Predicción temprana de la diabetes mediante modelos de aprendizaje supervisado utilizando variables clínicas y de estilo de vida de los pacientes

Brayan Alexis De Sousa Lora

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Antonio Jesús Tamayo Herrera, Ph. D.

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2026

Cita	(De Sousa Lora, 2025)
Referencia **Estilo APA 7 (2020)**	De Sousa Lora, B. A., (2026). *Detección anticipada de la diabetes usando modelos supervisados basados en variables clínicas y de estilo de vida de los pacientes* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Grupo de Investigación Intelligent Information Systems Lab In2Lab

Especialización en Analítica y Ciencia de Datos, Cohorte IX.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: Héctor Iván García García.

Decano: Elkin Libardo Ríos Ortiz.

Jefe departamento: Danny Alexandro Múnera Ramírez

Coordinadora del Programa: Maria Bernarda Salazar Sánchez

Dedicatoria

Dedico este trabajo a mi familia y a mi pareja, Paloma Aguirre, quienes con su apoyo, paciencia y motivación constante me acompañaron durante todo este proceso académico. Gracias por impulsarme a seguir adelante y por ser parte fundamental de este logro, que también les pertenece. También quiero dedicármelo a mí mismo, por el esfuerzo, la perseverancia y el compromiso para alcanzar esta meta.

Agradecimientos

Expreso mi más sincero agradecimiento a mi familia y a mi pareja, Paloma Aguirre, por su apoyo incondicional a lo largo de este camino. Asimismo, agradezco a los docentes y a la institución por compartir sus conocimientos, orientación y herramientas que contribuyeron significativamente a mi formación profesional y al desarrollo de este trabajo. Finalmente, agradezco a todas aquellas personas que, de una u otra forma, brindaron su apoyo y aportaron al cumplimiento de este objetivo académico.

Tabla de contenido

Resumen 9
Abstract 10
1. Introducción 11
1.1. Descripción del problema 11
1.2. Planteamiento del problema 13
1.3. Objetivo del estudio 14
1.3.1. Objetivo General 14
1.3.2. Objetivos Específicos 14
1.4. Metodología 15
1.4.1. Datos utilizados y su fuente 15
1.4.2. Algoritmos, modelos y técnicas aplicadas 15
1.4.3. Criterios de evaluación 16
1.5. Estructura 17
2. Materiales y Métodos 18
2.1. Descripción de los datos 18
2.2. Análisis exploratorio de datos (EDA) 20
2.3. Preprocesamiento y limpieza de datos 25
2.4. Métodos de análisis y modelado 26
2.5. Evaluación del modelo 38
2.5.1. Métricas de evaluación 38
2.5.2. Métodos de validación 38
2.5.3. Comparación entre modelos 39
3. Resultados y Discusión 56
3.1. Resultados ajustando el threshold para maximizar el F1-score 58
3.2. Resultados ajustando el threshold para maximizar el recall 59
3.3. Interpretación 62
3.4. Evaluación crítica 65
4. Conclusiones 67
Referencias bibliográficas 69
Anexos 70

Lista de Figuras

Figura 1. Frecuencia de instancias para variables binarias vs diagnóstico..... 21
Figura 2. Frecuencia de instancias para variables ordinales y continuas vs diagnóstico 22
Figura 3. Frecuencia de diagnóstico de Diabetes 23
Figura 4. Matriz de correlación 24
Figura 5. Selección secuencial de características (SFFS) 35
Figura 6. Varianza explicada acumulada en función del número de componentes (PCA).. 36
Figura 7. F1-score en función del número de componentes discriminantes (LDA) 37
Figura 8. Matriz de confusión de la Etapa 1 - Modelo SVM con kernel polinómico 56
Figura 9. Matriz de confusión de la Etapa 4-1 - Modelo SVM con kernel polinómico..... 57
Figura 10. Matriz de confusión de la Etapa 1 - Modelo XGBoost con threshold de 0.35 .. 59
Figura 11. Matriz de confusión de la Etapa 3 - Modelo Gradient Boosting con threshold de 0.3 59
Figura 12. Matriz de confusión de la Etapa 4-2 - Modelo Gradient Boosting con threshold de 0.3 60

Lista de Tablas

Tabla 1. Codificación de la Variable Age. 19
Tabla 2. Hiperparámetro – Regresión Logistica 27
Tabla 3. Hiperparámetro – Clasificador Bayesiano 27
Tabla 4. Hiperparámetro – k-Nearest Neighbors 28
Tabla 5. Hiperparámetro – Máquinas de Soporte Vectorial 29
Tabla 6. Hiperparámetro – Árboles de Decisión 30
Tabla 7. Hiperparámetro – Random Forest 31
Tabla 8. Hiperparámetro – AdaBoost 32
Tabla 9. Hiperparámetro – Gradient Boosting 33
Tabla 10. Hiperparámetro – Extreme Gradient Boosting 34
Tabla 11. Hiperparámetros óptimos y métrica de desempeño en la etapa 1 40
Tabla 12. Hiperparámetros óptimos y métrica de desempeño en la etapa 2 41
Tabla 13. Hiperparámetros óptimos y métrica de desempeño en la etapa 3 43
Tabla 14. Hiperparámetros óptimos y métrica de desempeño en la etapa 4 – proceso 1 44
Tabla 15. Hiperparámetros óptimos y métrica de desempeño en la etapa 4 – proceso 2 46
Tabla 16. Resultados en conjunto de prueba en la etapa 1 (sin threshold y F1-score) 47
Tabla 17. Resultados en conjunto de prueba en la etapa 1 (sin threshold y Recall) 48
Tabla 18. Resultados en conjunto de prueba en la etapa 1 (Threshold 0.35 y F1-score)..... 48
Tabla 19. Resultados en conjunto de prueba en la etapa 1 (Threshold 0.35 y Recall) 48
Tabla 20. Resultados en conjunto de prueba en la etapa 2 (sin threshold y F1-score) 49
Tabla 21. Resultados en conjunto de prueba en la etapa 2 (sin threshold y Recall) 49
Tabla 22. Resultados en conjunto de prueba en la etapa 2 (Threshold 0.35 y F1-score)..... 50
Tabla 23. Resultados en conjunto de prueba en la etapa 2 (Threshold 0.35 y Recall) 50
Tabla 24. Resultados en conjunto de prueba en la etapa 3 (sin threshold y F1-score) 51
Tabla 25. Resultados en conjunto de prueba en la etapa 3 (sin threshold y Recall) 51
Tabla 26. Resultados en conjunto de prueba en la etapa 3 (Threshold 0.3 y F1-score)..... 52
Tabla 27. Resultados en conjunto de prueba en la etapa 3 (Threshold 0.3 y Recall) 52
Tabla 28. Resultados en conjunto de prueba en la etapa 4 – p1 (sin threshold y F1-score) 53
Tabla 29. Resultados en conjunto de prueba en la etapa 4 - p1 (Threshold 0.3 y F1-score)53
Tabla 30. Resultados en conjunto de prueba en la etapa 4 – p1 (Threshold 0.3 y Recall) .. 54

Tabla 31. Resultados en conjunto de prueba en la etapa 4 – p2 (sin threshold y F1-score) 54
Tabla 32. Resultados en conjunto de prueba en la etapa 4 – p2 (Threshold 0.3 y F1-score) 55
Tabla 33. Resultados en conjunto de prueba en la etapa 4 – p2 (Threshold 0.3 y Recall) .. 55

Siglas, acrónimos y abreviaturas

ADA	American Diabetes Association
ADASYN	Adaptive Synthetic Sampling
AdaBoost	Adaptive Boosting
AHA	American Heart Association
BMI	Body Mass Index
BRFSS	Behavioral Risk Factor Surveillance System
CDC	Centers for Disease Control and Prevention
EDA	Exploratory Data Analysis
HDL	High-Density Lipoprotein
KNN	k-Nearest Neighbors
LDA	Linear Discriminant Analysis
LDL	Low-Density Lipoprotein
LIME	Local Interpretable Model-agnostic Explanations
LOF	Local Outlier Factor
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic
SFS	Sequential Feature Selection
SFFS	Sequential Forward Floating Selection
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

Resumen

Los avances en analítica de datos aplicada al sector salud permiten apoyar la detección temprana de enfermedades y la toma de decisiones preventivas. En particular, los modelos de aprendizaje automático identifican patrones en grandes volúmenes de información y generan predicciones útiles para el diagnóstico y seguimiento de enfermedades crónicas como la diabetes.

El objetivo de esta investigación fue implementar modelos de aprendizaje automático supervisado para predecir la presencia de diabetes a partir de variables demográficas, clínicas y de hábitos de vida. Para ello, se utilizó el conjunto de datos Behavioral Risk Factor Surveillance System (BRFSS) y se aplicaron etapas de análisis exploratorio, eliminación de valores atípicos, balanceo de clases, escalado de variables y técnicas de reducción de dimensión como SFFS, PCA y LDA.

Se evaluaron nueve algoritmos de clasificación, incluyendo Regresión Logística, Naive Bayes, k-Nearest Neighbors, Máquinas de Soporte Vectorial, Árboles de Decisión, Random Forest, AdaBoost, Gradient Boosting y XGBoost. Uno de los principales retos fue equilibrar el desempeño predictivo con la interpretabilidad del modelo. El modelo seleccionado fue Máquinas de Soporte Vectorial con kernel polinómico, entrenado con las variables originales, alcanzando un F1-score de 0.7656 y un recall de 0.8012. Los resultados demuestran que el aprendizaje automático constituye una herramienta efectiva para apoyar la detección temprana de diabetes.

Palabras clave: *diabetes, aprendizaje automático, detección temprana, F1-score y Recall.*

Abstract

Advances in healthcare analytics can support early disease detection and improve preventive decision-making. In particular, machine learning models can identify patterns in large volumes of data and generate useful predictions for the diagnosis and monitoring of chronic diseases such as diabetes.

The objective of this research was to design and implement supervised machine learning models to predict the presence of diabetes using demographic, clinical, and lifestyle variables. The Behavioral Risk Factor Surveillance System (BRFSS) dataset was used, and several preprocessing steps were applied, including exploratory data analysis, outlier removal, class balancing, feature scaling, and dimensionality reduction techniques such as Sequential Forward Floating Selection (SFFS), Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA).

Nine classification algorithms were evaluated, including Logistic Regression, Naive Bayes, k-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forest, AdaBoost, Gradient Boosting, and XGBoost. One of the main challenges was balancing predictive performance with model interpretability. The selected model was a polynomial-kernel Support Vector Machine trained on the original variables, achieving an F1-score of 0.7656 and a recall of 0.8012. The results demonstrate that machine learning is an effective tool for supporting early diabetes detection.

Keywords: diabetes, machine learning, early detection, F1-score and recall.

1. Introducción

1.1. Descripción del problema

La diabetes es una enfermedad crónica que se presenta cuando el organismo no produce suficiente insulina o no la utiliza de manera eficiente. Esta condición constituye uno de los principales problemas de salud pública a nivel mundial, debido a su alta prevalencia y a las complicaciones graves que puede generar, tales como enfermedades cardiovasculares, insuficiencia renal, pérdida de visión y alteraciones neurológicas. Diversos factores de riesgo, como una dieta inadecuada, el sedentarismo, el consumo de alcohol y el tabaquismo, se asocian significativamente con el desarrollo de la enfermedad (Oguntibeju, 2013). De acuerdo con la Organización Mundial de la Salud, la incidencia de la diabetes ha aumentado de forma considerable en las últimas décadas, posicionándose como una de las principales causas de mortalidad y discapacidad en el mundo (World Health Organization, 2023).

La diabetes genera diferentes implicaciones para la salud en general:

- La asociación americana de enfermedades cardiovasculares, (AHA por sus siglas en inglés), establece que la diabetes disminuye los niveles de colesterol HDL (también conocido como colesterol bueno), aumenta los niveles de triglicéridos y los niveles de colesterol LDL (también conocido como colesterol malo). Estos últimos incrementan el riesgo de padecer enfermedades cardiovasculares y derrames cerebrales (National Institute of Diabetes and Digestive and Kidney Diseases, s. f.).
- La diabetes tipo 2 puede aumentar el riesgo de demencia, como la enfermedad de Alzheimer o la esquizofrenia.
- Los síntomas de depresión son comunes en personas con diabetes tipo 1 y tipo 2. Tienen entre 2 y 3 veces más probabilidades de presentar depresión que las personas sin diabetes (Centers for Disease Control and Prevention, 2023).

Aunque la diabetes es una enfermedad crónica que en muchos casos no puede prevenirse completamente, la adopción de hábitos de vida saludables desempeña un papel fundamental en su control y manejo. Entre las principales recomendaciones se encuentran mantener una alimentación equilibrada, realizar actividad física de manera regular y controlar el peso corporal. Estas prácticas contribuyen a mejorar la sensibilidad a la insulina, regular los niveles de glucosa en sangre y reducir el riesgo de complicaciones asociadas a la enfermedad (American Diabetes Association, 2024).

Uno de los principales desafíos en el abordaje de la diabetes radica en que sus síntomas iniciales suelen ser leves o incluso imperceptibles, lo que dificulta su detección temprana. Como consecuencia, muchas personas desconocen que padecen la enfermedad hasta que se presentan complicaciones derivadas del daño prolongado causado por niveles elevados de

glucosa en la sangre. En este contexto, la detección temprana de la diabetes adquiere una importancia fundamental, ya que permite implementar intervenciones oportunas, mejorar el control glucémico y reducir el riesgo de desarrollar complicaciones graves como enfermedades cardiovasculares, daño renal, neuropatías y problemas visuales (Centers for Disease Control and Prevention, 2023).

