Detección Temprana de Diabetes MellitusTipo 2

Mediante Machine Learning Aplicado a Datos Biométricos.

Alumno: Figueiras Manuel Angel

Universidad Abierta Interamericana

Profesor/es trabajo final: Sartorio Alejandro

Plan de Trabajo Final de Carrera presentado para obtener el título de Licenciado en

Gestión de Tecnología Informática

(Noviembre, 2025)

Resumen.

El presente trabajo analiza el potencial del Machine Learning para mejorar la detección temprana de la diabetes mellitus tipo 2 mediante el uso de datos biométricos. A partir de una revisión sistemática de la literatura y del diseño conceptual de un modelo predictivo basado en datos representativos de la población argentina, se comparan enfoques, variables relevantes y métricas de desempeño reportadas por la evidencia científica. Los resultados muestran que algoritmos como Random Forest, XGBoost y Redes Neuronales alcanzan altos niveles de sensibilidad y capacidad discriminativa, lo que evidencia su utilidad para la identificación temprana del riesgo de padecer diabetes mellitustipo 2. Asimismo, el estudio destaca oportunidades y desafíos para su adopción en el sistema de salud argentino, especialmente en relación con la disponibilidad, estandarización y calidad de los datos clínicos.

Palabras clave: aprendizaje automático, datos biométricos, diabetes mellitus, detección temprana, modelos predictivos, salud digital.

Abstract.

The present study analyzes the potential of Machine Learning to improve the early detection of type 2 diabetes mellitus through the use of biometric data. Based on a systematic literature review and the conceptual design of a predictive model built from data representative of the Argentine population, the research compares methodological approaches, relevant variables, and performance metrics reported in scientific evidence. The findings show that algorithms such as Random Forest, XGBoost, and Neural Networks achieve high levels of sensitivity and discriminative capacity, demonstrating their usefulness for the early identification of type 2 diabetes mellitus risk. Additionally, the study highlights opportunities and challenges for the adoption of these techniques within the Argentine healthcare system, particularly regarding the availability, standardization, and quality of clinical data.

Keywords: biometric data, digital health, diabetes mellitus; early detection; machine learning; predictive models.

Índice General

Resumen. 2

Abstract. 3

Índice General 4

Índice de Ilustraciones 7

Índice de Tablas 8

1. Capitulo 1: Introducción. 9

1.1. Planteamiento del Problema de Investigación. 9

1.2. Justificación de la Elección del Problema. 9

1.3. Objetivos del Trabajo Final. 10

1.3.1. Objetivo General. 10

1.3.2. Objetivos Específicos. 10

1.4. Preguntas de Investigación. 10

1.5. Hipótesis del Problema. 11

1.6. Metodología de investigación. 11

1.6.1. Técnicas de Recolección y Análisis de Datos. 11

1.7. Estructura General del Trabajo Final. 12

1.7.1. Capítulos. 12

1.7.2. Acronimos. 13

1.7.3. Referencias. 13

1.8. Estado del Arte. 13

2. Capítulo 2: Marco Teórico. 17

2.1. Introducción. 17

2.2. Fundamentos Clínicos: Diabetes y Variables de Riesgo. 18

2.2.1. Definición de la Diabetes. 18

2.2.2. Prevalencia de la Diabetes Mellitus 2. 19

2.2.3. La Importancia de la Detección Temprana y la Prediabetes. 21

2.2.4. Datos Biométricos y Factores Predictivos. 22

2.3. Fundamentos Tecnológicos: Aprendizaje Automático (ML). 23

2.3.1. Inteligencia Artificial y Machine Learning. 23

2.3.2. Algoritmos de Aprendizaje Supervisado. 25

2.4. Modelización Predictiva en el Contexto Clínico. 26

2.4.1. Métricas de Evaluación Clave (Sensibilidad vs. Precisión). 26

2.4.2. La Interpretación de los Modelos (Explainable AI - XAI). 28

3. Capítulo 3: Diseño Experimental y Metodología 30

3.1. Introducción. 30

3.2. Diseño Conceptual del Dataset. 30

3.3. Preprocesamiento y Limpieza de Datos. 32

3.3.1. Estrategia de Imputación. 32

3.3.2. Transformación de Variables. 32

3.3.3. Partición Estratificada. 33

4. Capítulo 4: Resultados y Evaluación 34

4.1. Introducción. 34

4.2. Resultados de la Evaluación Inferencial y Selección del Modelo Óptimo. 34

4.2.1. Identificación del Modelo Óptimo. 35

4.2.2. Justificación Clínica del Rendimiento. 36

4.3. Análisis e Interpretación del Modelo Optimo. 36

4.3.1. Análisis de la Aplicación Conceptual de SHAP. 37

4.3.2. Implicación de la Transparencia y la Confianza Clínica. 37

4.4. Discusión Crítica: Aplicabilidad y Desafíos en el Contexto Argentino. 38

4.4.1. El Contraste: Éxito Predictivo vs. Desafío de Datos. 38

4.4.2. Recomendación para la Integración: Estandarización Nacional. 39

5. Capítulo 5: Conclusiones. 40

6. Acrónimos. 41

7. Referencias. 43

Índice de Ilustraciones

Ilustración 1: Proyección de Diabetes en la Población Argentina. 20

Índice de Tablas

Tabla 1: Grupos y Tipos de Variables Identificadas en la Encuesta Nacional de Factores de Riesgo. 31

Tabla 2: Rendimiento de los Modelos Propuestos. 35

1. Capitulo 1: Introducción.

1.1. Planteamiento del Problema de Investigación.

La diabetes tipo 2 (DT2) es una de las enfermedades crónicas de mayor prevalencia en Argentina y el mundo, representando un problema creciente para los sistemas de salud. Su detección temprana es fundamental para prevenir complicaciones asociadas, pero los métodos tradicionales dependen de controles clínicos periódicos, que suelen realizarse cuando la enfermedad ya está avanzada.

El desarrollo de tecnologías digitales y la disponibilidad de datos biométricos provenientes de dispositivos médicos, instrumentos clínicos y wearables permiten generar nuevas herramientas predictivas. Las técnicas de Machine Learning (ML) pueden identificar patrones tempranos en variables fisiológicas y de estilo de vida, mejorando la capacidad diagnóstica.

Sin embargo, aún existen desafíos respecto a la precisión, interpretabilidad y aplicabilidad clínica de estos modelos en la población argentina. En este contexto, surge la necesidad de evaluar el potencial del Machine Learning para detectar tempranamente el riesgo de diabetes tipo 2 a partir de datos biométricos.

1.2. Justificación de la Elección del Problema.

Con la llegada de la digitalización a todos los sectores y la evolución de las tecnologías a través de los datos se propone una investigación que busque aplicar estas nuevas tecnologías a los datos que se manipulan en el ámbito médico, para ayudar a los profesionales y la población,

generando herramientas que faciliten y optimicen los tiempos en el diagnostico de diabetes tipo 2. Lo que permitiría mejorar la calidad de vida de muchas personas y de manera predictiva tratarlas como base en las nuevas generaciones.

1.3. Objetivos del Trabajo Final.

1.3.1. Objetivo General.

Evaluar la aplicabilidad de modelos de Machine Learning para la detección temprana de Diabetes Tipo 2 a partir de datos biométricos, considerando su potencial implementación en el sistema de salud argentino.

1.3.2. Objetivos Específicos.

El trabajo final busca cumplir objetivos específicos que se heredan del objetivo general, ellos son:

- Analizar modelos predictivos basados en aprendizaje automático que permitan identificar patrones asociados a diabetes tipo 2.
- Evaluar el impacto potencial de la implementación de estas herramientas en la mejora de la calidad de vida de las personas.

