas recomendaciones siguiendo los lineamientos TRIPOD para reporte de modelos predictivos clínicos y documentando exhaustivamente las decisiones metodológicas.

A su vez, Kim y colaboradores [49] desarrollaron un modelo predictivo de DM2 a 1-2 años utilizando *Gradient Boosting* y Random Forest sobre variables clínicas amplias en población coreana, alcanzando alta precisión. Aunque su enfoque es prospectivo, el presente trabajo aporta valor mediante una arquitectura de doble modelo (A y B) con selección automática que se adapta a la disponibilidad real de exámenes en la práctica clínica colombiana.

Un estudio de gran escala basado en UK Biobank [50] empleó múltiples algoritmos de *machine learning* sobre más de 500,000 participantes, integrando HbA1c, IMC, GGT y otras variables, alcanzando AUC superiores a 0.85. El presente proyecto, aunque cuenta con un *dataset* menor (12.877 pacientes), aporta validación en población latinoamericana (subrepresentada en estudios internacionales) y alcanza un desempeño competitivo (AUC=0.806) sin utilizar HbA1c ni IMC como predictoras.

En el contexto latinoamericano, Sousa y colaboradores [51] aplicaron Regresión Logística y árboles de decisión en atención primaria en Brasil, alcanzando alta sensibilidad mediante variables de laboratorio y clínicas comunes. El presente proyecto se diferencia al implementar algoritmos avanzados (XGBoost) con interpretabilidad SHAP y *Decision Curve Analysis* (DCA),

técnicas escasamente reportadas en estudios latinoamericanos pero esenciales para evaluar la utilidad clínica neta del modelo.

En Colombia, Mejía y colaboradores [52] desarrollaron un modelo predictivo de DM2 con 10,889 usuarios del Sistema de Salud usando KNN, árboles de decisión y Random Forest, obteniendo un AUC de 0.61, basado principalmente en variables socioeconómicas y ambientales. Constituye el antecedente más cercano al presente trabajo en contexto colombiano, pero el presente proyecto alcanza un AUC sustancialmente superior (0.806) al utilizar variables bioquímicas con metodología más rigurosa, cubriendo así la brecha en la literatura nacional sobre predicción de DM2 mediante datos de laboratorio.

Steyerberg y colaboradores [34] establecieron el marco metodológico para evaluación de modelos predictivos clínicos, enfatizando que la discriminación es insuficiente y debe complementarse con análisis de calibración mediante *Brier Score* y curvas de calibración. El presente proyecto adopta esta recomendación, evaluando la calibración de ambos modelos y aplicando recalibración *post-hoc* mediante *Isotonic Regression* al Modelo B, que presentaba calibración deficiente.

Lindström y Tuomilehto [53] desarrollaron el FINDRISC, score clínico ampliamente validado internacionalmente con AUC entre 0.70 y 0.85, basado en variables auto-reportadas como IMC, perímetro abdominal y antecedentes familiares. Bernabé-Ortiz y colaboradores [54] validaron el FINDRISC en población peruana obteniendo AUC=0.71. El presente proyecto supera estos valores en población colombiana (AUC=0.806) utilizando exclusivamente variables de laboratorio rutinarias, lo que demuestra el valor agregado de los modelos basados en datos bioquímicos sobre los scores tradicionales.

Lundberg y colaboradores [55] aplicaron SHAP a modelos clínicos de hipoxemia perioperatoria, posicionando la interpretabilidad como componente esencial en sistemas predictivos en salud. Asi mismo, Yang y colaboradores [56] complementaron este abordaje aplicando SHAP a modelos de DM2 con análisis estratificado por subgrupos. El presente proyecto adopta ambas aproximaciones aplicando SHAP global, local y por cuartiles etarios, replicando el hallazgo de Yang sobre la mayor importancia relativa de variables lipídicas en pacientes jóvenes.

4. METODOLOGÍA

Este proyecto se desarrolló bajo la metodología *Cross-Industry Standard Process for Data Mining (*CRISP-DM, por sus siglas en inglés), ampliamente utilizada en el ámbito de la ciencia de datos por su estructura iterativa, flexible y orientada a resultados. Dicha metodología comprende seis fases que se adaptaron a los objetivos específicos del trabajo: comprensión del negocio y de los datos, preparación, modelado, evaluación, y despliegue.

4.1 VISIÓN GENERAL DEL FLUJO METODOLÓGICO

La Tabla 1 sintetiza la correspondencia entre las seis fases de CRISP-DM y las actividades específicas realizadas en el presente proyecto. Esta vista panorámica permite ubicar cada decisión metodológica en el contexto del proceso global del proyecto.

Tabla 1. Correspondencia entre fases CRISP-DM y actividades del proyecto

Fase CRISP-DM	Actividades principales
1. Comprensión del negocio	Definición del problema de tamizaje, criterios diagnósticos ADA.
2. Comprensión de los datos	Análisis exploratorio inicial, caracterización de variables, calidad de datos
3. Preparación de los datos	Limpieza, transformación pivot, ingeniería de características, definición del `TARGET`, selección formal de predictores, EDA
4. Modelado	División estratificada, entrenamiento de 3 algoritmos, optimización de hiperparámetros, análisis de estabilidad
5. Evaluación	Métricas clínicas, intervalos de confianza *Bootstrap*, DeLong/McNemar, calibración, DCA, SHAP, análisis por subgrupos, validación médica.
6. Despliegue	Persistencia de modelos, función de inferencia con selección automática.

4.2 FASE 1: COMPRENSIÓN DEL NEGOCIO

La fase de comprensión del negocio establece el problema a resolver, los objetivos del modelado y los criterios de éxito desde la perspectiva clínica. Esta fase es particularmente importante en aplicaciones de salud, donde el desempeño estadístico debe complementarse con utilidad clínica

práctica.

4.2.1 DEFINICIÓN DEL PROBLEMA CLÍNICO

La DM2 es una enfermedad crónica con alta prevalencia y subdiagnóstico en Colombia (entre 7% y 10% en población adulta), atribuible a la naturaleza asintomática de la enfermedad en sus etapas tempranas. Su detección temprana permite intervenciones preventivas que evitan complicaciones cardiovasculares, renales y oftalmológicas. El acceso a exámenes diagnósticos específicos (HbA1c, curva de tolerancia a la glucosa) es limitado en el primer nivel de atención colombiano, lo que motiva el desarrollo de herramientas de tamizaje basadas en exámenes de laboratorio rutinarios ya disponibles.

4.2.2 DEFINICIÓN DEL PROBLEMA DE *MACHINE LEARNING*

El problema se formuló como una tarea de clasificación binaria supervisada. La variable objetivo (`TARGET`) toma dos valores: 1 (Diabetes) y 0 (Normal), definidos según los criterios ADA 2026 [5]. Los predictores corresponden a variables demográficas y de laboratorio rutinario, excluyendo explícitamente glucosa (G06) y HbA1c (H10), ya que estas definen el `TARGET` y su inclusión generaría fuga de datos, invalidando el modelo para uso predictivo en escenarios donde dichos exámenes no estén disponibles.

4.3 FASE 2: COMPRENSIÓN DE LOS DATOS

Esta fase comprende el análisis exploratorio inicial del conjunto de datos, su caracterización estructural y la identificación de problemas de calidad que requieren tratamiento en la fase de preparación.

El conjunto de datos fue suministrado por IT Digital Health, empresa desarrolladora de software con operaciones en Colombia. Los datos corresponden a registros de exámenes de laboratorio rutinarios realizados a pacientes adultos en el período comprendido entre junio de 2025 y marzo de 2026, totalizando 339,101 filas y 17 columnas iniciales. La entrega de los datos se realizó mediante autorización formal de la empresa.

Es importante precisar la validez clínica de la fuente de los datos, los resultados de laboratorio que conforman el *dataset* provienen de laboratorios clínicos habilitados que operan bajo el sistema LIS administrado por IT Digital Health, los cuales se rigen por la normativa de habilitación de prestadores de servicios de salud vigente en Colombia. Si bien el presente proyecto no realizó una auditoría independiente del control de calidad analítico de dichos laboratorios, la validez clínica de los resultados se sustenta en su condición de prestadores habilitados y en operación regular dentro del sistema de salud colombiano, y no en una verificación adicional realizada por los autores.

4.3.1 ESTRUCTURA INICIAL DEL *DATASET*

El *dataset* original presenta estructura de formato largo, donde cada fila corresponde a un examen específico de un paciente. Las columnas principales del *dataset* original se agrupan en cuatro categorías funcionales: identificación del paciente (eliminada durante el proceso de anonimización), variables demográficas, datos de la orden médica y datos específicos de cada examen y analito. La Tabla 2 describe la estructura completa, incluyendo el tipo de dato y el significado de cada columna.

Tabla 2. Estructura inicial del *dataset*.

Nombre original	Tipo de dato	Significado
`IdentificationTypeCode`	object	Tipo de identificación del paciente
`IdentificationNumber`	object	Número de identificación del paciente, eliminado/anonimizado
`FirstName/SecondName`	object	Nombres del paciente, eliminados por anonimización
`FirstLastName/SecondLastName`	object	Apellidos del paciente, eliminados por anonimización
`BirthDate`	object	Fecha de nacimiento
`BiologicalSex`	object	`Sexo Biológico` registrado
`RequestNumber`	int64	Número de orden de laboratorio
`Observations`	object	Observaciones clínicas o del resultado
`RequestDate`	object	Fecha de solicitud del examen
`ExamCode`	object	Código del examen
`ExamName`	object	Nombre del examen
`AnalyteCode`	object	Código del analito
`AnalyteName`	object	Nombre del analito
`Results`	object	Resultado reportado del analito
`ResultDate`	object	Fecha de emisión del resultado

4.3.2 ANÁLISIS EXPLORATORIO PRELIMINAR

El análisis exploratorio inicial identificó las características y problemas de calidad que orientarían las decisiones de preparación: presencia de caracteres no numéricos (asteriscos) en la columna

Results indicando valores fuera de rango, heterogeneidad en la disponibilidad de exámenes por paciente (solo el 8% cuenta con perfil lipídico completo) y pacientes con múltiples órdenes médicas en el período del estudio

4.4 FASE 3: PREPARACIÓN DE LOS DATOS

La preparación de los datos constituye una fase crítica que determina directamente la calidad del modelado posterior. En este proyecto se aplicaron seis subprocesos secuenciales: limpieza, filtrado, codificación, transformación *pivot*, ingeniería de características y definición del TARGET.

4.4.1 LIMPIEZA, FILTRADO Y CODIFICACIÓN

Se procesó la columna Results para remover caracteres no numéricos (asteriscos) y se realizó la conversión a tipo numérico, descartando registros con valores no convertibles. La fecha de nacimiento (BirthDate) se convirtió a tipo datetime y se calculó la Edad como la diferencia entre la fecha del ingreso (RequestDate) y la fecha de nacimiento, expresada en años completos. Se filtró el *dataset* para retener únicamente registros de pacientes adultos (Edad ≥ 18 años), considerando que los criterios diagnósticos ADA aplican principalmente a esta población. La variable BiologicalSex se codificó numéricamente (0=Femenino, 1=Masculino) para su uso en modelos de *machine learning*. Los resultados se resumen en la Tabla 3.

Tabla 3. Resultados de preparación de los datos

Proceso	Resultado
Limpieza de columna Results	Valores no convertibles: 1.248 (0.37%)
Cálculo de Edad y filtrado de adultos	Registros adultos (≥18): 328,186 Rango etario: 18 - 126 años
Codificación de sexo biológico	0 (Femenino): 202508 1 (Masculino): 125678

4.4.2 ANONIMIZACIÓN

Tras el filtrado y codificación, se identificó que algunos pacientes presentaban más de una orden de laboratorio dentro del periodo analizado. Por esta razón, se conservó un identificador anonimizado del paciente (patient_hash) durante el procesamiento de datos, esta práctica garantiza el cumplimiento estricto de la Ley 1581 de 2012 sobre protección de datos personales y de los principios bioéticos de no maleficencia y respeto por la confidencialidad.

4.4.3 TRANSFORMACIÓN PIVOT DEL DATAFRAME

Se aplicó una transformación pivot del formato largo al formato amplio, generando una fila por orden de laboratorio y una columna por analito. Para evitar fuga de información por pacientes con múltiples órdenes, la partición entrenamiento/prueba se realizó posteriormente agrupando por el identificador anonimizado `patient_hash`, garantizando que las órdenes de un mismo paciente no quedaran simultáneamente en entrenamiento y prueba.

4.4.4 INGENIERÍA DE CARACTERÍSTICAS

Se construyeron variables derivadas con fundamento clínico, orientadas a capturar relaciones bioquímicas relevantes para la fisiopatología de la DM2:

Ratios lipídicos:

- `Índice Aterogénico (Colesterol Total/HDL)`: marcador integrador del riesgo cardiovascular global, propuesto por Castelli y colaboradores [13].
- `LDL/HDL Ratio`: cociente entre fracciones aterogénicas y protectoras.
- `TG/HDL Ratio`: marcador específico de resistencia a la insulina [57].

Indicadores binarios compuestos (basados en guías ATP III y KDIGO):

- `lipidic_flags`: contador de alteraciones lipídicas, suma de cuatro condiciones binarias (TG ≥ 150, HDL ≤ 40, LDL ≥ 130, CT ≥ 200), aproximación parcial al constructo de síndrome metabólico.

La inclusión de las variables derivadas se justifica por su uso establecido en la práctica clínica y por su capacidad de capturar información que las variables originales por separado no proporcionan eficientemente. Esta ingeniería de características es coherente con el principio metodológico de proporcionar al modelo representaciones que reflejan el conocimiento clínico previo. La Tabla 4 consolida las once variables consideradas en el estudio:

Tabla 4. Variables del estudio

Identificador Original	Variable clínica	Tipo	Justificación clínica
`Age`	`Edad` (años)	Demográfica	Factor de riesgo principal DM2 (ADA).
`BiologicalSex`	Sexo biológico	Demográfica	Factor de riesgo ADA/IDF; diferencias hormonales y de prevalencia.

C56	Creatinina (mg/dL)	Renal	Marcador de función renal, asociado a nefropatía diabética.
P65-1	Triglicéridos (mg/dL)	Lipídica	Resistencia a la insulina.
P65-2	Colesterol HDL (mg/dL)	Lipídica	Efecto protector documentado, síndrome metabólico.
P65-3	Colesterol LDL (mg/dL)	Lipídica	Riesgo cardiovascular asociado a DM2.
P65-4	Colesterol Total (mg/dL)	Lipídica	Marcador de dislipidemia global.
Indice_Aterogenico	CT/HDL ratio	Ratio derivado	Marcador integrador cardiometabólico.
lipidic_flags	Conteo criterios lipídicos alterados (0–4)	Indicador clínico	Síntesis lipídica.
G06	Glucosa	Variable objetivo	Usadas solo para definir TARGET; excluidas como predictores.
H10	/HbA1c	Variable objetivo	Usadas solo para definir TARGET; excluidas como predictores.

Nota. G06 y H10 definen el TARGET, pero no se incluyen como predictores para evitar fuga de datos.

4.4.5 DEFINICIÓN DEL TARGET

El TARGET binario se construyó aplicando los criterios diagnósticos de la ADA 2026 [5], adaptados a las variables disponibles en el *dataset*.

- TARGET = 1 (Diabetes): pacientes con G06 ≥ 126 mg/dL O H10 ≥ 6.5%.
- TARGET = 0 (Normal): pacientes con G06 < 100 mg/dL Y H10 < 5.7%.

La Tabla 5 presenta los umbrales de clasificación y el tamaño resultante de cada clase.

Tabla 5. Distribución del TARGET

Clase	Criterio Glucosa (G06)	Criterio HbA1c (H10)	N pacientes (%)
TARGET = 1 (Diabetes)	G06 ≥ 126 mg/dL	H10 ≥ 6.5%	8.806 (47.9%)
TARGET = 0 (Normal)	G06 < 100 mg/dL	H10 < 5.7%	9.572 (52.1%)
Excluidos (prediabetes)	100 ≤ G06 < 126 mg/dL	5.7% ≤ H10 < 6.5%	10.887 (excluidos)
Total *dataset* final	—	—	18.378 (100%)

***Nota*.** Pacientes con valores intermedios (prediabetes) fueron excluidos para garantizar clases bien diferenciadas. G06 y H10 no se incluyen como predictores para evitar fuga de datos.

La exclusión de 10.887 pacientes en estado de prediabetes (glucosa 100–125 mg/dL o HbA1c 5.7–6.4%) responde a una decisión metodológica deliberada, puesto que incluir pacientes con clasificación diagnóstica ambigua introduciría ruido en el entrenamiento y haría imposible la validación contra el estándar de referencia. El *dataset* final de 18.378 pacientes presenta un balance natural cercano al 50/50 (47.9% vs 52.1%) que favorece el entrenamiento de modelos sin necesidad de técnicas de sobre o submuestreo.

4.4.6 SELECCIÓN FORMAL DE PREDICTORES

Para garantizar una selección estadísticamente significativa del conjunto de predictores, se aplicó un esquema multinivel que combina cuatro métodos complementarios [58,59]:

- ANOVA F-test: evalúa la relación lineal de cada variable con el TARGET mediante análisis de varianza unidireccional.
- Información mutua: detecta dependencias no lineales entre cada variable y el TARGET, complementando el ANOVA F-test que asume linealidad.
- VIF iterativo: cuantifica la multicolinealidad entre predictores mediante el *Variance Inflation Factor,* eliminando iterativamente la variable con VIF más alto hasta que todas estén por debajo del umbral de 10.
- RFECV (Recursive *Feature Elimination* con Cross-Validation): elimina recursivamente variables y selecciona el subconjunto óptimo mediante validación cruzada estratificada.

4.5 ANALISIS EXPLORATORIO DE DATOS (EDA)

El EDA se realizó sobre el *dataset* estructurado después del pivote y la definición del TARGET, dado que las variables de interés clínico (perfil lipídico, ratios derivados, función renal) requieren

la estructura tabular y el `TARGET` binario para ser analizadas correctamente. El EDA se estructuró en tres niveles: análisis univariado (distribución de cada variable), análisis bivariado (relación con el `TARGET`) y análisis de multicolinealidad. A nivel metodológico, el EDA informó tres decisiones críticas:

- Selección de pruebas estadísticas: ninguna variable sigue distribución normal (*Shapiro-Wilk y Kolmogorov-Smirnov,* $p < 0.001$ en todos los casos), lo que fundamenta el uso exclusivo de pruebas no paramétricas (*Mann-Whitney U, Spearman, Kruskal-Wallis)* en el análisis bivariado.
- Conservación de valores atípicos: los *outliers* detectados corresponden a manifestaciones patológicas reales (hipercolesterolemia severa, enfermedad renal crónica, dislipidemia aterogénica), cuya eliminación comprometería la representatividad clínica del modelo.
- Arquitectura de doble modelo: la diferencia en disponibilidad de variables (`Edad` y `Creatinina` presentes en 18.378 pacientes vs perfil lipídico en solo 1.228) fundamenta el diseño de dos modelos con distintos conjuntos de predictores y distintos tamaños muestrales.

4.6 FASE 4: MODELADO

La fase de modelado comprende el diseño del esquema de entrenamiento, la selección de algoritmos y validación, la optimización de hiperparámetros y la calibración de probabilidades.

Dado que el *dataset* contiene múltiples registros por paciente, una partición aleatoria estándar permitiría que registros del mismo individuo aparecieran simultáneamente en los conjuntos de entrenamiento y prueba, constituyendo una fuga de datos (*data leakage*) que inflaría artificialmente las métricas de desempeño. Para evitar este sesgo, se implementó `GroupShuffleSplit` agrupado por identificador de paciente, garantizando que la totalidad de los registros de un mismo paciente pertenece exclusivamente a uno de los dos conjuntos, preservando así la independencia estadística entre entrenamiento y evaluación.

- Proporción: 80% entrenamiento/20% prueba, mantenida en ambos modelos.
- Modelo A: *Train*: 10.920 registros (9.412 pacientes) | *Test*: 2.730 registros (2.354 pacientes).
- Modelo B: *Train*: 899 registros (856 pacientes) | *Test*: 230 registros (215 pacientes).
- Verificación de no fuga: 0 pacientes comunes entre *train y test* en ambos modelos, verificado explícitamente antes del entrenamiento.

- Estratificación: distribución del `TARGET` preservada: Normal 52.1% / Diabetes 47.9% en entrenamiento; 50.9% / 49.1% en prueba.

La estrategia de transformación y escalamiento se diseñó de forma diferencial según el tipo de algoritmo, implementada dentro de un Pipeline de scikit-learn para garantizar que los parámetros de normalización se ajustan exclusivamente sobre el conjunto de entrenamiento y se aplican posteriormente en validación y prueba, evitando cualquier fuga de información. Los modelos basados en árboles de decisión (Random Forest y XGBoost) no requieren escalamiento dado que su naturaleza intrínsecamente invariante a transformaciones monótonas: las particiones recursivas del espacio de características operan mediante comparaciones relativas (mayor o menor que un umbral) que no se ven afectadas por la escala ni por la distribución de las variables. La Regresión Logística, en cambio, minimiza una función de pérdida convexa cuyo gradiente es sensible tanto a la escala como a la presencia de valores extremos, justificando la aplicación de `RobustScaler`, que utiliza la mediana y el rango intercuartílico como estadísticos de escala y de transformación logarítmica (log1p) sobre las variables con asimetría severa.

Antes de entrenar cualquier modelo de *machine learning*, se estableció un *baseline* mínimo mediante `DummyClassifier` con estrategia `stratified` (predice respetando la proporción de clases del entrenamiento). Este *baseline* representa el nivel de desempeño que cualquier modelo entrenado debe superar para justificar su uso.

4.6.1 ALGORITMOS EVALUADOS

Se evaluaron tres algoritmos de clasificación binaria para cada escenario, seleccionados por sus características complementarias frente al problema clínico:

- Regresión Logística (LR): modelo lineal interpretable que proporciona probabilidades calibradas de forma nativa y coeficientes directamente interpretables. Actúa como *baseline* paramétrico. Su sensibilidad a la escala y a la multicolinealidad justifica el uso de `RobustScaler` y la selección previa de *features* mediante VIF.
- Random Forest (RF): ensamble de árboles de decisión con muestreo aleatorio de *features* y *bootstrap*. Robusto a *outliers* y multicolinealidad, captura interacciones no lineales. Su principal limitación es la tendencia al *overfitting* en tamaños muestrales pequeños.
- XGBoost: *gradient boosting* con regularización L1/L2 integrada, manejo nativo de desbalance de clases y alta capacidad predictiva. Reconocido como el estado del arte en problemas de clasificación tabular con *datasets* de tamaño moderado.

La búsqueda de hiperparámetros se realizó mediante `GridSearchCV` con validación cruzada estratificada y agrupada por paciente (`StratifiedGroupKFold`, 5 *folds*), optimizando el

AUC-ROC como métrica de selección. El uso de validación cruzada agrupada garantiza que la selección de hiperparámetros no se contamina con información de pacientes presentes en la partición de validación.

Ambos modelos finales fueron calibrados mediante *Platt Scaling*, ajustando las probabilidades predichas para que correspondan a frecuencias reales de eventos. La calibración es especialmente relevante en el contexto clínico, donde los umbrales de decisión se aplican directamente sobre las probabilidades predichas y la interpretación del riesgo individual requiere que dichas probabilidades sean confiables.

4.7 FASE 5: EVALUACIÓN

La fase de evaluación se abordó mediante un esquema que incluye múltiples métricas clínicamente relevantes, intervalos de confianza, comparación estadística entre modelos, análisis de calibración, evaluación de utilidad clínica e interpretabilidad.

Para cuantificar la incertidumbre estadística asociada a las métricas se aplicó el método *bootstrap* [67] con 1,000 iteraciones de remuestreo con reemplazo. Los intervalos de confianza no paramétricos del 95% se calcularon mediante los percentiles 2.5% y 97.5% de la distribución *bootstrap*. Esta práctica es estándar en investigación clínica predictiva y reemplaza el reporte de métricas puntuales sin medida de incertidumbre.

Para determinar si las diferencias en AUC-ROC entre modelos son estadísticamente significativas se aplicó el test de DeLong [68], prueba estándar para comparar dos AUC obtenidas sobre el mismo conjunto de datos pareados.

Por otro lado, se realizó un análisis de la calibración, donde se evalúa si las probabilidades predichas reflejan las probabilidades reales de pertenencia a la clase positiva [70]. Se evaluó mediante: (i) Brier Score, que cuantifica el error cuadrático medio entre probabilidades predichas y observadas; y (ii) curvas de calibración, que visualizan la relación entre probabilidad predicha y observada por deciles de probabilidad. La Tabla 6 describe las métricas utilizadas y su relevancia clínica en el contexto del tamizaje de DM2.

Tabla 6. Métricas de evaluación y su relevancia clínica

Métrica	Fórmula	Relevancia clínica
AUC-ROC	Área bajo la curva ROC	Capacidad discriminativa global; independiente del umbral

Sensibilidad (*Recall*)	TP / (TP+FN)	Crítica en tamizaje: detectar diabéticos reales; minimizar FN
Especificidad	TN / (TN+FP)	Evitar falsos positivos; reducir carga de confirmación
VPN	TN / (TN+FN)	Seguridad del resultado negativo; confianza del descarte
VPP (Precisión)	TP / (TP+FP)	Valor del resultado positivo
F1-score	2·P·R/(P+R)	Balance entre precisión y sensibilidad
Brier Score	Media de (prob−real)²	Calibración de probabilidades predichas
AUC-PR	Área bajo curva P-R	Útil en clases desbalanceadas

Como otra medida adicional se tiene en cuenta, el *Decision Curve Analysis* [71], el cual evalúa la utilidad clínica neta del modelo a través de distintos umbrales de decisión, comparándolo con dos estrategias de referencia: "tratar a todos" y "no tratar a nadie".

La validación médica se abordó mediante la coherencia de los valores SHAP con la fisiopatología conocida de la DM2. Los métodos de interpretabilidad utilizados fueron MDI (*Mean Decrease in Impurity*) y SHAP (*SHapley Additive exPlanations)* [72], aplicados tanto a nivel global (importancia del conjunto de datos de prueba) como local (explicación de predicciones individuales). Para profundizar en la comprensión del comportamiento del modelo y validar que su capacidad discriminativa trasciende correlaciones demográficas evidentes, se aplicaron dos análisis complementarios:

- Estratificación por cuartiles etarios: el conjunto de prueba se dividió en cuatro cuartiles según `Edad` y se calculó AUC-ROC en cada subgrupo. Si el modelo solo capturase la correlación `Edad`-diabetes, el AUC en cuartiles homogéneos debería caer hasta valores cercanos al azar (0.5).
- Estudio de ablación: se entrenaron cuatro configuraciones del modelo eliminando subconjuntos de variables (modelo completo, sin `Edad`, solo demográficas, solo `Edad`), cuantificando el aporte incremental de cada conjunto al AUC.

4.8 FASE 6: DESPLIEGUE

La fase de despliegue materializa el valor práctico del modelo mediante una herramienta web de uso clínico. Se adoptó una arquitectura web desacoplada compuesta por: *backend* implementado en Python (responsable de la lógica de negocio, la inferencia del modelo y la seguridad), *frontend* desarrollado en Angular (responsable de la interfaz de usuario y captura de datos clínicos), comunicación mediante `API REST` con formato `JSON`, y persistencia de los modelos entrenados mediante archivos `joblib (.pkl)`. Esta arquitectura desacoplada facilita el mantenimiento, la escalabilidad y la integración futura con sistemas de información hospitalaria.

El *backend* implementa una lógica de selección automática que determina cuál modelo aplicar según los datos disponibles del paciente. Cuando se dispone de perfil lipídico completo más `Creatinina`, se activa el Modelo B. Cuando únicamente están disponibles los datos demográficos y `Creatinina`, se aplica el Modelo A. La respuesta de la API incluye la predicción binaria, la probabilidad calibrada, el modelo utilizado y la categorización del riesgo en tres niveles (bajo, moderado y alto).

La recomendación clínica personalizada se genera de forma dinámica mediante una consulta a una API de modelo de lenguaje de gran escala (LLM), que recibe como contexto el perfil del paciente, la probabilidad predicha, el nivel de riesgo y las variables más influyentes según los valores SHAP, produciendo una recomendación estructurada, contextualizada y adaptada a las características individuales del caso. Esta arquitectura garantiza que la recomendación no es un texto genérico predefinido sino una síntesis clínica coherente con el resultado específico del modelo. Todo el proceso es transparente para el usuario final, quien interactúa con un único formulario que se adapta dinámicamente a la información disponible, sin necesidad de seleccionar manualmente el modelo a aplicar.