La ciencia de datos juega un papel fundamental en el análisis de grandes volúmenes de información en el ámbito de la salud. En el caso de la diabetes, el uso de herramientas de ciencia de datos permite analizar múltiples factores de riesgo de manera simultánea, facilitando la identificación temprana de individuos con mayor probabilidad de desarrollar la enfermedad. Mediante técnicas de análisis estadístico, minería de datos y aprendizaje automático es posible identificar patrones ocultos en los datos e implementar modelos predictivos que apoyen la toma de decisiones basadas en evidencia (Provost & Fawcett, 2013; Han et al., 2012). En el caso de la diabetes, estas herramientas permiten analizar múltiples factores de riesgo de forma simultánea, facilitando la identificación temprana de individuos con mayor probabilidad de desarrollar la enfermedad y contribuyendo al diseño de estrategias más efectivas de prevención, diagnóstico y control (World Health Organization, 2023).

En este contexto, el análisis de datos relacionados con indicadores de salud permite identificar relaciones entre diferentes variables asociadas con la presencia o el riesgo de desarrollar diabetes, lo cual resulta clave para la implementación de modelos predictivos que apoyen el diagnóstico temprano y la gestión de la enfermedad.

Actualmente, la creciente disponibilidad de datos de salud provenientes de encuestas, registros médicos electrónicos y sistemas de información sanitaria ha generado nuevas oportunidades para la implementación de modelos analíticos que apoyen la gestión y prevención de enfermedades crónicas como la diabetes. El uso de técnicas de análisis de datos y aprendizaje automático permite identificar patrones, factores de riesgo y posibles predictores asociados a la enfermedad. Sin embargo, aún existen desafíos importantes, entre ellos la calidad y disponibilidad de los datos, la integración de diferentes fuentes de información, la interpretabilidad de los modelos y la implementación efectiva de estas herramientas en entornos clínicos o de salud pública (Obermeyer & Emanuel, 2016).

En este contexto, el análisis de datasets relacionados con indicadores de salud permite explorar cómo diferentes variables pueden influir en la aparición de la diabetes, aportando conocimiento valioso para la identificación de factores de riesgo y el desarrollo de soluciones basadas en datos que apoyen la toma de decisiones en el ámbito sanitario.

1.2. Planteamiento del problema

La diabetes constituye una de las enfermedades crónicas con mayor impacto en la salud pública a nivel mundial, debido a su alta prevalencia y a las complicaciones asociadas cuando no se detecta o controla oportunamente. Esta problemática afecta de manera significativa a los sistemas de salud, hospitales y clínicas, los cuales enfrentan un alto flujo de pacientes con diversas patologías, entre ellas aquellas relacionadas con trastornos metabólicos. En este contexto, la identificación temprana de personas con alto riesgo de desarrollar diabetes resulta fundamental para implementar estrategias de prevención, tratamiento y seguimiento que permitan reducir su impacto en la calidad de vida de la población (World Health Organization, 2023).

A pesar de los avances en el diagnóstico y tratamiento de esta enfermedad, en muchos casos su detección ocurre cuando ya se han desarrollado síntomas o complicaciones. Esta situación evidencia la necesidad de desarrollar herramientas analíticas capaces de procesar grandes volúmenes de datos de salud y detectar patrones asociados con la presencia o el riesgo de diabetes. En este sentido, el uso de información que integra indicadores de salud, hábitos de vida y variables demográficas de los pacientes representa una oportunidad para aplicar técnicas de ciencia de datos e implementar modelos predictivos que contribuyan a la identificación temprana de la enfermedad.

La motivación de este estudio surge de la creciente necesidad de aprovechar el gran volumen de datos disponibles en el ámbito de la salud para mejorar la comprensión y el manejo de enfermedades crónicas como la diabetes. Actualmente, los sistemas de salud generan grandes cantidades de información provenientes de encuestas poblacionales, registros médicos y sistemas de información clínica; sin embargo, una proporción considerable de estos datos no es explotada de manera integral para identificar patrones relevantes que faciliten la detección temprana de enfermedades. En este contexto, la ciencia de datos proporciona herramientas y metodologías que permiten analizar grandes conjuntos de datos, identificar relaciones entre múltiples variables e implementar modelos predictivos que apoyen el diseño de estrategias de prevención más efectivas y el fortalecimiento de los procesos de diagnóstico temprano (Provost & Fawcett, 2013).

La realización de este estudio se justifica por la necesidad de avanzar hacia enfoques predictivos en el ámbito de la salud, mediante el uso de herramientas analíticas que permitan mejorar la identificación temprana del riesgo de desarrollar diabetes. Esto contribuye al apoyo en la toma de decisiones preventivas y al fortalecimiento de los procesos de diagnóstico y seguimiento de los pacientes. La implementación de estrategias basadas en el análisis de datos facilita la identificación de factores asociados al desarrollo de la enfermedad, permitiendo diseñar acciones de prevención más efectivas y fortalecer las estrategias de salud pública orientadas a reducir la incidencia de la diabetes y sus complicaciones. En un contexto donde las enfermedades crónicas representan un desafío

creciente para los sistemas de salud, el uso de modelos predictivos basados en datos puede contribuir a optimizar los recursos disponibles y mejorar la calidad de vida de la población.

Adicionalmente, el análisis de información proveniente de indicadores de salud, hábitos de vida y variables demográficas permite comprender de manera más profunda la relación entre diferentes factores de riesgo y la aparición de la enfermedad. Esto facilita el desarrollo de herramientas analíticas que pueden ser utilizadas como apoyo en la evaluación del riesgo de diabetes, tanto por profesionales de la salud como en estudios epidemiológicos (World Health Organization, 2023).

No obstante, a pesar de la existencia de diversos estudios que aplican métodos estadísticos y técnicas de aprendizaje automático para la predicción de enfermedades crónicas, aún persisten desafíos importantes. Entre ellos se destacan la reducción adecuada de variables, la interpretabilidad de los modelos y la capacidad de generalizar los resultados en diferentes poblaciones. En muchos casos, los modelos implementados priorizan el desempeño predictivo, pero presentan limitaciones en términos de interpretabilidad, lo que dificulta su comprensión y aplicación en contextos clínicos por parte de los profesionales de la salud.

En este sentido, el análisis de datasets basados en indicadores de salud representa una oportunidad para explorar nuevas aproximaciones analíticas que permitan mejorar la identificación de factores asociados al riesgo de diabetes y apoyar la toma de decisiones basadas en datos. A través de este proyecto de monografía, se propone el desarrollo de una alternativa analítica orientada a generar resultados interpretables y de fácil comprensión, de modo que puedan ser utilizados de forma intuitiva tanto por el personal médico como por los pacientes involucrados en el proceso de evaluación del riesgo de diabetes.

1.3. Objetivo del estudio

1.3.1. Objetivo General

Diseñar e implementar modelos de aprendizaje automático supervisado para la predicción temprana de la diabetes, utilizando variables demográficas, clínicas y de hábitos de vida, con el fin de estimar la probabilidad de desarrollar la enfermedad y apoyar la toma de decisiones preventivas en salud.

1.3.2. Objetivos Específicos

- Analizar y caracterizar las variables demográficas, clínicas y de hábitos de vida para identificar su relación con la presencia de diabetes.
- Implementar modelos de aprendizaje automático supervisado, aplicando técnicas de preprocesamiento y reducción de dimensión, con el fin de mejorar la capacidad predictiva del modelo.

- Validar la capacidad del modelo para identificar de forma temprana el riesgo de diabetes, generando alertas que permitan apoyar la detección oportuna de la enfermedad y la toma de decisiones preventivas.

1.4. Metodología

La presente investigación se desarrolla bajo un enfoque cuantitativo, apoyado en métodos estadísticos y técnicas de aprendizaje automático, con el objetivo de analizar datos relacionados con factores de riesgo y construir modelos predictivos para la detección temprana de la diabetes.

1.4.1. Datos utilizados y su fuente

El estudio se basa en un conjunto de datos compuesto por variables demográficas, clínicas y hábitos de vida de los pacientes, las cuales han sido identificadas en la literatura médica como factores relevantes asociados al riesgo de desarrollar diabetes.

La fuente de los datos corresponde a la encuesta Behavioral Risk Factor Surveillance System (BRFSS), desarrollada por el Centers for Disease Control and Prevention (CDC) en el año 2015. Este conjunto de datos original contiene aproximadamente 441.456 registros. A partir de este, se realizó un proceso de limpieza y adaptación para su uso en la competición de Kaggle denominada *"Diabetes Prediction Competition (TFUG CHD Nov 2022)"*, obteniéndose un dataset final de aproximadamente 100.000 registros.

Este volumen de información proporciona suficiente representatividad y variabilidad, permitiendo el entrenamiento de modelos de aprendizaje supervisado y la división adecuada de los datos en conjuntos de entrenamiento y prueba, garantizando estabilidad en los resultados obtenidos.

1.4.2. Algoritmos, modelos y técnicas aplicadas

El procesamiento de los datos se llevó a cabo mediante un conjunto de técnicas orientadas a garantizar la calidad de la información y a optimizar el desempeño de los modelos predictivos.

A. Preprocesamiento de datos

- Se aplicaron las siguientes técnicas:
  - Limpieza de datos y tratamiento de valores faltantes
  - Detección y análisis de valores atípicos (outliers)
  - Análisis exploratorio de datos (EDA)
  - Estadística descriptiva para la caracterización de variables
  - Escalado de variables numéricas mediante normalización
  - Balanceo de clases
  - División del dataset en conjuntos de entrenamiento y prueba

- Adicionalmente, se implementaron técnicas de reducción de dimensión, con el fin de mejorar la eficiencia y el desempeño de los modelos:

    - Sequential Feature Selection (SFFS)
    - Análisis de Componentes Principales (PCA)
    - Análisis Discriminante Lineal (LDA)

B. Modelos de aprendizaje automático

- Se implementaron y evaluaron diferentes modelos de clasificación supervisada:

    - Regresión Logística
    - Clasificador Bayesiano
    - k-Nearest Neighbors (KNN)
    - Máquinas de Soporte Vectorial (SVM)
    - Árboles de Decisión
    - Random Forest
    - AdaBoost
    - Gradient Boosting
    - XGBoost

- Para cada modelo se realizó el ajuste de hiperparámetros con el fin de optimizar su desempeño.

1.4.3. Criterios de evaluación

El desempeño de los modelos se evaluó utilizando métricas de clasificación, entre las cuales se incluyen:

- Exactitud (Accuracy)
- Precisión (Precision)
- Sensibilidad (Recall)
- F1-score

La métrica principal utilizada fue el F1-score, debido a que permite evaluar de manera equilibrada la precisión y el recall, siendo especialmente útil en problemas donde existe desbalance entre clases.

Se empleó la matriz de confusión, con el fin de analizar el comportamiento de los modelos en términos de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Esto permitió una interpretación más detallada del desempeño de cada modelo y su capacidad para identificar correctamente los casos positivos de diabetes.

Adicionalmente, para los modelos que generan probabilidades de clasificación, se realizó el ajuste del umbral de decisión (threshold) con el objetivo de mejorar el equilibrio entre la precisión y el recall. Este proceso permitió optimizar el desempeño del modelo, especialmente en la identificación de casos positivos, lo cual es fundamental en problemas de salud donde la detección temprana es prioritaria.

1.5. Estructura

El presente estudio se encuentra estructurado en cuatro secciones principales, las cuales permiten desarrollar de manera ordenada el proceso de investigación y los resultados obtenidos:

- **Sección 1:** se presenta el contexto general del problema de investigación, el planteamiento del problema, la justificación del estudio y los objetivos general y específicos que orientan el desarrollo del trabajo. En esta sección también se expone la relevancia de la detección temprana de la diabetes y el potencial del aprendizaje automático como herramienta de apoyo para la toma de decisiones en salud.
- **Sección 2:** se describe de manera detallada el conjunto de datos utilizado, las variables consideradas y las diferentes etapas del proceso analítico. Se incluyen las actividades de limpieza y preprocesamiento de los datos, el tratamiento del desbalance de clases, las técnicas de reducción de dimensión, los algoritmos de aprendizaje automático implementados, la optimización de hiperparámetros y los criterios de validación y evaluación empleados para comparar el desempeño de los modelos.
- **Sección 3:** se presentan los principales hallazgos del estudio, incluyendo las métricas obtenidas por cada modelo y la comparación entre las diferentes estrategias evaluadas. Asimismo, se realiza la interpretación de los resultados en el contexto del problema de negocio, se analizan los factores que influyeron en el desempeño de los modelos y se contrastan los resultados con estudios previos reportados en la literatura científica.
- **Sección 4:** se resumen los hallazgos más relevantes del trabajo, el grado de cumplimiento de los objetivos planteados y las implicaciones prácticas del modelo seleccionado. También se exponen las principales limitaciones del estudio y se proponen posibles líneas de investigación futura para continuar fortaleciendo la capacidad predictiva y la aplicabilidad de este tipo de modelos en el ámbito de la salud.