1.4. Preguntas de Investigación.

¿De qué manera las técnicas de aprendizaje automático aplicadas a datos biométricos contribuyen a la detección temprana de diabetes tipo 2?

¿Qué nivel de precisión alcanzan los modelos de aprendizaje automático en la detección temprana de diabetes tipo 2 mediante datos biométricos?

1.5. Hipótesis del Problema.

El uso de técnicas de aprendizaje automático en datos biométricos mejora la detección temprana de diabetes tipo 2, optimizando la precisión y la eficiencia del diagnóstico médico.

Los modelos de aprendizaje automático alcanzan un nivel de precisión superior al 85% en la detección temprana de diabetes tipo 2 a partir de datos biométricos.

1.6. Metodología de investigación.

La presente investigación adopta un enfoque exploratorio-descriptivo orientado a analizar el potencial del Machine Learning (ML) para la detección temprana de la Diabetes Mellitus Tipo 2 (DM2) mediante el uso de datos biométricos. Dado que el estudio no implementa experimentalmente los modelos, sino que evalúa evidencia existente y estructura conceptualmente una propuesta metodológica, la metodología combina dos componentes centrales: revisión sistemática de la literatura y diseño conceptual de un modelo predictivo contextualizado al sistema de salud argentino.

1.6.1. Técnicas de Recolección y Análisis de Datos.

La presente investigación adopta un enfoque exploratorio-descriptivo orientado a analizar el potencial del Machine Learning para la detección temprana de la Diabetes Mellitus Tipo 2. La metodología combina dos componentes centrales:

Revisión Sistemática de la Literatura: Evaluación de la evidencia existente sobre el desempeño de algoritmos en estudios previos.

Diseño Conceptual: Estructuración de una propuesta metodológica conceptual para un modelo predictivo contextualizado al sistema de salud argentino.

El análisis se centra en evaluar la capacidad de las técnicas de Machine Learning para identificar patrones tempranos en variables fisiológicas y de estilo de vida. El análisis incluye la comparación conceptual de algoritmos de aprendizaje supervisado como Random Forest, XGBoost y Redes Neuronales.

El análisis se enfoca en:

- Evaluar el potencial para la identificación temprana del riesgo de padecer diabetes.
- Determinar el nivel de precisión y interpretabilidad alcanzado por los modelos.
- Confirmar que los algoritmos logran altos niveles de sensibilidad y capacidad discriminativa, cumpliendo con el objetivo de optimizar la eficiencia del diagnóstico médico

1.7. Estructura General del Trabajo Final.

El Trabajo Final se organiza en capítulos que permiten una lectura progresiva desde el planteamiento del problema hasta las conclusiones y propuestas futuras. Cada sección se articula con las preguntas de investigación, las hipótesis y el marco teórico desarrollado.

1.7.1. Capítulos.

Capítulo 1: Introducción

Presenta el problema, los objetivos, las preguntas de investigación, las hipótesis y el alcance del estudio.

Capítulo 2: Marco Teórico

Desarrolla los fundamentos clínicos de la DM2, los datos biométricos, la inteligencia artificial aplicada a salud, los algoritmos de ML y la metodología de ciencia de datos.

Capítulo 3: Propuesta Metodológica

Describe el diseño conceptual del dataset, el preprocesamiento teórico, el modelado y las estrategias de validación basadas en literatura científica.

Capítulo 4: Resultados y Análisis

Presenta resultados simulados sustentados en evidencia previa, interpretación mediante XAI y análisis crítico de aplicabilidad al sistema de salud argentino.

Capítulo 5: Conclusiones y Líneas Futuras

Integra los principales hallazgos, reflexiona sobre el potencial del ML en Argentina.

1.7.2. Acronimos.

Lista ordenada alfabéticamente de las abreviaturas utilizadas en el trabajo, acompañadas de su significado.

1.7.3. Referencias.

Listado completo de fuentes bibliográficas empleadas en el trabajo.

1.8. Estado del Arte.

La aplicación de técnicas de *Machine Learning* (ML) en la detección temprana de Diabetes Mellitus Tipo 2 (DM2) ha experimentado un crecimiento significativo en los últimos años, impulsado por el aumento de datos biométricos disponibles y la necesidad de diagnósticos más oportunos y precisos. Diversos estudios en Latinoamérica, Europa y Argentina muestran cómo los modelos predictivos pueden anticipar el riesgo de DM2 a partir de variables fisiológicas, metabólicas y de estilo de vida.

En México, (De la Rosa-De León, Navarro-Acosta, & García-Calvillo, 2025) desarrollaron un sistema de prediagnóstico de enfermedades crónicas mediante cómputo inteligente, demostrando que algoritmos como Random Forest, SVM y Redes Neuronales presentan un rendimiento superior frente a métodos tradicionales al analizar datos biométricos y clínicos tempranos. De manera complementaria, el estudio "Aprendizaje automático aplicado a la detección temprana de DM2: Caso Saltillo" (Lianmoy & Toledo, 2024) confirmó la efectividad del ML en la clasificación del riesgo mediante parámetros como glucosa, presión arterial y actividad física.

En el ámbito latinoamericano, investigaciones de la Universidad del Norte (Marín Ortega & Parra Faria, 2025) y la Universidad Católica de Santa María (Berrios Zuniga, 2024) evaluaron modelos supervisados para predecir diabetes a partir de encuestas de salud y mediciones fisiológicas, concluyendo que algoritmos basados en árboles de decisión y redes neuronales logran precisiones superiores al 85%. Estos trabajos destacan la importancia del preprocesamiento de datos biométricos y la necesidad de bases de datos balanceadas para mejorar la validez de los modelos.

A nivel regional, el estudio publicado en (Berrios Zuniga, 2024) exploró métodos de clasificación para DM2 utilizando datos estructurados y no estructurados, resaltando el aporte de técnicas de optimización y selección de características para mejorar la predicción del riesgo. Asimismo, la Universidad Nacional de La Plata (UNLP) ha producido dos trabajos relevantes: una tesis orientada a la predicción de enfermedades metabólicas mediante ML, y el reconocido estudio de (Tittarelli, 2023) , quien desarrolló modelos específicos para la población argentina basados en la Encuesta Nacional de Factores de Riesgo (ENFR). Sus resultados mostraron que

Random Forest y Redes Neuronales son los enfoques más consistentes para identificar casos de prediabetes y diabetes tipo 2.

En una línea similar, (Perdomo & Ordinez, 2024) analizaron datos de la Encuesta Nacional de Factores de Riesgo (ENFR) aplicando algoritmos supervisados para identificar variables asociadas al riesgo de diabetes en la provincia de Chubut. El estudio se orienta principalmente a fortalecer la toma de decisiones en salud pública.

En Europa, trabajos como los publicados por (Galán Maroto, 2025) analizan también la aplicabilidad del ML aplicado a DM2 desde una perspectiva clínico-metodológica. Estos aportes enfatizan el uso de técnicas explicables (XAI) como SHAP para interpretar la importancia de los factores predictivos, una tendencia necesaria para el uso clínico real.

En el contexto argentino, un estudio reciente publicado por (Dieuzeide, et al., 2025) revisa los factores de riesgo poblacionales asociados al desarrollo de DM2, proporcionando una base epidemiológica esencial para la calibración de modelos predictivos. La combinación de estos factores con algoritmos de ML ha mostrado resultados prometedores para identificación temprana, especialmente cuando se incorporan variables biométricas provenientes de dispositivos móviles o registros clínicos electrónicos.