La herramienta se diseñó como sistema de apoyo a la decisión, no como sistema autónomo de diagnóstico. En consecuencia, la respuesta de la API incluye explícitamente un *disclaimer* que enfatiza el carácter de apoyo de la herramienta y la necesidad de complementar las predicciones con el juicio clínico profesional y pruebas confirmatorias.

Todo el desarrollo se realizó en Google Colaboratory (Colab), entorno de ejecución basado en notebooks Jupyter en la nube. Esta elección se justificó por la disponibilidad de recursos computacionales gratuitos, la facilidad de colaboración entre los autores del proyecto y la integración nativa con Google Drive para almacenamiento de *datasets* y resultados intermedios. La Tabla 7 presenta las principales bibliotecas utilizadas.

Tabla 7. Bibliotecas y herramientas computacionales utilizadas

Biblioteca	Uso en el proyecto
`Python`	Lenguaje base de implementación
`pandas`	Manipulación y transformación de datos
`numpy`	Operaciones numéricas vectorizadas
`scikit-learn`	Modelos LR, RF, métricas, validación cruzada, `GridSearchCV`, calibración
`xgboost`	Implementación de XGBoost
`statsmodels`	Cálculo de VIF para análisis de multicolinealidad y test de McNemar
`shap`	Análisis de interpretabilidad del modelo
`matplotlib / seaborn`	Generación de visualizaciones
`scipy`	Pruebas estadísticas (Shapiro-Wilk, Mann-Whitney U)
`joblib`	Serialización de modelos entrenados (.pkl)

Para garantizar reproducibilidad, se aplicaron las siguientes prácticas: fijación de semillas aleatorias globales, almacenamiento intermedio de los conjuntos de datos procesados, documentación detallada de cada decisión metodológica en celdas *markdown* del *notebook*, y persistencia de los modelos finales con metadatos asociados. El *notebook* completo, está organizado por fases CRISP-DM y constituye un documento reproducible que permite replicar el proceso completo desde la carga de datos hasta la generación de resultados.

5. ANALISIS EXPLORATORIO DE DATOS

El análisis exploratorio de datos (EDA, por sus siglas en inglés) constituye una fase fundamental en cualquier proyecto de ciencia de datos. Este capítulo presenta el análisis aplicado al conjunto de datos consolidado tras la fase de preparación, organizado en tres niveles: análisis univariado, análisis bivariado y análisis correlacional incluyendo evaluación de multicolinealidad.

Este proceso permitió caracterizar estadísticamente la población del estudio y sus principales biomarcadores, identificar las variables con mayor capacidad discriminativa entre pacientes diabéticos y no diabéticos mediante pruebas estadísticas formales, y detectar problemas de calidad de datos como *outliers*, distribuciones no normales y multicolinealidad entre predictores, que orientan las decisiones metodológicas posteriores de modelado.

5.1 CARACTERIZACIÓN GENERAL DEL CONJUNTO DE DATOS

Como se observa en la Tabla 8, tras aplicar el proceso de preparación anterior se obtuvo un *dataset* consolidado de 18.378 pacientes con clasificación diagnóstica clara: 9.572 pacientes clasificados como Normal (52.1%) y 8.806 pacientes clasificados como Diabetes (47.9%). Esta distribución refleja un balance natural cercano a 50/50 que favorece el desempeño de los algoritmos de aprendizaje supervisado, evitando los problemas asociados a clases altamente desbalanceadas.

Tabla 8. Caracterización general del conjunto de datos consolidado

Característica	Valor	Observación
Total de pacientes	18.378	Con `TARGET` binario definido
Pacientes Normales (`TARGET`=0)	9.572 (52.1%)	GPA<100 Y HbA1c<5.7
Pacientes Diabéticos	8.806 (47.9%)	GPA≥126 O HbA1c≥6.5

(TARGET=1)		
Período de recolección	Junio 2025 - Marzo 2026	10 meses de datos
Rango etario	18 - 108 años	Población adulta

5.2 ANÁLISIS UNIVARIADO

El análisis univariado examina cada variable de manera individual, caracterizando su distribución mediante estadísticos descriptivos, evaluando su forma mediante medidas de asimetría y curtosis, y detectando valores atípicos. Adicionalmente, se aplican pruebas formales de normalidad para determinar el tipo de pruebas estadísticas (paramétricas o no paramétricas) apropiadas en el análisis bivariado posterior.

5.2.1 ESTADÍSTICOS DESCRIPTIVOS DE VARIABLES CONTINUAS

La Tabla 9 presenta los principales estadísticos descriptivos de las variables continuas del estudio. Estos valores caracterizan tanto la tendencia central (media, mediana) como la dispersión (desviación estándar, rango intercuartílico) y forma (asimetría, curtosis) de cada distribución, como se observa en la Figura 2.

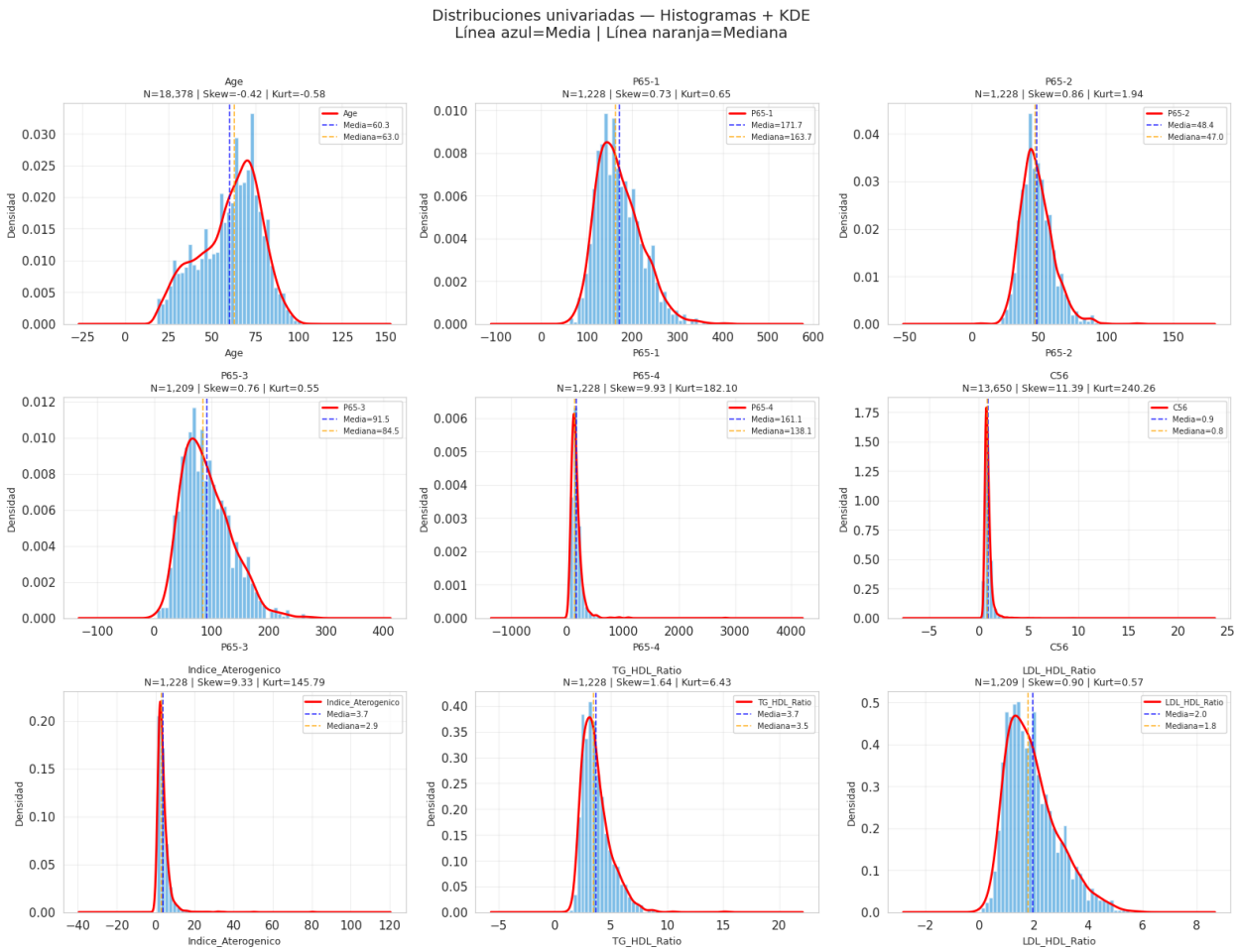


Figura 2. Distribuciones univariadas de variables continuas.

Nota. Histogramas de densidad con curva KDE superpuesta (línea roja). La línea azul discontinua indica la Media y la línea naranja discontinua indica la Mediana. Cada subgráfico muestra N, coeficiente de asimetría (*Skew*) y curtosis (*Kurt*).

Tabla 9. Estadísticos descriptivos de las variables continuas

Variable	N	Media	Mediana	Desv. Est.	Asimetría	Curtosis	Interp. Asimetría	Interp. Curtosis
Edad (años)	18.378	60.30	63.00	17.47	-0.418	-0.579	Simétrica	Mesocúrtica
Triglicéridos (mg/dL)	1.228	171.65	163.70	49.15	0.734	0.647	Mod. sesgada	Mesocúrtica

Colesterol HDL (mg/dL)	1.228	48.42	47.00	11.97	0.860	1.941	Mod. sesgada	Leptocúrtica
Colesterol LDL (mg/dL)	1.209	91.49	84.54	41.99	0.759	0.550	Mod. sesgada	Mesocúrtica
Colesterol Total (mg/dL)	1.228	161.07	138.14	123.61	9.930	182.102	Altamente sesgada	Leptocúrtica
Creatinina (mg/dL)	13.650	0.89	0.81	0.48	11.390	240.260	Altamente sesgada	Leptocúrtica
Índice Aterogénico	1.228	3.75	2.91	3.91	9.334	145.785	Altamente sesgada	Leptocúrtica
Ratio TG-HDL	1.228	3.70	3.45	1.28	1.636	6.429	Altamente sesgada	Leptocúrtica
Ratio LDL-HDL	1.209	1.95	1.78	0.95	0.898	0.569	Mod. sesgada	Mesocúrtica

*Nota***.** DE = Desviación Estándar. Asimetría e interpretación basadas en el coeficiente de Fisher. Curtosis exceso (*Fisher*): Mesocúrtica ≈ 0; Leptocúrtica > 1.

Los resultados de la Tabla 9 revelan los siguientes hallazgos:

- La `Edad` es la única variable con distribución aproximadamente simétrica (*skew* = -0.418, curtosis = -0.579), con una media de 60.3 años y mediana de 63.0 años. La ligera asimetría negativa refleja una concentración de pacientes en rangos etarios mayores, coherente con el perfil de una población que asiste a realizarse exámenes metabólicos de rutina.
- Las variables del perfil lipídico estándar: `Triglicéridos` (*skew* = 0.73), `Colesterol HDL` (*skew* = 0.86), `Colesterol LDL` (*skew* = 0.76) y `Ratio LDL-HDL` (*skew* = 0.90), presentan asimetría moderada con cola derecha, patrón típico en biomarcadores lipídicos donde una minoría de pacientes exhibe valores extremos asociados a dislipidemia severa. Sus curtosis se ubican en rangos mesocúrticos o levemente leptocúrticos, sin presencia de colas extremadamente pesadas.
- En contraste, cuatro variables exhiben asimetría severa con curtosis marcadamente leptocúrtica: el `Colesterol Total` (*skew* = 9.93, *kurt* = 182.10), la `Creatinina` (*skew* = 11.39, *kurt* = 240.26), el `Índice Aterogénico` (*skew* = 9.33, *kurt* = 145.79) y el `Ratio TG-HDL` (*skew* = 1.64, *kurt* = 6.43). Estos patrones son clínicamente

explicables: el Colesterol Total y el Índice Aterogénico presentan valores extremos en pacientes con hipercolesterolemia severa o síndromes metabólicos; la Creatinina concentra la mayoría de sus valores en el rango normal (0.6-1.2 mg/dL) con una cola derecha extensa correspondiente a pacientes con enfermedad renal crónica. La detección de estos patrones confirma la pertinencia de utilizar algoritmos robustos a *outliers* (Random Forest, XGBoost) y de aplicar RobustScaler en el modelo de Regresión Logística.

Finalmente, la diferencia en el tamaño muestral entre variables (18.378 observaciones para Edad, 13.650 para Creatinina, y 1.209-1.228 para el perfil lipídico) refleja la disponibilidad diferencial de cada examen en el sistema LIS y justifica la arquitectura de doble modelo adoptada: el Modelo A aprovecha la alta disponibilidad de datos demográficos y de Creatinina, mientras que el Modelo B explota el subconjunto de pacientes con panel lipídico completo.

5.2.2 PRUEBAS DE NORMALIDAD

Para determinar si las variables continuas siguen una distribución normal, se aplicaron dos pruebas estadísticas formales: la prueba de *Shapiro-Wilk* [73] y la prueba de *Kolmogorov-Smirnov* [74]. La hipótesis nula en ambas pruebas establece que la variable proviene de una distribución normal, siendo los valores p inferiores a 0.05 permiten rechazar dicha hipótesis. Dado que la prueba de *Shapiro-Wilk* pierde potencia con muestras grandes (n>5,000), se aplicó sobre una submuestra aleatoria de 5,000 observaciones cuando fue necesario, manteniendo la prueba de *Kolmogorov-Smirnov* sobre la totalidad de los datos. Como se observa en la Tabla 10.

Tabla 10. Resultados de pruebas de normalidad

Variable	N	*Shapiro-Wilk* p	KS p	Normalidad
Edad (años)	18.378	< 0.001	< 0.001	No
Triglicéridos (P65-1)	1.228	< 0.001	< 0.001	No
Colesterol HDL (P65-2)	1.228	< 0.001	< 0.001	No
Colesterol LDL (P65-3)	1.209	< 0.001	< 0.001	No
Colesterol Total (P65-4)	1.228	< 0.001	< 0.001	No
Creatinina (C56)	13.650	< 0.001	< 0.001	No
Índice	1.228	< 0.001	< 0.001	No

Aterogénico				
Ratio TG-HDL	1.228	< 0.001	< 0.001	No
Ratio LDL-HDL	1.209	< 0.001	< 0.001	No

Como muestra la Tabla 10, ninguna de las variables sigue una distribución normal (todos los p-valores son inferiores a 0.001, tanto en *Shapiro-Wilk* como en *Kolmogorov-Smirnov*). Este es un hallazgo esperable en datos clínicos reales según Altman y colaboradores [75], desde el punto de vista metodológico, este resultado tiene una implicación directa e irrevocable sobre el análisis bivariado posterior: no es válido aplicar pruebas paramétricas que asumen normalidad (prueba t de Student, ANOVA, correlación de Pearson). En consecuencia, el análisis bivariado se realizó íntegramente mediante pruebas no paramétricas: la prueba U de Mann-Whitney para comparar distribuciones entre grupos, y el coeficiente de correlación de Spearman para evaluar asociaciones entre variables, garantizando la validez estadística de las inferencias realizadas.

5.2.3 DETECCIÓN DE VALORES ATÍPICOS

Los valores atípicos (*outliers*) pueden afectar significativamente el desempeño de los modelos predictivos, particularmente aquellos basados en regresión lineal. Sin embargo, en datos clínicos reales muchos *outliers* son valores legítimos que reflejan condiciones patológicas relevantes. La decisión sobre cómo manejarlos requiere balancear la calidad estadística con la validez clínica.

Para la detección de *outliers* se aplicó el método del Rango Intercuartílico (IQR), que considera atípico cualquier valor que se ubique fuera del intervalo [Q1 - 1.5 × IQR, Q3 + 1.5 × IQR], donde Q1 y Q3 corresponden al primer y tercer cuartil respectivamente. La Figura 3 y Tabla 11 presentan el porcentaje de *outliers* detectados en cada variable.

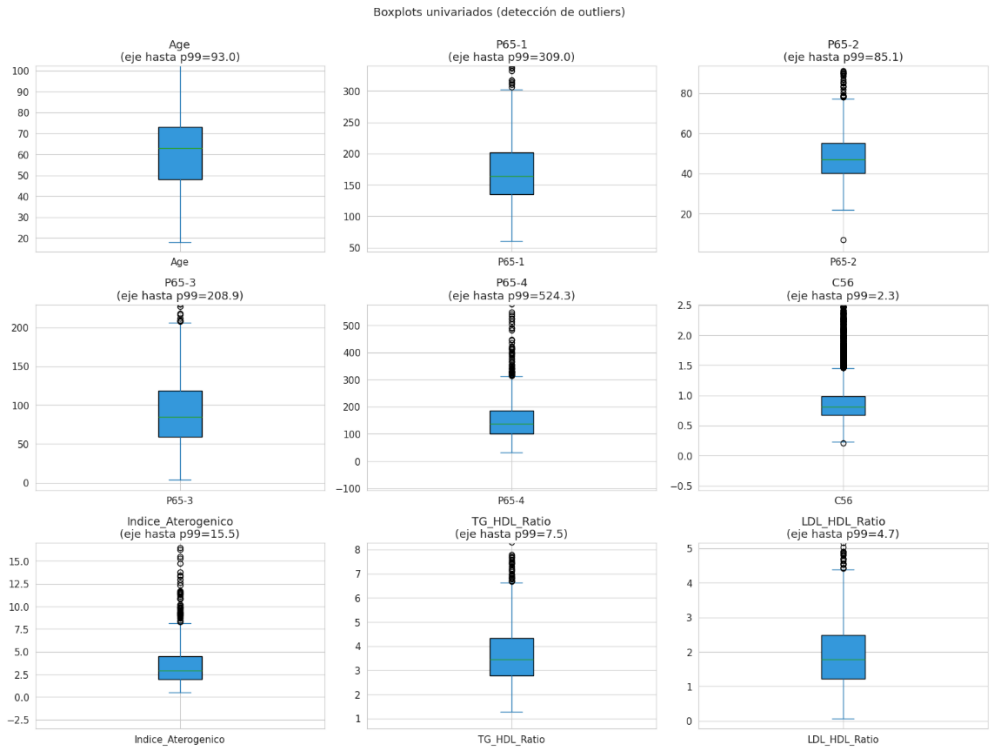


Figura 3. *Boxplots* de valores atípicos univariados.

Tabla 11. Detección de valores atípicos mediante método IQR

Variable	**N total**	**N *outliers***	**% *outliers***	**Límite inferior**	**Límite superior**	**Interpretación**
Edad (años)	18.378	0	0.00%	10.50	110.50	Sin *outliers*
Triglicéridos (mg/dL)	1.228	14	1.14%	34.29	303.05	Hipertrigliceridemia
Colesterol HDL (mg/dL)	1.228	26	2.12%	17.66	77.76	HDL muy bajo o muy alto
Colesterol LDL (mg/dL)	1.209	16	1.32%	-29.33	207.39	Dislipidemia severa
Colesterol Total (mg/dL)	1.228	63	5.13%	-24.62	312.84	Hipercolesterolemia severa
Creatinina (mg/dL)	13.650	653	4.78%	0.22	1.45	Daño renal / ERC

Índice Aterogénico	1.228	62	5.05%	-1.83	8.27	Dislipidemia aterogénica
Ratio TG-HDL	1.228	34	2.77%	0.45	6.65	Resistencia a la insulina
Ratio LDL-HDL	1.209	22	1.82%	-0.67	4.39	Perfil lipídico desfavorable

Nota: Método de detección: Rango Intercuartílico (IQR). Límites calculados como [Q1 − 1.5×IQR, Q3 + 1.5×IQR].

El análisis de valores atípicos mediante el método IQR revela un patrón consistente con la realidad clínica de la población estudiada. En conjunto, nueve variables fueron evaluadas y en ocho de ellas se detectaron *outliers*, con porcentajes que oscilan entre 1.14% y 5.13%. Un hallazgo central es que todos los valores extremos detectados corresponden a manifestaciones patológicas reales y no a errores de medición o registro.

La Edad es la única variable completamente libre de *outliers* (0%), lo cual es coherente con el rango definido de inclusión (18-108 años) y confirma la calidad del proceso de depuración de datos en la fase de preparación.

Las variables con mayor proporción de *outliers* son tres:

- Colesterol Total (5.13%, N=63): los 63 casos extremos superan el límite superior de 312.84 mg/dL, correspondiendo a pacientes con hipercolesterolemia severa o familiar. Su presencia es clínicamente relevante dado que esta condición es un factor de riesgo independiente para diabetes tipo 2 y síndrome metabólico.
- Índice Aterogénico (5.05%, N=62): valores superiores a 8.27 reflejan una relación CT/HDL extremadamente desfavorable, indicativa de dislipidemia aterogénica severa. Nótese que el límite inferior negativo (-1.83) es matemáticamente posible pero no tiene interpretación clínica real dado que ambos componentes del ratio son positivos.
- Creatinina (4.78%, N=653): con el mayor número absoluto de *outliers* del conjunto, los 653 casos superan el límite superior de 1.45 mg/dL, indicando deterioro de la función renal. Dado que el *dataset* cuenta con 13.650 observaciones de Creatinina, este 4.78% representa una proporción relevante de pacientes con nefropatía o enfermedad renal crónica, condición estrechamente asociada a la diabetes mellitus tipo 2.
- Las variables restantes (Triglicéridos (1.14%), HDL (2.12%), LDL (1.32%), Ratio TG-HDL (2.77%) y Ratio LDL-HDL (1.82%)) presentan porcentajes bajos consistentes con prevalencias esperadas de dislipidemia severa en población general.

Frente a estos hallazgos, se tomó la decisión metodológica de conservar todos los valores atípicos en el *dataset* por dos razones fundamentales. Primero, por validez clínica: los pacientes con valores extremos representan precisamente la población de mayor riesgo metabólico, cuya correcta clasificación constituye el objetivo central del modelo predictivo; eliminarlos comprometería la generalización del modelo a los casos más relevantes. Segundo, por robustez algorítmica: dos de los tres algoritmos seleccionados (Random Forest y XGBoost) son intrínsecamente robustos a *outliers* por su naturaleza basada en particiones recursivas del espacio de características. Para el modelo de Regresión Logística se aplicó `RobustScaler`, que utiliza la mediana y el IQR en lugar de la media y desviación estándar, mitigando el efecto de los valores extremos sin eliminarlos.

5.3 ANÁLISIS BIVARIADO

El análisis bivariado examina la relación entre cada variable predictora y la variable objetivo (`TARGET`), permitiendo identificar qué predictores presentan diferencias estadísticamente significativas entre pacientes diabéticos y no diabéticos. Dada la confirmada no normalidad de las variables continuas, se aplicaron pruebas no paramétricas: la prueba U de Mann-Whitney [76] para comparar la distribución de variables continuas entre grupos, y la prueba chi-cuadrado para variables categóricas.

Adicionalmente, para cada comparación se calculó el tamaño del efecto mediante Cohen's d [77] para variables continuas y Cramér's V [78] para variables categóricas. Esta práctica es metodológicamente importante porque en muestras grandes, diferencias estadísticamente significativas pueden corresponder a efectos clínicamente irrelevantes, el tamaño del efecto cuantifica la magnitud práctica de la diferencia. Los resultados del análisis bivariado se consolidan en la Tabla 12.

Tabla 12. Comparación de variables continuas entre pacientes normales y diabéticos

Variable	Med. Normal	Med. Diabetes	p-value (MW)	Sig.	Cohen's d	Efecto
`Edad`	53.0	69.00	< 0.001	Sí	0.963	Grande
`Colesterol HDL`	49.95	45.20	< 0.001	Sí	-0.384	Pequeño
`Colesterol LDL`	96.96	78.79	< 0.001	Sí	-0.353	Pequeño
`Triglicéridos`	118.85	147.01	< 0.001	Sí	0.343	Pequeño

Índice Aterogénico	2.38	3.18	< 0.001	Sí	0.308	Pequeño
Colesterol Total	173.75	158.50	< 0.001	Sí	-0.236	Pequeño
Ratio LDL-HDL	1.93	1.69	0.001	Sí	-0.181	Despreciable
Creatinina	0.81	0.81	0.593	No	0.064	Despreciable
Ratio TG-HDL	3.37	3.47	0.479	No	0.043	Despreciable

*Nota***.** Prueba U de Mann-Whitney bilateral. Nivel de significancia α = 0.05. Magnitud del efecto (Cohen's d): Grande ≥ 0.80; Mediano 0.50–0.79; Pequeño 0.20–0.49; Despreciable < 0.20. Las variables están ordenadas por magnitud absoluta del efecto descendente.

La Tabla 12 presenta los resultados del análisis bivariado mediante la prueba U de Mann-Whitney, evidenciando diferencias estadísticamente significativas en siete de las nueve variables evaluadas ($p < 0.05$). No obstante, la magnitud del efecto (cuantificada mediante el coeficiente d de Cohen) varía considerablemente entre variables, lo que permite establecer una jerarquía de relevancia clínica que trasciende la mera significancia estadística y orienta la selección de predictores hacia aquellas variables con mayor capacidad discriminativa real.

- La Edad es la variable con mayor capacidad discriminativa individual (Cohen's d = 0.963), clasificada como efecto grande según los criterios de Cohen. Los pacientes diabéticos presentan una mediana de 69 años frente a los 53 años de los pacientes normales, una diferencia de 16 años estadísticamente sólida ($p < 0.001$). Este resultado confirma el rol epidemiológico preponderante de la Edad como factor de riesgo establecido para DM2 según la ADA y la IDF, y anticipa que será la variable con mayor peso en el modelo predictivo.
- Las cinco variables del perfil lipídico presentan efectos de magnitud pequeña (entre 0.236 y 0.384), todos estadísticamente significativos ($p < 0.001$). El patrón de diferencias es clínicamente coherente con la fisiopatología de la DM2.
- Colesterol HDL (d = -0.384): los pacientes diabéticos presentan HDL más bajo (45.20 vs 49.95 mg/dL), consistente con la dislipidemia diabética típica donde el HDL disminuye por la resistencia a la insulina.
- Colesterol LDL (d = -0.353): el LDL es menor en diabéticos (78.79 vs 96.96 mg/dL). Este hallazgo, aparentemente paradójico, es explicable por el uso de estatinas en pacientes ya diagnosticados o con factores de riesgo cardiovascular conocidos, lo que reduce artificialmente el LDL medido.
- Triglicéridos (d = 0.343): mayor en diabéticos (147.01 vs 118.85 mg/dL), reflejo del aumento de VLDL y partículas remanentes asociado a la resistencia insulínica.

- Índice Aterogénico (d = 0.308): más elevado en diabéticos (3.18 vs 2.38), marcador integrador del riesgo cardiometabólico global que captura la relación desfavorable CT/HDL.
- Colesterol Total (d = -0.236): paradójicamente menores en diabéticos (158.50 vs 173.75 mg/dL), este hallazgo, aunque aparentemente contraintuitivo, podría explicarse por el uso previo de tratamiento hipolipemiante, especialmente estatinas, en pacientes con diagnóstico conocido de diabetes o con mayor riesgo cardiovascular. No obstante, esta interpretación debe asumirse como hipótesis, dado que el dataset no incluye información farmacológica que permita confirmarla.
- El Ratio LDL-HDL alcanza significancia estadística (p = 0.001, d = -0.181), pero su magnitud de efecto despreciable indica que la diferencia observada entre grupos, aunque real a nivel poblacional, tiene escasa utilidad práctica para discriminar casos individuales. Este hallazgo anticipa su posible exclusión durante la selección formal de predictores.
- La Creatinina (p = 0.593, d = 0.064) y el Ratio TG-HDL (p = 0.479, d = 0.043) no muestran diferencias estadísticamente significativas entre grupos diabéticos y normales en el análisis bivariado. Ambas medianas son prácticamente idénticas (Creatinina: 0.81 vs 0.81 mg/dL; Ratio TG-HDL: 3.47 vs 3.37).

En síntesis, el análisis bivariado confirma que la Edad es el predictor individual más potente, seguida por el perfil lipídico con efectos pequeños pero consistentes. La Creatinina, pese a no mostrar diferencias entre grupos de forma aislada, fue incluida en el modelo por su reconocida asociación con la nefropatía diabética y su disponibilidad universal en sistemas de laboratorio.

Para evaluar la asociación entre el Sexo Biológico y la presencia de diabetes se aplicó la prueba chi-cuadrado. La Tabla 13 presenta la distribución cruzada de pacientes por sexo y clase diagnóstica, mientras que la Tabla 13a consolida los estadísticos resultantes de la prueba. La Figura 4 complementa este análisis mediante la representación visual de la prevalencia de diabetes por sexo biológico, permitiendo apreciar de forma directa la escasa diferencia proporcional entre grupos.