2. Materiales y Métodos

2.1. Descripción de los datos

Los datos utilizados provienen de la encuesta pública Behavioral Risk Factor Surveillance System (BRFSS), desarrollada por el Centers for Disease Control and Prevention (CDC) en 2015. Este conjunto de datos, basado en encuestas de salud poblacional, fue posteriormente adaptado para la competición de Kaggle *"Diabetes Prediction Competition (TFUG CHD Nov 2022)"*, incluyendo aproximadamente 100.000 registros con variables demográficas, clínicas y de hábitos de vida (Kaggle, 2022).

El conjunto de datos se divide en dos secciones: un conjunto de entrenamiento (train) compuesto por 80.692 observaciones y 18 variables, incluyendo la variable objetivo Diabetes, y un conjunto de validación con 20.000 observaciones y 17 variables, en el cual no se incluye la variable objetivo. Todas las variables son de tipo numérico y corresponden a variables categóricas codificadas (binarias y ordinales) y variables continuas, relacionadas con características demográficas, condiciones clínicas y hábitos de vida de los pacientes.

El conjunto de datos está compuesto por variables de tipo binario, ordinal y continuo, las cuales permiten analizar de manera integral los factores asociados al riesgo de desarrollar diabetes. A continuación, se describen las variables empleadas:

a. Variables categóricas binarias (Corresponden a variables categóricas con dos posibles valores (0 y 1).
   - ***HighChol:*** Hace referencia a los niveles de colesterol en la sangre del paciente.
     0: Dentro del límite normal y 1: Mayor al límite o fuera del rango.
   - **Sex:** Corresponde con el género del paciente.
     0: Femenino y 1: Masculino.
   - ***CholCheck:*** Determina si el paciente ha realizado un control para medir sus niveles de colesterol en la sangre en los últimos 5 años.
     0: No y 1: Sí.
   - ***Smoker:*** Almacena la respuesta a la pregunta: ¿Ha fumado al menos 100 cigarrillos en toda su vida?
     0: No y 1: Sí.
   - ***HeartDiseaseorAttack:*** Hace referencia a la respuesta de la pregunta: ¿Sufre enfermedades coronarias o ha tenido infarto del miocardio?
     0: No y 1: Sí.
   - ***PhysActivity:*** Establece si el paciente ha practicado actividad física durante los últimos 30 días sin incluir la actividad inherente al trabajo.
     0: No y 1: Sí.
   - ***Fruits:*** Permite conocer si el paciente consume fruta una o más veces al día.
     0: No y 1: Sí.

- ***Veggies:*** Define si el paciente consume vegetales una o más veces al día.
  0: No y 1: Sí.
- ***HvyAlcoholConsump:*** Hace referencia a la siguiente pregunta dependiendo del género del paciente: para un adulto hombre, se pregunta si toma más de 14 bebidas alcohólicas por semana; para una mujer adulta, se pregunta si toma más de siete bebidas alcohólicas por semana.
  0: No y 1: Sí.
- ***DiffWalk:*** Almacena la respuesta de la siguiente pregunta: ¿tiene usted dificultades serias para caminar o subir escaleras?
  0: No y 1: Sí.
- ***Hypertension:*** Esta variable hace referencia a si el paciente padece de hipertensión.
  0: No y 1: Sí.
- ***Stroke:*** Establece si el paciente ha sufrido un derrame cerebral.
  0: No y 1: Sí.
- ***Diabetes:*** Esta variable categórica es la variable de respuesta dentro del conjunto de datos y hace referencia a si el paciente padece diabetes o no.
  0: No y 1: Sí.

b. Variables categóricas ordinales (Corresponden a variables categóricas con un orden definido).

- ***Age:*** Esta variable está relacionada con los rangos etarios de los pacientes que diligenciaron la encuesta. La codificación para este atributo de la base de datos es la siguiente:

Tabla 1. Codificación de la Variable Age.

Codificación	Rango de edad
1	18 – 24
2	25 – 29
3	30 – 34
4	35 – 39
5	40 – 44
6	45 – 49
7	50 – 54
8	55 – 59
9	60 – 64
10	65 – 69
11	70 – 74
12	75 – 79
13	80 o más

- ***GenHlth:*** Almacena la respuesta a la siguiente pregunta: ¿en una escala de uno a cinco: diría que en general su salud es: 1: excelente, 2: muy buena, 3: buena, 4: regular y 5: mala.

c. Variables continuas (Corresponden a variables numéricas que pueden tomar múltiples valores dentro de un rango).

- ***BMI:*** Variable numérica continua, hace referencia al índice de masa corporal, da información de los niveles de obesidad que tiene el paciente.
- ***PhysHlth:*** Variable numérica discreta que contabiliza el número de días (escala de 1 a 30) en los últimos 30 días en los que el paciente reportó problemas de salud física, incluyendo enfermedades o lesiones.
- ***MentHlth:*** Variable numérica discreta que contabiliza el número de días (escala de 1 a 30) en los últimos 30 días en los que el paciente reportó un estado de salud mental no favorable, incluyendo estrés, depresión y problemas emocionales.

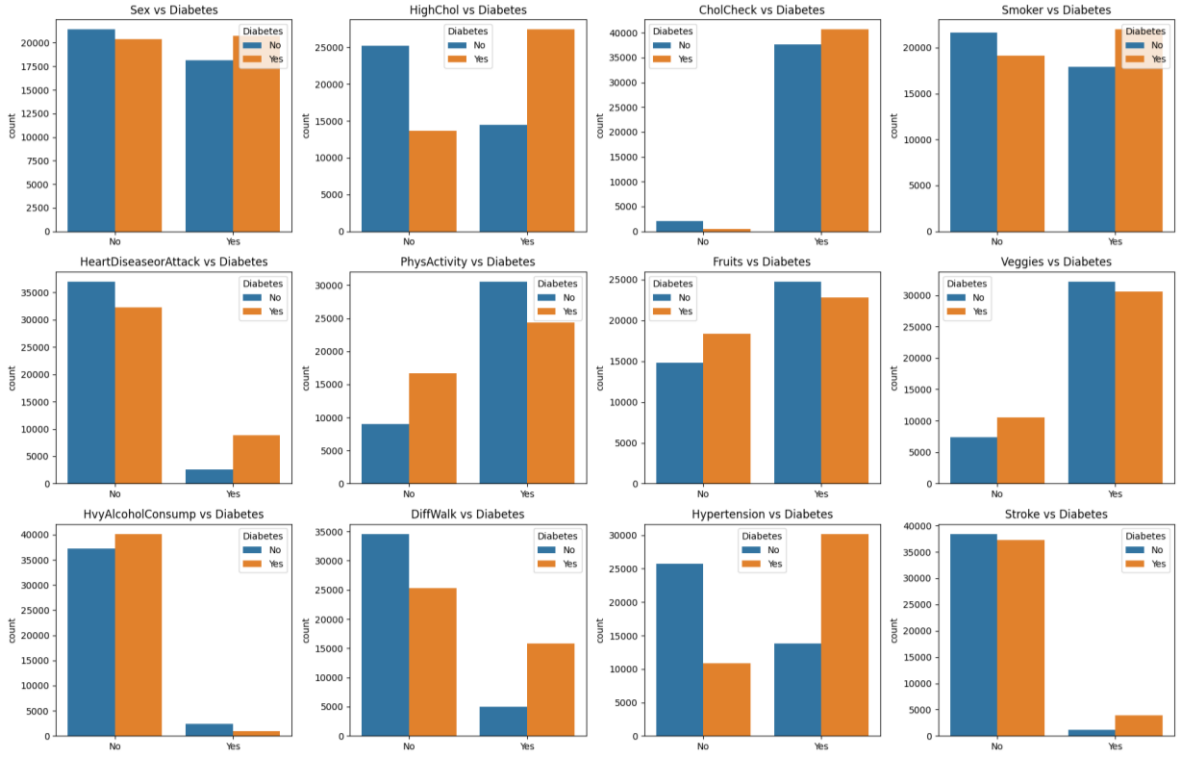
El conjunto de datos no presenta valores faltantes, lo que facilita su procesamiento y análisis. Sin embargo, se identifican posibles limitaciones inherentes a su naturaleza; al tratarse de datos provenientes de encuestas, existe la posibilidad de sesgo de respuesta, ya que la información es autorreportada por los participantes. Adicionalmente, se observa un ligero desbalance en la variable objetivo (Diabetes), lo cual puede afectar el desempeño de los modelos de clasificación, favoreciendo la clase mayoritaria. Este aspecto fue abordado mediante técnicas de balanceo de clases durante el preprocesamiento. Finalmente, se identificó la presencia de valores atípicos (outliers) en algunas variables numéricas, los cuales fueron tratados mediante técnicas de detección y eliminación, con el fin de mejorar la calidad de los datos y el rendimiento de los modelos.

2.2. Análisis exploratorio de datos (EDA)

a. Frecuencia de instancias para variables binarias vs diagnóstico

Se realizó la visualización de la frecuencia de instancias de las variables binarias en relación con la variable objetivo Diabetes. A través de gráficos de barras, se comparó la distribución de las categorías (0 y 1) para cada variable según el diagnóstico, lo que permitió identificar patrones y diferencias entre individuos con y sin diabetes.

Figura 1. Frecuencia de instancias para variables binarias vs diagnóstico



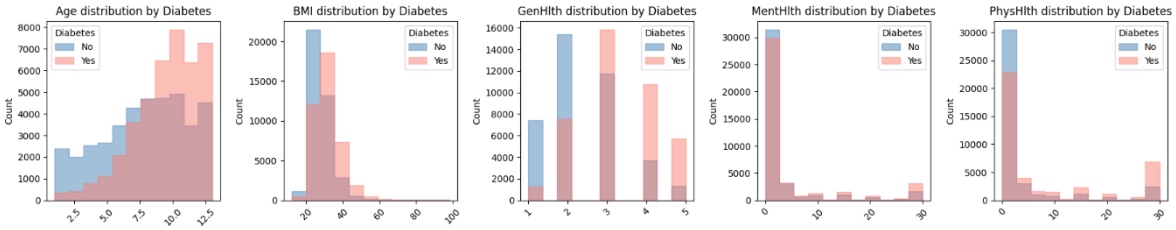
A partir de la visualización en la **Figura 1** se identifican patrones relevantes asociados al riesgo de la enfermedad:

- Variables como *HighChol* e *Hypertension* muestran una relación clara con la presencia de diabetes, evidenciando una mayor proporción de individuos con diagnóstico positivo en las categorías donde estas condiciones están presentes. Esto sugiere una fuerte asociación entre enfermedades cardiovasculares y el riesgo de diabetes.
- En variables como *HeartDiseaseorAttack* y *Stroke*, aunque la mayoría de los individuos se concentra en la categoría “No”, se observa que en la categoría “Sí” existe una mayor proporción de casos positivos de diabetes. Esto indica que, aunque estas condiciones son menos frecuentes en la población, su presencia está asociada a un mayor riesgo de desarrollar diabetes.
- En variables relacionadas con hábitos de vida, como *PhysActivity*, se observa que la inactividad física está asociada a una mayor proporción de casos de diabetes, mientras que la actividad física parece tener un efecto protector. Asimismo, variables como *Fruits* y *Veggies* indican que un menor consumo de alimentos saludables está relacionado con una mayor presencia de la enfermedad.
- La variable *DiffWalk* muestra una diferencia significativa, donde los individuos con dificultad para caminar presentan una mayor proporción de diabetes, lo cual puede estar asociado a condiciones de salud subyacentes.

b. Frecuencia de instancias para variables ordinales y continuas vs diagnóstico

Para las variables ordinales y continuas, se analizaron sus distribuciones en función del diagnóstico de diabetes mediante gráficos de frecuencia.

Figura 2. Frecuencia de instancias para variables ordinales y continuas vs diagnóstico



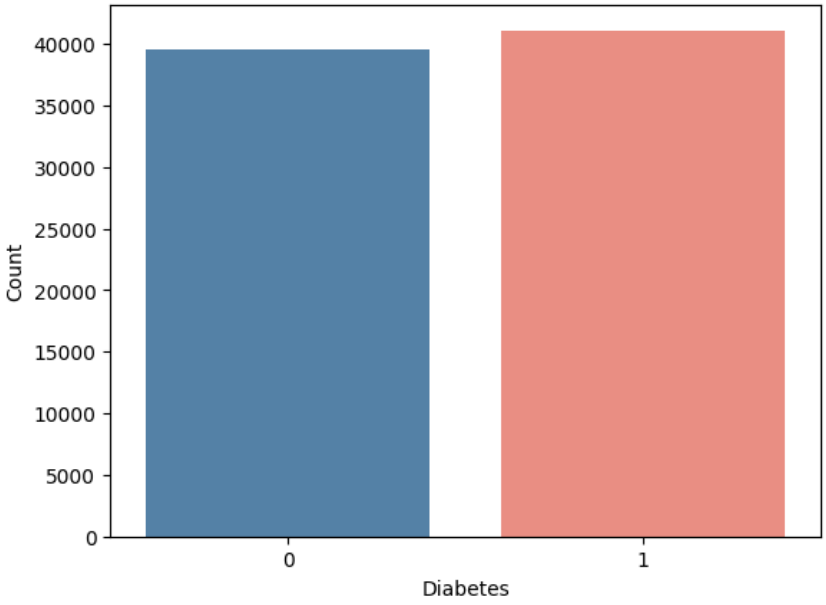
A partir de la visualización en la **Figura 2** se identifican patrones relevantes asociados al riesgo de la enfermedad:

- La variable *Age* evidencia una tendencia clara: los individuos con diabetes se concentran en rangos de edad más altos, lo que sugiere que el riesgo de desarrollar la enfermedad aumenta con la edad.
- En cuanto al *BMI*, se observa que las personas con diabetes presentan valores más elevados en comparación con aquellas sin la enfermedad, lo cual indica una fuerte relación entre el sobrepeso u obesidad y la presencia de diabetes.
- Respecto a la variable *GenHlth*, los individuos con diabetes tienden a reportar estados de salud más desfavorables (valores más cercanos a 4 y 5), mientras que quienes no presentan la enfermedad se concentran en categorías de mejor salud (1 y 2). Esto evidencia una asociación entre la percepción del estado de salud y la presencia de diabetes.
- Por otro lado, las variables *MentHlth* y *PhysHlth* muestran que los individuos con diabetes reportan una mayor cantidad de días con problemas de salud mental y física, respectivamente. En particular, se observa una mayor dispersión hacia valores altos, lo que indica una mayor afectación en su bienestar general.