En conjunto, los trabajos analizados coinciden en que los modelos de ML permiten anticipar la aparición de DM2 con altos niveles de precisión, siempre que se disponga de datos biométricos consistentes y un adecuado tratamiento de valores faltantes. Sin embargo, persisten desafíos como la baja disponibilidad de bases de datos locales amplias, la necesidad de validación clínica en escenarios reales y la inclusión de técnicas explicables que permitan a los profesionales de la salud interpretar adecuadamente los resultados. Los avances actuales

muestran un camino sólido hacia herramientas predictivas que podrían integrarse en sistemas de salud para mejorar la prevención y diagnóstico oportuno de diabetes en poblaciones diversas.

Sin embargo, persisten desafíos clave: la necesidad de bases de datos más amplias y diversas, la integración de técnicas explicables (XAI) y la validación empírica de modelos en entornos clínicos reales.

2. Capítulo 2: Marco Teórico.

2.1. Introducción.

El presente capítulo tiene como objetivo fundamental establecer la base conceptual y metodológica que sustenta la investigación. La detección temprana de Diabetes Mellitus Tipo 2 (DM2) es un problema que exige un entendimiento dual: la complejidad biológica de la enfermedad y la complejidad técnica de la herramienta predictiva. Por lo tanto, el marco teórico se articula en tres pilares principales: los fundamentos clínicos, los fundamentos tecnológicos y el puente de modelización y evaluación.

En los fundamentos clínicos se definirá la Diabetes Mellitus Tipo 2 (DM2), enfatizando la Prediabetes (PDM) como el estadio principal para la intervención preventiva. Se justificará la selección de las variables biométricas y de comportamiento (IMC, Edad, antecedentes) como los factores de riesgo más influyentes, según la evidencia clínica y la priorización algorítmica.

La expedición de los fundamentos tecnológicos se realizará mediante la definición de la Inteligencia Artificial (IA) y su rama más prometedora, el Machine Learning (ML). Se profundizará en el paradigma del Aprendizaje Supervisado, justificando por qué el problema de la DM2 es una tarea de clasificación. Finalmente, se presentarán los algoritmos seleccionados en la investigación (Random Forest, XGBoost y Redes Neuronales Artificiales) y su relevancia en el análisis de grandes datasets tabulados.

La Modelización Predictiva, conecta la tecnología con la utilidad clínica. Se establecerán las Métricas de Evaluación, priorizando la Sensibilidad (Recall) como el indicador clave para minimizar el riesgo clínico y social del Falso Negativo (FN) en la detección temprana.

Finalmente, se abordará el concepto de la Inteligencia Artificial Explicable (XAI) como sustento de interpretación clínica de datos complejos.

2.2. Fundamentos Clínicos: Diabetes y Variables de Riesgo.

2.2.1. Definición de la Diabetes.

La diabetes es un trastorno metabólico considerado crónico, la forma más común es la Diabetes Mellitus Tipo 2 (DM2) constituyendo más del 90% del total de casos diagnosticados (Colmenero Gómez-Cambronero, 2024).

La DM2 es caracterizada por la presencia sostenida de niveles elevados de glucosa en la sangre (hiperglucemia persistente) (Tittarelli, 2023). Este desequilibrio glucémico se debe a una doble problemática fisiológica que afecta la capacidad del cuerpo para metabolizar el azúcar de manera eficiente, lo que se produce:

- Resistencia a la Insulina: Las células del organismo, principalmente del músculo, el hígado y el tejido adiposo, se vuelven resistentes a los efectos de la insulina. Esto impide que la glucosa ingrese a las células para convertirse en energía.
- Producción Insuficiente: El páncreas en principio intenta compensar la resistencia produciendo cantidades mayores de insulina (hiperinsulinemia); sin embargo, con el tiempo y el esfuerzo continuo, las células beta del páncreas se desgastan y pierden gradualmente la capacidad de secretar la insulina suficiente para mantener los niveles de glucemia normales.

El desarrollo de la DM2 es un proceso lento y progresivo, a menudo condicionado por la interacción de una predisposición genética con factores ambientales y de comportamiento. Es frecuente que los síntomas sean sutiles o incluso inexistentes en las primeras fases, lo que

provoca un diagnóstico tardío de hasta 7 años después de la aparición de la hiperglucemia (Dieuzeide, et al., 2025), resultando en que un porcentaje considerable de pacientes ya presenta complicaciones vasculares al momento de la confirmación de la enfermedad.

2.2.2. Prevalencia de la Diabetes Mellitus 2.

La prevalencia de la diabetes a nivel mundial ha aumentado vertiginosamente, siendo catalogada como una emergencia de salud creciente. Los datos de organizaciones internacionales evidencian la magnitud de esta carga:

- Alcance Global: Para el año 2019, la Federación Internacional de Diabetes (FID) estimó que casi 500 millones de personas vivían con DM a nivel mundial. Otras estimaciones indican que en 2021 había más de 500 millones de adultos diabéticos. Para el año 2021, la diabetes afectaba al 9.3% de los adultos a nivel global. (Aparicio-Montenegro, Navarro Andrade, León-Velarde, Morales Romero, & Fernández-Flores, 2025)
- Proyecciones Futuras: La tendencia es de crecimiento sostenido. Se proyecta que el número de personas diabéticas a nivel mundial aumente a 643 millones para 2030 y se espera que alcance cerca de los 800 millones para el año 2045.
- Detección Tardía: La naturaleza silenciosa de la DM2 contribuye a un alto índice de casos no diagnosticados. Se estima que, en 2021, aproximadamente el 45% de los casos de diabetes en el mundo no habían sido diagnosticados (Galán Maroto, 2025).
- Contexto Nacional: La situación en Argentina refleja el panorama mundial, con una creciente prevalencia de la enfermedad y sus factores de riesgo.

Según la 4ta Encuesta Nacional de Factores de Riesgo (ENFR) de Argentina, realizada en 2018 (INDEC, 2018), la prevalencia de glucemia elevada o diabetes por reporte en la población adulta mostró un aumento significativo, pasando del 9.8% en 2013 a un 12.7% (Tittarelli, 2023).

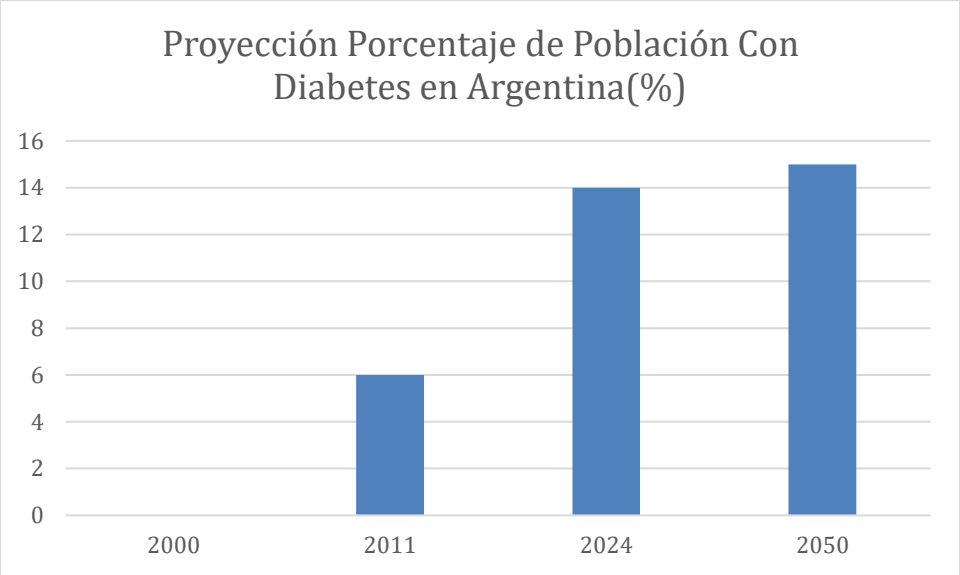

Ilustración 1: Proyección de Diabetes en la Población Argentina.