Tabla 13. Distribución cruzada Sexo y TARGET

Sexo	Normal	Diabetes	Total
Femenino (0)	6.333 (52.8%)	5.654 (47.2%)	11.987
Masculino (1)	3.239 (50.7%)	3.152 (49.3%)	6.391
Total	9.572	8.806	18.378

Tabla 13a. Resultados de la prueba Chi-cuadrado (Sexo y `TARGET`)

Estadístico	Valor
Chi-cuadrado (χ^2)	7.6464
p-value	0.005688
Significativo ($\alpha = 0.05$)	Sí
Cramér's V	0.0190
Tamaño del efecto	Pequeño (efecto despreciable en la práctica)
Validez de la prueba	Válida — todas las frecuencias esperadas ≥ 5
Chi-cuadrado (χ^2)	7.6464
p-value	0.005688

*Nota***.** Prueba Chi-cuadrado de Pearson bilateral. Cramér's V: efecto pequeño = 0.10–0.29; despreciable < 0.10.

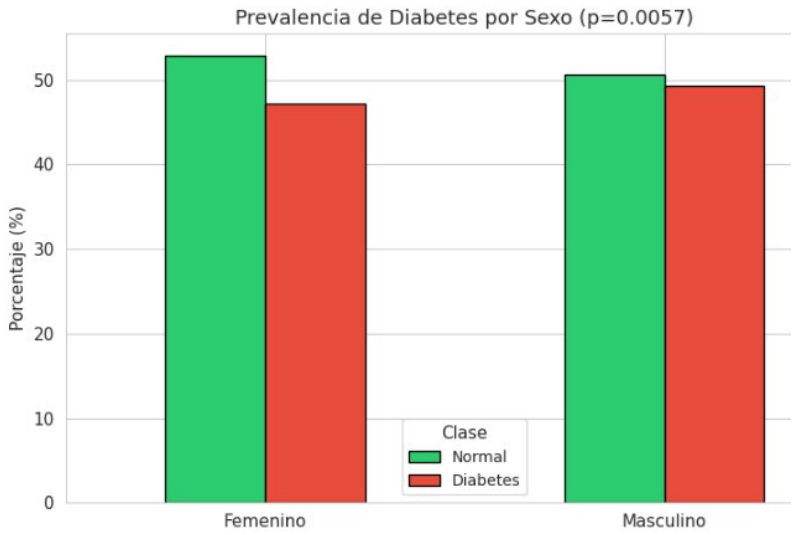


Figura 4. Prevalencia de Diabetes por Sexo Biológico.

Nota. Barras de porcentaje por categoría de sexo. Verde = Normal (`TARGET`=0); Rojo = Diabetes (`TARGET`=1).

El análisis bivariado de la variable categórica `Sexo Biológico` frente al `TARGET` revela una situación metodológicamente relevante: la asociación es estadísticamente significativa pero

prácticamente irrelevante en términos de magnitud. La prueba Chi-cuadrado confirma la existencia de una asociación real entre el `Sexo Biológico` y la presencia de diabetes (χ^2 = 7.6464, p = 0.006), resultado válido dado que todas las frecuencias esperadas superan el umbral de 5, lo que descarta que la distribución observada sea atribuible al azar.

Sin embargo, la magnitud de esta asociación es despreciable. El Cramér's V = 0.019 se ubica muy por debajo del umbral de efecto pequeño (V = 0.10), indicando que el `Sexo Biológico` explica menos del 0.04% de la varianza del `TARGET` (V^2 = 0.00036). Como se aprecia en la Figura 4, las proporciones de diabéticos y normales son casi idénticas en ambos sexos, con una diferencia de apenas 2.1 puntos porcentuales entre mujeres (47.2% de diabetes) y hombres (49.3% de diabetes).

A pesar de su bajo poder discriminativo individual, el `Sexo Biológico` fue retenido en ambos modelos por decisión clínica fundamentada, la ADA y la IDF lo reconocen como factor de riesgo modificador, y su exclusión comprometería la comparabilidad con estudios previos y la representatividad del modelo en escenarios reales de tamizaje.

La Tabla 14 y la Figura 5 presentan los hallazgos del análisis bivariado de la variable `lipidic_flags`, construida como conteo ordinal de criterios lipídicos alterados simultáneamente (valores 0–4). La prueba de Kruskal-Wallis evidencia una asociación estadísticamente significativa con el `TARGET` (p = 0.030); no obstante, el patrón de distribución revela una relación no lineal y clínicamente compleja que merece análisis detallado.

Tabla 14. Distribución de lipidic_flags y TARGET

`Lipidic flags`	**N**	**Normal (%)**	**Diabetes (%)**	**Interpretación clínica**
0 criterios alterados	44	29.5%	**70.5%**	Sin alteración lipídica
1 criterio alterado	388	38.6%	61.4%	Alteración lipídica leve
2 criterios alterados	523	33.9%	66.1%	Alteración lipídica moderada
3 criterios alterados	234	26.1%	73.9%	Alteración lipídica severa - mayor prevalencia de diabetes del grupo ordinal
4 criterios alterados	40	40.0%	60.0%	Dislipidemia total - tratamiento previo reduce la

				señal discriminativa
Kruskal-Wallis H	**p = 0.0296**	**Significativo: Sí (α = 0.05)**	**-**	**-**

*Nota***.** `lipidic_flags` = número de criterios lipídicos alterados simultáneamente (0–4): HDL bajo, LDL alto, `Triglicéridos` y Colesterol Total altos. N estimado sobre el subconjunto con panel lipídico completo (N total = 1.228). Prueba de Kruskal-Wallis para comparación de más de dos grupos.

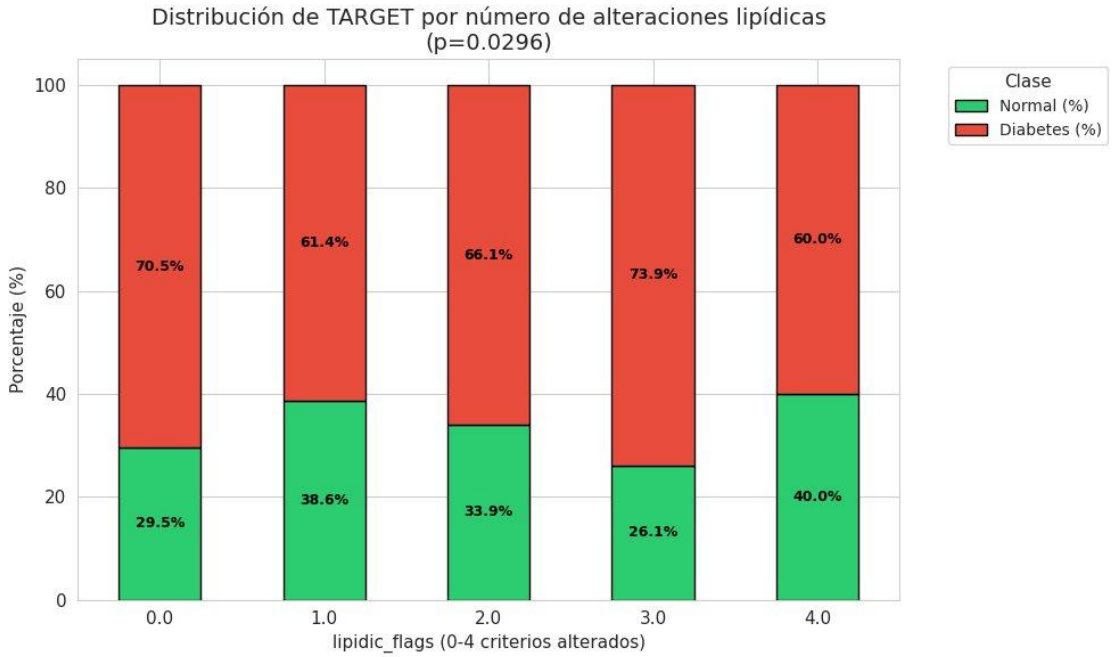


Figura 5. Distribución de TARGET por número de alteraciones lipídicas.

Nota. Barras apiladas por categoría de `lipidic_flags`. Verde = Normal (`TARGET`=0); Rojo = Diabetes (`TARGET`=1). p-value obtenido mediante prueba de Kruskal-Wallis (p = 0.0296).

El hallazgo más llamativo es la prevalencia de diabetes en el grupo de 0 criterios alterados (70.5%), la segunda más alta de todos los grupos y superior incluso a los grupos con 1 y 2 alteraciones lipídicas. Este resultado, aparentemente paradójico, tiene una explicación clínica coherente, los pacientes sin ninguna alteración lipídica detectable no son necesariamente los más sanos del conjunto, sino que en muchos casos corresponden a pacientes diabéticos bajo tratamiento farmacológico intensivo que han logrado normalizar su perfil lipídico. En este escenario, el tratamiento exitoso enmascara precisamente la señal que `lipidic_flags` intenta capturar,

invirtiendo el gradiente esperado entre ausencia de alteración lipídica y menor riesgo de diabetes.

El patrón más consistente con la fisiopatología esperada se observa en el grupo de 3 criterios alterados simultáneamente, que registra la mayor prevalencia de diabetes (73.9%) y la menor proporción de pacientes normales (26.1%). Este hallazgo confirma que la acumulación severa de dislipidemia múltiple sí se asocia con mayor riesgo de DM2, en concordancia con el concepto de síndrome metabólico y con la evidencia internacional que vincula la coexistencia de alteraciones lipídicas con resistencia a la insulina.

En consecuencia, `lipidic_flags` no puede tratarse como una variable con relación monótona con el `TARGET`. Su comportamiento no lineal (con el grupo de 0 alteraciones mostrando mayor prevalencia de diabetes que los grupos de 1 y 2), limita su utilidad como predictor lineal independiente, pero no invalida su inclusión en el modelo: los algoritmos basados en árboles de decisión (Random Forest, XGBoost) capturan naturalmente estas no linealidades sin transformaciones adicionales.

5.4 ANÁLISIS CORRELACIONAL Y MULTICOLINEALIDAD

El análisis correlacional examina las relaciones entre las variables predictoras, persiguiendo dos objetivos: identificar predictores que comparten información (lo cual orienta la selección final de variables) y detectar problemas de multicolinealidad que podrían afectar la interpretabilidad de modelos lineales. Dada la confirmada no normalidad de las variables, se utilizó el coeficiente de correlación de Spearman, que evalúa relaciones monótonas sin asumir distribución normal.

5.4.1 MATRIZ DE CORRELACIÓN DE SPEARMAN

La Figura 6 presenta la representación visual de la matriz de correlación de Spearman aplicada al conjunto de variables predictoras, cuyos patrones revelan relaciones tanto esperables como metodológicamente relevantes para el modelado. La Tabla 15 consolida los criterios de interpretación de la magnitud del coeficiente de correlación según las convenciones propuestas por Mukaka:

Tabla 15. Regla para interpretar el tamaño de un coeficiente de correlación

Tamaño de la correlación	Interpretación
0,90 a 1,00 (-0,90 a -1,00)	Correlación positiva (negativa) muy alta
0,70 a 0,90 (-0,70 a -0,90)	Alta correlación positiva (negativa)

0,50 a 0,70 (-0,50 a -0,70)	Correlación positiva (negativa) moderada
0,30 a 0,50 (-0,30 a -0,50)	Baja correlación positiva (negativa)
0,00 a 0,30 (0,00 a -0,30)	Correlación insignificante

*Nota***.** Tomado de [79]

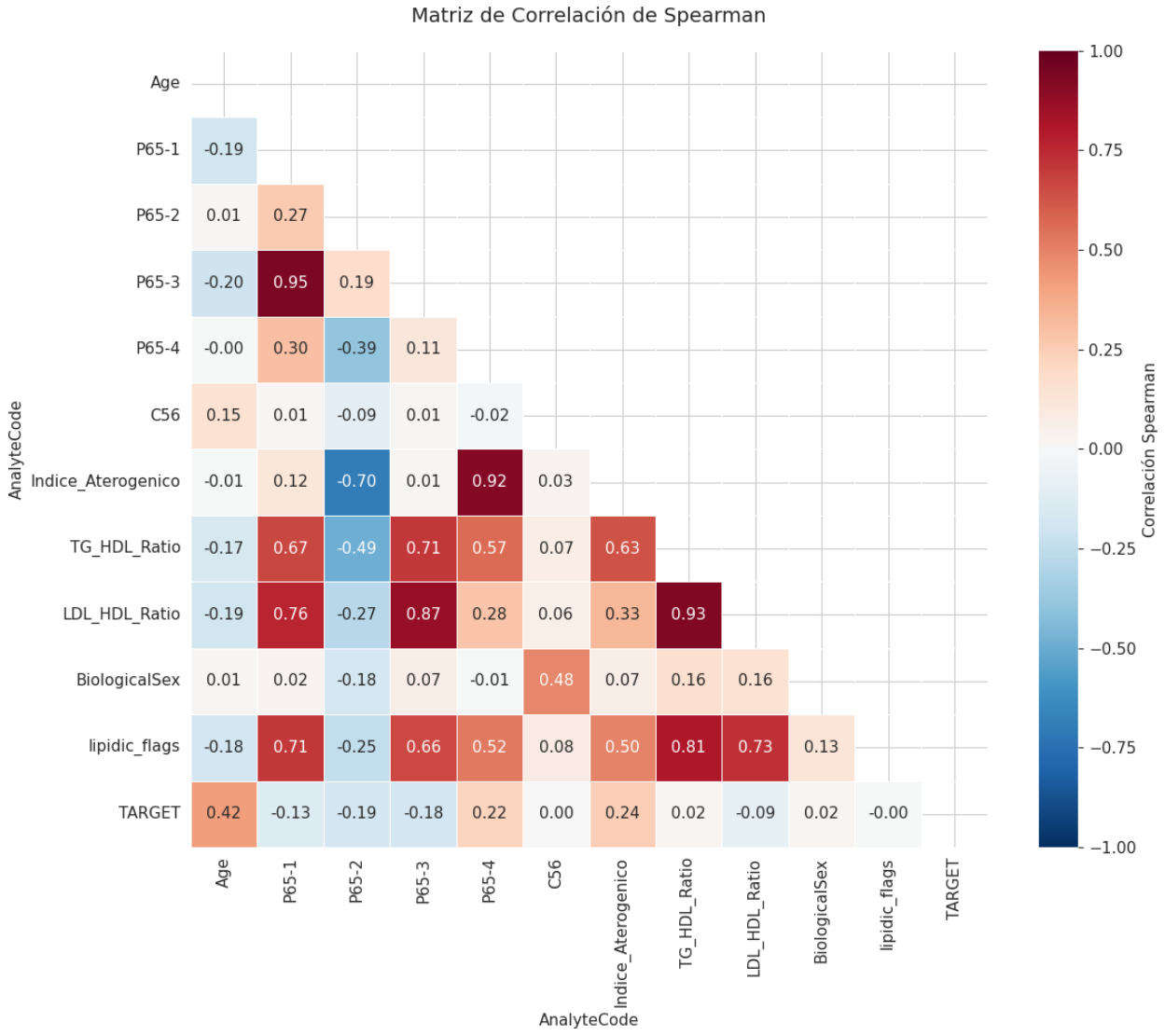


Figura 6. Matriz de correlación de Spearman entre variables.

Nota. Matriz triangular inferior. Escala de color: rojo = correlación positiva, azul = correlación negativa; la intensidad representa la magnitud. Solo se muestran las celdas con correlación distinta de la diagonal.

Como se observa en la Figura 6, el análisis de correlación de Spearman revela que la `Edad` es el predictor individual más fuertemente asociado con el `TARGET` (r = 0.42), seguida por las variables del perfil lipídico con correlaciones bajas (r entre 0.13 y 0.24), mientras que la `Creatinina` y `lipidic_flags` presentan correlaciones virtualmente nulas con la variable objetivo de forma individual.

Entre predictores, se identifican dos bloques de multicolinealidad con implicaciones directas sobre la selección de variables. El primero y más crítico agrupa las variables lipídicas brutas y sus ratios derivados, con correlaciones entre ellas de hasta r = 0.95 y VIF superiores a 10^6, configurando un escenario de redundancia extrema que haría inestables los coeficientes de cualquier modelo lineal. El segundo bloque involucra a `lipidic_flags` con ese mismo conjunto lipídico; sin embargo, su VIF de 3.69 (notablemente inferior al de sus componentes individuales) confirma su utilidad como variable de síntesis que captura la información del perfil lipídico sin heredar la redundancia que caracteriza a las variables que lo conforman.

5.4.2 ANÁLISIS DE MULTICOLINEALIDAD MEDIANTE VIF

El Factor de Inflación de la Varianza (VIF, por sus siglas en inglés) cuantifica el grado en que la varianza de los coeficientes de un modelo de regresión se infla como consecuencia de la colinealidad entre predictores. Valores de VIF superiores a 10 indican multicolinealidad problemática o grave que compromete la estabilidad e interpretación de los coeficientes estimados [80]. Como se observa en la Tabla 16, los valores VIF iniciales calculados sobre el conjunto completo de variables revelan una separación binaria y extrema entre dos grupos: cuatro variables con VIF cercano a la unidad (indicativo de ausencia de colinealidad) y siete variables lipídicas con valores que oscilan entre 7.74×10^5 y 1.31×10^7, superando en cuatro a siete órdenes de magnitud el umbral bibliográfico de referencia.

Tabla 16. Análisis VIF inicial sobre el conjunto de variables candidatas

Variable	VIF	Interpretación
`Edad` (`Age`)	1.08	Sin colinealidad problemática
`Creatinina` (C56)	1.14	Sin colinealidad problemática
Sexo (`BiologicalSex`)	1.19	Sin colinealidad problemática
`lipidic_flags`	3.69	Multicolinealidad moderada
`Colesterol HDL` (P65-2)	7.74×10^5	Multicolinealidad grave
`Colesterol Total` (P65-4)	1.32×10^6	Multicolinealidad grave

Índice Aterogénico	1.94×10^6	Multicolinealidad grave
Ratio LDL-HDL	6.11×10^6	Multicolinealidad grave
Colesterol LDL (P65-3)	9.87×10^6	Multicolinealidad grave
Ratio TG-HDL	1.01×10^7	Multicolinealidad grave
Triglicéridos (P65-1)	1.31×10^7	Multicolinealidad grave

Nota. VIF calculado sobre el conjunto completo de variables antes de la selección de predictores. VIF=1: no hay multicolinealidad, VIF 1-5: multicolinealidad moderada, VIF ≥ 5: alta multicolinealidad, VIF ≥ 10, multicolinealidad grave.

El análisis del VIF sobre el conjunto completo de variables confirma y cuantifica con precisión el patrón de multicolinealidad identificado en la matriz de correlación Spearman, revelando una separación binaria y extrema entre dos grupos de variables, como se observa en la Figura 7.

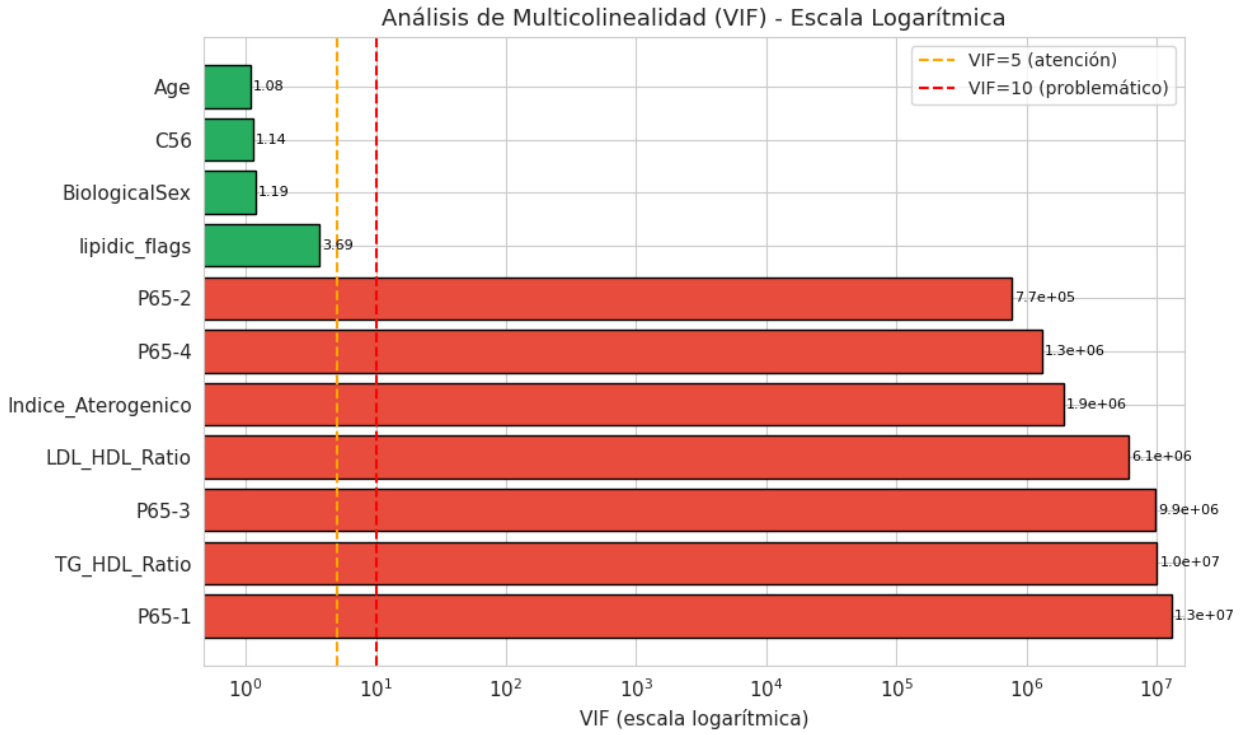


Figura 7. Análisis de Multicolinealidad (VIF) - Escala Logarítmica.

Nota. Escala logarítmica en el eje X. Línea naranja discontinua = VIF 5 (atención); línea roja discontinua = VIF 10 (problemático). Barras verdes = sin colinealidad; barras rojas = colinealidad alta.

Cuatro variables presentan VIF cercano a la unidad, confirmando ausencia de multicolinealidad

con el resto del conjunto:

- La `Edad` (VIF = 1.08), prácticamente ortogonal a todos los demás predictores
- La `Creatinina` (VIF = 1.14), cuya medición de función renal no correlaciona con el perfil lipídico en este *dataset*
- El `Sexo Biológico` (VIF = 1.19), sin colinealidad detectable respecto a ninguna otra variable
- `Lipidic_flags` (VIF = 3.69), que, aunque correlaciona con el perfil lipídico por construcción, mantiene su VIF muy por debajo del umbral problemático gracias a su función de síntesis: al consolidar cuatro variables lipídicas en un único conteo ordinal de criterios alterados, elimina la redundancia interna del bloque sin perder su información clínica.

En contraste, las siete variables restantes presentan VIF superiores a 10^5, lo que descarta su uso simultáneo en cualquier modelo lineal o basado en gradiente de descenso, dado que la matriz de diseño resultante estaría próxima a ser singular. Los algoritmos basados en particiones recursivas (Random Forest y XGBoost) son intrínsecamente insensibles a este problema, ya que en cada nodo seleccionan un subconjunto aleatorio de variables y toman decisiones de corte individuales que no se ven afectadas por la correlación entre predictores.

Este análisis fundamenta de forma definitiva la arquitectura de doble modelo adoptada: el Modelo A opera exclusivamente sobre las tres variables sin colinealidad (`Edad`, `Sexo Biológico` y `Creatinina`), garantizando estabilidad y cobertura universal; el Modelo B incorpora adicionalmente el bloque lipídico representado por `lipidic_flags` y las variables individuales con mayor aporte SHAP, prescindiendo de aquellas con VIF extremo que introducirían inestabilidad sin incrementar el poder predictivo.

Es importante precisar que el análisis VIF no fue utilizado como un criterio automático y aislado de eliminación de variables para todos los modelos, sino como una herramienta diagnóstica para identificar redundancia estadística entre predictores y orientar la interpretación del conjunto de variables. Aunque los modelos basados en árboles, como Random Forest y XGBoost, son menos sensibles a la multicolinealidad que los modelos lineales, la presencia de predictores altamente redundantes puede afectar la estabilidad interpretativa de la importancia de variables y dificultar la lectura clínica del modelo. Por esta razón, el VIF se empleó principalmente para depurar la selección de predictores, controlar la redundancia del bloque lipídico y justificar la construcción de modelos con conjuntos de variables diferenciados, sin asumir que toda variable con alta colinealidad debía eliminarse de manera automática.

6. EVALUACIÓN DEL RENDIMIENTO DE LOS MODELOS DE *MACHINE LEARNING*

Este capítulo presenta los resultados del proceso de modelado y evaluación de los modelos de aprendizaje automático desarrollados para el tamizaje de Diabetes mellitus tipo 2. Dado el contexto clínico del problema, donde las consecuencias de un error diagnóstico tienen implicaciones directas sobre la salud del paciente, la evaluación se estructura en cinco dimensiones complementarias: estrategia de validación y pruebas, métricas estadísticas de desempeño, comparación entre modelos, interpretabilidad del modelo y validación médica.

6.1 ESTRATEGIA DE VALIDACIÓN Y PRUEBAS

6.1.1 DISEÑO DE LA PARTICIÓN *TRAIN/TEST*

Dado que el *dataset* contiene múltiples registros por paciente, se aplicó una partición agrupada por identificador de paciente mediante `GroupShuffleSplit` (80% entrenamiento / 20% prueba), garantizando que ningún paciente aparece simultáneamente en ambos conjuntos y eliminando el riesgo de fuga de datos. El Modelo A comprende 10.920 registros en entrenamiento y 2.730 en prueba; el Modelo B, restringido al subconjunto con panel lipídico completo, cuenta con 899 registros en entrenamiento y 230 en prueba. Esta asimetría en el tamaño muestral tiene implicaciones directas sobre la estabilidad de los modelos y la amplitud de los intervalos de confianza, como se documenta en las secciones siguientes.

6.1.2 EVALUACIÓN INICIAL - COMPARATIVA DE LOS SEIS MODELOS ENTRENADOS

Previo a la selección del modelo final para cada escenario, se entrenaron y evaluaron tres algoritmos de clasificación por escenario (Regresión Logística (LR), Random Forest (RF) y XGBoost) totalizando seis modelos evaluados en el conjunto de prueba. La Tabla 17 consolida los resultados completos sobre el conjunto de prueba (umbral = 0.50) junto con los *baselines* de referencia.

Tabla 17. Resultados en conjunto de prueba

Escenario	Algoritmo	AUC-ROC	Sens.	Esp.	VPN	VPP	F1	Brier	TP/FP/FN/TN
Baseline	`DummyClassifier`	0.492	0.493	0.493	—	—	0.476	—	—
Modelo A	Regresión Logística	0.716	0.690	0.632	0.679	0.644	0.666	0.212	*925/512/415/878*

Modelo A	Random Forest	0.757	0.869	0.516	0.804	0.634	0.733	0.197	*1165/673/175/717*
Modelo A	**XGBoost**	**0.754**	**0.845**	**0.548**	**0.786**	**0.643**	**0.731**	**0.197**	***1133/629/207/761***
Baseline	`DummyClassifier`	0.521	0.600	0.600	—	—	0.716	—	—
Modelo B	Regresión Logística	0.650	0.679	0.523	0.391	0.783	0.727	0.228	*112/31/53/34*
Modelo B	XGBoost	0.681	0.648	0.569	0.390	0.793	0.713	0.223	*107/26/58/39*
Modelo B	**Random Forest (regulado)**	**0.701**	**0.806**	**0.477**	**0.492**	**0.809**	**0.786**	**0.209**	***126/30/39/35***

Nota. Sens. = Sensibilidad; Esp. = Especificidad; VPN = Valor Predictivo Negativo; VPP = Valor Predictivo Positivo. TP/FP/FN/TN = conteos absolutos en el conjunto de prueba. Modelo A: N test = 2.730. Modelo B: N test = 230.

Los resultados permiten establecer varias conclusiones comparativas antes de la selección final:

- Superioridad de modelos no lineales en Modelo A: XGBoost (AUC=0.754) y Random Forest (AUC=0.757) superan ampliamente a la Regresión Logística (AUC=0.716), con una diferencia de 0.038-0.041 puntos AUC que resultó estadísticamente significativa ($p<0.001$ en comparación *bootstrap* con corrección Benjamini-Hochberg). La LR mejora el *baseline* en 0.224 puntos AUC pero sacrifica sensibilidad (0.690 vs 0.845–0.869 de los modelos de árbol), penalizando la capacidad de detección que es crítica en el contexto de tamizaje.
- Equivalencia estadística XGBoost vs Random Forest en Modelo A: la diferencia de AUC entre XGBoost (0.754) y Random Forest (0.757) es de apenas 0.002 puntos, estadísticamente no significativa ($p=0.481$). La selección de XGBoost como modelo final se fundamentó en criterios secundarios: menor *gap* de aprendizaje (0.022 vs mayor en RF a umbral 0.5) y mayor estabilidad entre semillas ($\sigma=0.012$ vs σ ligeramente mayor en RF).
- Perfil diferencial de errores entre algoritmos en Modelo A: Random Forest maximiza la sensibilidad (0.869) a costa de la especificidad más baja (0.516), generando 673 falsos positivos. XGBoost ofrece un balance más conservador (Sen=0.845, Esp=0.548) con 629 falsos positivos, perfil más adecuado para tamizaje primario donde los falsos positivos

serán confirmados mediante criterios ADA. La Regresión Logística presenta el mejor balance Sen/Esp (0.690/0.632) pero el AUC más bajo, indicando que su frontera de decisión lineal no captura la complejidad del problema.