En conjunto, estos resultados sugieren que factores como la edad avanzada, un mayor índice de masa corporal y un estado de salud general deteriorado están significativamente asociados con la presencia de diabetes, lo cual es consistente con la literatura médica.

c. Distribución de la variable objetivo Diabetes

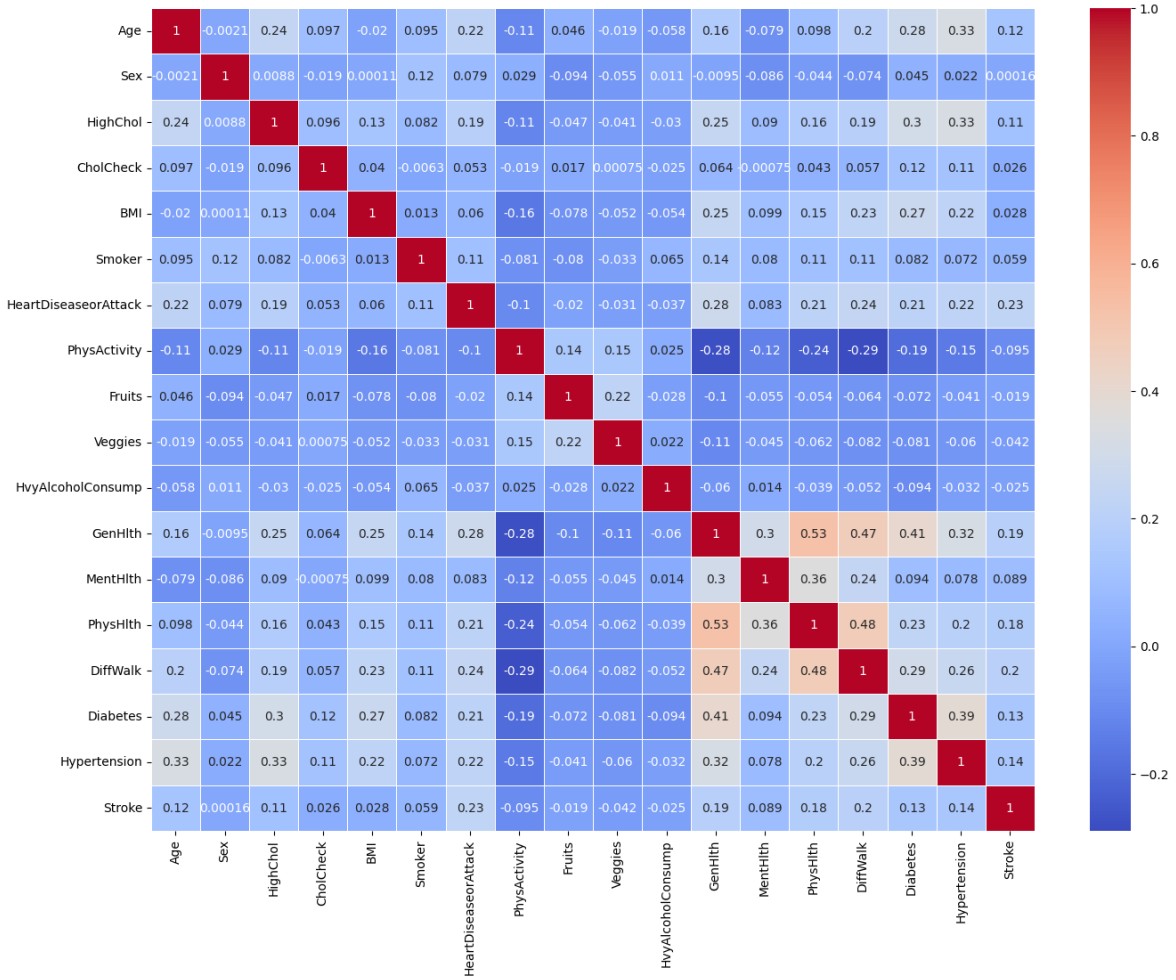
Figura 3. Frecuencia de diagnóstico de Diabetes



Se analizó la distribución de la variable objetivo *Diabetes*, observándose una proporción relativamente equilibrada entre las clases 0 (sin diabetes) y 1 (con diabetes), aunque con una ligera mayor frecuencia de casos positivos. Si bien el desbalance no es extremadamente pronunciado, se decidió aplicar una técnica de balanceo de clases mediante undersampling sobre la clase mayoritaria, con el fin de mejorar la capacidad de los modelos para identificar correctamente los casos de diabetes. Esta decisión se fundamenta en la importancia de reducir el sesgo del modelo hacia la clase dominante y en priorizar métricas como el recall y el F1-score, especialmente relevantes en problemas de salud donde la correcta detección de casos positivos es crítica.

d. Correlación entre variables

Figura 4. Matriz de correlación



A partir de la matriz de correlación, se identifican las variables con mayor relación con la variable objetivo *Diabetes*. En general, las correlaciones observadas son de magnitud baja a moderada, lo cual es esperado en problemas de salud donde múltiples factores influyen en la aparición de la enfermedad. Entre las variables con mayor correlación positiva con *Diabetes* se destacan *GenHlth*, *Hypertension*, *BMI* y *Age*, lo que sugiere que un peor estado de salud general, la presencia de hipertensión, un mayor índice de masa corporal y una mayor edad están asociados con un mayor riesgo de desarrollar diabetes.

Por otro lado, variables como *PhysActivity*, *Fruits* y *Veggies* presentan correlaciones negativas, indicando que hábitos de vida saludables, como la actividad física y una alimentación balanceada, podrían estar asociados con una menor probabilidad de padecer la enfermedad. Asimismo, se observa que no existen correlaciones extremadamente altas entre las variables independientes, lo cual sugiere una baja multicolinealidad en el conjunto de datos.

2.3. Preprocesamiento y limpieza de datos

a. División de los datos en conjuntos de entrenamiento, validación y prueba

El primer paso consistió en dividir el conjunto de datos en subconjuntos de entrenamiento y prueba, utilizando una proporción de 90% para entrenamiento y 10% para prueba. Como ya se mencionó anteriormente, ya se cuenta con un conjunto de validación independiente, previamente definido. Esta separación se realizó al inicio del proceso con el fin de evitar fugas de información (*data leakage*) y garantizar que el conjunto de prueba represente datos no vistos durante el entrenamiento. De esta manera, se busca evaluar el desempeño de los modelos en condiciones más cercanas a un escenario real.

Es importante destacar que el conjunto de prueba conserva la distribución original de los datos, incluyendo la posible presencia de valores atípicos, lo cual permite analizar el comportamiento del modelo frente a situaciones reales.

b. Manejo de valores nulos y outliers

El conjunto de datos no presenta valores nulos, lo cual facilita su procesamiento y evita la necesidad de aplicar técnicas de imputación.

En cuanto a los valores atípicos, se aplicó el algoritmo Local Outlier Factor (LOF) para su identificación en las variables numéricas. Este método permite detectar observaciones que presentan un comportamiento significativamente diferente al de sus vecinos más cercanos, basándose en la densidad local de los datos. Para la implementación del método, se definió el número de vecinos (k) como la raíz cuadrada del número total de observaciones, lo cual permite un balance adecuado entre sensibilidad y estabilidad en la detección de outliers. Asimismo, se estableció un parámetro de contaminación del 10%, indicando la proporción estimada de valores atípicos en el conjunto de datos. A partir de este proceso, se identificaron las observaciones consideradas como outliers, las cuales fueron posteriormente eliminadas del conjunto de datos con el fin de mejorar la calidad de la información y el desempeño de los modelos de aprendizaje automático.

c. Balanceo de clases

Debido a que el conjunto de datos presenta un ligero desbalance en la variable objetivo, se aplicó una técnica de undersampling (submuestreo) sobre la clase mayoritaria en el conjunto de entrenamiento, con el fin de equilibrar la proporción de clases.

Para ello, se aplicó la técnica de RandomUnderSampler el cual reduce aleatoriamente el número de observaciones de la clase mayoritaria hasta igualarlo con la clase minoritaria. Lo anterior permite que los modelos aprendan de manera más equilibrada y evitando un sesgo hacia la clase dominante.

Es importante destacar que este proceso se aplicó exclusivamente al conjunto de entrenamiento, mientras que el conjunto de prueba conserva la distribución original de los datos, con el fin de evaluar el desempeño del modelo en condiciones más cercanas a un entorno real.

d. Transformaciones aplicadas (escalado, normalización, codificación de variables categóricas)

Con el fin de preparar los datos para el entrenamiento de los modelos de aprendizaje automático, se aplicaron diferentes transformaciones orientadas a mejorar la calidad y el desempeño del modelo.

En primer lugar, se realizó el escalado de las variables numéricas ordinales y continuas mediante el uso de MinMaxScaler, el cual transforma los valores a un rango entre 0 y 1. Esta normalización permite que las variables mantengan su relación de orden y facilita el entrenamiento de algoritmos sensibles a la escala de los datos.

En cuanto a la codificación de variables, no fue necesario aplicar técnicas adicionales, ya que todas las variables del conjunto de datos se encontraban previamente representadas en formato numérico. Las variables binarias ya estaban codificadas en valores 0 y 1, mientras que variables ordinales como *Age* y *GenHlth* conservaron su estructura original. Por esta razón, no se consideró necesario aplicar técnicas como One-Hot Encoding a estas variables, dado que los modelos pueden trabajar directamente con valores numéricos y, en este caso, mantener su codificación permite preservar la relación de orden entre las categorías.

2.4. Métodos de análisis y modelado

i. Algoritmos utilizados y configuraciones

Todos los modelos que se consideraron durante este proyecto son de aprendizaje automático supervisado y se seleccionaron por exhibir un buen desempeño en bases de datos de clasificación binaria. Los hiperparámetros óptimos para todos los algoritmos fueron seleccionados luego de haber aplicado una búsqueda en malla o grid search. Adicionalmente, se utilizó el hiperparámetro *random state* como una constante para que la aleatoriedad del modelo no sea un factor influyente a la hora de replicar los resultados. A continuación, se realiza una descripción de cada uno de los modelos empleados:

a. Regresión Logística: modelo estadístico lineal que permite estimar la probabilidad de ocurrencia de un evento binario, siendo ampliamente utilizado por su interpretabilidad y eficiencia. Se usaron los diferentes solucionadores: ‘liblinear’, ‘lbfgs’, ‘newton-cg’ y ‘saga’.

Este modelo se implementó con penalización L2 (regularización Ridge), la cual añade un término de penalización a los coeficientes del modelo con el objetivo de evitar el sobreajuste (*overfitting*) y mejorar la capacidad de generalización.

Finalmente, se establecen los rangos sobre los cuales fueron evaluados los hiperparámetro para este modelo a través de la siguiente tabla:

Tabla 2. Hiperparámetro – Regresión Logistica

Hiperparámetro	Descripción	Rango
Penalty	Define qué tipo de penalización se aplica al modelo para evitar el sobreajuste.	[L2]
Solver	Es aquel criterio que permite definir bajo qué metodología se resolverá el problema de optimización de la regresión.	[liblinear, lbfgs, newton-cg y saga]
C	Esta cantidad representa el inverso de la fuerza de regularización del modelo.	[0.1, 0.5, 1, 2, 5]

b. Clasificador Bayesiano: el cual se basa en el teorema de Bayes y asume independencia entre las variables, permitiendo realizar predicciones rápidas con un bajo costo computacional. En particular, se evaluaron tres variantes del modelo:

- *Gaussian Naive Bayes:* adecuado para variables continuas, asumiendo una distribución normal de los datos.
- *Bernoulli Naive Bayes:* orientado a variables binarias.
- *Multinomial Naive Bayes:* comúnmente utilizado en datos de conteo o frecuencia.

Finalmente, se establecen los rangos sobre los cuales fueron evaluados los hiperparámetro para este modelo a través de la siguiente tabla:

Tabla 3. Hiperparámetro – Clasificador Bayesiano

Hiperparámetro	Descripción	Rango
Modelo	Variantes del clasificador bayesiano (Naive Bayes)	[GaussianNB, BernoulliNB y MultinomialNB]

c. **k-Nearest Neighbors (KNN)**: es un método de clasificación basado en la similitud entre instancias. Este modelo no realiza un proceso de entrenamiento explícito, sino que clasifica nuevas observaciones en función de las k instancias más cercanas en el conjunto de entrenamiento.

   El funcionamiento del algoritmo consiste en calcular la distancia entre una nueva observación y todos los datos del conjunto de entrenamiento, seleccionando los k vecinos más cercanos. Posteriormente, la clase asignada corresponde a la categoría más frecuente entre dichos vecinos.