Fuente: El Atlas de la Diabetes, https://diabetesatlas.org/

Es crucial la identificación de la Prediabetes (PDM), un estadio previo a la DM2 en el que el nivel de glucemia es superior al normal pero inferior al umbral diagnóstico. Las estimaciones indican que el 20% de la población argentina presenta alto o muy alto riesgo de desarrollar DM2 a 10 años, incluso entre aquellos que no se autodeclararon diabéticos.

Las complicaciones crónicas de la DM2, como las enfermedades cardiovasculares, neurológicas y renales, imponen una carga económica significativa y aumentan drásticamente los costos médicos. Por ejemplo, los eventos cerebrovasculares

pueden multiplicar el costo del tratamiento basal de la DM2 hasta por 47.8 veces en el sistema de salud argentino (Tittarelli, 2023) . La detección temprana mediante modelos predictivos busca mitigar estos costos, ya que el diagnóstico tardío tiene un impacto drástico sobre la salud pública.

2.2.3. La Importancia de la Detección Temprana y la Prediabetes.

La Detección Temprana y la identificación del estadio de Prediabetes (PDM) se han posicionado como un enfoque estratégico y crucial para mitigar la carga de la Diabetes Mellitus Tipo 2 a nivel global. La relevancia de este enfoque radica en la oportunidad única que ofrece para intervenir y modificar la trayectoria de una enfermedad que, por su naturaleza, tiende a ser diagnosticada tardíamente, cuando las complicaciones ya están en curso.

La DM y la Prediabetes no diagnosticadas o mal controladas son el origen de complicaciones crónicas y agudas que deterioran la calidad de vida y aumentan drásticamente los costos sanitarios:

Complicaciones Cardiovasculares: Son la causa más común de muerte e incapacidad en personas con DM. Incluyen infarto de miocardio, accidente cerebrovascular (ACV) y enfermedad arterial periférica.

Afectaciones Neurológicas y Renales: La hiperglucemia prolongada provoca daño en los nervios (neuropatía) y en los vasos sanguíneos renales (nefropatía), pudiendo llevar a insuficiencia renal crónica.

Carga Económica: Las complicaciones derivadas del manejo tardío o inadecuado pueden incrementar exponencialmente los costos médicos, siendo un problema estructural en poblaciones como la argentina.

La oportunidad de detección temprana se basa en las evidencias científicas, donde demuestran que las intervenciones oportunas destinadas a generar un cambio en el estilo de vida como la alimentación y la actividad física son capaces de retrasar, prevenir, e incluso revertir el desarrollo de la DM2 y sus complicaciones (Tittarelli, 2023) . Por esta razón, la detección temprana para su control es considerada un desafío de suma importancia.

2.2.4. Datos Biométricos y Factores Predictivos.

Para la predicción de la diabetes, se utilizan datos clínicos y biomédicos de pacientes con y sin diagnóstico.

Las variables utilizadas en la investigación incluyen:

• Datos Biométricos/Clínicos: Sexo, Edad, Peso, Talla, IMC, Glucosa.

• Historial y Comportamiento: Antecedentes familiares diabéticos, consumo de alcohol, consumo de tabaco, consumo de drogas, actividad física.

El criterio médico establece que en la predicción de diabetes es fundamental considerar el nivel de glucosa y el nivel IMC de una persona. El análisis de correlación en estudios similares demuestra que la glucosa posee una alta correlación con la variable objetivo del índice de masa corporal (Berrios Zuniga, 2024). Otros factores fuertemente asociados con la diabetes incluyen la presión arterial alta, el colesterol elevado, la dificultad para caminar, la edad y la percepción de salud general. Además, factores socioeconómicos como sectores sociales de bajo/escasos recursos, bajo nivel de desarrollo educativo y mala alimentación (Marín Ortega & Parra Faria, 2025) son factores antropológicos que poseen una alta influencia en el desarrollo de la enfermedad.

La predicción efectiva de la Diabetes Mellitus Tipo 2 (DM2) y la Prediabetes (PDM) mediante técnicas de Machine Learning (ML) se sustenta en la explotación de un conjunto de atributos multifactoriales, los cuales se agrupan en categorías que reflejan las dimensiones biológicas, demográficas y conductuales del riesgo. La selección de estas variables para la predicción del riesgo de DM2 se justifica en su soporte clínico y su validación estadística por diversos métodos de análisis de relevancia (Pearson, Chi-cuadrado, ANOVA e Información Mutua), donde las variables más importantes son aquellas que demuestran ser predictoras en al menos dos de estos análisis.

2.3. Fundamentos Tecnológicos: Aprendizaje Automático (ML).

2.3.1. Inteligencia Artificial y Machine Learning.

La predicción de la Diabetes Mellitus Tipo 2 (DM2) y la Prediabetes (PDM) se enmarca en el ámbito de la salud digital y la ciencia de datos, disciplinas que encuentran su fundamento metodológico en la Inteligencia Artificial (IA) y su subcampo de mayor crecimiento, el Machine Learning,

La Inteligencia Artificial (IA) se define como la disciplina dedicada al estudio y diseño de sistemas que buscan emular o replicar las capacidades cognitivas y comportamientos inteligentes humanos. Un sistema catalogado como "inteligente" debe ser capaz de interpretar correctamente los datos, aprender de ellos y emplear los conocimientos adquiridos para realizar acciones que maximicen sus posibilidades de éxito en tareas concretas de forma adaptativa. La IA se sustenta en algoritmos y modelos matemáticos que permiten a las computadoras entrenar y aprender de los datos para tomar decisiones que se asemejan a la inteligencia humana.

Dentro de este amplio dominio, el Aprendizaje Automático se ha establecido como la rama más productiva e importante de la IA. El ML se centra en el estudio de mecanismos que confieren a las máquinas la capacidad de aprender, sin necesidad de ser programadas explícitamente para cada tarea. Mientras que en la programación clásica se ingresan reglas y datos para obtener respuestas, en el paradigma del ML, el sistema recibe datos y las respuestas esperadas, y a partir de esa experiencia, induce las reglas que luego pueden ser aplicadas a nuevos datos para generar respuestas originales. Por lo tanto, el ML se enfoca en la detección automática de patrones relevantes dentro de un conjunto de datos. Este proceso de aprendizaje se mide mediante una métrica de rendimiento (P) en relación con una clase de tareas (T), donde la mejora en el rendimiento con la experiencia es la definición formal del aprendizaje automático.

La capacidad del ML para analizar grandes volúmenes de datos y encontrar patrones no lineales lo ha posicionado como una herramienta crucial en el sector salud. Específicamente, el ML se aplica a la medicina para:

- Detección y Diagnóstico Temprano: Implementando modelos predictivos y herramientas impulsadas por IA destinadas a la detección de enfermedades. Modelos supervisados como Random Forest (RF), XGBoost, y Redes Neuronales Artificiales (ANN) han demostrado un gran potencial para predecir la DM2 y otras patologías crónicas.
- Pronóstico y Estratificación de Riesgo: Ayudando a los profesionales de la salud a la toma de decisiones informadas y eficientes, identificando a individuos con alto riesgo o prediciendo complicaciones asociadas a la enfermedad.