- Modelo B con convergencia hacia Random Forest: en el escenario lipídico los tres algoritmos presentan AUC más cercanos entre sí (0.650-0.701), sin diferencias estadísticamente significativas tras corrección Benjamini-Hochberg, atribuible al menor tamaño muestral de prueba (n=230). Random Forest regulado (AUC=0.701) lidera numéricamente y presenta el VPP más alto del conjunto (0.809), mientras que la Regresión Logística muestra el VPN más bajo (0.391), limitando severamente la seguridad del resultado negativo. XGBoost ocupa una posición intermedia (AUC=0.681) con métricas inferiores a Random Forest en las dimensiones clínicamente prioritarias.
- Impacto de la regularización en el Modelo B: la versión regulada de Random Forest (`max_depth`=5, `min_samples_leaf`=5) redujo el *gap* de *overfitting* de 0.162 a 0.146 con una pérdida de AUC de solo 0.002 en validación cruzada (0.702 → 0.700), considerada estadísticamente irrelevante. Esta regularización es la razón por la que el AUC del mejor Modelo B (0.701) es ligeramente inferior al AUC CV reportado en la Tabla de hiperparámetros (0.702): el *gap* residual de 0.146 confirma que la limitación es estructural al tamaño muestral y no corregible con regularización adicional.

A partir de estos resultados, se seleccionaron XGBoost como Modelo A final y Random Forest regulado como Modelo B final. Las secciones siguientes presentan la evaluación detallada de estos dos modelos seleccionados, incluyendo intervalos de confianza al 95% mediante *bootstrap*, análisis de calibración, *Decision Curve Analysis* e interpretabilidad mediante SHAP.

6.1.3 ANÁLISIS DE ESTABILIDAD CON MÚLTIPLES SEMILLAS

Para evaluar la robustez de los modelos ante variaciones en la inicialización aleatoria, ambos fueron entrenados con 10 semillas distintas, permitiendo cuantificar la variabilidad del AUC debida al azar computacional y no a las características del modelo. La Figura 8 muestra la distribución de AUC resultante para cada algoritmo y escenario.

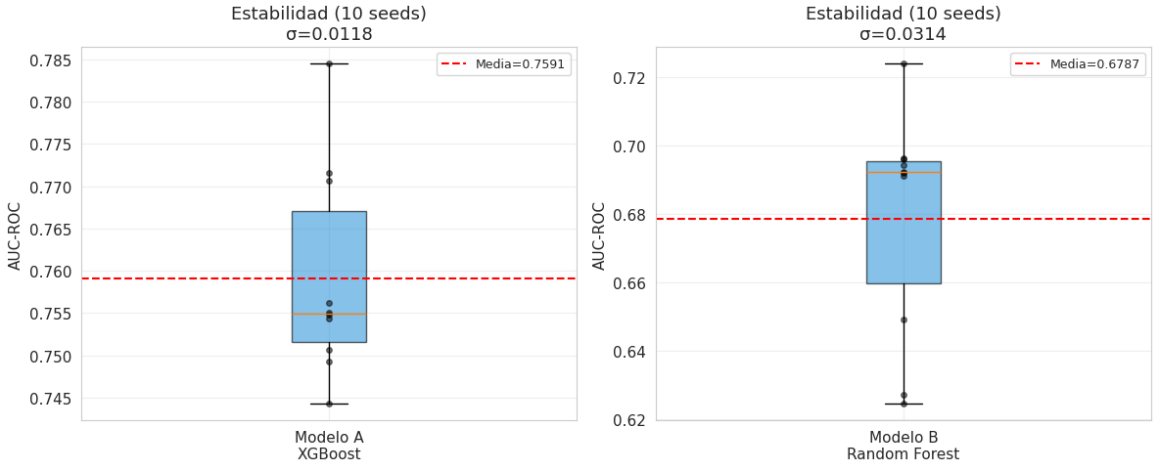


Figura 8. Análisis de estabilidad.

Nota. Cajas y bigotes de AUC obtenido con n=10 semillas distintas para Modelo A (XGBoost) y Modelo B (Random Forest regulado).

El Modelo A (XGBoost) obtuvo AUC = 0.759 ± 0.012 con un rango de [0.744-0.784], evidenciando alta estabilidad y reproducibilidad (σ < 0.02). El Modelo B (Random Forest regulado) presentó AUC = 0.679 ± 0.031 con un rango de [0.625-0.724], con mayor dispersión atribuible al tamaño muestral reducido (n ≈ 900), aunque dentro de un margen aceptable (σ < 0.05).

6.1.4 CURVAS DE APRENDIZAJE Y ANÁLISIS DE *OVERFITTING*

Las curvas de aprendizaje se calcularon mediante validación cruzada estratificada y agrupada por paciente (`StratifiedGroupKFold`), evaluando el AUC de entrenamiento y validación en función del tamaño muestral. Las Figuras 9 y 10 muestran los resultados antes y después de aplicar regularización explícita al Modelo B.

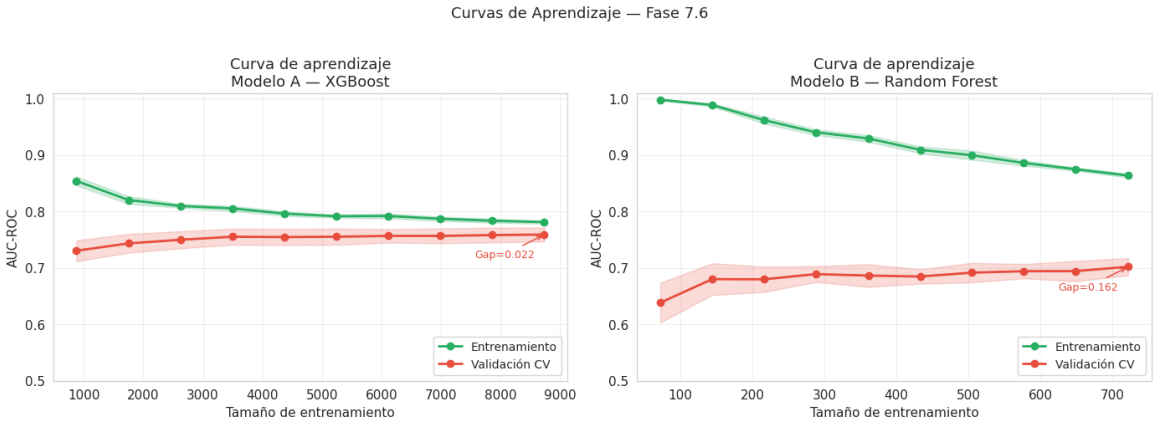


Figura 9. Curvas de aprendizaje - Modelos A y B.

Nota. AUC de entrenamiento (línea sólida) y validación (línea discontinua) en función del número de registros de entrenamiento. *Gap* = diferencia entre curvas en el extremo derecho.

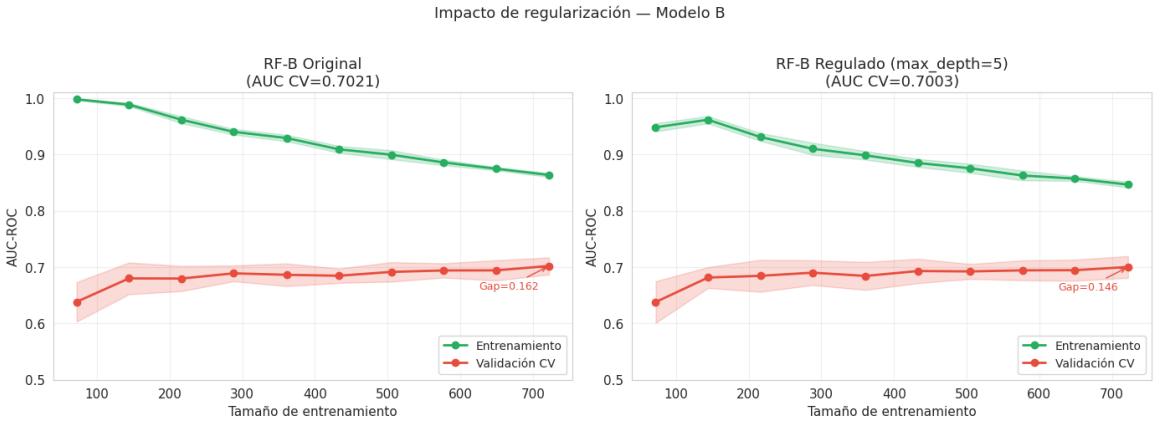


Figura 10. Curvas de aprendizaje - Modelo B tras regularización

Nota. Comparación de curvas de aprendizaje del Modelo B antes y después de regularización explícita.

Para el Modelo A (XGBoost), el *gap train*-validación es de 0.022, por debajo del umbral aceptable de 0.10, confirmando ausencia de *overfitting*. El AUC de validación se estabiliza en torno a 0.77, nivel esperable dado que el modelo opera con únicamente tres *features* (`Edad`, `Sexo Biológico` y `Creatinina`); con este conjunto reducido de variables, este nivel de discriminación representa el techo natural del modelo en el *dataset* actual. La curva de validación mantiene una pendiente positiva leve al alcanzar los 9.000 registros de entrenamiento, indicando que el modelo no ha llegado a su límite de rendimiento y se beneficiaría de datos adicionales en el futuro.

Para el Modelo B (Random Forest regulado), el *gap* inicial de 0.162: originado por la alta varianza asociada al tamaño muestral reducido (n ≈ 900 total) y la profundidad libre de los árboles, se redujo a 0.146 tras imponer restricciones explícitas de regularización (`max_depth`=5, `min_samples_leaf`=5), con una pérdida de AUC de solo 0.002 en validación cruzada, estadísticamente irrelevante. La banda de varianza amplia en la curva de validación, especialmente en los *folds* iniciales, confirma la inestabilidad estructural derivada del tamaño muestral: con aproximadamente 700 registros en entrenamiento y múltiples *features*, los árboles de decisión sin restricciones de profundidad tienen suficientes grados de libertad para memorizar el conjunto de entrenamiento. El *gap* residual de 0.146 supera el umbral problemático de 0.10 y no es corregible con regularización adicional, constituyendo una limitación estructural inherente al tamaño del subconjunto lipídico.

En consecuencia, el Modelo B debe interpretarse como un modelo complementario de valor exploratorio y no como una herramienta de tamizaje primario independiente. Aunque incorpora información bioquímica adicional del perfil lipídico y alcanza un VPP elevado, su tamaño muestral reducido, la amplitud de sus intervalos de confianza y el gap de aprendizaje residual limitan la estabilidad de sus estimaciones. Por tanto, sus resultados deben ser considerados preliminares y requieren confirmación mediante una cohorte externa con mayor número de pacientes con panel lipídico completo.

6.2 MÉTRICAS ESTADÍSTICAS DE DESEMPEÑO

La Tabla 18 consolida las métricas de desempeño de los modelos seleccionados sobre el conjunto de prueba reservado, con intervalos de confianza al 95% obtenidos mediante *bootstrap* (n = 1.000 remuestreos) y umbral de clasificación fijo de 0.50.

Tabla 18. Métricas de desempeño en conjunto de prueba con IC 95% *Bootstrap*

Métrica	**A - XGBoost (mejor A)**	**A - RF**	**B - RF regulado (mejor B)**	***Baseline* A**	***Baseline* B**	**Métrica**
AUC-ROC	0.754 [0.736-0.772]	0.756 [0.739-0.774]	0.701 [0.621-0.776]	0.492	0.521	AUC-ROC

Sensibilidad	0.845 [0.826-0.864]	0.869 [0.850–0.887]	0.806 [0.699-0.828]	0.493	0.600	Sensibilidad
Especificidad	0.548 [0.521-0.573]	0.516 [0.491-0.543]	0.477 [0.410-0.661]	0.493	0.600	Especificidad
VPN	0.786 [0.759–0.811]	0.804 [0.776-0.830]	0.492 [0.360-0.600]	—	—	VPN
VPP	0.643 [0.621-0.665]	0.634 [0.612-0.655]	0.809 [0.747-0.865]	—	—	VPP
F1-score	0.730 [0.712-0.747]	0.733 [0.716-0.750]	0.786 [0.732-0.834]	0.476	0.716	F1-score
Brier Score	0.197 [0.190-0.203]	0.197 [0.191-0.203]	0.209 [0.194-0.226]	—	—	*Brier Score*
Gap aprendizaje	0.022	—	0.146	—	—	*Gap* aprendizaje
N prueba	2.730	2.730	230	2.730	230	N prueba

Nota. IC 95% calculado mediante *bootstrap* n=1.000. Umbral = 0.50. Modelo A: N test=2.730. Modelo B: N test=230. *Baseline* = `DummyClassifier` estratificado.

El Modelo A alcanza un AUC-ROC de 0.754 [IC95%: 0.736-0.772], superando el *baseline* en 0.262 puntos. En el contexto del tamizaje primario, las métricas más relevantes son la sensibilidad y el VPN. La sensibilidad de 0.845 [0.826-0.864] indica que el modelo detecta correctamente el 84.5% de los pacientes diabéticos reales. En términos absolutos, de los 1.340 diabéticos presentes en el conjunto de prueba, 1.133 son correctamente clasificados (207 falsos negativos). El VPN de 0.786 [0.759-0.811] garantiza que un resultado negativo descarta la enfermedad con una probabilidad del 78.6%, parámetro crítico en tamizaje: un resultado negativo suficientemente seguro evita derivaciones innecesarias al nivel diagnóstico.

La especificidad de 0.548 [0.521-0.573] es moderada, lo que implica que el 45.2% de los pacientes sanos son clasificados como positivos (629 falsos positivos). Esta especificidad moderada es aceptable y esperada en un modelo de tamizaje primario que prioriza no perder casos reales sobre evitar falsos positivos, los cuales serán confirmados o descartados mediante criterios ADA

en el nivel diagnóstico. El *Brier Score* de 0.197 [0.190-0.203], por debajo del umbral de referencia de 0.20, confirma una calibración adecuada de las probabilidades predichas.

El Modelo B opera sobre el subconjunto de 1.228 pacientes con panel lipídico completo (N test = 230) y presenta un perfil de métricas cualitativamente distinto al Modelo A. El AUC-ROC de 0.701 [IC95%: 0.621-0.776] supera el *baseline* en 0.180 puntos, aunque el intervalo de confianza considerablemente más amplio que el del Modelo A refleja la menor precisión del estimador con n = 230. La sensibilidad de 0.806 [0.699–0.828] es ligeramente inferior a la del Modelo A, mientras que el VPP de 0.809 [0.747–0.865] constituye la métrica más destacada del Modelo B, indicando que nueve de cada diez resultados positivos corresponden a diabéticos reales en este subconjunto.

El VPN de 0.492 [0.360-0.600] es el punto débil más crítico del Modelo B: en el subconjunto lipídico, con una prevalencia de diabetes del 71.7% en el conjunto de prueba, un resultado negativo no descarta la enfermedad con la seguridad suficiente para prescindir de confirmación diagnóstica. Este comportamiento es esperado en una subpoblación con alta prevalencia de DM2 y perfil metabólico alterado, y refuerza el rol del Modelo B como complemento confirmatorio del Modelo A, no como herramienta de descarte independiente. El *Brier Score* de 0.209 [0.194-0.226], ligeramente por encima del umbral de 0.20, es consistente con el *overfitting* estructural mencionado anteriormente.

La principal limitación operativa del Modelo B es su bajo valor predictivo negativo, lo cual implica que un resultado negativo no permite descartar con seguridad la presencia de DM2. Por esta razón, el Modelo B no debe utilizarse como herramienta de descarte diagnóstico, sino como un complemento cuando se dispone de perfil lipídico completo, especialmente útil para aumentar la confianza en resultados positivos. Esta interpretación es coherente con su mayor VPP, pero también con su menor estabilidad estadística debido al tamaño reducido del subconjunto lipídico.

6.2.1 CURVAS ROC Y PRECISIÓN-*RECALL*

Como se observa en la Figura 11, las curvas ROC y Precisión-*Recall* permiten comparar el comportamiento de los tres algoritmos evaluados en cada escenario a lo largo de todo el espacio de umbrales, complementando los valores puntuales reportados en la Tabla 18.

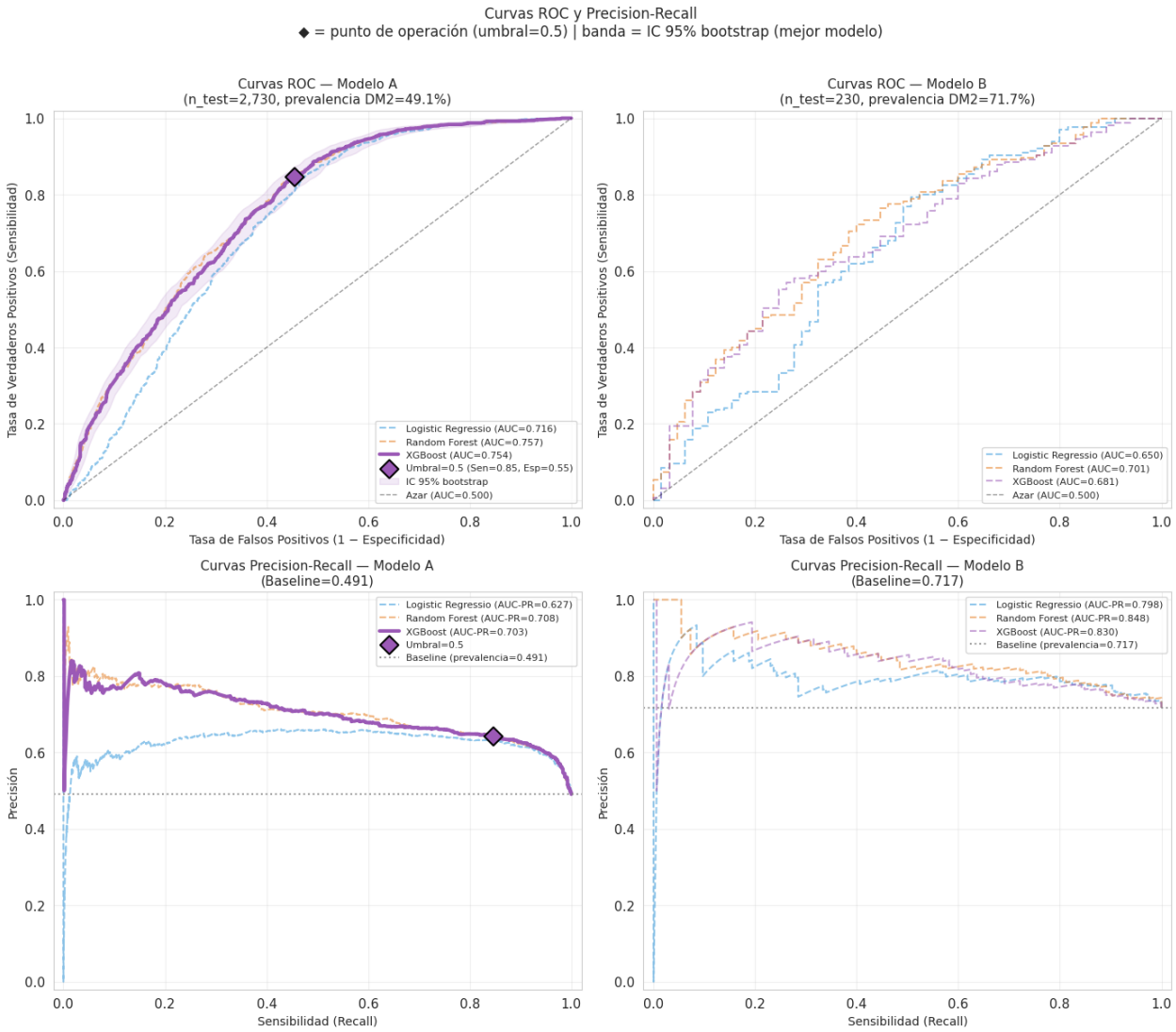


Figura 11. Curvas ROC y Precisión-*Recall* - comparativa de algoritmos.

Nota. Panel superior: Curvas ROC (AUC-ROC). Panel inferior: Curvas Precisión-*Recall* (AUC-PR). Izquierda: Modelo A (N test=2.730). Derecha: Modelo B (N test=230). LR=*Logistic Regression*, RF=Random Forest, XGB=XGBoost.

En el Modelo A, las curvas ROC de XGBoost y Random Forest son prácticamente superpuestas en todo el rango de tasas de falsos positivos. Ambas se separan consistentemente de la Regresión Logística, que muestra una curva inferior en todo el espacio de umbrales, evidenciando que su frontera de decisión lineal es insuficiente para capturar la complejidad de las relaciones entre las variables predictoras y el `TARGET`. En el Modelo B, las tres curvas ROC son notablemente más próximas entre sí, lo que refleja la mayor incertidumbre asociada al tamaño muestral reducido (N test = 230) y es consistente con la ausencia de diferencias estadísticamente significativas entre

algoritmos documentada en la comparación *bootstrap* con corrección Benjamini-Hochberg.

Las curvas de Precisión-*Recall* aportan información complementaria especialmente relevante para el Modelo B, donde la prevalencia de diabetes en el conjunto de prueba alcanza el 71.7%. En este escenario de alta prevalencia, el AUC-PR del Modelo B (0.848 para Random Forest regulado) es sustancialmente más alto que el del Modelo A (0.703 para XGBoost), no porque el Modelo B sea intrínsecamente superior, sino porque en poblaciones con mayor prevalencia de la clase positiva la curva PR se desplaza naturalmente hacia valores más altos. Esta distinción es relevante porque comparar AUC-PR entre modelos entrenados en subpoblaciones con distintas prevalencias no permite concluir superioridad directa de un escenario sobre el otro.

6.2.2 MATRICES DE CONFUSIÓN

Como se observa en la Figura 12, las matrices de confusión permiten traducir las métricas de desempeño a conteos absolutos de pacientes, facilitando la interpretación clínica directa de los errores de clasificación.

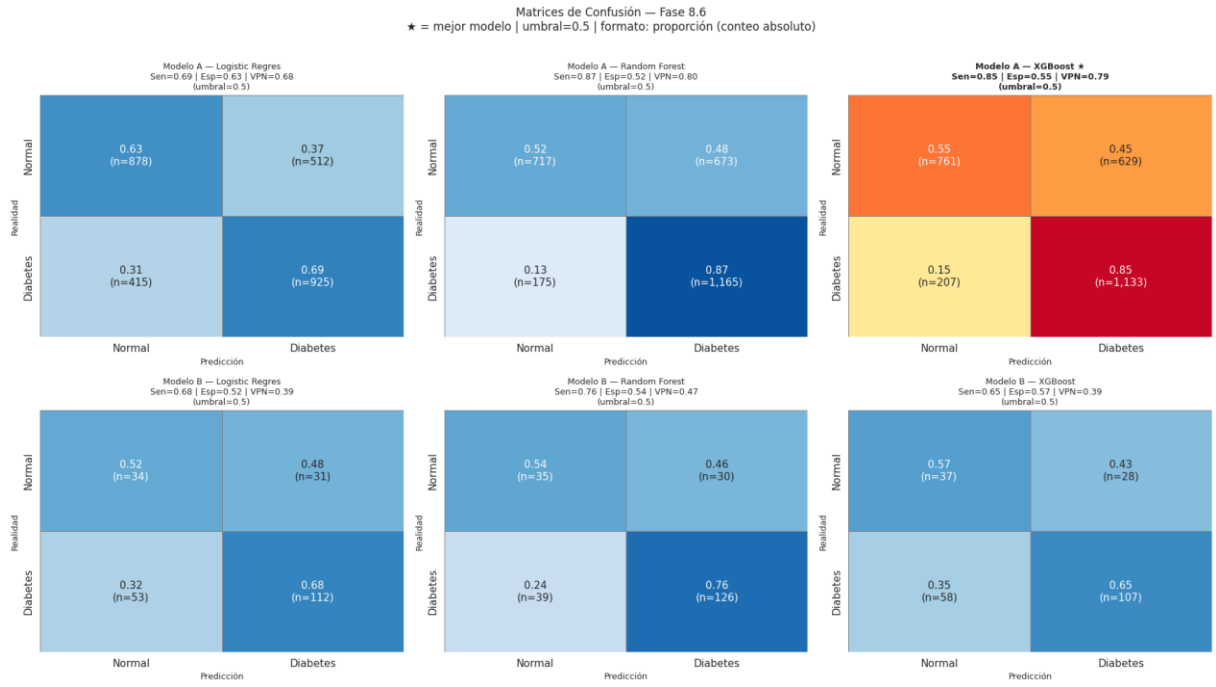


Figura 12. Matrices de confusión - Modelo A (XGBoost) y Modelo B (RF regulado)

Nota. Conteos absolutos de Verdaderos Positivos (TP), Falsos Positivos (FP), Falsos Negativos (FN) y Verdaderos Negativos (TN). Umbral = 0.50. Modelo A: N test=2.730. Modelo B: N test=230.

Para el Modelo A (XGBoost), sobre un conjunto de prueba de 2.730 pacientes: TP = 1.133, FP = 629, FN = 207 y TN = 761. El volumen de falsos positivos (629 pacientes sanos clasificados como diabéticos) es la consecuencia esperada de priorizar la sensibilidad en un modelo de tamizaje primario, estos 629 pacientes serán derivados a confirmación diagnóstica mediante criterios ADA, donde la mayoría será correctamente descartada sin intervención terapéutica. Los 207 falsos negativos representan un hallazgo relevante, siendo diabéticos no detectados que permanecerán sin diagnóstico, y su perfil demográfico (`Edad` media de 51 años) identifica a los pacientes jóvenes con DM2 de inicio temprano como el principal punto ciego del Modelo A.

Para el Modelo B (Random Forest regulado), sobre un conjunto de prueba de 230 pacientes: TP = 126, FP = 30, FN = 39 y TN = 35. El reducido número de falsos positivos (30) se refleja directamente en el VPP de 0.809, el más alto de todos los modelos evaluados en el subconjunto lipídico, nueve de cada diez resultados positivos corresponden a diabéticos reales, lo que otorga al Modelo B una alta precisión en la confirmación del diagnóstico cuando se dispone del perfil lipídico completo. Sin embargo, los 39 falsos negativos sobre apenas 165 diabéticos reales en el conjunto de prueba representan una tasa de error del 23.6%, considerablemente superior a la del Modelo A (15.4%), y son consistentes con el VPN bajo de 0.492 documentado en la Tabla 18. Este patrón refleja el punto ciego estructural del Modelo B, pacientes con DM2 cuyo perfil lipídico ha sido farmacológicamente normalizado no generan la señal lipídica que el modelo utiliza para la detección.

6.2.3 CALIBRACIÓN DE PROBABILIDADES

Como se aprecia en la Figura 13, los diagramas de fiabilidad *(reliability diagrams*) permiten evaluar la correspondencia entre las probabilidades predichas por cada modelo y las frecuencias reales de diabetes observadas en el conjunto de prueba, donde una curva perfectamente calibrada coincidiría con la diagonal.