   Finalmente, se establecen los rangos sobre los cuales fueron evaluados los hiperparámetro para este modelo a través de la siguiente tabla:

Tabla 4. Hiperparámetro – k-Nearest Neighbors

Hiperparámetro	Descripción	Rango
k	Es el hiperparámetro que permite establecer el número de vecinos a utilizar de forma predeterminada.	[1, 10, 50, 100, 200, 300]
Metric	Es aquel criterio que permite calcular la distancia entre puntos del algoritmo.	[euclidean, manhattan y chebyshev]

d. **Máquinas de Soporte Vectorial (SVM)**: modelo que busca encontrar el hiperplano óptimo que separa las clases en el espacio de características, maximizando la distancia (margen) entre los puntos más cercanos de cada clase, conocidos como vectores de soporte.

   Este modelo es especialmente efectivo en problemas de clasificación binaria y en conjuntos de datos con alta dimensionalidad, ya que permite construir fronteras de decisión tanto lineales como no lineales mediante el uso de funciones kernel.

   Finalmente, se establecen los rangos sobre los cuales fueron evaluados los hiperparámetro para este modelo a través de la siguiente tabla:

Tabla 5. Hiperparámetro – Máquinas de Soporte Vectorial

Hiperparámetro	Descripción	Rango
Kernel	Establece el tipo de solver que se utilizará en el algoritmo.	[linear, poly, rbf y sigmoid]
C	Es la regularización, controla el equilibrio entre error y generalización.	[0.001, 0.01, 0.1, 0.3, 0.5, 1, 2, 3, 4, 5, 7, 10]
degree	Controla el grado del polinomio usado para separar los datos. Solo en kernel 'poly'.	[2, 3]
gamma	Define qué tan lejos llega la influencia de cada punto. Solo en kernel 'rbf' y 'sigmoid'.	['scale', 'auto', 0.01, 0.1, 1]

e. Árbol de Decisión: modelo que utiliza una estructura jerárquica en forma de árbol para tomar decisiones basadas en las características de los datos. Además, es capaz de capturar relaciones no lineales entre las variables.

El modelo funciona dividiendo recursivamente el conjunto de datos en subconjuntos más pequeños mediante reglas de decisión, con el objetivo de maximizar la separación entre clases. Cada nodo del árbol representa una condición sobre una variable, y las hojas corresponden a la clase final asignada.

Finalmente, se establecen los rangos sobre los cuales fueron evaluados los hiperparámetro para este modelo a través de la siguiente tabla:

Tabla 6. Hiperparámetro – Árboles de Decisión

Hiperparámetro	Descripción	Rango
ccp_alpha	Parámetro de poda que permite simplificar el árbol eliminando ramas poco relevantes.	[1e-6, 1e-5, 1e-4, 0.001, 0.01, 0.1, 1, 10, 100]
max_depth	Limita la profundidad máxima del árbol, evitando estructuras demasiado complejas.	[None]
min_samples_split	Define el número mínimo de muestras necesarias para dividir un nodo.	[2]
min_samples_leaf	Establece el número mínimo de muestras que debe tener un nodo hoja.	[1]

f. Random Forest: método basado en la combinación de múltiples árboles de decisión. Este modelo pertenece a las técnicas de ensamble (ensemble learning) y tiene como objetivo mejorar la capacidad de generalización y reducir el sobreajuste presente en un único árbol de decisión.

El funcionamiento de Random Forest consiste en construir múltiples árboles a partir de diferentes subconjuntos de datos generados mediante muestreo aleatorio (bagging), así como en seleccionar subconjuntos aleatorios de variables en cada división. Posteriormente, las predicciones de todos los árboles se combinan mediante votación para determinar la clase final. Este enfoque permite reducir la varianza del modelo y mejorar su estabilidad, logrando un mejor desempeño frente a datos no lineales y con ruido.

Finalmente, se establecen los rangos sobre los cuales fueron evaluados los hiperparámetro para este modelo a través de la siguiente tabla:

Tabla 7. Hiperparámetro – Random Forest

Hiperparámetro	Descripción	Rango
n_estimators	Cantidad de árboles de decisión.	[100, 120]
max_features	Número de variables consideradas en cada división.	[5, 7, 9, 11]
max_depth	Profundidad máxima de los árboles.	[3, 5, 10, 15, 20]
criteria	Criterio de división utilizado en los nodos.	['gini', 'entropy']

g. AdaBoost: método basado en técnicas de ensamble (boosting), cuyo objetivo es mejorar el desempeño del modelo combinando múltiples clasificadores débiles.

A diferencia de métodos como Random Forest, que construyen los modelos de forma independiente, AdaBoost entrena los clasificadores de manera secuencial. En cada iteración, el modelo asigna mayor peso a aquellas observaciones que fueron clasificadas incorrectamente en iteraciones anteriores, obligando al modelo a enfocarse en los casos más difíciles.

Generalmente, AdaBoost utiliza árboles de decisión poco profundos (conocidos como stumps) como clasificadores base, lo que permite construir un modelo final más robusto a partir de la combinación de múltiples modelos simples.

Finalmente, se establecen los rangos sobre los cuales fueron evaluados los hiperparámetro para este modelo a través de la siguiente tabla:

Tabla 8. Hiperparámetro – AdaBoost

Hiperparámetro	Descripción	Rango
n_estimators	Número de clasificadores débiles utilizados en el ensamble.	[5, 10, 20, 30, 40, 50, 100, 150, 200]
learning_rate	Tasa de aprendizaje que controla la contribución de cada clasificador.	[0.01, 0.05, 0.1, 0.3, 0.5]
algorithm	Este parámetro controla el algoritmo base utilizado para construir los clasificadores débiles.	['SAMME']
base_estimators	Árboles de decisión con diferentes profundidades.	[DecisionTreeClassifier(max_depth=1), DecisionTreeClassifier(max_depth=2), DecisionTreeClassifier(max_depth=3)]

h. Gradient Boosting: técnica de ensamble basada en boosting que construye modelos de manera secuencial con el objetivo de minimizar una función de error. A diferencia de AdaBoost, que ajusta los pesos de las observaciones, Gradient Boosting entrena cada nuevo modelo para corregir los errores residuales del modelo anterior mediante un proceso de optimización basado en gradiente.

En este enfoque, cada modelo se construye sobre los errores del anterior, lo que permite mejorar progresivamente el desempeño del modelo final. Generalmente, se utilizan árboles de decisión poco profundos como estimadores bases, lo que facilita capturar relaciones no lineales en los datos.

Finalmente, se establecen los rangos sobre los cuales fueron evaluados los hiperparámetro para este modelo a través de la siguiente tabla:

Tabla 9. Hiperparámetro – Gradient Boosting

Hiperparámetro	Descripción	Rango
n_estimators	Número de árboles en el modelo.	[50, 100, 150, 200]
max_features	Número de variables consideradas en cada división.	['sqrt', 'log2']
max_depht	Profundidad de los árboles base.	[2, 3, 5]
subsample	Proporción de datos utilizada para entrenar cada árbol.	[0.5, 0.8, 1]
learning_rate	Tasa de aprendizaje.	[0.01, 0.05, 0.1, 0.3, 0.5]

i. **Extreme Gradient Boosting**: Versión optimizada del Gradient Boosting que incorpora mejoras en eficiencia computacional, regularización y manejo de datos, lo que permite obtener un alto desempeño en problemas de clasificación.

   Al igual que Gradient Boosting, XGBoost construye modelos de manera secuencial, donde cada nuevo árbol se entrena para corregir los errores residuales del modelo anterior. Sin embargo, este algoritmo introduce mecanismos adicionales como regularización (L1 y L2), control de complejidad y optimización paralela, lo que mejora la capacidad de generalización y reduce el riesgo de sobreajuste.

   Finalmente, se establecen los rangos sobre los cuales fueron evaluados los hiperparámetro para este modelo a través de la siguiente tabla:

Tabla 10. Hiperparámetro – Extreme Gradient Boosting

Hiperparámetro	Descripción	Rango
n_estimators	Número de árboles en el modelo.	[100, 150, 200]
max_depht	Profundidad de los árboles base.	[3, 5]
learning_rate	Tasa de aprendizaje.	[0.01, 0.05, 0.1]
subsample	Proporción de datos utilizada en cada iteración.	[0.8, 1]
colsample_bytree	Proporción de variables utilizadas por árbol.	[0.8, 1]
gamma	Controla la ganancia mínima necesaria para realizar una división.	[0, 0.1, 0.2]
reg_alpha (L1)	Penalización que puede llevar coeficientes a cero.	[0, 0.1, 1]
reg_lambda (L2)	Penalización que reduce la magnitud de los coeficientes.	[1, 5, 10]

ii. Técnicas de reducción de dimensionalidad

Con el fin de mejorar el desempeño de los modelos y analizar el impacto de las variables en la predicción, se implementaron diferentes estrategias de reducción de dimensionalidad de manera progresiva.

a. En una primera etapa, los modelos fueron entrenados utilizando el conjunto completo de variables disponibles, con el objetivo de establecer una línea base de desempeño.

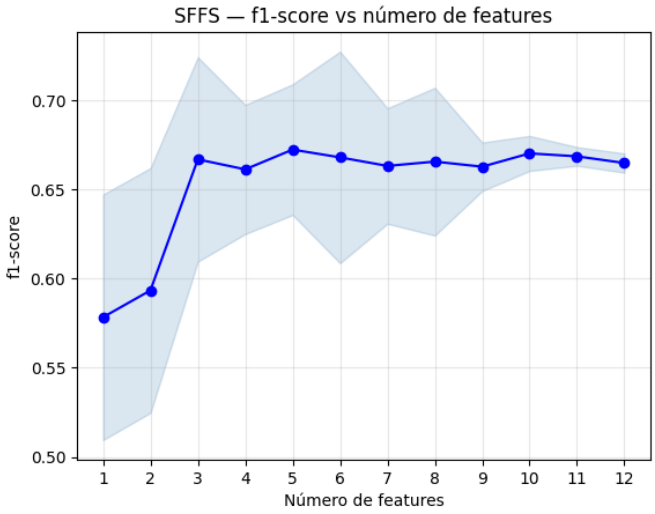
b. En una segunda etapa, se aplicaron métodos de Selección Secuencial de Características (SSF - Sequential Feature Selection), los cuales consisten en procedimientos iterativos que permiten seleccionar subconjuntos de variables relevantes en función de su contribución al desempeño del modelo (Raschka, s. f.).

   En particular, se utilizó la técnica de Selección Secuencial hacia Adelante (SFFS - Sequential Forward Feature Selection), la cual inicia con un conjunto vacío de variables y, en cada iteración, incorpora la característica que genera la mayor mejora en el desempeño del modelo.

   Proceso:

   - Se utilizó un modelo base de k-Nearest Neighbors (KNN) con $k = 4$, evaluando diferentes subconjuntos de variables mediante validación cruzada de 5 particiones (*cross-validation*). El criterio de evaluación utilizado fue el F1-score, dado que permite equilibrar precisión y *recall*.
   - El proceso de selección consistió en evaluar subconjuntos de características en un rango entre 5 y 12 variables, incorporando en cada iteración la característica que mejoraba el desempeño del modelo. Adicionalmente, el enfoque *floating* permite eliminar variables previamente seleccionadas si dejan de aportar al rendimiento, haciendo el proceso más flexible y eficiente.
   - Para analizar los resultados, se generó una gráfica del F1-score en función del número de características, lo que permitió identificar el subconjunto óptimo de variables.

Figura 5. Selección secuencial de características (SFFS)



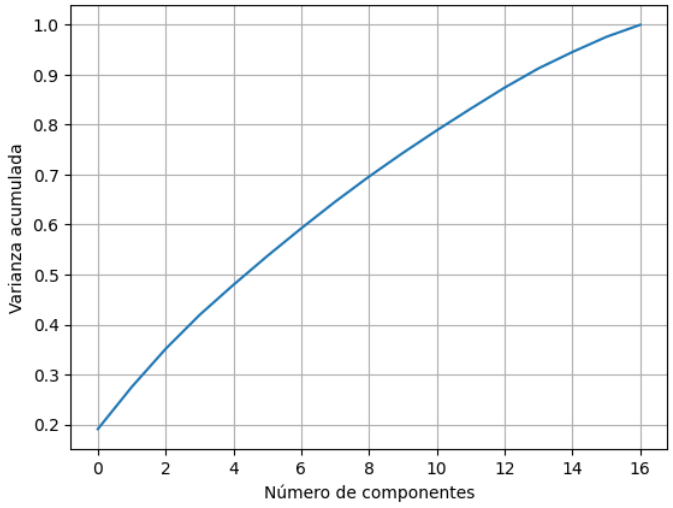
- La selección final se realizó considerando el menor número de características que alcanzara un F1-score alto, buscando un equilibrio entre desempeño del modelo y simplicidad, favoreciendo así la interpretabilidad y reduciendo el riesgo de sobreajuste. De acuerdo con la Figura 5, se seleccionaron 5 características, logrando reducir el conjunto original de 17 a 5 variables: *HighChol*, *CholCheck*, *GenHlth*, *PhysHlth* e *Hypertension*. Este resultado evidencia que un subconjunto reducido de variables puede mantener un desempeño competitivo, optimizando la complejidad del modelo.

c. En una tercera etapa, se empleó el Análisis de Componentes Principales (PCA), técnica de reducción de dimensionalidad que transforma las variables originales en un conjunto de componentes no correlacionados, conservando la mayor cantidad de varianza posible.