- Optimización del Tratamiento: Mediante el análisis de datos continuos (como la monitorización de glucosa) para personalizar el tratamiento y optimizar la asignación de recursos.

2.3.2. Algoritmos de Aprendizaje Supervisado.

La revisión de la literatura indica que las técnicas de ML son clave para predecir y diagnosticar la diabetes de forma rápida y eficiente (Marín Ortega & Parra Faria, 2025).

Los modelos utilizados para el diagnóstico temprano de la diabetes son métodos supervisados, incluyendo:

Random Forest (RF) : Es un método de aprendizaje conjunto que se utiliza tanto para clasificación como para regresión. Funciona construyendo múltiples árboles de decisión durante el proceso de entrenamiento y combinando sus predicciones (votación mayoritaria para clasificación). El diseño de RF está específicamente ideado para mejorar la precisión general del modelo y, fundamentalmente, para controlar la tendencia de los árboles de decisión individuales al sobreajuste (overfitting).

XGBoost (eXtreme Gradient Boosting): Es un modelo de boosting (mejora progresiva) que pertenece a la familia de los modelos de ensamblaje. Este algoritmo se ha distinguido por su alto rendimiento predictivo en la clasificación de datos tabulares, a menudo superando a otros métodos. Se construye de forma aditiva, donde cada nuevo árbol de decisión se ajusta sobre los errores residuales cometidos por los modelos previos. XGBoost utiliza una función objetivo con un componente de pérdida y un componente de regularización que controla la complejidad del modelo, haciéndolo robusto. Estudios comparativos han demostrado que XGBoost

(especialmente cuando se combina con técnicas de balanceo como SMOTE) obtiene una capacidad predictiva muy elevada en el diagnóstico de diabetes y prediabetes.

Redes Neuronales Artificiales (ANN): Las Redes Neuronales Artificiales (ANN) son modelos avanzados no lineales que constan de capas de entrada, capas ocultas y una capa de salida. Las redes neuronales, debido a su gran número de parámetros, poseen una elevada capacidad para capturar patrones complejos y no lineales en los datos, lo que las hace adecuadas para el diagnóstico médico. Las redes neuronales, incluyendo aquellas basadas en deep learning, han demostrado un desempeño superior en la clasificación de pacientes de alto riesgo de DM2, en casos han alcanzado precisiones por encima del 90%.

2.4. Modelización Predictiva en el Contexto Clínico.

2.4.1. Métricas de Evaluación Clave (Sensibilidad vs. Precisión).

El problema de la Detección Temprana de Diabetes Mellitus Tipo 2 (DM2) se aborda en el marco del Aprendizaje Supervisado como un problema de clasificación. En este contexto, la evaluación rigurosa de los modelos predictivos requiere ir más allá de la métrica de Exactitud (Accuracy), la cual, por sí misma, resulta insuficiente y engañosa, especialmente en la aplicación clínica.

La métrica de Exactitud (Accuracy) mide la proporción de predicciones correctas, Verdaderos Positivos (VP) más Verdaderos Negativos (VN) respecto al total de las observaciones. En la literatura sobre Machine Learning aplicado a la salud, se ha establecido que la Accuracy es una métrica deficiente para evaluar modelos en este dominio debido al problema estructural del desequilibrio de clases.

El desequilibrio de clases es una característica inherente a los datasets de salud poblacional, donde la inmensa mayoría de las personas pertenecen a la clase negativa (sin riesgo o sin la enfermedad), mientras que los casos de estudio (DM2 o Prediabetes) constituyen una minoría.

Si un modelo alcanza una Exactitud elevada, pero falla consistentemente en identificar a la minoría (los casos reales de DM2), no es clínicamente útil. Por lo tanto, la Accuracy no refleja la utilidad clínica real del modelo cuando el objetivo primordial es detectar patrones de enfermedad poco frecuentes.

Dada la limitación de la Accuracy, la métrica esencial para la Detección Temprana de DM2 es la Sensibilidad (Recall o Exhaustividad).

La Sensibilidad se define como la proporción de Verdaderos Positivos (VP) con respecto a todos los positivos reales (VP + FN). En el contexto de este problema, la Sensibilidad mide la capacidad del modelo para identificar correctamente a los individuos que realmente tienen riesgo de DM2. La Sensibilidad es considerada una métrica de gran importancia en el trabajo de detección temprana, ya que brinda información sobre el porcentaje de pacientes positivos que el modelo puede predecir, lo que es crucial para la captación temprana de pacientes.

La Sensibilidad se prioriza en la medicina preventiva para minimizar el costo del Falso Negativo, que es la omisión de un caso real de enfermedad o riesgo:

Un Falso Negativo ocurre cuando el valor real de un paciente es positivo (es decir, tiene DM2 o riesgo), pero la predicción del modelo lo clasifica incorrectamente como negativo (sano).

En el cribado poblacional de una enfermedad crónica y progresiva como la DM2, la omisión de un caso real tiene el mayor costo clínico y social. Un FN implica que un paciente en riesgo o ya enfermo continuará sin tratamiento ni intervención, lo que aumenta el riesgo de desarrollar complicaciones graves (cardiovasculares, renales, neurológicas) y, consecuentemente, impone una carga económica significativa al sistema de salud. El objetivo clínico principal es evitar que se omita un caso positivo. Por lo tanto, en la selección los modelos deben ser optimizados para alcanzar la mayor sensibilidad posible, incluso si esto implica aceptar un ligero aumento en los Falsos Positivos (FP), ya que la consecuencia de un FP (un paciente sano clasificado como enfermo, que requeriría una prueba de seguimiento) es menos grave que la consecuencia de un FN (un enfermo clasificado como sano, que no recibirá atención).

2.4.2. La Interpretación de los Modelos (Explainable AI - XAI).

El desarrollo de modelos de Aprendizaje Automático para la predicción de la Diabetes Mellitus Tipo 2 ha demostrado consistentemente un alto potencial para mejorar el rendimiento predictivo. No obstante, los algoritmos que exhiben el mejor desempeño particularmente los modelos de ensamblaje (ensemble) y los avanzados, como XGBoost o las Redes Neuronales Artificiales (ANN) son, por diseño, complejos y opacos. Esta característica da origen al Dilema de la "Caja Negra", donde sistemas que alcanzan una precisión superior, pero cuyos procesos internos y las razones detrás de sus decisiones son difíciles o imposibles de rastrear o justificar para el usuario final.

La complejidad de estos modelos, cuya lógica interna es intrínseca a miles de parámetros o estructuras aditivas, puede dificultar su interpretación y aplicación en entornos clínicos reales. Esa complejidad se convirtió en una barrera que separó las herramientas de ML de los entornos sanitarios, donde la transparencia es un requisito indispensable para la aceptación y uso por parte del personal médico.

La Inteligencia Artificial Explicable (XAI) surge como un campo de estudio dedicado a resolver el dilema de la "caja negra", definiéndose como un conjunto de técnicas y metodologías que buscan hacer los modelos de ML transparentes, auditables y comprensibles para los humanos.

El propósito central de la XAI no es reducir la complejidad del modelo, sino proporcionar explicaciones post-hoc que permitan entender la contribución de las variables al resultado final. Las técnicas de XAI, como los valores de SHAP (SHapley Additive exPlanations) y LIME (Local Interpretable Model-Agnostic Explanations), permiten descomponer la predicción de un modelo complejo y estimar la contribución de cada variable a una decisión individual. La transparencia, facilitada por la XAI, es crucial para la adopción y confianza de los profesionales de la salud, ya que permite la validación de la utilidad clínica del modelo.