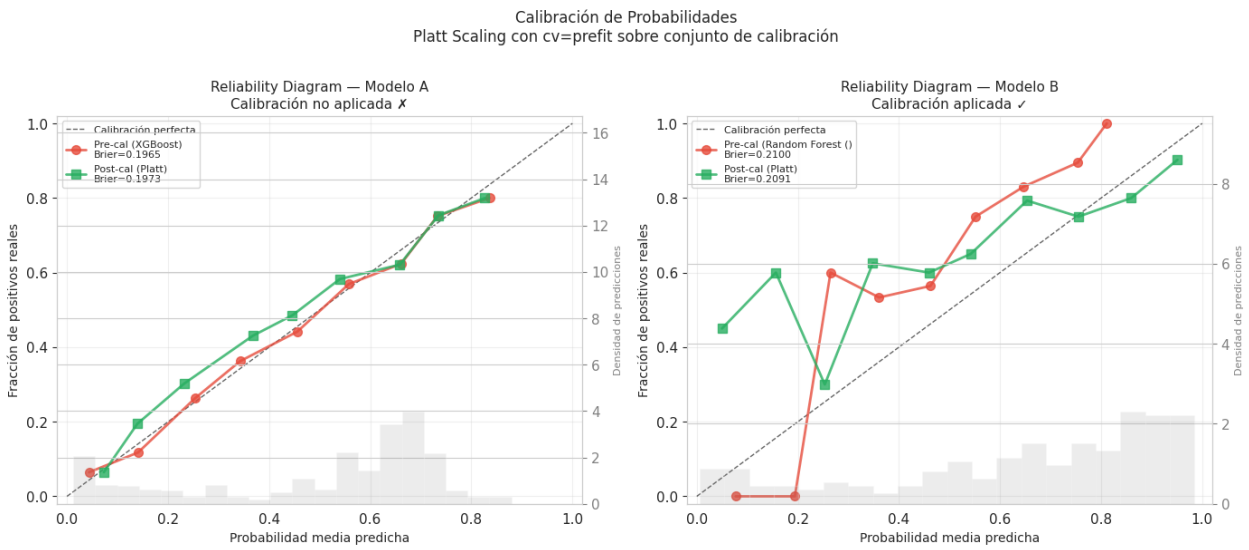


Figura 13. Diagramas de fiabilidad.

Nota. Probabilidad predicha (eje X) vs fracción de positivos observados (eje Y). La línea diagonal perfecta indica calibración perfecta. Rojo = pre-calibración; verde = post-calibración.

Para el Modelo A (XGBoost), la calibración mediante Platt Scaling produjo una mejora consistente en todo el rango de probabilidades, con la curva post-calibración siguiendo la diagonal de referencia de forma más ajustada que la curva pre-calibración. La mejora en el *Brier Score* fue marginal (0.0009), lo que indica que el modelo base ya presentaba una calibración razonablemente buena antes de la corrección, y que el Platt Scaling afinó principalmente el comportamiento en los rangos extremos de probabilidad.

Para el Modelo B (Random Forest regulado), la situación es cualitativamente distinta. La curva pre-calibración muestra un comportamiento errático en el rango de probabilidades bajas (0.0-0.3), característico del *overfitting* estructural documentado en las curvas de aprendizaje, el modelo asigna probabilidades muy bajas a pacientes que en realidad son diabéticos, generando los falsos negativos identificados anteriormente. La calibración Platt Scaling mejora parcialmente este comportamiento, pero la curva post-calibración sigue siendo irregular en ese rango, confirmando que la raíz del problema es estructural al tamaño muestral reducido y no corregible exclusivamente mediante técnicas de calibración. Esta limitación tiene una implicación clínica directa, las probabilidades predichas por el Modelo B en el rango bajo (< 0.30) deben interpretarse como estimaciones de riesgo individual, siendo el modelo más confiable para la clasificación binaria positivo/negativo que para la estratificación de riesgo continua basada en probabilidades absolutas.

6.3 COMPARACIÓN DE MODELOS

6.3.1 COMPARACIÓN ESTADÍSTICA ENTRE ALGORITMOS

La Tabla 19 presenta los resultados de la comparación estadística entre pares de algoritmos dentro de cada escenario, realizada mediante *bootstrap* de diferencias de AUC (n = 2.000 remuestreos) con corrección múltiple de Benjamini-Hochberg (FDR) para controlar el error tipo I en comparaciones múltiples.

Tabla 19. Comparación estadística de AUC entre algoritmos

Comparación	AUC 1	AUC 2	Δ AUC	p corr.	Conclusión	Comparación
[A] LR vs RF	0.716	0.756	−0.041	**<0.001*** **	RF significativamente superior	[A] LR vs RF
[A] LR vs XGBoost	0.716	0.754	−0.038	**<0.001*** **	XGBoost significativamente superior	[A] LR vs XGBoost
[A] RF vs XGBoost	0.756	0.754	+0.002	0.481 ns	Rendimiento equivalente	[A] RF vs XGBoost
[B] LR vs RF	0.650	0.701	−0.051	0.095 ns	Tendencia RF>LR; no sig. tras corrección BH	[B] LR vs RF
[B] LR vs XGBoost	0.650	0.681	−0.031	0.272 ns	Sin diferencia significativa	[B] LR vs XGBoost
[B] RF vs XGBoost	0.701	0.681	+0.020	0.121 ns	Sin diferencia significativa	[B] RF vs XGBoost
Mejor A vs Mejor B (n=53)	0.654	0.716	−0.061	0.519 ns	Tendencia B>A; potencia insuficiente	Mejor A vs Mejor B (n=53)

Nota. *Bootstrap* n=2.000. Corrección múltiple Benjamini-Hochberg (FDR). *** p<0.001; ns = no significativo. Comparación A vs B sobre n=53 registros comunes en test.

En el Modelo A, tanto XGBoost como Random Forest superan significativamente a la Regresión Logística (ΔAUC = -0.038 y -0.041 respectivamente, p < 0.001 en ambos casos), confirmando que

la complejidad no lineal de los modelos de árbol captura patrones que la frontera de decisión lineal de la LR no puede representar. Sin embargo, XGBoost y Random Forest no presentan diferencia estadísticamente significativa entre sí (ΔAUC = 0.002, p = 0.481), lo que implica que la selección de XGBoost como modelo final no se sustenta en superioridad estadística directa sino en criterios secundarios, menor *gap* de aprendizaje (0.022 vs valores superiores en RF) y mayor estabilidad entre semillas (σ = 0.012).

En el Modelo B, ninguna de las tres comparaciones alcanza significancia estadística tras la corrección Benjamini-Hochberg, con p-values que oscilan entre 0.095 y 0.272. Este resultado no implica equivalencia real entre los algoritmos, sino potencia estadística insuficiente con n = 230 en el conjunto de prueba para detectar diferencias del orden observado (ΔAUC entre 0.020 y 0.051). Random Forest regulado mantiene la ventaja numérica en todas las comparaciones y es seleccionado como Modelo B final por su mejor perfil en las métricas clínicamente prioritarias (AUC = 0.701, VPP = 0.809).

Respecto a la comparación directa entre el Modelo A y el Modelo B, realizada sobre los 53 registros presentes simultáneamente en ambos conjuntos de prueba, el Modelo B muestra una ventaja numérica de ΔAUC = -0.061 (0.716 vs 0.654), pero esta diferencia no alcanza significancia estadística (p = 0.519). La potencia de esta prueba es de aproximadamente 9.4%, requiriéndose al menos 191 registros comunes para alcanzar una potencia de 0.80. La evidencia disponible muestra una tendencia favorable al Modelo B en el subconjunto con perfil lipídico completo, pero el tamaño muestral es insuficiente para sostener esta afirmación estadísticamente.

6.3.2 *DECISION CURVE ANALYSIS* — UTILIDAD CLÍNICA NETA

Como se observa en la Figura 14, el *Decision Curve Analysis* evalúa la utilidad clínica neta (NB, por sus siglas en inglés) de ambos modelos en función del umbral de probabilidad de decisión clínica, comparándolos contra las estrategias de referencia de tratar a todos los pacientes o no tratar a ninguno.

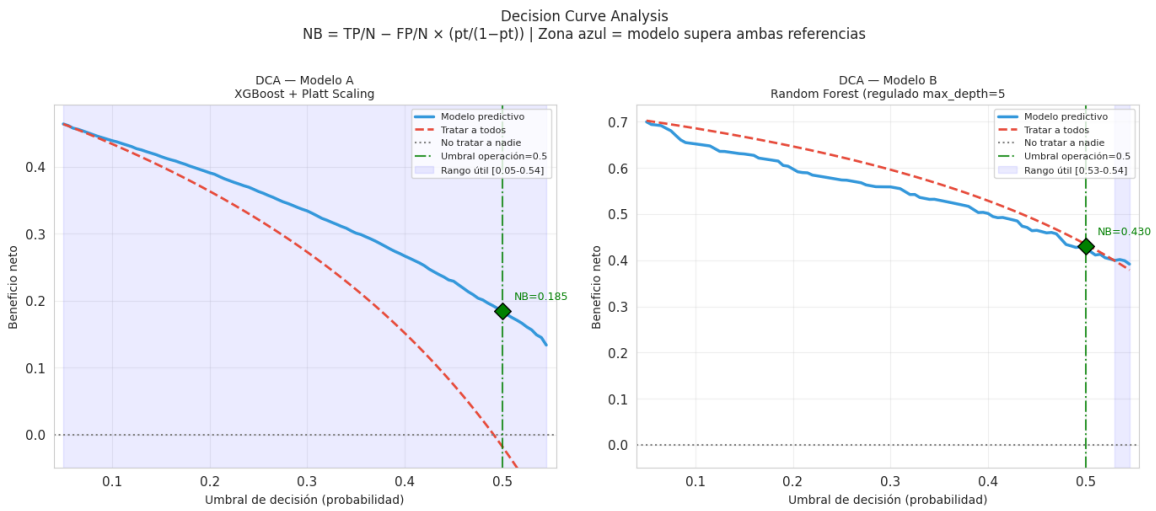


Figura 14. ***Decision Curve Analysis*** **(DCA)**

Nota. Beneficio neto en función del umbral de probabilidad de decisión clínica. Línea gris = 'tratar a todos'; línea horizontal = 'no tratar a nadie'. Área de utilidad clínica neta positiva indica que el modelo aporta valor sobre las referencias.

El Modelo A presenta un perfil de utilidad clínica sólido y sostenido, aporta beneficio neto superior a ambas referencias en prácticamente todo el rango clínicamente relevante (umbrales 0.05-0.54). En el punto de operación seleccionado (umbral = 0.50), el NB = 0.185 indica que, por cada 1.000 pacientes tamizados, el modelo identifica 185 diabéticos verdaderos adicionales netos sobre la estrategia de no intervención. Este resultado es clínicamente significativo, publicable como evidencia de utilidad práctica, y es robusto dado que se mantiene positivo en un rango amplio de umbrales de decisión, lo que implica que el modelo conserva su utilidad clínica incluso si el médico tratante ajusta el umbral de derivación.

El Modelo B presenta un NB = 0.430 en umbral = 0.50, numéricamente superior al del Modelo A. Esto se debe a dos razones: por un lado, opera sobre una subpoblación con prevalencia de diabetes del 71.7% en el conjunto de prueba, fruto del sesgo de selección del subconjunto lipídico, lo que infla el NB absoluto de forma no comparable directamente con el Modelo A. Por otro lado, el rango de utilidad real del Modelo B es extremadamente estrecho, fuera del intervalo de umbrales 0.53-0.54, la curva del Modelo B cae por debajo de la referencia de tratar a todos, lo que implica que en la mayoría de los escenarios de decisión clínica el modelo no aporta utilidad neta adicional. Esta fragilidad del perfil DCA del Modelo B es coherente con su *overfitting* estructural y refuerza su caracterización como herramienta complementaria y confirmatoria, no como instrumento de tamizaje primario independiente.

6.3.3 ANÁLISIS EXPLORATORIO DE UMBRALES CLÍNICOS

Como se aprecia en la Figura 15, el análisis exploratorio de umbrales permite examinar el comportamiento de las métricas de clasificación en función del punto de corte aplicado a las probabilidades predichas, en el rango de 0.20 a 0.70. Es importante señalar que este análisis es exploratorio sobre el conjunto de prueba, la selección definitiva del umbral clínico requerirá validación prospectiva y una definición explícita del costo relativo de falsos positivos y falsos negativos dependiendo el contexto clínico.

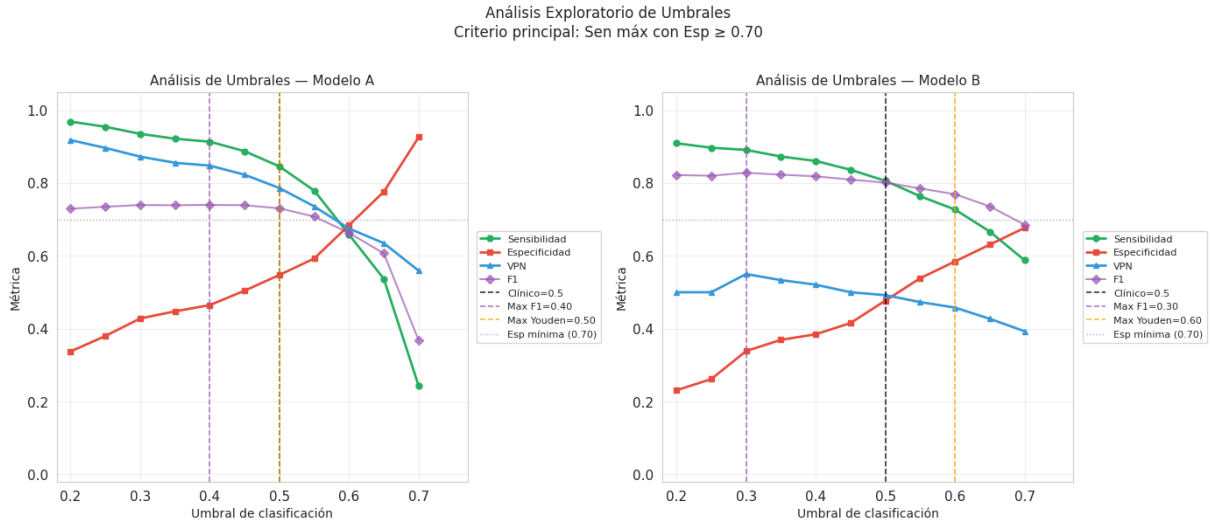


Figura 15. Análisis exploratorio de umbrales

Nota. Métricas de clasificación en función del umbral de probabilidad (rango 0.20-0.70). Línea vertical discontinua = umbral clínico seleccionado (0.50).

Para el Modelo A, el cruce entre las curvas de sensibilidad y especificidad se produce en torno al umbral de 0.65–0.70, mientras que el índice de Youden se maximiza en umbral = 0.50, coincidiendo exactamente con el umbral clínico seleccionado. Esta convergencia valida la elección con umbral = 0.50, la sensibilidad es de 0.845 y la especificidad de 0.548, configurando el perfil de tamizaje primario que prioriza la detección de casos reales sobre la reducción de falsos positivos. Si se priorizara el F1-score, el umbral óptimo sería 0.40, elevando la sensibilidad a 0.913 a costa de reducir la especificidad a 0.465, opción que incrementaría la carga de confirmaciones diagnósticas, pero reduciría el número de diabéticos no detectados.

Para el Modelo B, el índice de Youden sugiere un umbral óptimo de 0.60, ligeramente superior al umbral actual de 0.50. Con el umbral actual, el Modelo B presenta sensibilidad alta, pero especificidad relativamente baja (~0.48), lo que genera un volumen de falsos positivos que podría

reducirse desplazando el umbral hacia 0.60 con una mejora en el balance sensibilidad-especificidad, aunque a costa de una reducción en la tasa de detección. Esta exploración queda pendiente de validación prospectiva, dado que el tamaño del conjunto de prueba (n = 230) introduce incertidumbre suficiente para que las curvas de umbral no sean determinantes para una decisión clínica definitiva.

6.4 INTERPRETABILIDAD DEL MODELO

6.4.1 IMPORTANCIA DE VARIABLES - MDI Y SHAP GLOBAL

Como se observa en la Figura 16, la importancia MDI (*Mean Decrease in Impurity*) revela estructuras de decisión fundamentalmente distintas entre ambos modelos, coherentes con sus conjuntos de variables y sus roles clínicos complementarios.

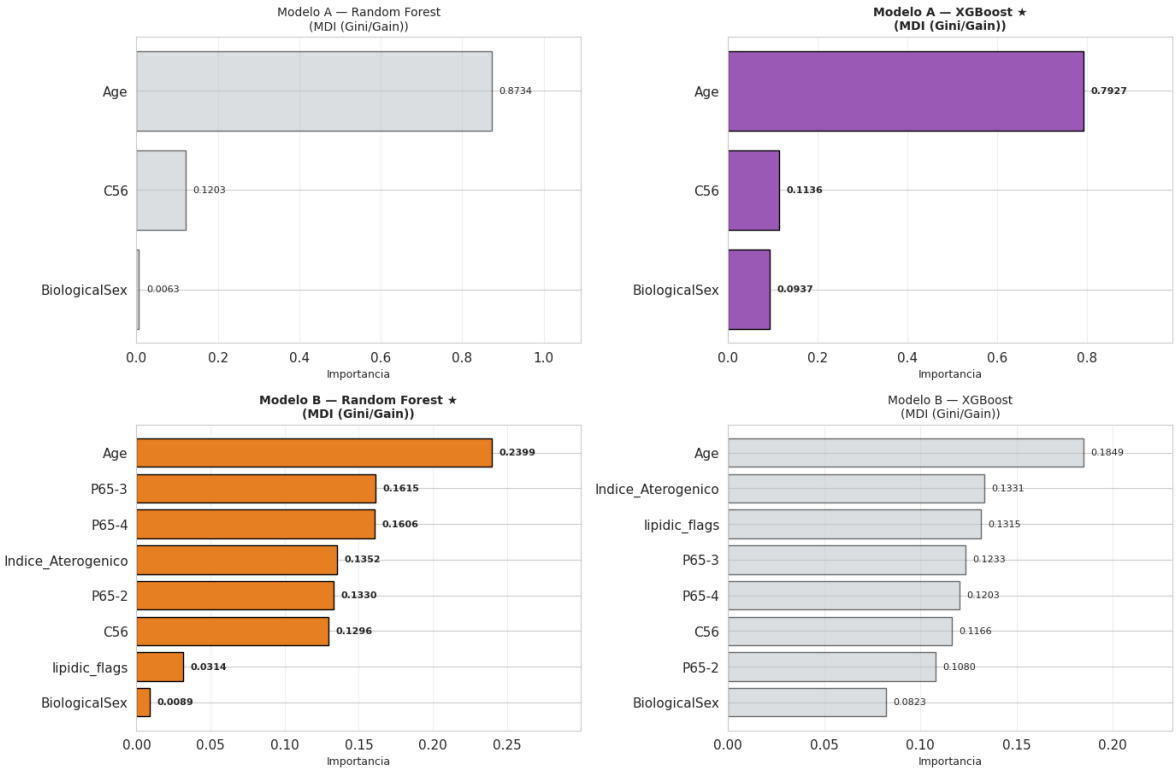


Figura 16. Importancia de variables MDI.

Nota. Importancia relativa (%) calculada mediante MDI. Panel superior: Modelo A. Panel inferior: Modelo B. Las variables están ordenadas de mayor a menor importancia.

El Modelo A (XGBoost) es esencialmente un modelo etario, la `Edad` concentra el 79.3% de la

importancia MDI, posicionándose como la señal dominante sobre la que el modelo construye prácticamente toda su capacidad discriminativa. La `Creatinina` aporta el 11.4% (reflejo de su asociación con la nefropatía diabética y la función renal) y el `Sexo Biológico` el 9.4% restante. Esta concentración de importancia en una sola variable no es una limitación del algoritmo sino una consecuencia directa del conjunto de variables disponibles, con solo tres variables, la `Edad` es la que aporta más información independiente sobre el `TARGET`, como confirman su Cohen's d de 0.963 y su correlación Spearman de 0.42, las más altas del conjunto.

El Modelo B (Random Forest regulado) presenta una distribución de importancia cualitativamente distinta, ninguna variable supera el 25% de importancia MDI, evidenciando un modelo de decisión distribuida donde el perfil lipídico completo (LDL, `Colesterol Total`, HDL e `Índice Aterogénico`) aporta conjuntamente el 59.1% del poder discriminativo. La reducción de la importancia de la `Edad` del 79.3% en el Modelo A al 24% en el Modelo B, cuantifica con precisión el valor informativo adicional del perfil lipídico, cuando se dispone de datos bioquímicos lipídicos, otros factores explican el 76% del poder discriminativo del modelo, reduciendo sustancialmente la dependencia en el factor etario y permitiendo capturar señales metabólicas que la `Edad` por sí sola no puede representar.

6.4.2 SHAP GLOBAL - *BEESWARM PLOTS*

Como se observa en las Figuras 17 y 18, los *beeswarm plots* de SHAP complementan la importancia MDI al revelar no solo qué variables contribuyen más a la predicción, sino también la dirección y magnitud del efecto para cada paciente individual del conjunto de prueba.

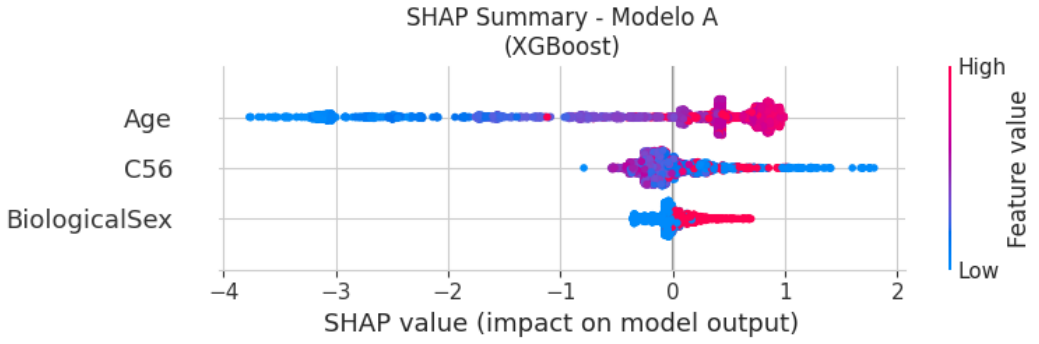


Figura 17. SHAP global - Modelo A (XGBoost)

Nota. *Beeswarm plot de* valores SHAP. Cada punto representa un paciente del conjunto de prueba. Eje X = valor SHAP (impacto en la predicción). Color rosa = valor alto de la variable. Color azul = valor bajo. Las variables están ordenadas por importancia SHAP media absoluta.

En el Modelo A, el patrón SHAP de la Edad es el más informativo y claro del conjunto: valores altos de Edad (puntos rosas) generan SHAP positivos de +2 a +3, incrementando sustancialmente la probabilidad predicha de DM2, mientras que valores bajos (puntos azules) producen SHAP fuertemente negativos de -3 a -4, reflejando la marcada protección que confiere la juventud contra el diagnóstico de diabetes. La distribución es asimétrica hacia la izquierda, con los puntos azules extendiéndose más en la dirección negativa, lo que cuantifica visualmente que la protección de la Edad joven es más pronunciada que el incremento de riesgo de la Edad avanzada. La Creatinina presenta un patrón menos intuitivo, valores muy altos de Creatinina (rosa) muestran en algunos casos SHAP negativos, comportamiento atribuible a una interacción no lineal con la Edad (pacientes con insuficiencia renal severa tienden a ser adultos mayores en quienes otros factores compensan la señal), o a colinealidad residual entre ambas variables en el extremo superior del rango. El Sexo Biológico muestra una distribución bimodal compacta centrada en torno a cero, confirmando cuantitativamente su bajo poder discriminativo individual y justificando retrospectivamente su inclusión por criterios clínicos y no estadísticos.

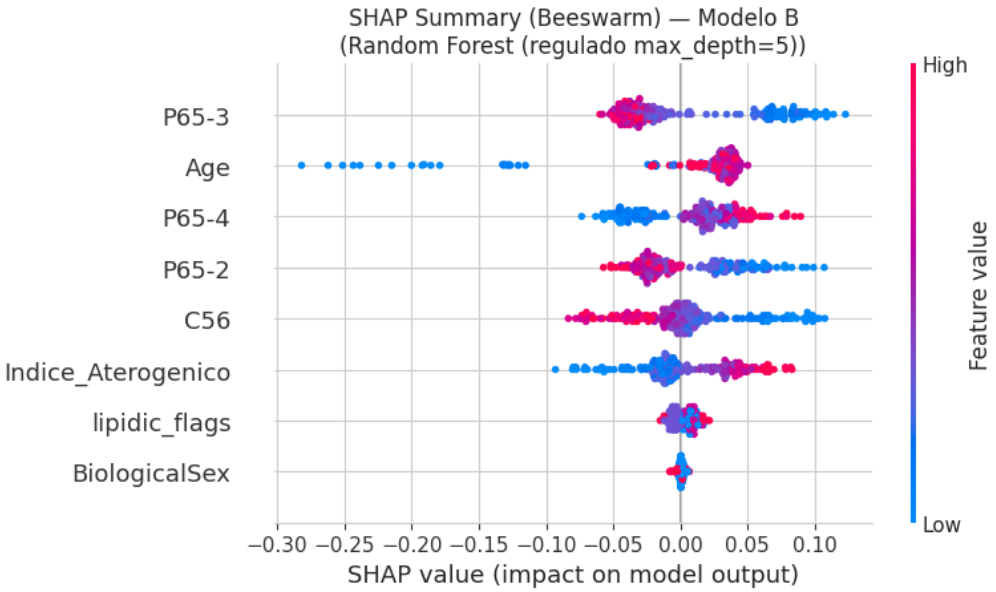


Figura 18. SHAP global - Modelo B (RF regulado)

Nota. Beeswarm plot de valores SHAP para el Modelo B. La distribución más compacta de Age refleja la competencia con el perfil lipídico. LDL (P65-3) en primer lugar confirma su rol dominante en el contexto del síndrome metabólico.

En el Modelo B, el hallazgo SHAP más relevante desde el punto de vista clínico es el posicionamiento del LDL (P65-3) en primer lugar, valores altos de LDL (rosa) generan SHAP positivos con alta dispersión, indicando que LDL elevado aumenta el riesgo predicho de DM2 con una variabilidad individual considerable según el contexto lipídico de cada paciente, hallazgo consistente con la literatura sobre síndrome metabólico y resistencia a la insulina. La `Edad` asciende al segundo lugar en importancia SHAP (frente al tercero en MDI) con una distribución más compacta que en el Modelo A, reflejo de que aquí la señal etaria compite con información lipídica que explica parte de la varianza que en el Modelo A la `Edad` acapara en solitario. El HDL (P65-2) presenta un SHAP medio firmado negativo (-0.034), clínicamente coherente con su papel protector documentado previamente, los puntos azules correspondientes a HDL bajo se sitúan a la derecha del cero (aumentan el riesgo predicho), mientras que los puntos rosas de HDL alto se desplazan hacia la izquierda (reducen el riesgo), representando visualmente el efecto protector del HDL sobre el riesgo de DM2 que la literatura clínica internacional ha establecido consistentemente.

6.4.3 ANÁLISIS DE DESEMPEÑO POR SUBGRUPOS ETARIOS

Como se aprecia en la Figura 19 y la Tabla 20, el análisis de desempeño por subgrupos etarios revela un patrón crítico con implicaciones directas para la aplicabilidad clínica de ambos modelos.

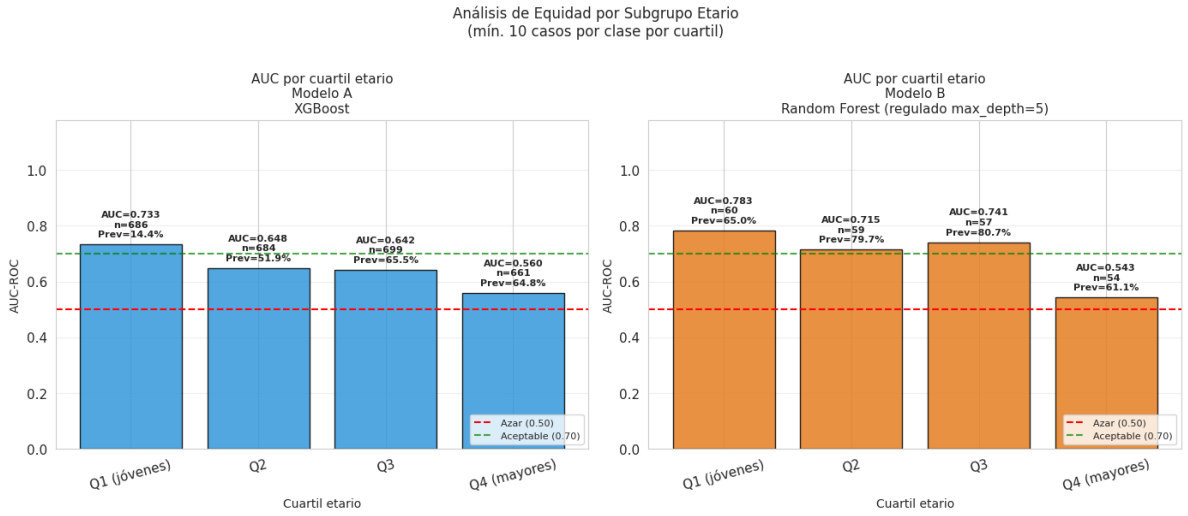


Figura 19. AUC por cuartil etario - Modelo A y Modelo B

Nota. AUC de discriminación calculado separadamente en cada cuartil etario del conjunto de prueba. Panel izquierdo: Modelo A (N test=2.730). Panel derecho: Modelo B (N test=230). Línea roja discontinua = AUC=0.70 (umbral de referencia).