Proceso:

- Para determinar el número óptimo de componentes, se analizó mediante una gráfica la varianza explicada acumulada.

Figura 6. Varianza explicada acumulada en función del número de componentes (PCA)



- A partir de la curva generada, se seleccionaron 12 componentes principales, ya que estos permiten conservar más del 85% de la varianza total del conjunto de datos.
- Posteriormente, se realizó la transformación del conjunto de entrenamiento utilizando este número de componentes, reduciendo la dimensionalidad del problema y manteniendo la mayor parte de la información relevante para el entrenamiento de los modelos.

d. Finalmente, se aplicó el Análisis Discriminante Lineal (LDA), el cual permite reducir la dimensionalidad maximizando la separación entre clases, siendo especialmente útil en problemas de clasificación supervisada. Se aplicaron dos implementaciones:

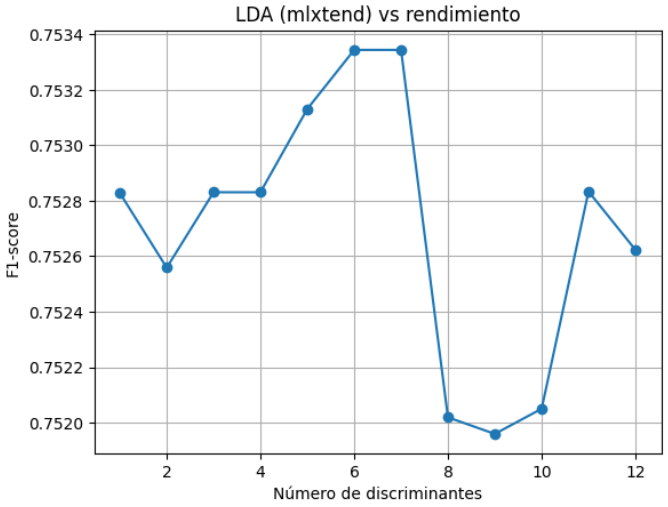
Proceso 1:

- Se utilizó la versión de la librería scikit-learn, donde se obtuvo un único componente discriminante. Esto se debe a que, en problemas de clasificación binaria, LDA solo permite generar un número máximo de componentes igual al número de clases menos uno.

Proceso 2:

- Se utilizó la implementación de la librería *mlxtend*, la cual permite evaluar múltiples dimensiones discriminantes. En este caso, se analizaron diferentes valores de componentes (de 1 a 12), entrenando un modelo de Regresión Logística sobre los datos transformados y evaluando su desempeño mediante el F1-score (Raschka, s. f.).

Figura 7. F1-score en función del número de componentes discriminantes (LDA)



- A partir de la curva generada, se seleccionaron 6 discriminantes, ya que este valor presenta uno de los mayores F1-score, evidenciando un mejor desempeño del modelo en comparación con otras configuraciones.
- Posteriormente, se realizó la transformación del conjunto de entrenamiento utilizando este número de discriminantes, reduciendo la dimensionalidad del problema y manteniendo una adecuada capacidad de separación entre las clases, lo que favorece el desempeño de los modelos de clasificación.

2.5. Evaluación del modelo

2.5.1. Métricas de evaluación

Para evaluar el desempeño de los modelos de clasificación, se utilizaron métricas orientadas a medir la calidad de las predicciones en términos de clasificación binaria.

Las métricas principales empleadas fueron:

- *Accuracy (Exactitud):* mide la proporción de predicciones correctas respecto al total de observaciones. Aunque es una métrica general, puede resultar engañosa en escenarios con desbalance de clases, ya que no distingue adecuadamente entre los distintos tipos de error.
- *Precision (Precisión):* mide la proporción de predicciones positivas correctas respecto al total de predicciones positivas realizadas por el modelo.
- *Recall (Sensibilidad):* evalúa la capacidad del modelo para identificar correctamente los casos positivos, es decir, la proporción de verdaderos positivos sobre el total de positivos reales.
- *F1-score:* corresponde a la media armónica entre precisión y recall, proporcionando una medida balanceada del desempeño del modelo, especialmente útil en escenarios con desbalance de clases.

En este estudio, el F1-score se estableció como la métrica principal de evaluación, debido a que permite considerar simultáneamente la precisión y la capacidad de detección de casos positivos, lo cual es fundamental en el contexto de detección temprana de diabetes.

Adicionalmente, se utilizó la matriz de confusión, la cual permite analizar de manera detallada los resultados de clasificación, identificando los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos generados por cada modelo.

Finalmente, se realizó un proceso de ajuste del umbral de decisión (threshold) sobre las probabilidades predichas por los modelos, con el objetivo de optimizar el equilibrio entre precisión y recall. Este ajuste permitió comparar el desempeño de los modelos bajo diferentes puntos de decisión, seleccionando aquellos valores de threshold que maximizan el F1-score y mejoran la capacidad de detección de la enfermedad.

2.5.2. Métodos de validación

La métrica de negocio de este proyecto corresponde a la selección del modelo de aprendizaje supervisado que, tras pasar por las etapas de entrenamiento, validación y prueba, presenta el mejor desempeño en términos de F1-score.

Para ello, se definió el siguiente procedimiento:

- En primer lugar, el conjunto de datos fue dividido en un 90% para implementación del modelo y un 10% reservado como conjunto de prueba final, el cual no fue utilizado durante las etapas de entrenamiento ni validación.
- Posteriormente, el 90% destinado a la implementación fue dividido internamente en 80% para entrenamiento y 20% para validación, permitiendo evaluar el desempeño de los modelos durante el proceso de ajuste.
- Los modelos seleccionados fueron entrenados utilizando múltiples combinaciones de hiperparámetros mediante una búsqueda en malla (*GridSearchCV*), con el objetivo de identificar las configuraciones óptimas para cada algoritmo.
- Una vez obtenidos los mejores hiperparámetros, los modelos fueron evaluados sobre el conjunto de validación, comparando su desempeño mediante el F1-score.
- Finalmente, los modelos con mejor desempeño en la etapa de validación fueron seleccionados y evaluados sobre el conjunto de prueba (10%), con el fin de estimar su capacidad de generalización en datos no vistos.

2.5.3. Comparación entre modelos

2.5.3.1. Etapa 1

Recordemos que en esta etapa los modelos fueron entrenados utilizando el conjunto completo de variables disponibles, con el objetivo de establecer una línea base de desempeño.

Para los modelos de Máquinas de Soporte Vectorial (SVM), AdaBoost, Gradient Boosting y XGBoost, se emplearon submuestras del conjunto de datos, debido a su alto costo computacional, con el fin de optimizar los tiempos de entrenamiento sin comprometer significativamente la calidad de los resultados.

Posteriormente, se realizó un proceso de búsqueda de hiperparámetros, el cual permitió identificar las configuraciones óptimas para cada modelo. Como resultado, se obtuvieron los mejores valores de hiperparámetros junto con el F1-score más alto alcanzado en la etapa de validación, los cuales se presentan a continuación:

Tabla 11. Hiperparámetros óptimos y métrica de desempeño en la etapa 1

Modelo	Hiperparámetros	Valores óptimos	Métrica de desempeño F1-score
Regresión Logística	Penalty	*l2*	0.7682
	C	1	
	Solver	*saga*	
Clasificador Bayesiano	Modelo	*GaussianNB*	0.7123
KNN	k	80	0.7448
	Metric	*manhattan*	
SVM-1	Kernel	*linear*	0.7535
	C	0.3	
SVM-2	Kernel	*poly*	0.7605
	C	5	
	degree	2	
SVM-3	Kernel	*rbf*	0.7569
	C	5	
	gamma	*auto*	
Árbol de Decisión	ccp_alpha	0.0001	0.7520
	max_depht	*None*	
	min_samples_split	2	
	min_samples_leaf	1	
Random Forest	n_estimators	100	0.7557
	max_features	5	
	max_depth	10	
	criteria	*gini*	
AdaBoost	n_estimators	50	0.7544
	learning_rate	0.1	
	algorithm	*SAMME*	
	base_estimators	*max_depth=3*	
Gradient Boosting	n_estimators	100	0.7568
	max_features	*log2*	
	max_depth	5	
	subsample	0.5	
	learning_rate	0.05	
XGBoost	n_estimators	100	0.7592
	max_depth	5	
	learning_rate	0.5	
	subsample	0.8	
	colsample_bytree	1	
	gamma	0.1	
	reg_alpha	0.1	
	reg_lambda	5	

2.5.3.2. Etapa 2

Recordemos que en esta etapa se aplicaron métodos de Selección Secuencial de Características (SSF - Sequential Feature Selection) y los modelos fueron entrenados utilizando 5 variables: *HighChol*, *CholCheck*, *GenHlth*, *PhysHlth* e *Hypertension*.

Para los modelos de Máquinas de Soporte Vectorial (SVM), AdaBoost, Gradient Boosting y XGBoost, se emplearon submuestras del conjunto de datos, debido a su alto costo computacional, con el fin de optimizar los tiempos de entrenamiento sin comprometer significativamente la calidad de los resultados.

Posteriormente, se realizó un proceso de búsqueda de hiperparámetros, el cual permitió identificar las configuraciones óptimas para cada modelo. Como resultado, se obtuvieron los mejores valores de hiperparámetros junto con el F1-score más alto alcanzado en la etapa de validación, los cuales se presentan a continuación:

Tabla 12. Hiperparámetros óptimos y métrica de desempeño en la etapa 2

Modelo	**Hiperparámetros**	**Valores óptimos**	**Métrica de desempeño**
			F1-score
Regresión Logística	Penalty	*l2*	0.7207
	C	1	
	Solver	*lbfs*	
Clasificador Bayesiano	Modelo	*GaussianNB*	0.7302
KNN	k	360	0.7301
	Metric	*euclidean*	
SVM-1	Kernel	*linear*	0.7056
	C	0.1	
SVM-2	Kernel	*poly*	0.7256
	C	0.1	
	degree	3	
SVM-3	Kernel	*rbf*	0.7278
	C	2	
	gamma	*scale*	
Árbol de Decisión	ccp_alpha	0.001	0.7439
	max_depht	*None*	
	min_samples_split	2	
	min_samples_leaf	1	
Random Forest	n_estimators	100	0.7426
	max_features	5	
	max_depth	5	
	criteria	*gini*	
AdaBoost	n_estimators	20	0.7431
	learning_rate	0.1	
	algorithm	*SAMME*	
	base_estimators	*max_depth=3*	

Gradient Boosting	n_estimators	50	0.7403
	max_features	*sqrt*	
	max_depth	3	
	subsample	0.8	
	learning_rate	0.1	
XGBoost	n_estimators	150	0.7403
	max_depth	3	
	learning_rate	0.05	
	subsample	1	
	colsample_bytree	1	
	gamma	0.2	
	reg_alpha	0.1	
	reg_lambda	1	

2.5.3.3. Etapa 3

Recordemos que en esta etapa se aplicó el Análisis de Componentes Principales (PCA), técnica de reducción de dimensionalidad y los modelos fueron entrenados con una transformación de 12 componentes principales.

Para los modelos de Máquinas de Soporte Vectorial (SVM), AdaBoost, Gradient Boosting y XGBoost, se emplearon submuestras del conjunto de datos, debido a su alto costo computacional, con el fin de optimizar los tiempos de entrenamiento sin comprometer significativamente la calidad de los resultados.

Posteriormente, se realizó un proceso de búsqueda de hiperparámetros, el cual permitió identificar las configuraciones óptimas para cada modelo. Como resultado, se obtuvieron los mejores valores de hiperparámetros junto con el F1-score más alto alcanzado en la etapa de validación, los cuales se presentan a continuación:

Tabla 13. Hiperparámetros óptimos y métrica de desempeño en la etapa 3

Modelo	Hiperparámetros	Valores óptimos	Métrica de desempeño F1-score
Regresión Logística	Penalty	*l2*	0.7463
	C	1	
	Solver	*lbfs*	
Clasificador Bayesiano	Modelo	*GaussianNB*	0.7516
KNN	k	110	0.7600
	Metric	*chebyshev*	
SVM-1	Kernel	*linear*	0.7503
	C	0.01	
SVM-2	Kernel	*poly*	0.7456
	C	1	
	degree	3	
SVM-3	Kernel	*rbf*	0.7630
	C	2	
	gamma	0.01	
Árbol de Decisión	ccp_alpha	0.1	0.7607
	max_depht	*None*	
	min_samples_split	2	
	min_samples_leaf	1	
Random Forest	n_estimators	100	0.7622
	max_features	5	
	max_depth	5	
	criteria	*gini*	
AdaBoost	n_estimators	30	0.7614
	learning_rate	0.5	
	algorithm	*SAMME*	
	base_estimators	*max_depth=2*	
Gradient Boosting	n_estimators	100	0.7610
	max_features	*sqrt*	
	max_depth	5	
	subsample	1	
	learning_rate	0.01	
XGBoost	n_estimators	100	0.7570
	max_depth	3	
	learning_rate	0.1	
	subsample	1	
	colsample_bytree	0.8	
	gamma	0.2	
	reg_alpha	0	
	reg_lambda	1	

2.5.3.4. Etapa 4 – Proceso 1

Recordemos que en esta etapa se aplicó el Análisis Discriminante Lineal (LDA), técnica de reducción de dimensionalidad con la versión de la librería *scikit-learn* y los modelos fueron entrenados con una transformación de un único componente discriminante.