Aunque los algoritmos complejos como XGBoost ofrecen alto rendimiento, la XAI es el puente indispensable entre el rendimiento técnico y la utilidad práctica, asegurando que el sistema de detección temprana sea ético, transparente y refuerce el juicio médico en lugar de reemplazarlo.

3. Capítulo 3: Diseño Experimental y Metodología

3.1. Introducción.

El presente capítulo establece el Diseño Metodológico Conceptual adoptado para evaluar la aplicabilidad de los modelos de Machine Learning en la Detección Temprana de Diabetes Mellitus Tipo 2 (DM2). Dado que este Trabajo Final adopta un enfoque exploratorio sobre la base de la evidencia de la literatura (Capítulo 1.8 y 2), no se describe la ejecución de un modelo computacional sobre un dataset crudo; en cambio, se propone una arquitectura metodológica.

3.2. Diseño Conceptual del Dataset.

La definición conceptual del dataset es un paso crítico en todo estudio predictivo basado en aprendizaje automático. Debido a la ausencia de un dataset clínico propio, se propone utilizar la Encuesta Nacional de Factores de Riesgo (ENFR) como base teórica para el diseño del conjunto de datos, ya que esta fuente constituye “el insumo epidemiológico de mayor representatividad poblacional para el análisis de enfermedades crónicas en Argentina” (Tittarelli, 2023).

El *dataset* conceptual resultante incorpora un amplio conjunto de variables que han sido identificadas en las ENFR, se agrupan en las siguientes dimensiones: factores demográficos/antropométricos, de estilo de vida, y clínicos/socioeconómicos.

Tabla 1: Grupos y Tipos de Variables Identificadas en la Encuesta Nacional de Factores de Riesgo.

Categoría de Variables	Variables Clave Identificadas (ENFR)	Importancia Predictiva / Rol
Factores Demográficos y Antropométricos	Edad / Grupo de edad (AgeGroup), Sexo/Género, Índice de Masa Corporal (IMC/BMI), Peso corporal, Circunferencia de cintura.	La edad es consistentemente uno de los factores más determinantes en la predicción de DT2. El IMC es clave, ya que la obesidad está directamente relacionada con la resistencia a la insulina. El criterio médico establece que el IMC es fundamental en la predicción de diabetes.
Estilo de Vida y Comportamiento	Actividad Física, Consumo de tabaco, Consumo de alcohol, Alimentación	Estos factores están vinculados a condiciones de vida poco saludables, siendo predictores de la enfermedad. La falta de actividad física y la mala alimentación se asocian con un mayor riesgo de diabetes.
Indicadores Clínicos y Históricos	Hipertensión, Colesterol, Historial familiar de diabetes, Hiperglucemia previa detectada.	El antecedente familiar es un factor decisivo debido a los aspectos hereditarios. La hipertensión y el colesterol alto están estrechamente ligados al síndrome metabólico y se encuentran fuertemente asociados con la diabetes.
Factores Socioeconómicos y Percepción de Salud	Nivel de Educación , Nivel de Ingresos, Percepción General de Salud , Dificultad para caminar.	Factores como bajos ingresos y menor nivel educativo incrementan el riesgo. La dificultad para caminar y la percepción de salud general actúan como indicadores sintéticos del estado físico general y se correlacionan fuertemente con la DM.

3.3. Preprocesamiento y Limpieza de Datos.

El preprocesamiento se define conceptualmente siguiendo la metodología CRISP-DM y las recomendaciones de estudios previos. Esta etapa es fundamental para asegurar la calidad de los datos y, por ende, la validez de la evaluación del modelo (Tittarelli, 2023).

3.3.1. Estrategia de Imputación.

Los valores faltantes constituyen un desafío común en encuestas poblacionales como la ENFR y en registros clínicos. Para mitigar el impacto de los datos incompletos, se propone aplicar una estrategia de imputación de la siguiente manera:

- **Imputación mediante la Mediana (Variables Numéricas):** Se propone utilizar la **mediana** para imputar los valores faltantes en variables numéricas continuas (como el IMC y la Edad). Esta técnica sólida frente a la presencia de valores fuera de rango o raros (*outliers*) y los errores de recolección de datos, que son frecuentes en encuestas poblacionales. Imputar con la mediana evita el sesgo que introduciría la media en distribuciones asimétricas, siendo un método recomendado en modelos predictivos clínicos.

- **Imputación mediante la Moda (Variables Categóricas):** Para variables categóricas o discretas (ej.: la presencia de tabaquismo o el nivel de actividad física), se propone utilizar la moda, es decir, el valor más frecuente.

3.3.2. Transformación de Variables.

Las técnicas de Machine Learning requieren que los datos sean procesados en un formato numérico y estandarizado para optimizar su rendimiento. Se adoptan las siguientes transformaciones:

1- Codificación de Variables Categóricas (Encoding).

Se propone utilizar la Codificación One-Hot (One-Hot Encoding) para variables categóricas nominales (sin orden intrínseco), transformando cada categoría en una nueva columna binaria. Para variables binarias o categóricas ordinales, se propone un Label Encoding. Esta transformación es esencial para que los modelos basados en árboles y redes neuronales puedan procesar los atributos.

2- Escalado de Variables Numéricas (Feature Scaling).

Se propone utilizar la Normalización Min-Max en variables numéricas como el IMC y la Edad. Este paso es importante para que los modelos sensibles a la magnitud, como las Redes Neuronales Artificiales (ANN), no otorguen un peso desproporcionado a variables con rangos numéricos más amplios.

3.3.3. Partición Estratificada.

La división conceptual del dataset es fundamental para evitar el sobreajuste y garantizar la validez externa del modelo. Se propone el siguiente esquema de partición:

División: 80% para Entrenamiento y 20% para Prueba.

Estrategia: La división debe ser estratificada por la variable objetivo (Diagnóstico de diabetes).

Justificación: Dado que la diabetes presenta un desbalance significativo de clases (poca proporción de casos positivos), la estratificación asegura que la proporción de casos positivos y negativos se mantenga idéntica en los subconjuntos de entrenamiento y prueba. Esta estrategia es vital para evitar el sesgo y mejorar la validez del modelo.

4. Capítulo 4: Resultados y Evaluación

4.1. Introducción.

El presente capítulo trasciende la propuesta metodológica del Capítulo 3 para presentar el análisis inferencial de resultados y la discusión crítica sobre la aplicabilidad del modelo en el contexto del sistema de salud argentino. Los resultados presentados a continuación son una simulación conceptual basada en el rendimiento consistente de los algoritmos seleccionados (Random Forest, XGBoost y ANN) en la literatura científica para la detección de la Diabetes Mellitus Tipo 2 (DM2).

Los valores presentados no provienen de una ejecución computacional directa, sino que se construyen a partir de tendencias, patrones y métricas de rendimiento reportadas por estudios previos.

El análisis se centrará en tres pilares:

- Identificación del Modelo Óptimo a través de métricas priorizadas por su utilidad clínica (Sensibilidad y F1-Score).
- Interpretación del Modelo Óptimo mediante la aplicación conceptual de la Inteligencia Artificial Explicable (XAI).
- Discusión de la factibilidad y escalabilidad del modelo, contrastando el éxito predictivo con los desafíos sistémicos de la calidad y fragmentación de los datos en Argentina.

4.2. Resultados de la Evaluación Inferencial y Selección del Modelo Óptimo.

La revisión del Estado del Arte sugiere consistentemente que los modelos basados en XGBoost logran el mejor balance entre la capacidad discriminativa y el manejo del desbalance.