Tabla 20. AUC por cuartil etario - Análisis de subgrupos

Cuartil	Rango etario	N (A)	N (B)	AUC Modelo A	AUC Modelo B	Cuartil
Q1 - Jóvenes	18-50(A) / 20-61(B)	686	60	0.733	0.783	Q1 - Jóvenes
Q2	51-64(A) / 62-69(B)	684	59	0.648	0.715	Q2
Q3	65-74(A) / 70-75(B)	699	57	0.643	0.741	Q3
Q4 - Mayores	75-106(A) / 76-93(B)	661	54	0.560	0.543	Q4 - Mayores

Nota. Cuartiles calculados separadamente por modelo. N reducido en Modelo B (54–60 por cuartil).

El hallazgo más relevante es el colapso del AUC en el cuartil de adultos mayores (Q4) en ambos modelos, con 0.560 en el Modelo A y 0.543 en el Modelo B, valores prácticamente equivalentes al azar. En el Modelo A, este deterioro es estructuralmente esperado, dado que la `Edad` concentra el 79.3% de la importancia MDI, dentro de un grupo etario homogéneo donde todos los pacientes tienen `Edad`es similares (75-106 años), la señal dominante desaparece y el modelo pierde su principal mecanismo de discriminación. En el Modelo B, el colapso tiene una causa adicional, en adultos muy mayores, con tratamiento farmacológico intensivo y la acumulación de comorbilidades homogenizan el perfil lipídico independientemente del estado diabético. Este hallazgo tiene una implicación clínica directa sobre los dos modelos, ya que no deben utilizarse como herramienta de tamizaje primario en adultos mayores de 75 años sin criterios diagnósticos adicionales, y el desarrollo de un modelo específico para este subgrupo, que incorpore otras variables clínicas y constituya una línea de trabajo futura.

En contraste, el Modelo B supera consistentemente al Modelo A en los cuartiles Q1, Q2 y Q3 (AUC 0.715–0.783 vs 0.560–0.733), confirmando que el perfil lipídico mantiene poder discriminativo en adultos jóvenes y de mediana `Edad` con independencia del factor etario. Este resultado cuantifica el valor complementario del Modelo B precisamente donde el Modelo A es más vulnerable, en pacientes jóvenes con DM2 de inicio temprano, donde la `Edad` baja suprime la predicción del Modelo A, el perfil lipídico alterado puede capturar la señal metabólica que la demografía no proporciona. En conjunto, estos resultados sugieren que la mejora del desempeño dentro de grupos etarios homogéneos requerirá la incorporación de variables adicionales (antropométricas,

de estilo de vida o de trayectoria temporal) que aporten información discriminativa independiente de la Edad.

6.4.4 ESTUDIO DE ABLACIÓN

La Figura 20, muestra el estudio de ablación, que cuantifica el aporte incremental de cada bloque de variables al AUC del Modelo B mediante validación cruzada, permitiendo descomponer la contribución independiente de cada componente del conjunto de predictores.

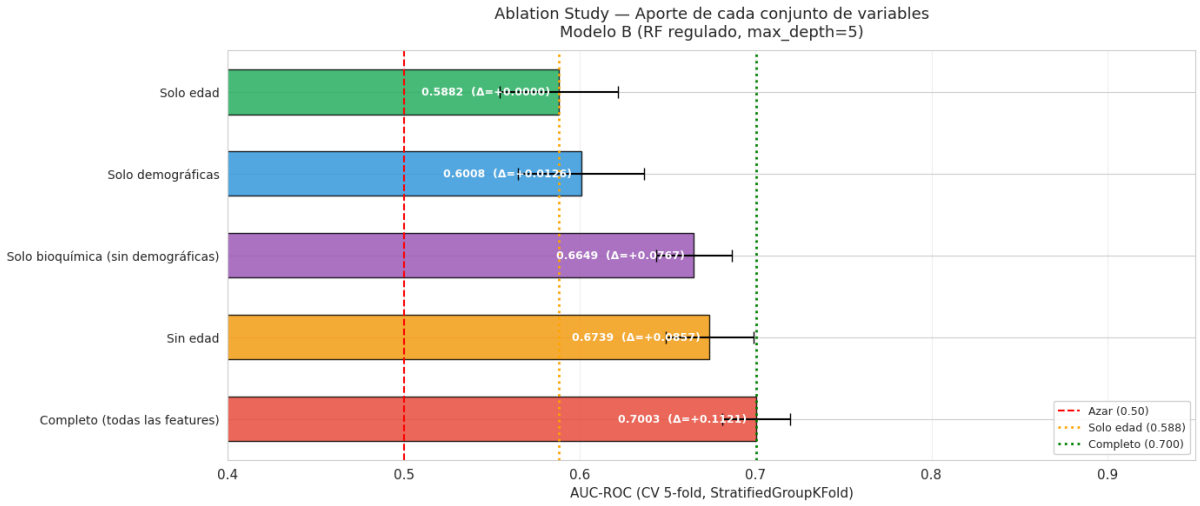


Figura 20. Estudio de ablación.

Nota. AUC medio en validación cruzada para configuraciones de variables en el Modelo B. Las barras muestran el AUC de cada configuración y el aporte incremental de cada bloque adicional.

El hallazgo central es que el perfil lipídico aporta +0.112 puntos AUC sobre la configuración del Modelo A, confirmando que la información bioquímica lipídica no es redundante con la demográfica y justifica cuantitativamente la existencia del Modelo B como complemento. Más aún, la bioquímica lipídica en ausencia de variables demográficas (AUC = 0.665) supera a la demografía sola (AUC = 0.601), posicionando el perfil lipídico como el predictor dominante del Modelo B y no como un complemento secundario de la Edad.

El Sexo Biológico aporta solo +0.013 AUC marginal sobre la combinación de Edad y bioquímica, confirmando lo documentado en el análisis bivariado y en los valores SHAP, su inclusión se justifica por consenso clínico ADA y comparabilidad con estudios previos, no por aporte estadístico en este *dataset* específico. Este resultado es coherente con el Cramér's V de 0.019 y la importancia SHAP de 0.002 reportados anteriormente.

Un hallazgo importante es que el modelo completo es el único que alcanza el umbral de AUC = 0.70 considerado clínicamente útil (AUC = 0.700), mientras que ninguna configuración parcial lo supera de forma independiente. Esto implica que la combinación de todos los predictores disponibles es una condición necesaria para que el Modelo B alcance un desempeño clínicamente aceptable, y que la exclusión de cualquier bloque de variables (incluso el de menor aporte individual como el Sexo Biológico) compromete el cruce de ese umbral de referencia. Este resultado refuerza la decisión de mantener el conjunto completo de ocho variables en el Modelo B, incluso aquellas con aporte marginal estadísticamente modesto.

6.4.5 SHAP *DEPENDENCE PLOTS* - INTERACCIONES CON LA `EDAD`

Como se aprecia en la Figura 21, los SHAP *dependence plots* permiten examinar cómo el efecto de la `Edad` sobre la predicción varía en función de los valores de las variables lipídicas, revelando la presencia o ausencia de interacciones entre predictores en el Modelo B.

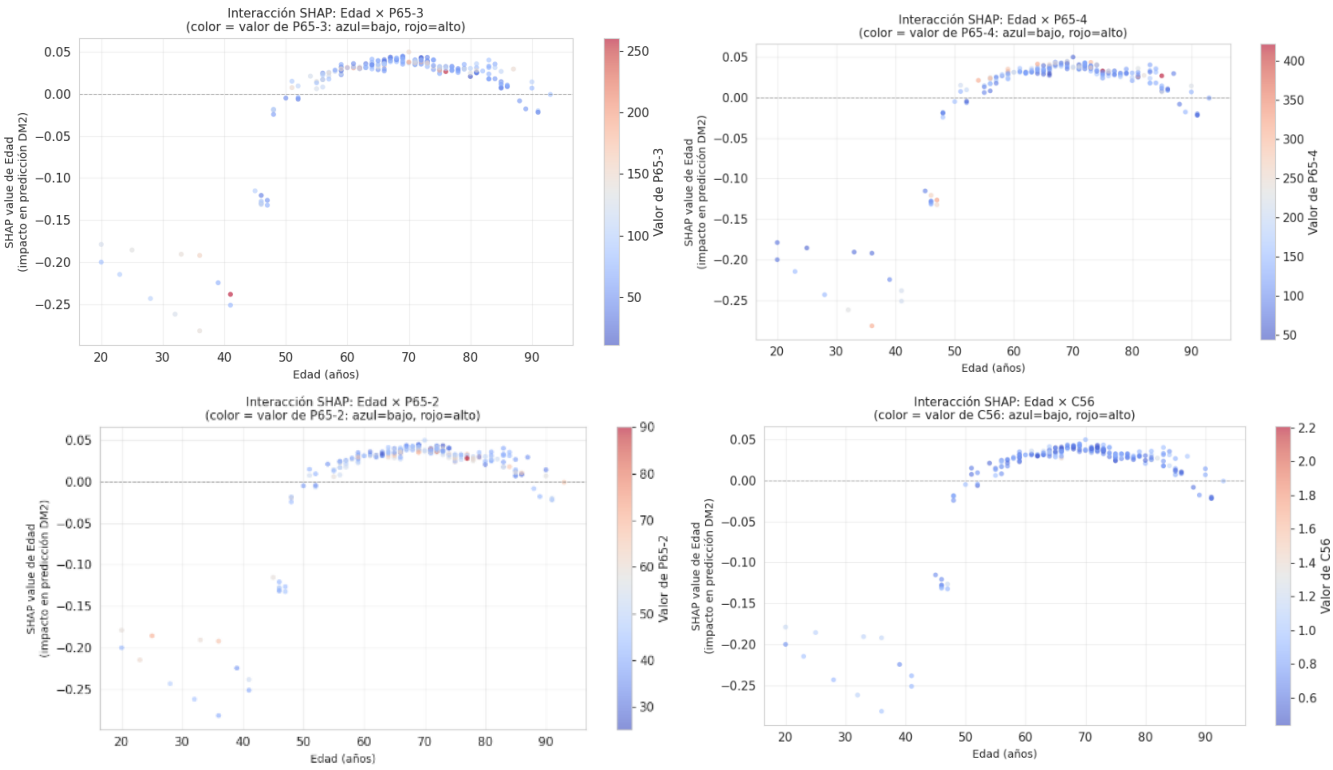


Figura 21. SHAP *dependence plots.*

Nota. Cada punto representa un paciente. Eje X = valor de la `Edad`, eje Y = valor SHAP de la `Edad`. El color indica el valor de la variable lipídica de interacción (rosa = alto; azul = bajo). Una separación de colores indica interacción entre variables.

Todos los gráficos comparten la misma forma de campana asimétrica en los valores SHAP de la Edad, fuertemente negativos en menores de 50 años (la juventud actúa como factor protector marcado que domina la predicción), positivos en el rango 55-75 años (la Edad madura incrementa el riesgo predicho) y con aplanamiento progresivo a partir de los 80 años, reflejo de un efecto de saturación donde la Edad muy avanzada deja de aportar información discriminativa adicional, consistente con el colapso del AUC en Q4, tal como se documentó anteriormente.

Respecto a las interacciones específicas con cada variable lipídica, los resultados muestran efectos predominantemente aditivos e independientes. La interacción Edad-LDL es débil a moderada y sutil, sin que el nivel de LDL modifique de forma apreciable el patrón etario. La interacción Edad-Colesterol Total es igualmente débil, con una posible atenuación leve en pacientes jóvenes con CT muy elevado que no altera la interpretación global. La interacción Edad-Creatinina es prácticamente inexistente, confirmando que ambas variables actúan como predictores ortogonales con efectos completamente aditivos, coherente con su baja correlación de Spearman. La interacción más notable, aunque igualmente sutil, es la de Edad-HDL en mayores de 60 años, donde HDL alto puede amplificar levemente el efecto positivo de la Edad sobre el riesgo predicho, resultado aparentemente contraintuitivo dado el papel protector del HDL. Una interpretación plausible es que en adultos mayores con HDL alto, otros factores de riesgo dominantes (no capturados directamente por las variables del modelo) son los verdaderos determinantes del diagnóstico, y el HDL elevado coexiste con ellos sin contrarrestarlos.

La ausencia generalizada de interacciones fuertes entre la Edad y el perfil lipídico tiene una implicación interpretativa relevante, el Modelo B captura fundamentalmente efectos aditivos, donde cada variable contribuye de forma independiente al riesgo predicho sin modificar sustancialmente el efecto de las demás. Esta propiedad simplifica la interpretación clínica del modelo y refuerza su transparencia como herramienta de apoyo a la decisión médica.

6.4.6 ANÁLISIS DE ERRORES - PERFIL DE FALSOS NEGATIVOS Y FALSOS POSITIVOS

Las Figura 22a y 22b, caracterizan el análisis de errores que permite caracterizar con precisión los perfiles clínicos de los pacientes donde cada modelo falla sistemáticamente, información esencial para definir los límites de aplicabilidad de cada herramienta y diseñar el algoritmo de tamizaje secuencial.

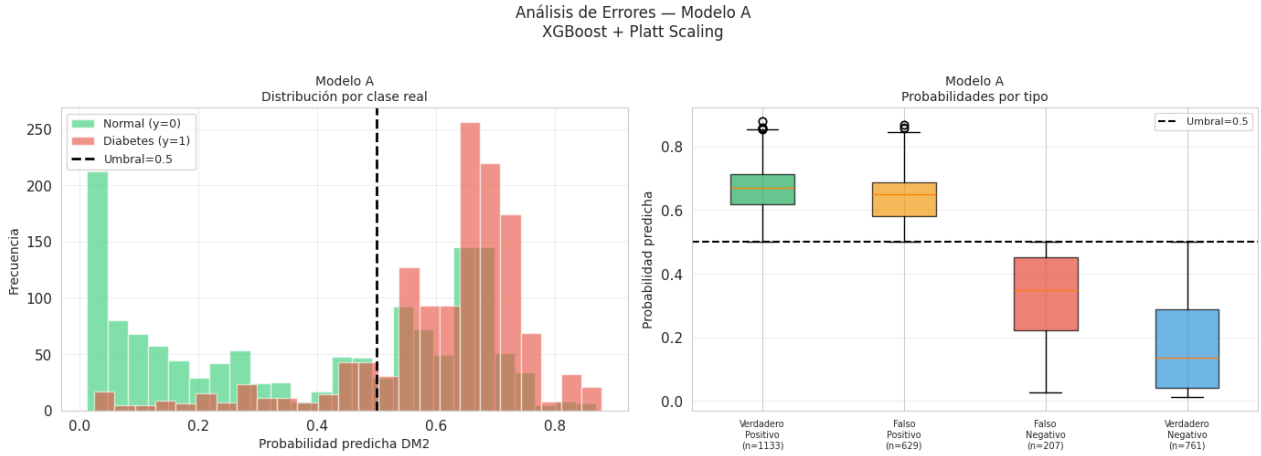


Figura 22a. Distribución de FN y FP - Modelo A

Nota. Comparación de características clínicas (`Edad`, `Creatinina`, variables lipídicas) entre Verdaderos Positivos (TP), Falsos Negativos (FN), Verdaderos Negativos (TN) y Falsos Positivos (FP).

En el Modelo A, los falsos negativos (N = 207, `Edad` media = 51 años) corresponden predominantemente a pacientes jóvenes con DM2 de inicio temprano, cuya `Edad` baja genera valores SHAP fuertemente negativos que suprimen la predicción a pesar de la diabetes real. Los falsos positivos (N = 629, `Edad` media = 71 años) representan un hallazgo, adultos mayores sanos cuya `Edad` elevada genera SHAP positivos que elevan la predicción por encima del umbral, aunque no exista diabetes. En consecuencia, en pacientes mayores de 70 años un resultado positivo del Modelo A debe confirmarse sistemáticamente mediante criterios ADA antes de cualquier intervención.

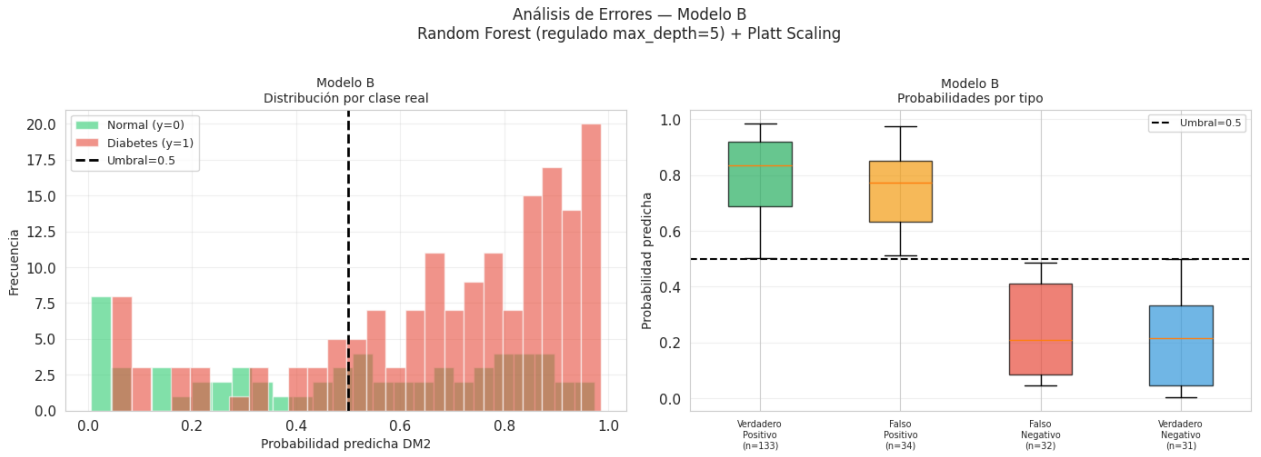


Figura 22b. Distribución de FN y FP - Modelo B

Nota. Comparación de características clínicas (`Edad`, `Creatinina`, variables lipídicas) entre

Verdaderos Positivos (TP), Falsos Negativos (FN), Verdaderos Negativos (TN) y Falsos Positivos (FP).

En el Modelo B, los falsos negativos (N = 39, `Edad` media = 61.7 años) definen el punto ciego estructural del modelo: diabéticos cuyo perfil lipídico ha sido farmacológicamente normalizado, y que por tanto no generan la señal bioquímica lipídica sobre la que el modelo basa su predicción. Este error es inherente al diseño del modelo y no corregible mediante ajuste de hiperparámetros, dado que la información necesaria para detectar estos casos (el tratamiento farmacológico activo) no está disponible en el *dataset* inicial. Los falsos positivos (N = 30) corresponden a pacientes con dislipidemia severa sin DM2, cuyo perfil lipídico alterado simula el patrón metabólico de un paciente diabético, generando clasificaciones positivas incorrectas que el alto VPP del modelo (0.809) limita, pero no elimina.

El hallazgo más relevante de este análisis es que los perfiles de error de ambos modelos son complementarios, los falsos negativos del Modelo A son jóvenes con DM2 donde el perfil lipídico del Modelo B podría capturar la señal metabólica ausente en la demografía, y los falsos negativos del Modelo B son diabéticos con lípidos normalizados donde la `Edad` y la `Creatinina` del Modelo A mantienen parte de su poder discriminativo. Esta complementari`edad` fundamenta el algoritmo de tamizaje secuencial, donde la combinación de ambos modelos permite cubrir los puntos ciegos individuales de cada uno.

6.5 VALIDACIÓN CLÍNICA E INTERPRETABILIDAD MÉDICA

La validación clínica realizada en este proyecto debe entenderse como una validación preliminar de plausibilidad clínica e interpretabilidad médica, no como una validación prospectiva formal ni como una evaluación definitiva en escenarios asistenciales reales. Esta aproximación se fundamentó en la coherencia entre las variables seleccionadas y la fisiopatología conocida de la Diabetes Mellitus tipo 2, y en la dirección de los efectos observados mediante valores SHAP. Bajo este enfoque, el objetivo fue verificar que los patrones aprendidos por los modelos fueran clínicamente razonables y no artefactos puramente estadísticos.

Como parte de esta validación se exploró inicialmente la posibilidad de contrastar el desempeño del Modelo B contra una versión adaptada del FINDRISC (*Finnish Diabetes Risk Score*), uno de los instrumentos de tamizaje de DM2 más utilizados y validados internacionalmente. Sin embargo, dado que el FINDRISC original requiere variables antropométricas y de antecedentes (IMC, perímetro abdominal, actividad física y antecedentes familiares) no disponibles en el *dataset*

utilizado, cualquier versión simplificada construida únicamente con variables bioquímicas no reproduce el instrumento validado ni constituye una comparación estadísticamente válida frente a dicho score de referencia. Por esta razón, esta comparación se excluyó del análisis de validación clínica del presente trabajo. La contrastación formal del Modelo B frente al FINDRISC completo queda planteada como parte de la validación prospectiva externa descrita en los trabajos futuros, una vez se disponga de las variables antropométricas necesarias para una comparación íntegra y metodológicamente sólida.

No obstante, esta validación de plausibilidad clínica no sustituye una validación médica formal con profesionales de la salud ni una validación prospectiva externa. La revisión de plausibilidad clínica permite establecer que las variables y la dirección de los efectos son consistentes con la literatura, pero no permite concluir por sí sola que la herramienta mejora la toma de decisiones clínicas en escenarios reales. Por tanto, antes de cualquier implementación asistencial, se requiere una fase posterior de validación con médicos, basada en casos clínicos estructurados, evaluación de concordancia con juicio experto, análisis de usabilidad y medición del impacto potencial sobre decisiones de tamizaje.

7. IMPLEMENTACIÓN DE HERRAMIENTA GRÁFICA

Como parte del proyecto se diseñó y desarrolló una herramienta web denominada DiaPredict MD, orientado a la consulta de modelos predictivos para el apoyo al tamizaje de riesgo de Diabetes Mellitus tipo 2. A continuación, se documentan los requerimientos funcionales y no funcionales, la arquitectura de microservicios adoptada, el entorno de desarrollo, la integración del modelo predictivo, las pruebas realizadas y la documentación técnica generada.

7.1 DEFINICIÓN DE REQUERIMIENTOS FUNCIONALES, NO FUNCIONALES Y ARQUITECTURA

7.1.1 REQUERIMIENTOS FUNCIONALES

Se definieron siete requerimientos funcionales (RF) que delimitan las capacidades del sistema, organizados en *sprints* incrementales según metodología ágil. La Tabla 21 presenta su descripción y estado de implementación.

Tabla 21. Requerimientos funcionales del sistema DiaPredict MD

ID	Nombre	Descripción
RF-1	Autenticación de usuarios	El sistema permitirá registro e inicio de sesión mediante correo y contraseña, con generación de *token* JWT y diferenciación por roles (médico, administrador).
RF-2	Captura clínica	El sistema permitirá ingresar datos demográficos (sexo, `Edad`) y resultados de laboratorio rutinarios (`Creatinina`, perfil lipídico) mediante formulario validado.
RF-3	Predicción del riesgo	El sistema seleccionará automáticamente entre Modelo A (tamizaje básico) o Modelo B (tamizaje completo) según los datos disponibles, retornando probabilidad y categoría de riesgo (bajo, medio, alto).
RF-4	Recomendaciones LLM	El sistema generará recomendaciones clínicas contextualizadas mediante un modelo de lenguaje (OpenAI GPT-4o-mini), con anonimización previa de datos sensibles del paciente.
RF-5	Interpretabilidad SHAP	El sistema mostrará la importancia relativa de cada variable en la predicción mediante valores SHAP normalizados como porcentajes, presentados visualmente al usuario.
RF-6	Retroalimentación médica	El sistema permitirá al usuario clínico calificar la utilidad del análisis (Útil, Poco claro, Incorrecto) y registrar observaciones libres,

		alimentando el ciclo de mejora continua.
RF-7	Trazabilidad y auditoría	El sistema registrará en bitácora cada análisis realizado, incluyendo usuario, fecha, datos de entrada, predicción, modelo aplicado y *feedback* recibido, soportando auditoría retrospectiva.

7.1.2 REQUERIMIENTOS NO FUNCIONALES

Los requerimientos no funcionales (RNF) definen los atributos de calidad del sistema más allá de su funcionalidad explícita. La Tabla 22 sintetiza los principales RNF abordados, agrupados por categoría según la norma ISO/IEC 25010.

Tabla 22. Requerimientos no funcionales del sistema

Categoría	**Atributo**	**Implementación**
Seguridad	Autenticación	JWT, contraseñas bcrypt (12 rounds)
Privacidad	Anonimización de PII	Eliminación de identificadores antes del envío al LLM externo
Mantenibilidad	Separación de responsabilidades	Arquitectura de microservicios con *backend*, ml-service y *frontend* desacoplados.
Portabilidad	Despliegue reproducible	Contenerización con Docker Compose, configuración por variables de entorno
Usabilidad	Interfaz *responsive*	*Frontend* Angular adaptable a escritorio.
Confiabilidad	*Healthchecks* y alta disponibilidad	Verificación automatizada de salud en todos los servicios Docker.

7.1.3 ARQUITECTURA DEL SISTEMA

Se adoptó una arquitectura de microservicios distribuida en cuatro componentes desacoplados, comunicados mediante una red Docker virtualizada. Esta decisión se justifica por la separación de responsabilidades, la escalabilidad independiente de cada servicio y la facilidad de mantenimiento. La Figura 23 representa el diagrama de componentes del sistema.

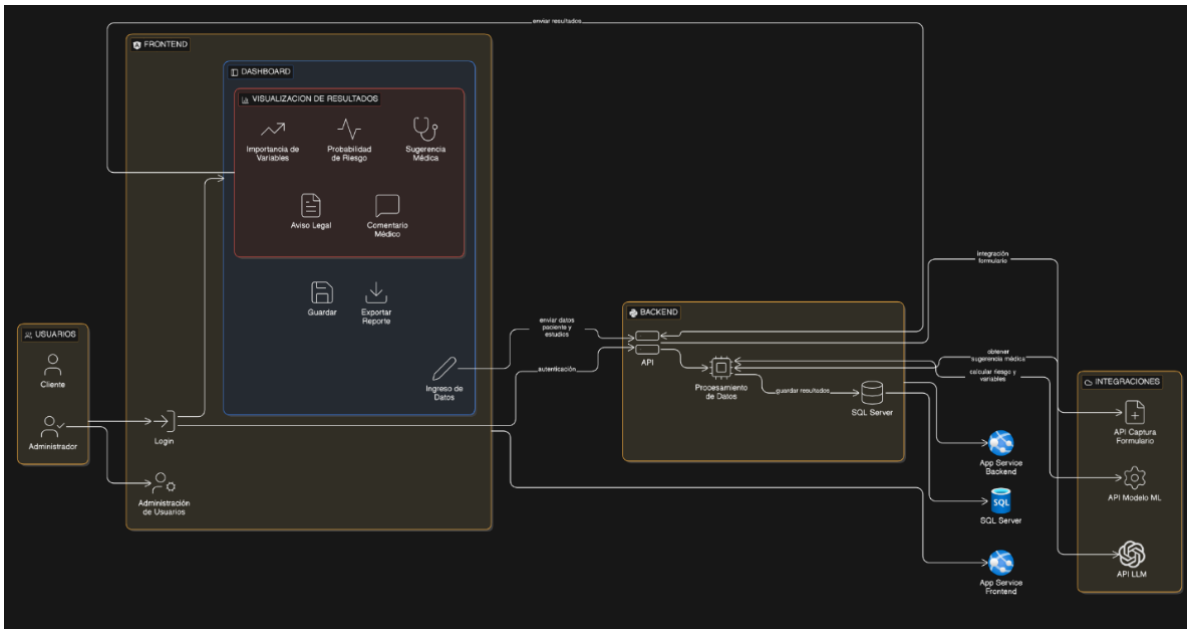


Figura 23. Arquitectura de microservicios del sistema DiaPredict MD.

Los cuatro componentes principales son:

1. *Frontend* Angular servido mediante Nginx, responsable de la interfaz de usuario.
2. *Backend* API desarrollado en FastAPI, encargado de la autenticación, gestión de usuarios, persistencia de pacientes y orquestación de los servicios externos.
3. ML Service desarrollado en FastAPI, microservicio aislado dedicado a la inferencia del modelo y al cálculo de explicaciones SHAP
4. Base de datos SQL Server 2022 para almacenamiento persistente. Adicionalmente, el *Backend* se comunica con la API de OpenAI para generar recomendaciones clínicas mediante LLM.