Para los modelos de Máquinas de Soporte Vectorial (SVM), AdaBoost, Gradient Boosting y XGBoost, se emplearon submuestras del conjunto de datos, debido a su alto costo computacional, con el fin de optimizar los tiempos de entrenamiento sin comprometer significativamente la calidad de los resultados.

Posteriormente, se realizó un proceso de búsqueda de hiperparámetros, el cual permitió identificar las configuraciones óptimas para cada modelo. Como resultado, se obtuvieron los mejores valores de hiperparámetros junto con el F1-score más alto alcanzado en la etapa de validación, los cuales se presentan a continuación:

Tabla 14. Hiperparámetros óptimos y métrica de desempeño en la etapa 4 – proceso 1

Modelo	**Hiperparámetros**	**Valores óptimos**	**Métrica de desempeño**
			F1-score
Regresión Logística	Penalty	*l2*	0.7528
	C	1	
	Solver	*lbfs*	
Clasificador Bayesiano	Modelo	*GaussianNB*	0.7600
KNN	k	280	0.7602
	Metric	*euclidean*	
SVM-1	Kernel	*linear*	0.7585
	C	4	
SVM-2	Kernel	*poly*	0.7773
	C	0.3	
	degree	3	
SVM-3	Kernel	*rbf*	0.7699
	C	4	
	gamma	0.01	
Árbol de Decisión	ccp_alpha	0.01	0.7692
	max_depht	*None*	
	min_samples_split	2	
	min_samples_leaf	1	
Random Forest	n_estimators	100	0.7681
	max_features	5	
	max_depth	3	
	criteria	*entropy*	
AdaBoost	n_estimators	30	0.7771
	learning_rate	0.5	
	algorithm	*SAMME*	
	base_estimators	*max_depth=1*	

Gradient Boosting	n_estimators	100	0.7580
	max_features	*sqrt*	
	max_depth	3	
	subsample	0.5	
	learning_rate	0.1	
XGBoost	n_estimators	100	0.7585
	max_depth	5	
	learning_rate	0.05	
	subsample	1	
	colsample_bytree	0.8	
	gamma	0.2	
	reg_alpha	1	
	reg_lambda	5	

2.5.3.5. Etapa 4 – Proceso 2

Recordemos que en esta etapa se aplicó el Análisis Discriminante Lineal (LDA), técnica de reducción de dimensionalidad con la versión de la librería *mlxtend* y los modelos fueron entrenados con una transformación 6 componentes discriminantes.

Para los modelos de Máquinas de Soporte Vectorial (SVM), AdaBoost, Gradient Boosting y XGBoost, se emplearon submuestras del conjunto de datos, debido a su alto costo computacional, con el fin de optimizar los tiempos de entrenamiento sin comprometer significativamente la calidad de los resultados.

Posteriormente, se realizó un proceso de búsqueda de hiperparámetros, el cual permitió identificar las configuraciones óptimas para cada modelo. Como resultado, se obtuvieron los mejores valores de hiperparámetros junto con el F1-score más alto alcanzado en la etapa de validación, los cuales se presentan a continuación:

Tabla 15. Hiperparámetros óptimos y métrica de desempeño en la etapa 4 – proceso 2

Modelo	Hiperparámetros	Valores óptimos	Métrica de desempeño F1-score
Regresión Logística	Penalty	*l2*	0.7533
	C	1	
	Solver	*lbfs*	
Clasificador Bayesiano	Modelo	*GaussianNB*	0.7503
KNN	k	270	0.7611
	Metric	*euclidean*	
SVM-1	Kernel	*sigmoid*	0.7601
	C	0.01	
	gamma	0.01	
SVM-2	Kernel	*poly*	0.7694
	C	0.01	
	degree	3	
SVM-3	Kernel	*rbf*	0.7664
	C	7	
	gamma	0.01	
Árbol de Decisión	ccp_alpha	0.1	0.7692
	max_depht	*None*	
	min_samples_split	2	
	min_samples_leaf	1	
Random Forest	n_estimators	100	0.7684
	max_features	5	
	max_depth	3	
	criteria	*entropy*	
AdaBoost	n_estimators	30	0.7771
	learning_rate	0.05	
	algorithm	*SAMME*	
	base_estimators	*max_depth=1*	
Gradient Boosting	n_estimators	50	0.7666
	max_features	*sqrt*	
	max_depth	2	
	subsample	0.5	
	learning_rate	0.01	
XGBoost	n_estimators	100	0.7631
	max_depth	3	
	learning_rate	0.05	
	subsample	0.8	
	colsample_bytree	1	
	gamma	0.2	
	reg_alpha	1	
	reg_lambda	5	

En cada una de las etapas del estudio, se seleccionaron los cinco modelos con mejor desempeño, de acuerdo con el F1-score obtenido en el conjunto de validación. Estos modelos fueron posteriormente evaluados sobre el conjunto de prueba independiente (10%), con el objetivo de estimar su capacidad de generalización frente a datos no vistos. Este procedimiento permitió realizar una comparación objetiva del desempeño final entre los modelos seleccionados.

Los resultados se presentan en tablas ordenadas tanto por F1-score como por recall. En particular, el recall se analiza de manera explícita debido a su importancia en el contexto del problema, ya que permite evaluar la capacidad del modelo para identificar correctamente los casos positivos (personas con diabetes), lo cual es fundamental en escenarios de detección temprana donde minimizar los falsos negativos resulta prioritario.

Adicionalmente, se realizó un ajuste del umbral de decisión (threshold) sobre las probabilidades predichas por los modelos, evaluando diferentes valores con el fin de maximizar F1-score y recall, priorizando la detección de casos positivos. Este proceso permite adaptar el comportamiento del modelo, incluso a costa de una posible reducción en la precisión, en favor de una mayor sensibilidad en la identificación de la enfermedad**.**

- **Aplicando los 5 mejores modelos de la etapa 1.**

Tabla 16. Resultados en conjunto de prueba en la etapa 1 (sin threshold y F1-score)

Modelo	**Accurracy**	**Precision**	**Recall**	**F1-score**
SVM-poly	0.7479	0.7330	0.8012	0.7656
SVM-rbf	0.7454	0.7325	0.7947	0.7623
Random Forest	0.7483	0.7435	0.7785	0.7606
SVM-lineal	0.7482	0.7452	0.7747	0.7596
XGBoost	0.7474	0.7435	0.7761	0.7594

Tabla 17. Resultados en conjunto de prueba en la etapa 1 (sin threshold y Recall)

Modelo	Accurracy	Precision	Recall	F1-score
SVM-poly	0.7479	0.7330	0.8012	0.7656
SVM-rbf	0.7454	0.7325	0.7947	0.7623
Árbol de Decisión	0.7425	0.7324	0.7858	0.7581
AdaBoost	0.7443	0.7368	0.7814	0.7585
Random Forest	0.7483	0.7435	0.7785	0.7606

Tabla 18. Resultados en conjunto de prueba en la etapa 1 (Threshold 0.35 y F1-score)

Modelo	Accurracy	Precision	Recall	F1-score
XGBoost	0.7411	0.6907	0.8982	0.7809
Gradient Boosting	0.7411	0.6932	0.8900	0.7793
Regresión Logística	0.7410	0.6951	0.8832	0.7779
Random Forest	0.7340	0.6834	0.8987	0.7764
Árbol de Decisión	0.7332	0.6846	0.8912	0.7743

Tabla 19. Resultados en conjunto de prueba en la etapa 1 (Threshold 0.35 y Recall)

Modelo	Accurracy	Precision	Recall	F1-score
Random Forest	0.7340	0.6834	0.8987	0.7764
XGBoost	0.7411	0.6907	0.8982	0.7809
Árbol de Decisión	0.7332	0.6846	0.8912	0.7743
Gradient Boosting	0.7411	0.6932	0.8900	0.7793
KNN	0.7202	0.6733	0.8842	0.7645

- **Aplicando los 5 mejores modelos de la etapa 2.**

Tabla 20. Resultados en conjunto de prueba en la etapa 2 (sin threshold y F1-score)

Modelo	Accurracy	Precision	Recall	F1-score
AdaBoost	0.7297	0.7072	0.8087	0.7545
Random Forest	0.7297	0.7075	0.8077	0.7543
Árbol de Decisión	0.7247	0.6974	0.8200	0.7537
XGBoost	0.7278	0.7089	0.7978	0.7507
Gradient Boosting	0.7274	0.7094	0.7949	0.7497

Tabla 21. Resultados en conjunto de prueba en la etapa 2 (sin threshold y Recall)

Modelo	Accurracy	Precision	Recall	F1-score
Árbol de Decisión	0.7247	0.6974	0.8200	0.7537
AdaBoost	0.7297	0.7072	0.8087	0.7545
Random Forest	0.7297	0.7075	0.8077	0.7543
XGBoost	0.7278	0.7089	0.7978	0.7507
Gradient Boosting	0.7274	0.7094	0.7949	0.7497

Tabla 22. Resultados en conjunto de prueba en la etapa 2 (Threshold 0.35 y F1-score)

Modelo	Accurracy	Precision	Recall	F1-score
Regresión Logística	0.7192	0.6807	0.8538	0.7575
Random Forest	0.7182	0.6791	0.8557	0.7573
KNN	0.7142	0.6729	0.8632	0.7563
Árbol de Decisión	0.7132	0.6717	0.8639	0.7558
XGBoost	0.7121	0.6699	0.8666	0.7557

Tabla 23. Resultados en conjunto de prueba en la etapa 2 (Threshold 0.35 y Recall)

Modelo	Accurracy	Precision	Recall	F1-score
Gradient Boosting	0.7100	0.6664	0.8721	0.7555
XGBoost	0.7121	0.6699	0.8666	0.7557
Árbol de Decisión	0.7132	0.6717	0.8639	0.7558
KNN	0.7142	0.6729	0.8632	0.7563
Clasificador Bayesiano	0.7147	0.6752	0.8569	0.7553

- **Aplicando los 5 mejores modelos de la etapa 3.**

Tabla 24. Resultados en conjunto de prueba en la etapa 3 (sin threshold y F1-score)

Modelo	Accurracy	Precision	Recall	F1-score
SVM-rbf	0.7453	0.7309	0.7981	0.7630
Random Forest	0.7402	0.7194	0.8104	0.7622
AdaBoost	0.7412	0.7232	0.8039	0.7614
Gradient Boosting	0.7426	0.7275	0.7978	0.7610
Árbol de Decisión	0.7314	0.7014	0.8309	0.7607

Tabla 25. Resultados en conjunto de prueba en la etapa 3 (sin threshold y Recall)

Modelo	Accurracy	Precision	Recall	F1-score
Árbol de Decisión	0.7314	0.7014	0.8309	0.7607
Random Forest	0.7402	0.7194	0.8104	0.7622
AdaBoost	0.7412	0.7232	0.8039	0.7614
SVM-rbf	0.7453	0.7309	0.7981	0.7630
Gradient Boosting	0.7426	0.7275	0.7978	0.7610

Tabla 26. Resultados en conjunto de prueba en la etapa 3 (Threshold 0.3 y F1-score)

Modelo	Accurracy	Precision	Recall	F1-score
Regresión Logística	0.7238	0.6714	0.9054	0.7710
XGBoost	0.7200	0.6654	0.9151	0.7705
Random Forest	0.7167	0.6619	0.9168	0.7688
KNN	0.7166	0.6617	0.9170	0.7687
SVM-rbf	0.7453	0.7309	0.7981	0.7630

Tabla 27. Resultados en conjunto de prueba en la etapa 3 (Threshold 0.3 y Recall)

Modelo	Accurracy	Precision	Recall	F1-score
Gradient Boosting	0.6373	0.5883	0.9793	0.7350
AdaBoost	0.6915	0.6338	0.9464	0.7592
KNN	0.7166	0.6617	0.9170	0.7687
Random Forest	0.7167	0.6619	0.9168	0.7688
XGBoost	0.7200	0.6654	0.9151	0.7705

- **Aplicando los 5 mejores modelos de la etapa 4 – proceso 1.**

Tabla 28. Resultados en conjunto de prueba en la etapa 4 – p1 (sin threshold y F1-score)

Modelo	Accurracy	Precision	Recall	F1-score
SVM-poly	0.7316	0.6772	0.9122	0.7773
AdaBoost	0.7419	0.6983	0.8758	0.7771
SVM-rbf	0.7479	0.7249	0.8207	0.7699
Árbol de Decisión	0.7508	0.7336	0.8084	0.7692
Random Forest	0.7505	0.7350	0.8043	0.7681

Los resultados en conjunto de prueba en la etapa 4 – p1 (sin threshold y Recall) son los mismos ilustrados en la Tabla 28.