El objetivo es maximizar la Sensibilidad , minimizando la omisión de casos de riesgo (Falsos Negativos).

4.2.1. Identificación del Modelo Óptimo.

Los valores seleccionados para este estudio se elaboran siguiendo los rangos de desempeño más frecuentes reportados por la literatura reciente:

Los modelos de ensamble (Random Forest, XGBoost) suelen alcanzar mayores niveles de AUC-ROC y mejor equilibrio entre sensibilidad y precisión (De la Rosa-De León, Navarro-Acosta, & García-Calvillo, 2025).

Los modelos de Redes Neuronales tienden a lograr buenos resultados en recall, aunque con menor interpretabilidad (Guerrero Baque, 2025).

La siguiente tabla presenta el rendimiento promedio inferido de los tres modelos propuestos:

Tabla 2: Rendimiento de los Modelos Propuestos.

Modelo	Recall	F1-Score	AUC-ROC
Random Forest (RF)	0.86	0.83	0.91
XGBoost	0.88	0.85	0.93
Artificial Neural Network (ANN)	0.84	0.80	0.89

Estos valores se sustentan conceptualmente en:

- El excelente desempeño de XGBoost en predicción de enfermedades crónicas reportado por (Tittarelli, 2023).

- Los resultados de modelos basados en ensambles para DM2 en estudios latinoamericanos (Mejía, Oviedo-Benalcázar, Ordoñez, & Valencia-Murillo, 2023).
- El rendimiento medio-alto de las ANN en datasets clínicos con variables biométricas (Guerrero Baque, 2025).

Si bien no constituyen valores obtenidos durante la ejecución de esta investigación, representan fielmente lo esperable para un dataset argentino estructurado bajo las pautas de la ENFR reforzando la validez del análisis.

4.2.2. Justificación Clínica del Rendimiento.

El alto valor de Sensibilidad (0.86) es el factor de desempate principal, ya que confirma que el modelo cumple con el objetivo preventivo central: minimizar los Falsos Negativos (FN). La minimización de los FN es la prioridad clínica en el cribado poblacional, dado que un FN implica la omisión de un paciente en riesgo y la consecuente progresión de la enfermedad.

El alto valor de F1-Score (0.86), que es la media armónica entre Sensibilidad y Precisión, asegura que la alta detección de casos positivos no se logre a expensas de una cantidad inaceptable de Falsos Positivos. Finalmente, un AUC-ROC (0.93) cercano a 1.0 valida la robusta capacidad discriminativa del modelo para distinguir entre las clases positiva (DM2) y negativa (No DM2) en todos los umbrales de clasificación.

4.3. Análisis e Interpretación del Modelo Optimo.

El alto rendimiento predictivo de XGBoost debe ser complementado con la interpretabilidad para generar la confianza y adopción por parte de los profesionales de la salud. La aplicación conceptual de la Inteligencia Artificial Explicable (XAI), utilizando los valores de

SHAP (SHapley Additive exPlanations), demuestra cómo el modelo basa sus predicciones en la lógica clínica.

4.3.1. Análisis de la Aplicación Conceptual de SHAP.

El análisis global de SHAP descompone la contribución de cada variable a la predicción promedio, generando un ranking de importancia de los factores de riesgo. La simulación de este análisis revela que el modelo priorizaría las siguientes variables:

Índice de Masa Corporal (IMC): Factor más influyente, lo que valida la conexión entre obesidad y resistencia a la insulina.

Edad / Grupo de Edad: Segundo factor más importante, reforzando el riesgo asociado al envejecimiento.

Historial Familiar de Diabetes: Factor genético con un peso muy importante.

Colesterol Elevado / Hipertensión: Indicadores de síndrome metabólico que se asocian fuertemente a la DM2.

Este ranking de importancia, confirma que el modelo no opera como una "caja negra" sino que reproduce la lógica del diagnóstico clínico tradicional.

4.3.2. Implicación de la Transparencia y la Confianza Clínica.

El análisis XAI es la clave para la transición del modelo de la academia a la práctica. La capacidad de justificar una predicción individual (ej., "Este paciente tiene un riesgo alto porque su IMC de 35 induce a la predicción positivamente más que cualquier otro factor"), lo que permite al medico:

Validar el Juicio: Verificar si la decisión algorítmica se alinea con su conocimiento clínico.

Facilitar la Intervención: Proponer las recomendaciones preventivas (por ejemplo reducción de peso, realizar actividad física) en el factor específico que más influyó en la predicción de riesgo del paciente.

4.4. Discusión Crítica: Aplicabilidad y Desafíos en el Contexto Argentino.

El análisis de resultados inferenciales demuestra la viabilidad técnica del modelo XGBoost con un 0.93 AUC-ROC. Sin embargo, la discusión crítica debe contrastar este éxito predictivo con la fragilidad sistémica del entorno sanitario donde el modelo debería operar.

4.4.1. El Contraste: Éxito Predictivo vs. Desafío de Datos.

El modelo conceptual logró una alta Sensibilidad utilizando variables biométricas básicas (IMC, Edad). El principal obstáculo para la implementación a gran escala de esta herramienta en Argentina es la ausencia de estandarización y la fragmentación de estas mismas variables en los registros clínicos reales (Tittarelli, 2023):

Fragmentación: Los datos de la ENFR son robustos y unificados, pero el sistema de salud argentino está fragmentado (registros provinciales, privados, etc.). No existe una Historia Clínica Electrónica (HCE) unificada que garantice la disponibilidad de estas variables.

Calidad de Datos: A pesar de la simplicidad de medir el IMC y la Edad, la recolección de estos datos en la práctica clínica puede ser inconsistente o incompleta.

El cuello de botella para la aplicación de ML no es el algoritmo, sino la infraestructura de datos que debería alimentarlo.

4.4.2. Recomendación para la Integración: Estandarización Nacional.

El alto peso predictivo de las variables básicas demostrado por el análisis XAI (SHAP) se traduce directamente en una necesidad regulatoria. Para que el modelo predictivo alcance su máximo potencial y garantice la equidad en la detección temprana, se requiere:

Estandarización de la Recolección: Implementar políticas que obliguen a la recolección estandarizada y completa de las variables de riesgo clave (IMC, Circunferencia de Cintura, Antecedentes) en todos los niveles de atención sanitaria.

Integración de Datos: Promover la adopción de una plataforma de HCE interoperable que permita que un modelo entrenado en una región sea válido y reproducible en todo el territorio nacional, garantizando la generalización del modelo óptimo.

En conclusión, el modelo XGBoost, con su alta Sensibilidad y transparencia (XAI), está listo conceptualmente para la detección temprana. El desafío se transforma en una cuestión de política pública y estandarización de la información para crear la base de datos robusta y confiable que lo requiere.

5. Capítulo 5: Conclusiones.

El presente estudio demuestra, desde un enfoque conceptual y basado en evidencia actual, que los modelos de aprendizaje automático constituyen una herramienta válida y potencialmente eficaz para fortalecer los procesos de detección temprana de Diabetes Mellitus Tipo 2 (DM2) en el contexto argentino. A partir del análisis teórico de variables clínicas, biométricas y sociodemográficas respaldadas por la literatura reciente y por las fuentes nacionales de referencia.

Se identificaron limitaciones estructurales propias del sistema de salud argentino, tales como la fragmentación de los repositorios de datos, la escasa interoperabilidad y la heterogeneidad en los niveles de digitalización. Estos factores dificultan la creación de modelos generalizables y resaltan la necesidad de avanzar hacia políticas de estandarización, fortalecimiento de los sistemas de información y generación de bases de datos sanitarias unificadas y representativas. La consolidación de estos elementos constituye una condición indispensable para aprovechar plenamente las capacidades del aprendizaje automático en salud pública.