El *frontend* tiene entradas separadas hacia el *Backend* (autenticación, gestión clínica) y hacia el ML Service (predicciones), permitiendo una distribución de carga más eficiente y un acoplamiento mínimo entre los componentes.

7.2 ENTORNO DE DESARROLLO

El *stack* tecnológico se seleccionó privilegiando madurez, comunidad activa, rendimiento y compatibilidad con buenas prácticas de seguridad. La Tabla 23 detalla las tecnologías empleadas por capa.

Tabla 23. *Stack* tecnológico del sistema

Capa	Tecnología
Frontend	`Angular`
Backend	`Python 3.13 + FastAPI`
Validación	Pydantic v2
Base de datos	SQL Server 2022
ML	`scikit-learn` + XGBoost + SHAP
LLM	OpenAI API (GPT-4o-mini)
Seguridad	JWT + bcrypt + slowapi
Runtime	Docker Desktop + Nginx

El repositorio se estructuró como un *monorepo* que agrupa *backend*, ml-service, scripts SQL, configuración Nginx y documentación. La orquestación de los cuatro servicios se materializó mediante Docker Compose, lo que permite un despliegue completo del entorno con un solo comando (docker compose up). Los servicios se comunican a través de una red Docker virtual privada y se inicializan respetando dependencias mediante *healthchecks*.

7.3 INTEGRACIÓN DEL MODELO PREDICTIVO

La integración del modelo se materializó mediante un microservicio independiente (ml-service) implementado en FastAPI, que expone tres *endpoints*: /health (verificación de salud), /predict (predicción del riesgo) y /explain (cálculo de valores SHAP).

Los modelos finales (Modelo A y Modelo B con calibración) se persistieron en formato *pickle* mediante joblib y se cargan una sola vez al inicio del servicio, evitando la sobrecarga de carga repetida en cada petición.

El sistema tiene la capacidad de selección automática del modelo apropiado según los datos disponibles del paciente. Cuando el formulario incluye los cuatro componentes lipídicos (`Triglicéridos`, HDL, LDL y `Colesterol Total`), se calculan las variables derivadas (`Índice Aterogénico` y `lipidic_flags`) y se aplica el Modelo B; en caso contrario, se aplica el Modelo A.

La respuesta del *endpoint* /predict incluye no solo la probabilidad y la clasificación, sino también el modelo aplicado y el nivel de confianza, otorgando transparencia al usuario clínico sobre qué modelo se utilizó.

7.3.2 CATEGORIZACIÓN DE RIESGO E INTERPRETABILIDAD MEDIANTE SHAP

La probabilidad continua se categoriza en tres niveles de riesgo según umbrales clínicamente justificados: bajo (<0.34), medio (0.34–0.66) y alto (>0.66). Los umbrales bajo, medio y alto se definieron como una estrategia operativa de visualización para el prototipo, no como puntos de corte clínicos definitivos. Estos cortes permiten comunicar la probabilidad estimada en tres categorías comprensibles para el usuario. Su validación y ajuste deberán realizarse en estudios posteriores con profesionales clínicos, considerando sensibilidad, especificidad, beneficio neto y consecuencias de falsos positivos y falsos negativos.

El *endpoint* /explain devuelve los valores SHAP normalizados como porcentajes, ordenados por magnitud absoluta de impacto. Estos valores se utilizan en el *frontend* para mostrar barras horizontales que comunican al médico qué variables tuvieron mayor peso en la predicción específica del paciente, dando lugar a una explicabilidad local accesible.

7.3.3 RECOMENDACIONES LLM CON ANONIMIZACIÓN

Para enriquecer la respuesta clínica, el *Backend* integra un módulo de generación de recomendaciones mediante OpenAI GPT-4o-mini. Antes de enviar los datos al servicio externo, se aplica un proceso de anonimización que elimina cualquier identificador del paciente, conservando únicamente las variables numéricas y categóricas necesarias para contextualizar la sugerencia. Esta práctica garantiza el cumplimiento de la Ley 1581 de 2012 [39] sobre protección de datos personales.

7.4 PRUEBAS DE USABILIDAD Y VALIDACIÓN

7.4.1 PRUEBAS FUNCIONALES MEDIANTE CASOS DE USO

Las pruebas funcionales del sistema se realizaron mediante la ejecución de dos casos de uso representativos que verifican la selección automática de modelo, la generación de la recomendación clínica mediante LLM y la categorización diferenciada de riesgo. Como se ilustra en las Figuras 24 y 25, el sistema presenta una interfaz de acceso seguro mediante autenticación de usuario y un *dashboard* principal que centraliza el formulario de captura clínica, adaptándose dinámicamente a la información disponible del paciente para determinar qué modelo aplicar sin intervención del usuario.

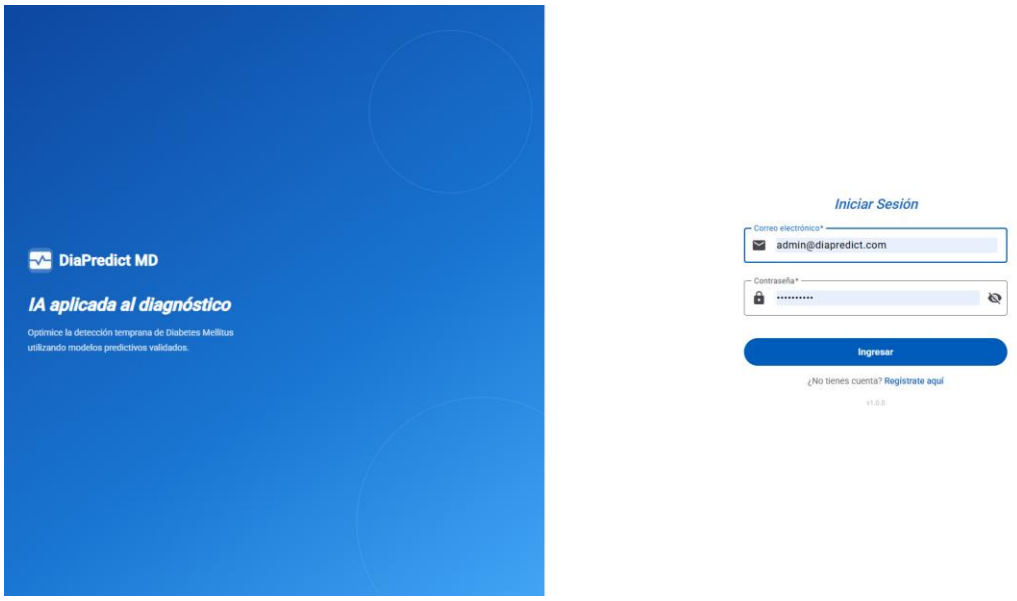


Figura 24. Pantalla de inicio de sesión del sistema DiaPredict MD.

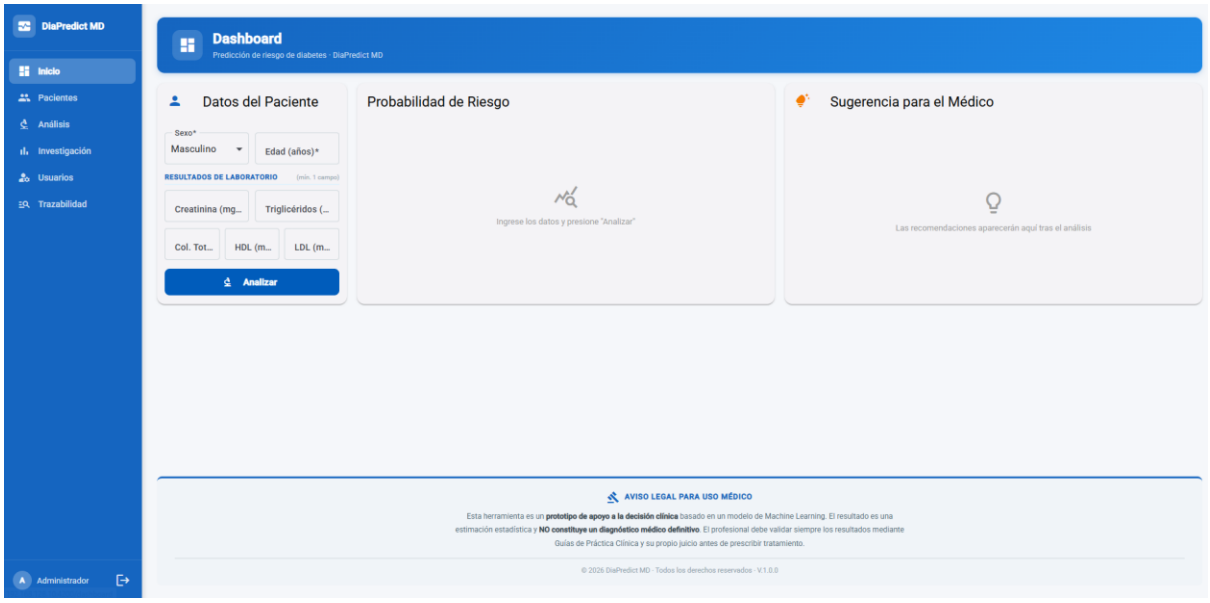


Figura 25. *Dashboard* principal con formulario de captura clínica.

El primer caso de uso corresponde a un paciente masculino de 50 años con datos básicos disponibles (`Creatinina` = 1.2 mg/dL). Al no disponer de panel lipídico completo, el sistema selecciona automáticamente el Modelo A y retorna una probabilidad de diabetes del 27%, categorizada como riesgo bajo. La descomposición SHAP presentada en la interfaz refleja la

dominancia esperada de la Edad (70%), seguida por la Creatinina (24%) y el Sexo Biológico(6%). La Figura 26 muestra el resultado completo generado por el sistema, incluyendo la recomendación clínica contextualizada producida por el LLM a partir del perfil individual del paciente.

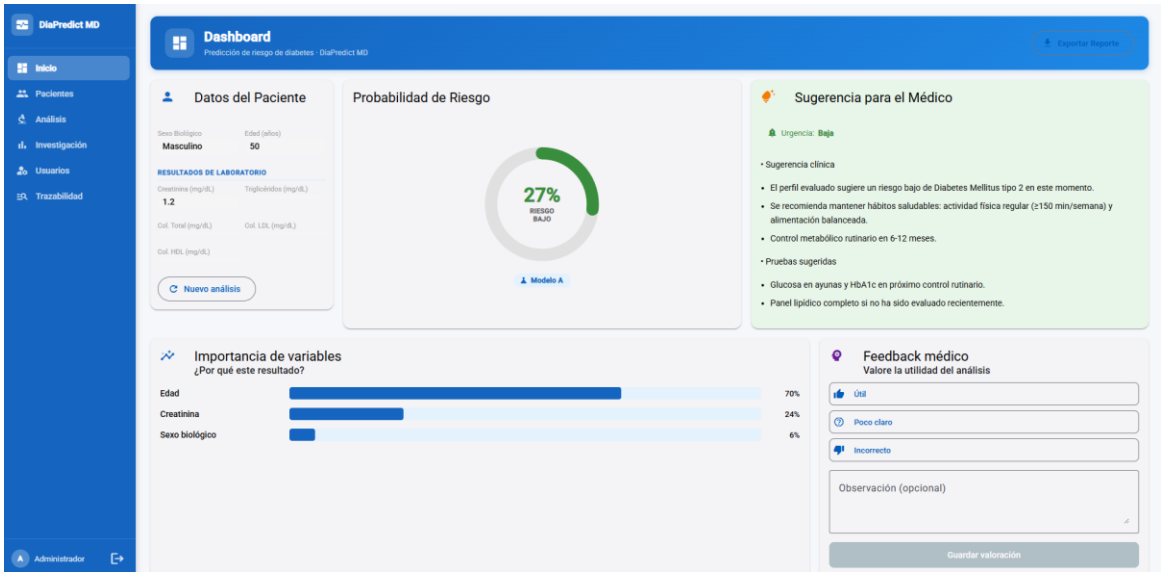


Figura 26. Resultado de análisis con Modelo A: paciente con riesgo bajo (27%)

El segundo caso de uso corresponde a una paciente femenina de 65 años con perfil bioquímico completo (Creatinina = 1.1 mg/dL, Triglicéridos = 220 mg/dL, Colesterol Total = 240 mg/dL, LDL = 38 mg/dL, HDL = 145 mg/dL). La disponibilidad del panel lipídico activa automáticamente el Modelo B, que retorna una probabilidad de diabetes del 57%, categorizada como riesgo moderado. La descomposición SHAP refleja la mayor riqueza informativa del Modelo B, la Edad aporta el 30%, el Índice Aterogénico el 27%, la Creatinina el 19% y el HDL el 17%, distribución equilibrada que contrasta con la concentración etaria del Modelo A y confirma el aporte clínico del perfil lipídico en la predicción individual. La Figura 27 presenta el resultado completo. Es pertinente señalar que los valores lipídicos de este caso (LDL = 38 mg/dL y HDL = 145 mg/dL) corresponden a un perfil atípico que el sistema procesa correctamente, generando una predicción y una recomendación coherentes con los valores individuales a través del componente LLM.

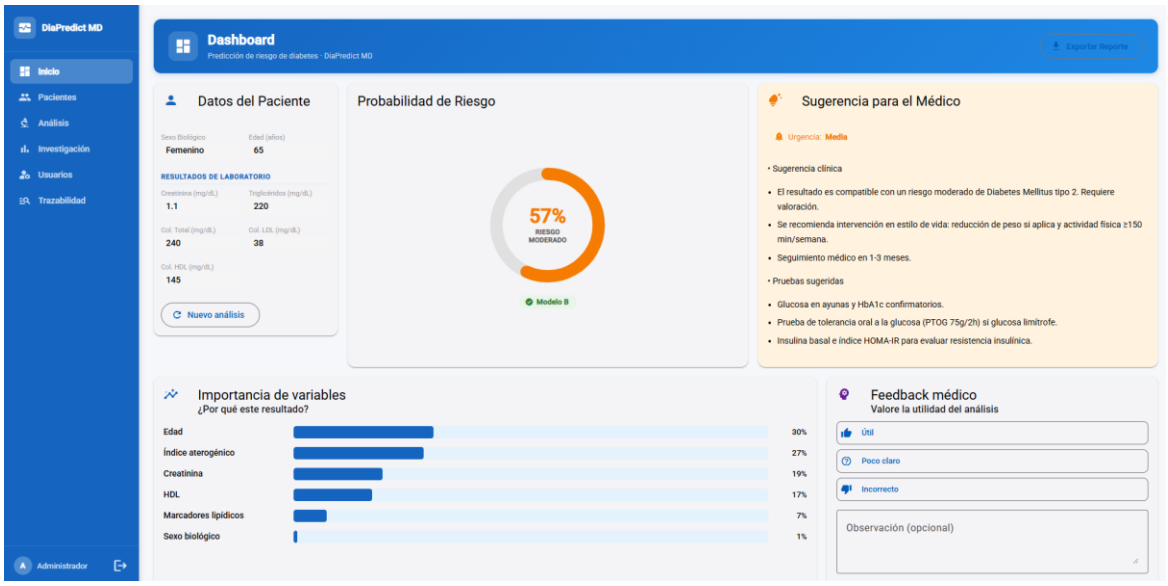


Figura 27. Resultado de análisis con Modelo B: paciente con riesgo moderado (57%).

Ambos casos de uso confirman el correcto funcionamiento de los cuatro componentes centrales del sistema: la selección automática del modelo según disponibilidad de datos, la generación de recomendaciones clínicas contextualizadas mediante LLM, la visualización de la importancia de variables mediante valores SHAP individuales, y el componente de retroalimentación médica integrado en cada análisis.

7.4.2 VALIDACIÓN

La validación del sistema se abordó en esta etapa desde una perspectiva funcional y de plausibilidad clínica preliminar. En primer lugar, se verificó que la herramienta ejecutara correctamente los flujos principales: autenticación, captura de variables, selección automática del modelo, generación de predicción, visualización de importancia de variables mediante SHAP y emisión de recomendaciones contextualizadas mediante LLM. En segundo lugar, se revisó la coherencia clínica general de las salidas del sistema frente a criterios diagnósticos y factores de riesgo documentados en la literatura.

Sin embargo, esta fase no constituye una validación clínica formal en entorno asistencial. La evaluación con profesionales médicos mediante casos clínicos estructurados, escalas de utilidad, medición de concordancia y análisis prospectivo queda planteada como una etapa posterior del proyecto. Esta precisión es importante porque delimita el alcance real del prototipo y evita atribuirle un nivel de evidencia clínica superior al que efectivamente fue alcanzado en esta fase académica.

7.5 DOCUMENTACIÓN TÉCNICA Y MANUAL DE USUARIO

7.5.1 DOCUMENTACIÓN TÉCNICA

El proyecto incluye documentación técnica formal organizada en la carpeta docs/ del repositorio, cubriendo los aspectos arquitectónicos, operativos y contractuales del sistema. La Tabla 24 sintetiza los principales documentos generados, que en conjunto garantizan la trazabilidad, mantenibilidad y reproducibilidad del sistema en entornos de producción o continuación del desarrollo.

Tabla 24. Documentación técnica del proyecto

Documento	Contenido
01_arquitectura.md	Descripción detallada de los componentes, flujos de información y decisiones arquitectónicas
02_plan_sprints.md	Plan de los 8 *sprints* ejecutados con objetivos, requerimientos abordados y entregables
03_modelo_datos.md	Especificación del modelo relacional en SQL Server con entidades, atributos y relaciones
04_contratos_api.md	Especificación de *endpoints* REST con request/response, códigos de estado y ejemplos
08_despliegue.md	Guía operativa de despliegue, configuración de variables y verificación post-despliegue

7.5.2 MANUAL DE USUARIO

Se desarrolló un manual de usuario orientado al profesional médico (ver Anexo 1), estructurado para guiar al clínico a través de cada etapa de interacción con el sistema sin requerir conocimientos técnicos previos. El manual cubre el control documental y el acceso al sistema mediante credenciales, con posibilidad de registro para nuevos usuarios. La sección de análisis describe cómo ingresar los resultados de laboratorio disponibles, con validación de los datos en el formulario antes de su procesamiento. La interpretación de resultados explica al usuario el significado de la probabilidad predicha, la categoría de riesgo (bajo, moderado, alto), las recomendaciones contextualizadas generadas por el LLM y la descomposición de importancia de variables mediante SHAP. Finalmente, el módulo de retroalimentación permite al médico calificar la utilidad del análisis (Útil, Poco claro, Incorrecto) y registrar observaciones libres, alimentando el sistema de mejora continua.

El sistema declara explícitamente su carácter de prototipo de apoyo a la decisión clínica,

incluyendo un aviso legal que enfatiza tres aspectos irrenunciables: la herramienta no constituye un diagnóstico médico definitivo, el resultado debe complementarse con pruebas confirmatorias y juicio clínico profesional, y las recomendaciones se basan en guías de práctica clínica generales que deben adaptarse al contexto específico de cada paciente. Esta declaración explícita delimita el alcance del sistema y establece la responsabilidad clínica en el profesional tratante, no en la herramienta.

8.CONCLUSIONES Y TRABAJOS FUTUROS

8.1. CONCLUSIONES

En relación con el primer objetivo específico, que consistió en identificar y seleccionar las variables clínicas más relevantes provenientes de exámenes de laboratorio rutinarios, se concluye que la implementación de un proceso de selección multimétodo (ANOVA F-test, *Mutual Information*, VIF iterativo y RFECV) permitió identificar un subconjunto robusto y clínicamente justificado de predictores para cada escenario de tamizaje. El hallazgo más determinante fue el análisis VIF, que reveló valores superiores a 10^5 en siete de las once variables lipídicas disponibles, fundamentando la arquitectura de doble modelo como solución necesaria y metodológicamente sólida. En consecuencia, el análisis VIF no implica eliminar variables como `Edad`, Sexo o `Creatinina`; por el contrario, confirma su estabilidad como predictores independientes. La principal advertencia metodológica se concentra en el bloque lipídico, donde la combinación de variables originales y ratios derivados genera multicolinealidad extrema. Por esta razón, el uso del perfil lipídico dentro del Modelo B requiere una selección cuidadosa de variables, evitando redundancias innecesarias y privilegiando predictores con respaldo clínico, bajo solapamiento estadístico y aporte demostrado al desempeño del modelo. El Modelo A quedó configurado con tres variables universalmente disponibles en cualquier sistema LIS (`Edad`, `Sexo Biológico` y `Creatinina`), mientras que el Modelo B incorporó adicionalmente variables del perfil lipídico con VIF controlado. La retención de variables como `BiologicalSex` (F=1.034, p=0.309) y `lipidic_flags` (r=-0.0006 con el `TARGET`) evidencia que la selección de predictores en contextos clínicos puede y debe integrar criterios respaldados por guías ADA junto con consideraciones de multicolinealidad, produciendo modelos más estables, interpretables y clínicamente válidos que los obtenidos por criterios puramente estadísticos.

Respecto al segundo objetivo, que buscaba implementar modelos de aprendizaje automático capaces de estimar la probabilidad de pertenencia a la clase de riesgo asociada a DM2, se concluye que los seis modelos entrenados (Regresión Logística, Random Forest y XGBoost para cada escenario) superaron ampliamente el *baseline* de clasificación aleatoria (AUC=0.492 y 0.521 respectivamente), confirmando la viabilidad del enfoque y el potencial de los datos de laboratorio rutinario como fuente predictiva. El Modelo A seleccionó XGBoost con calibración *Platt Scaling*, alcanzando AUC-ROC de 0.754 [IC95%: 0.736-0.772], sensibilidad de 0.845 y VPN de 0.786 sobre 2.730 pacientes de prueba, superando los criterios de éxito preestablecidos. El Modelo B seleccionó Random Forest regulado, alcanzando AUC=0.701 y VPP=0.809 sobre 230 pacientes, con un aporte independiente del perfil lipídico cuantificado en +0.112 AUC sobre la configuración

del Modelo A. La estabilidad de ambos modelos con 10 semillas distintas (σ=0.012 y σ=0.031 respectivamente) garantiza la reproducibilidad de los resultados. En conjunto, los modelos implementados demuestran que variables de laboratorio rutinario, sin requerir encuestas ni equipos especializados, son suficientes para construir herramientas de tamizaje útiles para la detección temprana de DM2 en la población colombiana.

En relación con el tercer objetivo, que consistió en evaluar el desempeño de los modelos mediante métricas estadísticas pertinentes y validarlos para garantizar su precisión, confiabilidad e interpretabilidad clínica, se concluye que la evaluación fue integral y multidimensional. Las métricas de desempeño (AUC-ROC, sensibilidad, especificidad, VPN, VPP, F1-score y *Brier Score*) se calcularon con intervalos de confianza al 95% mediante *bootstrap* (n=1.000 remuestreos), y la comparación estadística entre algoritmos se realizó mediante *bootstrap* de diferencias de AUC con corrección de *Benjamini-Hochberg,* estableciendo la equivalencia entre XGBoost y Random Forest (ΔAUC=0.002, p=0.481) y la superioridad de ambos sobre Regresión Logística (p<0.001). El *Decision Curve Analysis* confirmó un beneficio neto de 185 casos adicionales detectados por cada 1.000 pacientes tamizados con el Modelo A. La validación clínica se fundamentó en la coherencia entre los valores SHAP y la fisiopatología conocida de la DM2, en el Modelo A la `Edad` concentra el 79.3% de la importancia con el patrón de riesgo etario esperado por la ADA, en el Modelo B el LDL lidera con efecto positivo, el HDL presenta efecto protector negativo y el `Índice Aterogénico` es positivo, todas las direcciones coherentes con la literatura internacional. El análisis de subgrupos etarios identificó áreas de mejora precisas por el desempeño limitado en adultos mayores de 75 años (AUC=0.560 y 0.543), que orientan directamente las líneas de desarrollo futuro y fundamentan el uso combinado de ambos modelos como estrategia de tamizaje secuencial en la herramienta web. Como recomendación explícita derivada de esta evaluación, se considera indispensable que, antes de cualquier uso clínico real de los modelos, sus resultados y las recomendaciones generadas sean revisados y validados formalmente por profesionales de la salud mediante un proceso de validación externa con casos clínicos reales, que confirme la concordancia entre las predicciones del modelo y el juicio médico experto antes de su implementación asistencial.

Finalmente, frente al cuarto objetivo, que buscaba presentar una herramienta que permita representar gráficamente los resultados del modelo predictivo, se concluye que DiaPredict MD materializa de forma completa y funcional los resultados del proceso de modelado, traduciendo la complejidad estadística en información accionable para el profesional médico. La herramienta implementa selección automática del modelo según la disponibilidad de datos del paciente, visualización de la probabilidad de riesgo con categorización en tres niveles (bajo, moderado,

alto), descomposición de la importancia de variables mediante valores SHAP individuales para cada predicción, y generación de recomendaciones clínicas contextualizadas mediante un modelo de lenguaje de gran escala. La validación funcional mediante casos de uso representativos confirmó la correcta operación de la selección automática y la coherencia de las descomposiciones SHAP individuales con los patrones globales documentados. La herramienta declara explícitamente su carácter de prototipo de apoyo a la decisión clínica, complementario al juicio del profesional médico y a los criterios diagnósticos ADA, estableciendo desde su diseño el marco ético y clínico apropiado para su uso responsable.

En conjunto, los hallazgos de este proyecto demuestran que la combinación de rigor metodológico en la selección de variables, algoritmos de aprendizaje automático calibrados y evaluación estadística robusta permite construir herramientas de tamizaje de DM2 clínicamente válidas a partir de datos ya disponibles en los sistemas de información de laboratorio, sin costos adicionales para el sistema de salud ni para el paciente. Los modelos desarrollados no buscan reemplazar el juicio clínico ni los criterios diagnósticos de la ADA, sino ampliar la capacidad del sistema de salud para identificar pacientes en riesgo antes de que el daño orgánico sea irreversible, contribuyendo concretamente a la reducción de la brecha diagnóstica de DM2 en Colombia.

No obstante, los resultados deben interpretarse en el marco de las limitaciones del estudio, las cuales no invalidan los hallazgos, sino que trazan con precisión la hoja de ruta del desarrollo futuro. Desde el punto de vista de los datos, el *dataset* proviene de una fuente única y carece de variables antropométricas (IMC, perímetro abdominal y antecedentes familiares) cuya incorporación representa una oportunidad concreta de incrementar sustancialmente el poder predictivo del Modelo A. Desde el punto de vista metodológico, el tamaño muestral de prueba del Modelo B (n=230) invita a su confirmación mediante validación prospectiva externa; la naturaleza transversal de ambos modelos abre la posibilidad de explotar la trayectoria temporal de los marcadores bioquímicos mediante arquitecturas longitudinales; y la ausencia de análisis de equidad algorítmica establece un paso necesario y valioso antes de cualquier implementación a escala poblacional. Estas limitaciones, lejos de ser obstáculos, definen las condiciones bajo las cuales los modelos pueden aplicarse con confianza clínica hoy y señalan el camino hacia versiones progresivamente más completas, equitativas y generalizables del sistema desarrollado.

8.2. TRABAJOS FUTUROS

Las limitaciones identificadas no representan obstáculos al valor del proyecto sino oportunidades concretas de desarrollo que ampliarán progresivamente su alcance clínico.

La extensión más inmediata y de mayor impacto es la validación prospectiva externa mediante una cohorte de pacientes con panel lipídico completo en instituciones distintas a las del entrenamiento, que permitirá confirmar la generalización del Modelo B y alcanzar la potencia estadística necesaria para la comparación formal contra FINDRISC y contra el Modelo A. En la misma dirección, el análisis de subgrupos etarios abre una línea de investigación prioritaria: el comportamiento de ambos modelos en adultos mayores de 75 años señala una oportunidad de especialización, desarrollando un enfoque diferenciado que incorpore variables de fragilidad, polifarmacia y función cognitiva que capturan mejor la complejidad clínica del envejecimiento y que trascienden el perfil bioquímico rutinario actual.

Desde el punto de vista de los datos disponibles, la integración con la historia clínica electrónica abre la posibilidad de incorporar IMC, perímetro abdominal y antecedentes familiares, habilitando una comparación justa con el FINDRISC original e incrementando sustancialmente el poder predictivo del Modelo A, cuya importancia actualmente concentrada en la `Edad` refleja la ausencia de variables antropométricas más que la irrelevancia de otros factores de riesgo. Complementariamente, explotar la naturaleza temporal del *dataset* LIS permitiría capturar la trayectoria de evolución de los marcadores bioquímicos entre visitas, incorporando la dinámica de progresión de la resistencia insulínica como señal predictiva y superando la limitación de corte transversal de los modelos actuales.

En cuanto a la cobertura diagnóstica, la expansión hacia un modelo de tres clases que incluya la prediabetes como estadio intermedio representa la extensión clínicamente más valiosa a largo plazo, las intervenciones de estilo de vida en este estadio reducen la progresión a DM2 en el 58% de los casos según el *Diabetes Prevention Program*, y los 10.887 pacientes con prediabetes excluidos del *dataset* actual proveen el conjunto de entrenamiento natural para esta extensión, convirtiendo una decisión metodológica del presente estudio en el punto de partida de una investigación futura de alto impacto preventivo. Antes de cualquier implementación a escala poblacional, el análisis de equidad algorítmica por grupos demográficos se presenta como una oportunidad de garantizar que la herramienta amplíe el acceso al diagnóstico de forma equitativa, sin reproducir ni ampliar las desigualdades existentes en el sistema de salud.

9. REFERENCIAS BIBLIOGRÁFICAS

[1] P. Aschner, «Epidemiología de la diabetes en Colombia», Revista Española de Geriatría y Gerontología, vol. 45, Supl. 1, pp. 8–11, 2010

[2] D. A. Quinde Pulla et al., «Diabetes mellitus tipo 2: rol de la insulinorresistencia y la hiperinsulinemia en la etiopatogenia», Zenodo, 2022

[3] L. R. Fernández-Lázaro, A. García Hernández, R. Caballero-García, D. Villaverde-Gutiérrez y C. Mielgo-Ayuso, «Actividad física y riesgo de diabetes tipo 2 en estudiantes universitarios», Medicina, vol. 56, no. 10, p. 496, 2020.

[4] A. G. Tabák, C. Herder, W. Rathmann, E. J. Brunner y M. Kivimäki, «Prediabetes: un estado de alto riesgo para el desarrollo de diabetes», The Lancet, vol. 379, no. 9833, pp. 2279–2290, 2012.

[5] American Diabetes Association Professional Practice Committee for Diabetes, "2. Diagnosis and classification of diabetes: Standards of Care in Diabetes—2026," Diabetes Care, vol. 49, Suppl. 1, pp. S27–S49, Jan. 2026, doi: 10.2337/dc26-S002.

[6] R. A. DeFronzo, "From the triumvirate to the ominous octet: A new paradigm for the treatment of type 2 diabetes mellitus," Diabetes, vol. 58, no. 4, pp. 773–795, Apr. 2009, doi: 10.2337/db09-9028.

[7] D. M. Nathan, J. B. Buse, M. B. Davidson, E. Ferrannini, R. R. Holman, R. Sherwin, and B. Zinman, "Medical manAgement of hyperglycemia in type 2 diabetes: A consensus algorithm for the initiation and adjustment of therapy," Diabetes Care, vol. 32, no. 1, pp. 193–203, Jan. 2009, doi: 10.2337/dc08-9025.

[8] A. D. Mooradian, "Dyslipidemia in type 2 diabetes mellitus," Nature Clinical Practice Endocrinology & Metabolism, vol. 5, no. 3, pp. 150–159, Mar. 2009, doi: 10.1038/ncpendmet1066.

[9] M. Adiels, S.-O. Olofsson, M.-R. Taskinen, and J. Borén, "Overproduction of very low-density lipoproteins is the hallmark of the dyslipidemia in the metabolic syndrome," Arteriosclerosis, Thrombosis, and Vascular Biology, vol. 28, no. 7, pp. 1225–1236, Jul. 2008, doi: 10.1161/ATVBAHA.107.160192.

[10] National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, "Third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III): Final report," Circulation, vol. 106, no. 25, pp. 3143–3421, Dec. 2002.

[11] National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, "Third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III): Final report," Circulation, vol. 106, no. 25, pp. 3143–3421, Dec. 2002.

[12] R. M. Krauss, "Lipids and lipoproteins in patients with type 2 diabetes," Diabetes Care, vol. 27, no. 6, pp. 1496–1504, Jun. 2004, doi: 10.2337/diacare.27.6.1496.

[13] W. P. Castelli, "Cholesterol and lipids in the risk of coronary artery disease—The Framingham Heart Study," Canadian Journal of Cardiology, vol. 4, Suppl. A, pp. 5A–10A, Jul. 1988.

[14] J. Millán, X. Pintó, A. Muñoz, M. Zúñiga, J. Rubiés-Prat, L. F. Pallardo, L. Masana, A. Mangas,

A. Hernández-Mijares, P. González-Santos, J. F. Ascaso, and J. Pedro-Botet, “Lipoprotein ratios: Physiological significance and clinical usefulness in cardiovascular prevention,” Vascular Health and Risk ManAgement, vol. 5, pp. 757–765, 2009.
[15] T. McLaughlin, F. Abbasi, K. Cheal, J. Chu, C. Lamendola, and G. Reaven, “Use of metabolic markers to identify overweight individuals who are insulin resistant,” Annals of Internal Medicine, vol. 139, no. 10, pp. 802–809, Nov. 2003, doi: 10.7326/0003-4819-139-10-200311180-00007.
[16] R. Z. Alicic, M. T. Rooney, and K. R. Tuttle, “Diabetic kidney disease: Challenges, progress, and possibilities,” Clinical Journal of the American Society of Nephrology, vol. 12, no. 12, pp. 2032–2045, Dec. 2017, doi: 10.2215/CJN.11491116.
[17] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group, “KDIGO 2024 clinical practice guideline for the evaluation and manAgement of chronic kidney disease,” Kidney International, vol. 105, no. 4S, pp. S117–S314, Apr. 2024.
[18] International Diabetes Federation, IDF Diabetes Atlas, 11.ª ed., Bruselas, Bélgica: IDF, 2025. [En línea]. Disponible en: https://diabetesatlas.org. Accedido: 20 de mayo de 2025.
[19] Ministerio de Salud y Protección Social, «Día Mundial de la Diabetes», minsalud.gov.co, 14 de noviembre de 2022. [En línea]. Disponible en: https://www.minsalud.gov.co/Paginas/dia-mundial-de-la-diabetes.aspx. Accedido: 24 de mayo de 2025.
[20] Secretaría de Salud e Instituto Nacional de Salud Pública, Boletín de Práctica Médica Efectiva: Diabetes mellitus tipo 2 (DM2), vol. PME, México, agosto de 2006. [En línea]. Disponible en: https://www.insp.mx/resources/imAges/stories/Centros/ciss/pme/diabetes.pdf. Accedido: 27 de mayo de 2025. 24
[21] Diabetes Prevention Program Research Group, “Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study,” The Lancet Diabetes & Endocrinology, vol. 3, no. 11, pp. 866–875, Nov. 2015.
[22] E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” Nature Medicine, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.
[23] Oracle, «Desmitificando el aprendizaje automático: una guía completa». [En línea]. Disponible en: https://www.oracle.com/co/artificial-intelligence/machine-learning/what-is-machine-learning/. Accedido: 25 de mayo de 2025.
[24] A. L. Beam and I. S. Kohane, “Big data and machine learning in health care,” JAMA, vol. 319, no. 13, pp. 1317–1318, Apr. 2018, doi: 10.1001/jama.2017.18391
[25] MIT Sloan, «Machine learning, explained | MIT Sloan». [En línea]. Disponible en: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained. Accedido: 27 de mayo de 2025.
[26] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.
[27] Y. Wang et al., “A high-performance model for prediction of type 2 diabetic retinopathy using machine learning,” Scientific Reports, vol. 14, 2024
[28] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning

Research, vol. 12, pp. 2825–2830, 2011.
[29] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[30] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[31] IBM, «Aprendizaje automático (machine learning)». [En línea]. Disponible en: https://www.ibm.com/mx-es/topics/machine-learning. Accedido: 29 de mayo de 2025.
[32] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” PLOS ONE, vol. 10, no. 3, Art. no. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.
[33] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” Radiology, vol. 143, no. 1, pp. 29–36, Apr. 1982, doi: 10.1148/radiology.143.1.7063747.
[34] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, “Assessing the performance of prediction models: A framework for traditional and novel measures,” Epidemiology, vol. 21, no. 1, pp. 128–138, Jan. 2010, doi: 10.1097/EDE.0b013e3181c30fb2.
[35] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.
[36] A. J. Vickers and E. B. Elkin, “Decision curve analysis: A novel method for evaluating prediction models,” Medical Decision Making, vol. 26, no. 6, pp. 565–574, Nov.–Dec. 2006, doi: 10.1177/0272989X06295361.
[37] Asamblea Nacional Constituyente, Constitución Política de Colombia. Bogotá, Colombia: Gaceta Constitucional No. 116, 1991.
[38] Congreso de la República de Colombia, Ley Estatutaria 1581 de 2012, "Por la cual se dictan disposiciones generales para la protección de datos personales". Bogotá, Colombia: Diario Oficial No. 48.587, 18 de octubre de 2012.
[39] Ministerio de Comercio, Industria y Turismo, Decreto 1377 de 2013, "Por el cual se reglamenta parcialmente la Ley 1581 de 2012". Bogotá, Colombia: Diario Oficial No. 48.834, 27 de junio de 2013.
[40] T. L. Beauchamp y J. F. Childress, Principles of Biomedical Ethics, 8th ed. New York, NY, USA: Oxford University Press, 2019.
[41] World Health Organization, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva, Switzerland: WHO Press, 2021. [En línea]. Disponible: https://www.who.int/publications/i/item/9789240029200
[42] X. Li, F. Ding, L. Zhang, et al., “Interpretable machine learning method to predict the risk of pre-diabetes using national-wide cross-sectional data: evidence from CHNS,” BMC Public Health, vol. 25, no. 1145, 2025.
[43] D. Lindholm, P. Jernberg, M. Melhus, et al., “Machine learning for prediction of incident diabetes: The Swedish National Diabetes Register,” PLOS ONE, vol. 15, no. 10, e0238708, 2020.

[44] S. Khedkar, N. Patil, and A. Waghmare, “Efficient prediction of diabetes using machine learning classifiers and oversampling techniques,” Materials Today: Proceedings, vol. 47, no. 9, pp. 1248–1255, 2021.
[45] J. Cardozo et al., “Detection of type 2 diabetes using laboratory data and machine learning,” Frontiers in Artificial Intelligence, vol. 5, 2022.
[46] R. Kopitar et al., “Early detection of type 2 diabetes mellitus using machine learning-based prediction models,” Scientific Reports, vol. 10, no. 1, 2020. 25
[47] Y. Zhang et al., “Community-based early warning prediction model for diabetes using big data mining,” Primary Care Diabetes, vol. 17, no. 3, 2023.
[48] D. Silva et al., “Machine learning models for the prediction of type 2 diabetes: a systematic review,” Diabetology & Metabolic Syndrome, vol. 13, 2021.
[49] S. Kim et al., “Prediction of future diabetes using machine learning techniques,” Diabetes & Metabolism Journal, vol. 47, no. 3, 2023.
[50] M. Choudhury et al., “Top 10 predictors of type 2 diabetes in the UK Biobank cohort identified by machine learning,” Scientific Reports, vol. 14, 2024.
[51] B. S. Sousa, A. N. Santos, L. L. de Oliveira, et al., “Risk prediction model for Type 2 Diabetes using clinical and laboratory variables: a population-based study in Brazil,” Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 15, no. 2, pp. 589–595, 2021.
[52] J. A. Mejía, M. A. Oviedo-Benalcázar, J. A. Ordoñez, y J. F. Valencia-Murillo, “Aprendizaje automático aplicado a la predicción de diabetes mellitus, utilizando información socioeconómica y ambiental de usuarios del sistema de salud,” Rev. Fac. Nac. Salud Pública, vol. 41, no. 2, e06, 2023.
[53] J. Lindström and J. Tuomilehto, "The diabetes risk score: a practical tool to predict type 2 diabetes risk," Diabetes Care, vol. 26, no. 3, pp. 725-731, 2003.
[54] A. Bernabé-Ortiz, et al., "Diagnostic accuracy of the Finnish Diabetes Risk Score (FINDRISC) for undiagnosed T2DM in Peruvian population," Primary Care Diabetes, vol. 12, no. 6, pp. 517-525, 2018.
[55] S. M. Lundberg, et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56-67, 2020.
[56] J. Yang, et al., "SHAP-based interpretation of XGBoost predictions for diabetes risk in Chinese adults," Frontiers in Endocrinology, vol. 13, 2022.
[57] T. McLaughlin, F. Abbasi, K. Cheal, J. Chu, C. Lamendola, and G. Reaven, “Use of metabolic markers to identify overweight individuals who are insulin resistant,” *Annals of Internal Medicine*, vol. 139, no. 10, pp. 802–809, Nov. 2003, doi: 10.7326/0003-4819-139-10-200311180-00007.
[58] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York, NY, USA: John Wiley & Sons, 1980.
[59] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003.
[60] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2013.
[61] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[62] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[63] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 1995, vol. 2, pp. 1137–1143.
[64] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[65] X. Bouthillier et al., “Accounting for variance in machine learning benchmarks,” Proceedings of Machine Learning and Systems, vol. 3, pp. 747–769, 2021.
[66] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” PLOS ONE, vol. 10, no. 3, Art. no. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.
[67] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993
[68] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach,” Biometrics, vol. 44, no. 3, pp. 837–845, Sep. 1988, doi: 10.2307/2531595.
[69] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentAges,” Psychometrika, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/BF02295996.
[70] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, “Assessing the performance of prediction models: A framework for traditional and novel measures,” Epidemiology, vol. 21, no. 1, pp. 128–138, Jan. 2010, doi: 10.1097/EDE.0b013e3181c30fb2.
[71] A. J. Vickers and E. B. Elkin, “Decision curve analysis: A novel method for evaluating prediction models,” Medical Decision Making, vol. 26, no. 6, pp. 565–574, Nov.–Dec. 2006, doi: 10.1177/0272989X06295361.
[72] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 2017, pp. 4765–4774.
[73] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality: Complete samples,” Biometrika, vol. 52, no. 3/4, pp. 591–611, Dec. 1965, doi: 10.2307/2333709.
[74] F. J. Massey, Jr., “The Kolmogorov-Smirnov test for goodness of fit,” Journal of the American Statistical Association, vol. 46, no. 253, pp. 68–78, Mar. 1951, doi: 10.1080/01621459.1951.10500769.
[75] D. G. Altman and J. M. Bland, “Parametric v non-parametric methods for data analysis,” BMJ, vol. 338, p. a3167, Jan. 2009, doi: 10.1136/bmj.a3167.
[76] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” The Annals of Mathematical Statistics, vol. 18, no. 1, pp. 50–60, Mar. 1947, doi: 10.1214/aoms/1177730491.
[77] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ, USA:

Lawrence Erlbaum Associates, 1988.
[78] H. Cramér, Mathematical Methods of Statistics. Princeton, NJ, USA: Princeton University Press, 1946.
[79] M. M. Mukaka, “A guide to appropriate use of correlation coefficient in medical research,” Malawi Medical Journal, vol. 24, no. 3, pp. 69–71, Sep. 2012.
[80] D. A. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York, NY, USA: John Wiley & Sons, 1980.

ANEXOS

Anexo 1. Permiso de la institución para acceso y uso de datos de pacientes.

Bogotá, 02 de junio de 2025

IT HEALTH

Gerencia General

A QUIEN CORRESPONDA:

En mi calidad de Gerente de IT HEALTH, yo, José Palencia Téllez, por medio de la presente hago constar que la institución otorga el aval y la autorización formal a los estudiantes Natalia Alejandra Macana Díaz, identificada con cédula de ciudadanía No. 1.032.490.745, e Iván René Valderrama Corredor, identificado con cédula de ciudadanía No. 1.050.200.857, para el uso y tratamiento de datos provenientes del sistema LIS, correspondientes al periodo comprendido entre junio de 2025 y marzo de 2026.

Esta autorización se concede con el fin exclusivo de desarrollar el trabajo de investigación para optar al título de Magíster en Ciencia de Datos, titulado: "MODELO DE *MACHINE LEARNING* PARA LA PREDICCIÓN DE DIABETES MELLITUS TIPO 2, BASADO EN DATOS DE EXÁMENES DE LABORATORIO RUTINARIOS".

El uso de dicha base de datos queda sujeto a las siguientes consideraciones:

- Confidencialidad: Los investigadores deberán garantizar el anonimato de los pacientes, asegurando que no se divulgará información que permita su identificación directa o indirecta, en cumplimiento de la Ley 1581 de 2012 y demás normas aplicables sobre protección de datos personales.
- Uso académico: Los datos serán utilizados estrictamente con fines académicos, investigativos, de modelado predictivo y análisis estadístico, dentro del marco del trabajo de grado mencionado.
- Ética y responsabilidad: El proyecto deberá seguir los lineamientos éticos aplicables para el manejo responsable de información clínica, garantizando la reserva, seguridad y uso adecuado de los datos suministrados.

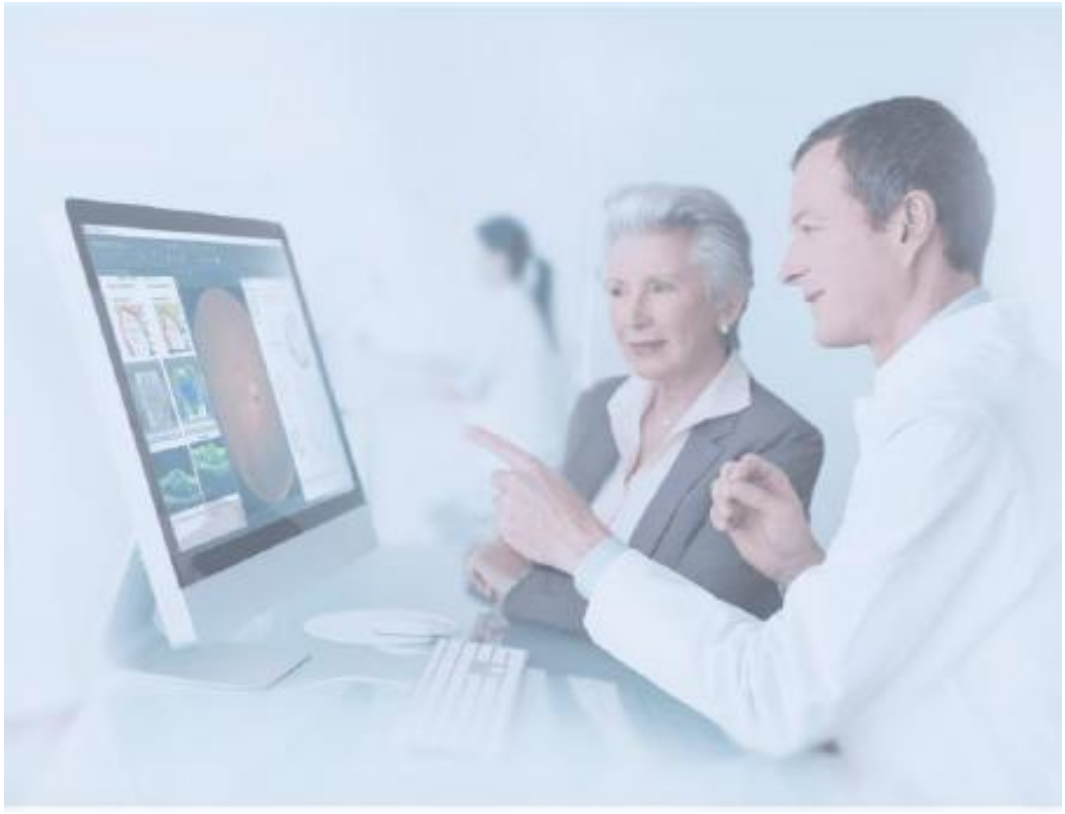
Reconocemos la importancia de este estudio para fortalecer el uso de herramientas analíticas y modelos de *machine learning* en el apoyo a la identificación temprana del riesgo de Diabetes Mellitus Tipo 2, a partir de datos derivados de exámenes de laboratorio rutinarios.

Se expide la presente constancia a solicitud de los interesados, para los fines académicos correspondientes.

Atentamente,

José Palencia Téllez
Gerente IT HEALTH

Anexo 2. Manual de Usuario



MANUAL DE USUARIO

DIAPREDICT MD

Versión del manual 1.0.0
01 – Mayo -2026

Índice

1 | INTRODUCCIÓN

El presente manual tiene como objetivo proporcionar una guía clara y detallada sobre el uso de DiaPredict MD, un aplicativo diseñado para optimizar la gestión de los procesos de diagnóstico preventivo de Diabetes Mellitus. Esta herramienta ha sido desarrollada para atender las necesidades del personal médico, facilitando el almacenamiento, procesamiento y administración de información clave mediante modelos predictivos de Machine Learning.

DiaPredict MD cuenta con diversos módulos que permiten gestionar diferentes áreas dentro de la consulta clínica.

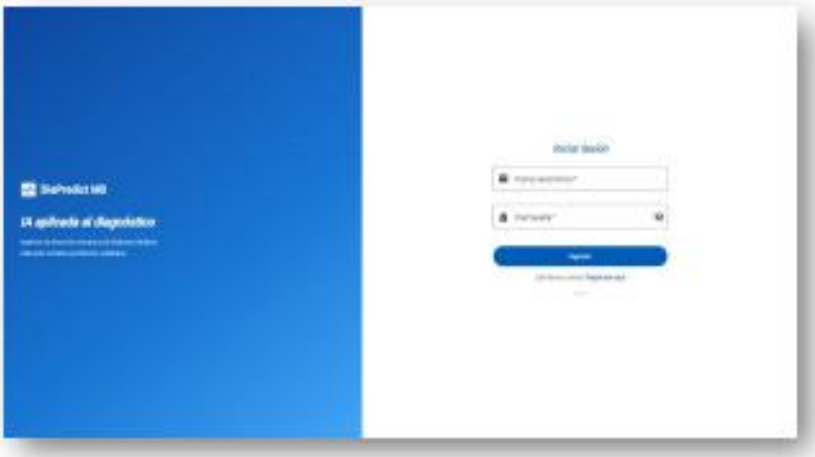
2 | ACCESO AL SISTEMA

Gestión de credenciales

Este módulo permite al profesional de la salud autenticarse de forma segura en la plataforma, asegurando la trazabilidad de cada consulta realizada.

- Registro de usuario: Para nuevos usuarios, el sistema permite la creación de una cuenta profesional a través del enlace "Regístrate aquí".

- Inicio de sesión: Ingrese su correo y contraseña. El sistema cuenta con encriptación de seguridad.

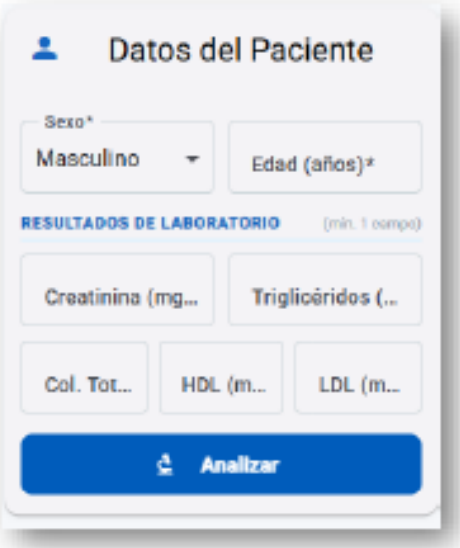


3 | ANÁLISIS

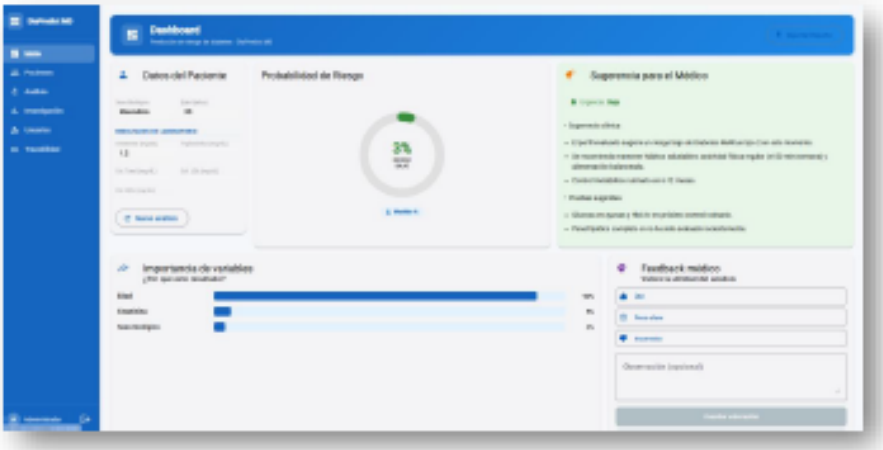
Formulario clínico

Este módulo permite registrar y configurar las variables clínicas del paciente.

- Datos del paciente: Sexo biológico y Edad (Variables obligatorias).
- Resultados de laboratorio: Ingrese valores de Creatinina, Triglicéridos, Colesterol HDL, Colesterol LDL y Colesterol Total.



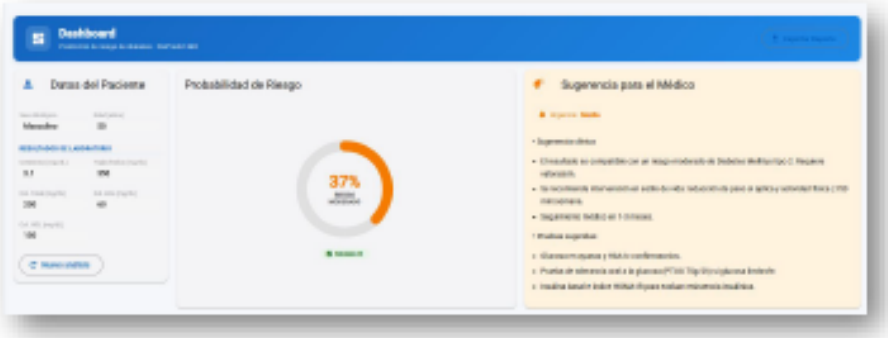
- Botón Analizar: Ejecuta el modelo una vez cargados los datos.



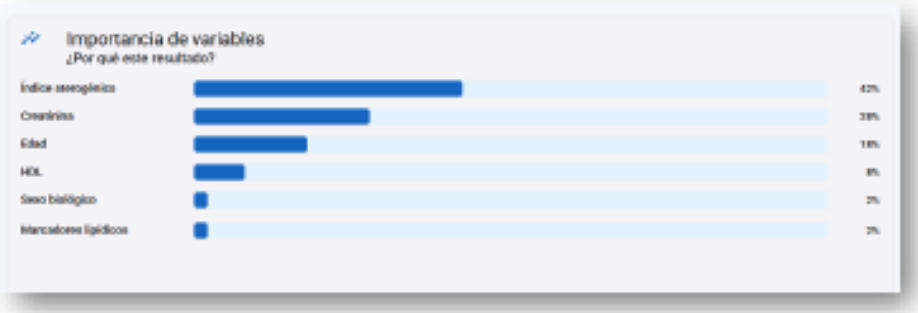
4 | INTERPRETACIÓN DE RESULTADOS

Soporte a la Decisión Clínica

- Probabilidad de Riesgo: Gráfico radial que indica el nivel porcentual de riesgo.
- Sugerencia Médica: Recomendaciones clínicas automatizadas según guías de práctica.



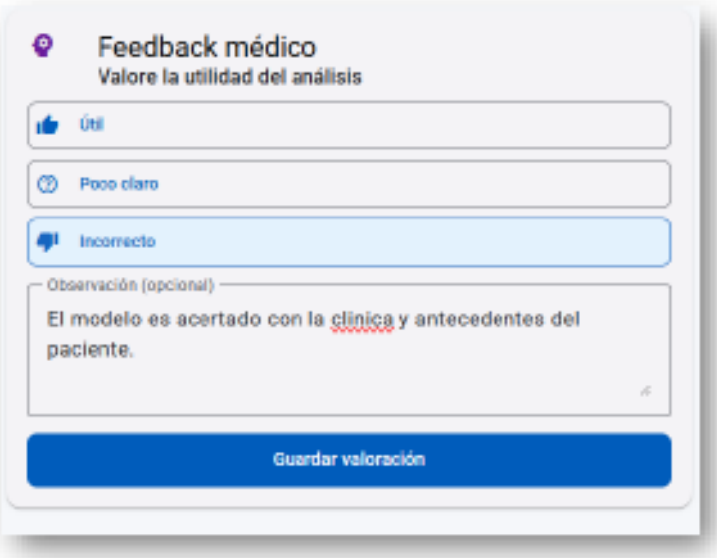
- Importancia de Variables: Gráfico de barras que explica el peso de cada dato en el diagnóstico.



5 | FEEDBACK Y REPORTES

Mejora Continua

- Valoración: Seleccione Útil, Poco claro o Incorrecto.



- Exportar: Descargue el reporte en formato PDF para la historia clínica.

6 | HISTORIAL DE ACTUALIZACIONES

Versión	Fecha	Descripción del cambio
1.0.0	01/05/2026	Creación inicial del manual para DiaPredict MD.