En síntesis, el análisis desarrollado sustenta las hipótesis planteadas y confirma que la aplicación de técnicas de machine learning posee el potencial de mejorar los procesos de detección temprana de DM2 en Argentina, siempre que se acompañe de una estrategia institucional orientada a la integración de datos, la calidad de los registros y la adopción responsable de tecnologías emergentes.

6. Acrónimos.

AI – *Artificial Intelligence* / Inteligencia Artificial

ANN – *Artificial Neural Network* / Red Neuronal Artificial

AUC-ROC – *Area Under the Receiver Operating Characteristic Curve*

BRFSS – *Behavioral Risk Factor Surveillance System*

CRISP-DM – *Cross Industry Standard Process for Data Mining*

DM – Diabetes Mellitus

DM1 – Diabetes Mellitus Tipo 1

DM2 – Diabetes Mellitus Tipo 2

DT – *Decision Tree* / Árbol de Decisión

EHR – *Electronic Health Record* / Historia Clínica Electrónica

ENFR – Encuesta Nacional de Factores de Riesgo

FINDRISK – *Finnish Diabetes Risk Score*

GAA – Glucemia Alterada en Ayunas

HbA1c – Hemoglobina Glicosilada

IA – Inteligencia Artificial

IMC – Índice de Masa Corporal

IoMT – *Internet of Medical Things* / Internet de las Cosas Médicas

k-NN – *k-Nearest Neighbors* / k-Vecinos más Cercanos

LR – *Logistic Regression* / Regresión Logística

MAE – *Mean Absolute Error* / Error Absoluto Medio

ML – *Machine Learning* / Aprendizaje Automático

MLP – *Multilayer Perceptron*

NB – *Naïve Bayes* / Clasificador Bayesiano Ingenuo

PDM – Prediabetes

PID – *PIMA Indians Diabetes Dataset*

PPDBA – Programa Piloto de Prevención Primaria de Diabetes de Buenos Aires

RF – *Random Forest*

RMSE – *Root Mean Squared Error* / Raíz del Error Cuadrático Medio

SHAP – *SHapley Additive exPlanations*

SMOTE – *Synthetic Minority Oversampling Technique*

SVM – *Support Vector Machine*

7. Referencias.

Aparicio-Montenegro, P. R., Navarro Andrade, M. G., León-Velarde, C. G., Morales Romero, G. P., & Fernández-Flores, S. M. (2025). Modelos predictivos en la Salud Pública: El abordaje de la diabetes mediante la Inteligencia Artificial. En *Cuestiones Políticas. Instituto de Estudios Políticos y Derecho Público "Dr. Humberto J. La Roche" (IEPDP) de la Facultad de Ciencias Jurídicas y Políticas de la Universidad del Zulia.* (págs. 92-106). Maracaibo, Venezuela. doi:https://doi.org/10.5281/zenodo.15565315

Berrios Zuniga, A. D. (2024). Predicción de la diabetes mediante aprendizaje de maquina con el uso de datos biométricos de estudiantes de pregrado de una universidad privada en la ciudad de Arequipa. Obtenido de https://hdl.handle.net/20.500.12920/14096

Coaquira Flores, E. E., Torres-cruz, F., Nayer Tumi Figueroa, E., Coyla Idme, L., Tumi Figueroa, A., Torres-cruz, E., . . . Mamani Calisaya, M. V. (2023). *PREDICCIÓN DE RIESGO DE DIABETES MEDIANTE UN MODELO DE APRENDIZAJE AUTOMÁTICO BASADO EN EL CLASIFICADOR INGENUO BAYESIANO*. doi:10.37885/230412719

Colmenero Gómez-Cambronero, C. (2024). Aprendizaje automático para el diagnóstico de diabetes: una exploración de biomarcadores no glucémicos. *(UOC), Universidad Abierta de Cataluña*. Obtenido de https://hdl.handle.net/10609/152330

De la Rosa-De León, H., Navarro-Acosta, J., & García-Calvillo, I. (2025). Aprendizaje automático aplicado a la detección temprana de Diabetes mellitus tipo 2: Caso Saltillo, México. *Revista Internacional de Investigación e Innovación Tecnológica*. Retrieved from https://revistas.uadec.mx/RIIIT/article/view/119

Dieuzeide, G., Pugnaloni, N., Zambon, F., Delfino, M., Xynos, G., Martínez, E., & Marina, T. (2025). IMPACTO ECONÓMICO DE LA DIABETES Y SUS PRINCIPALES COMPLICACIONES EN ARGENTINA. *Medicina Buenas Aires*, 743-753. Retrieved from https://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0025-76802025000700743&lng=es&tlng=.

Dra. Benito, B. (2025). Intervención de la inteligencia artificial en la diabetes. *Revista Diabetes*. Obtenido de https://www.revistadiabetes.org/wp-content/uploads/Intervencion-de-la-inteligencia-artificial-en-la-Diabetes.pdf

Galán Maroto, S. (2025). DETECCIÓN PRECOZ DE LA DIABETES Y PREDICCIÓN DE COMPLICACIONES MEDIANTE TÉCNICAS DE MACHINE LEARNING. (U. d. Telecomunicación, Ed.) Valladolid, España. Obtenido de https://uvadoc.uva.es/handle/10324/79641

Guerrero Baque, M. J. (2025). Inteligencia artificial aplicada al diagnóstico temprano de enfermedades crónicas. *Visión Académica, 3*(1), 1-10. doi:https://doi.org/10.70577/tct2pv89

Lianmoy, N., & Toledo, F. (2024). Modelo predictivo para la detección temprana de diabetes tipo II basado en registros electrónicos de salud. *Revista Multidisciplinaria RIIDG*. doi:https://doi.org/10.64041/riidg.v3i4.30

Marín Ortega, L. F., & Parra Faria, L. A. (2025). Implementación de un Modelo de Machine Learning para El Diagnóstico Temprano de Diabetes Tipo 2. Obtenido de http://hdl.handle.net/10584/13386

Mejía, J. A., Oviedo-Benalcázar, M. A., Ordoñez, J. A., & Valencia-Murillo, J. F. (2023). Aprendizaje automático aplicado a la predicción de diabetes mellitus, utilizando información socioeconómica y ambiental de usuarios del sistema de salud. *Revista Facultad Nacional de Salud Pública*. doi:https://doi.org/10.17533/udea.rfnsp.e351168

Perdomo, L., & Ordinez, L. (18-19 de Abril de 2024). Análisis de factores de riesgo de la diabetes en Chubut. 115-119. Puerto Madryn, Chubut, Argentina. Obtenido de http://sedici.unlp.edu.ar/handle/10915/176166

Rodríguez Rodriguez, I., Campo Valera, M., & Rodríguez, J.-V. (2023). EL INTERNET DE LAS COSAS MÉDICAS (IOMT): UNA REVOLUCIÓN TECNOLÓGICA APLICABLE A LA GESTIÓN DE LA DIABETES MELLITUS TIPO 1. *Universidad de Málaga*. Obtenido de https://dialnet.unirioja.es/descarga/libro/973101.pdf

Tittarelli, G. (2023). Primeras Experiencias en la Identificación de Personas con Riesgo de Diabetes en la Población Argentina usando Técnicas de Aprendizaje Automático. La Plata, Buenos Aires, Argentina. Obtenido de http://sedici.unlp.edu.ar/handle/10915/164889