Tabla 29. Resultados en conjunto de prueba en la etapa 4 - p1 (Threshold 0.3 y F1-score)

Modelo	Accurracy	Precision	Recall	F1-score
Random Forest	0.7324	0.6787	0.9098	0.7774
SVM-poly	0.7316	0.6772	0.9122	0.7773
Regresión Logística	0.7321	0.6782	0.9103	0.7773
AdaBoost	0.7321	0.6786	0.9088	0.7770
Clasificador Bayesiano	0.7321	0.6789	0.9078	0.7768

Tabla 30. Resultados en conjunto de prueba en la etapa 4 – p1 (Threshold 0.3 y Recall)

Modelo	Accurracy	Precision	Recall	F1-score
Árbol de Decisión	0.7068	0.6476	0.9414	0.7674
XGBoost	0.7295	0.6742	0.9158	0.7767
KNN	0.7288	0.6736	0.9158	0.7763
Gradient Boosting	0.7277	0.6732	0.9134	0.7751
SVM-poly	0.7316	0.6772	0.9122	0.7773

- **Aplicando los 5 mejores modelos de la etapa 4 – proceso 2.**

Tabla 31. Resultados en conjunto de prueba en la etapa 4 – p2 (sin threshold y F1-score)

Modelo	Accurracy	Precision	Recall	F1-score
AdaBoost	0.7419	0.6983	0.8758	0.7771
SVM-poly	0.7431	0.7138	0.8345	0.7694
Árbol de Decisión	0.7508	0.7336	0.8084	0.7692
Random Forest	0.7507	0.7348	0.8053	0.7684
Gradient Boosting	0.7503	0.7373	0.7983	0.7666

Los resultados en conjunto de prueba en la etapa 4 – p2 (sin threshold y Recall) son los mismos ilustrados en la Tabla 31.

Tabla 32. Resultados en conjunto de prueba en la etapa 4 – p2 (Threshold 0.3 y F1-score)

Modelo	Accurracy	Precision	Recall	F1-score
Regresión Logística	0.7313	0.6773	0.9110	0.7770
AdaBoost	0.7321	0.6786	0.9088	0.7770
XGBoost	0.7287	0.6735	0.9158	0.7762
Random Forest	0.7278	0.6729	0.9148	0.7755
KNN	0.7256	0.6705	0.9160	0.7743

Tabla 33. Resultados en conjunto de prueba en la etapa 4 – p2 (Threshold 0.3 y Recall)

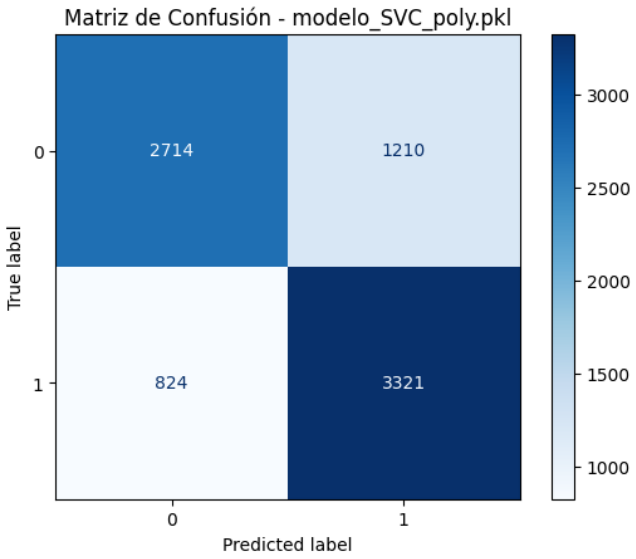
Modelo	Accurracy	Precision	Recall	F1-score
Gradient Boosting	0.5137	0.5137	1.0000	0.6787
KNN	0.7256	0.6705	0.9160	0.7743
XGBoost	0.7287	0.6735	0.9158	0.7762
Random Forest	0.7278	0.6729	0.9148	0.7755
Regresión Logística	0.7313	0.6773	0.9110	0.7770

3. Resultados y Discusión

Luego de realizar cinco etapas de experimentación con diferentes estrategias de reducción de dimensionalidad y algoritmos de aprendizaje supervisado, se obtuvieron resultados relevantes para la predicción del riesgo de diabetes.

En la Etapa 1 (Tabla 16), en la cual se utilizaron todas las variables originales del conjunto de datos, el mejor desempeño se obtuvo con el modelo de Máquinas de Soporte Vectorial (SVM) con kernel polinómico, alcanzando un F1-score de 0.7656.

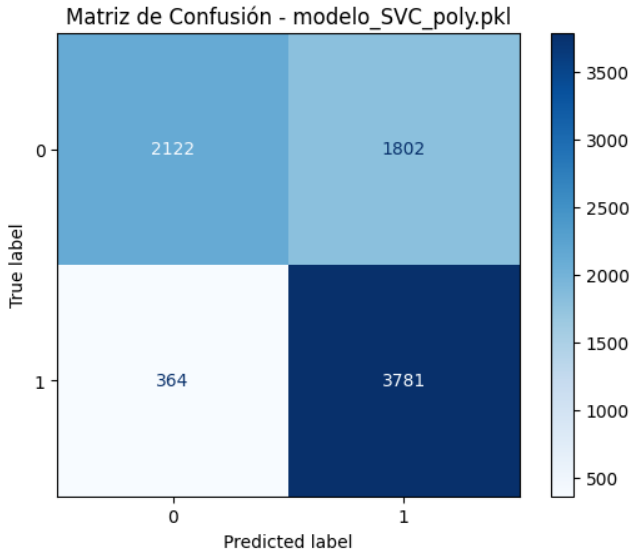
Figura 8. Matriz de confusión de la Etapa 1 - Modelo SVM con kernel polinómico



Posteriormente, con el propósito de evaluar si era posible mejorar este resultado, se aplicaron distintas estrategias de reducción de dimensionalidad en las etapas siguientes. En la Etapa 2 (Tabla 20), correspondiente a la Selección Secuencial hacia Adelante con Retroceso (SFFS), se obtuvo un F1-score máximo de 0.7545 con el modelo de AdaBoost. En la Etapa 3 (Tabla 24), utilizando Análisis de Componentes Principales (PCA), se alcanzó un F1-score de 0.7630 con el modelo de Máquinas de Soporte Vectorial (SVM) con kernel rbf. Finalmente, en las etapas 4-1 (Tabla 28) y 4-2 (Tabla 31), basadas en Análisis Discriminante Lineal (LDA), se obtuvieron F1-score de 0.7773 con el modelo de Máquinas de Soporte Vectorial (SVM) con kernel polinómico y 0.7771 con el modelo de AdaBoost, respectivamente.

Estos resultados evidencian que el mejor desempeño global en términos de F1-score se obtuvo en la Etapa 4-1 con el modelo de Máquinas de Soporte Vectorial (SVM) con kernel polinómico (0.7773). Estos resultados muestran que la estrategia con mejor desempeño fue LDA, al incrementar ligeramente el valor del F1-score respecto al modelo base.

Figura 9. Matriz de confusión de la Etapa 4-1 - Modelo SVM con kernel polinómico



Adicionalmente, se realizó un análisis complementario ordenando los modelos de acuerdo con la métrica de recall, con el fin de evaluar su capacidad para identificar correctamente los casos positivos (personas con diabetes). Esta métrica resulta especialmente relevante en escenarios de detección temprana, donde minimizar los falsos negativos es prioritario, ya que un caso no detectado puede retrasar la intervención médica oportuna.

En la Etapa 1 (Tabla 17), en la cual se utilizaron todas las variables originales del conjunto de datos, el mejor desempeño se obtuvo con el modelo de Máquinas de Soporte Vectorial (SVM) con kernel polinómico, alcanzando un recall de 0.8012. En la Etapa 2 (Tabla 21), correspondiente a la Selección Secuencial hacia Adelante con Retroceso (SFFS), se obtuvo un recall máximo de 0.8200 con el modelo Árbol de Decisión. En la Etapa 3 (Tabla 25), utilizando Análisis de Componentes Principales (PCA), se alcanzó un recall de 0.8309 con el modelo Árbol de Decisión. Finalmente, en las etapas 4-1 (Tabla 28) y 4-2 (Tabla 31), basadas en Análisis Discriminante Lineal (LDA), se obtuvieron valores de recall de 0.9122 con el modelo de Máquinas de Soporte Vectorial (SVM) con kernel polinómico y 0.8758 con el modelo de AdaBoost, respectivamente.

Estos resultados evidencian que el mejor desempeño global en términos de Recall se obtuvo en la Etapa 4-1 con el modelo de Máquinas de Soporte Vectorial (SVM) con kernel polinómico (0.9122). Como se evidencia en la Figura 9, las estrategias basadas en LDA lograron los valores más altos de recall, incrementando de forma significativa la capacidad de detección de casos positivos respecto al modelo base (Figura 8).

En los análisis presentados anteriormente, los resultados fueron obtenidos utilizando un umbral de decisión (threshold) igual a 0.5 sobre las probabilidades estimadas por los modelos. Este valor corresponde al criterio predeterminado en la mayoría de los algoritmos de clasificación, según el cual una observación es asignada a la clase positiva cuando la probabilidad predicha es mayor o igual al 50%. No obstante, ajustar el threshold permite adaptar el comportamiento del modelo a las necesidades específicas del problema, en lugar de depender del valor estándar. En este estudio, dicha estrategia resulta especialmente pertinente para analizar los comportamientos.

Con base en lo anterior, se evaluaron diferentes valores de threshold para cada una de las etapas del estudio, con el fin de identificar aquellos que permitieran maximizar el desempeño de los modelos según las métricas de interés. Una vez definido el threshold óptimo en cada caso, se repitieron los dos análisis globales previamente descritos: el primero ordenando los modelos de acuerdo con el F1-score, y el segundo priorizando la métrica de recall.

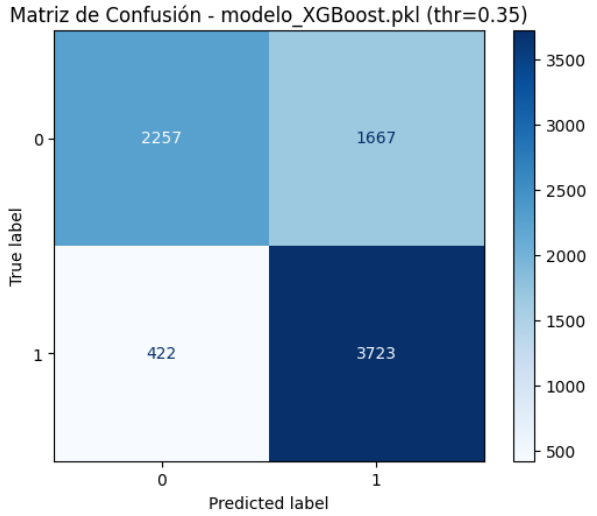
3.1. Resultados ajustando el threshold para maximizar el F1-score

En la Etapa 1 (Tabla 18), el mejor resultado se obtuvo con el modelo de XGBoost, utilizando un threshold de 0.35, con un F1-score de 0.7809. En la Etapa 2 (Tabla 22), correspondiente a SFFS, el mejor desempeño fue alcanzado por la Regresión Logística con threshold de 0.35, obteniendo un F1-score de 0.7575. En la Etapa 3 (Tabla 26), basada en Análisis de Componentes Principales (PCA), la Regresión Logística alcanzó un F1-score de 0.7710 con threshold de 0.30.

Finalmente, en las etapas basadas en Análisis Discriminante Lineal (LDA), se obtuvieron F1-score de 0.7774 en la etapa 4-1 (Tabla 29), mediante un modelo de Random Forest con threshold de 0.30, y de 0.7770 en la etapa 4-2 (Tabla 32), mediante Regresión Logística con threshold de 0.30.

Estos resultados evidencian que el mejor desempeño global en términos de F1-score se obtuvo en la Etapa 1 con XGBoost (0.7809), lo que demuestra que el ajuste del threshold permitió mejorar el equilibrio entre precisión y recall respecto a los resultados obtenidos con el umbral estándar de 0.5 (Figura 8).

Figura 10. Matriz de confusión de la Etapa 1 - Modelo XGBoost con threshold de 0.35



3.2. Resultados ajustando el threshold para maximizar el recall

Al priorizar la métrica de recall, en la Etapa 1 (Tabla 19) el mejor resultado se obtuvo con Random Forest, utilizando un threshold de 0.35, alcanzando un recall de 0.8987. En la Etapa 2 (Tabla 23), el modelo de Gradient Boosting obtuvo un recall de 0.8721 con threshold de 0.35. En la Etapa 3 (Tabla 27), Gradient Boosting alcanzó un recall de 0.9793 con threshold de 0.30.

Finalmente, en las etapas basadas en LDA, se obtuvieron valores de recall de 0.9414 en la etapa 4-1 (Tabla 30), mediante un modelo de Árbol de Decisión con threshold de 0.30, y de 1.0000 en la etapa 4-2 (Tabla 33), mediante Gradient Boosting con threshold de 0.30.

Figura 11. Matriz de confusión de la Etapa 3 - Modelo Gradient Boosting con threshold de 0.3

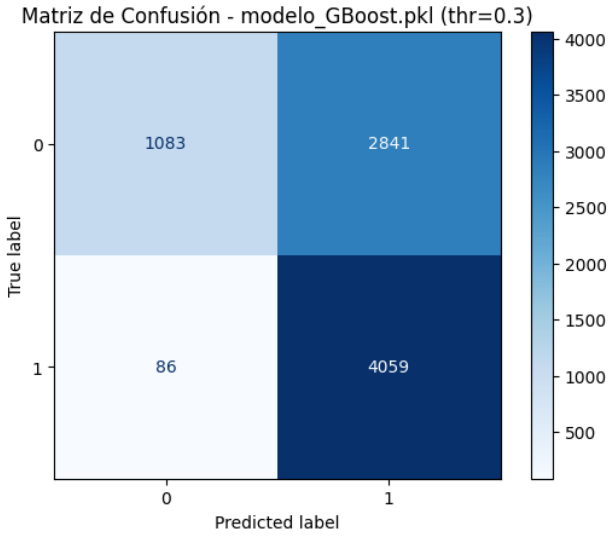
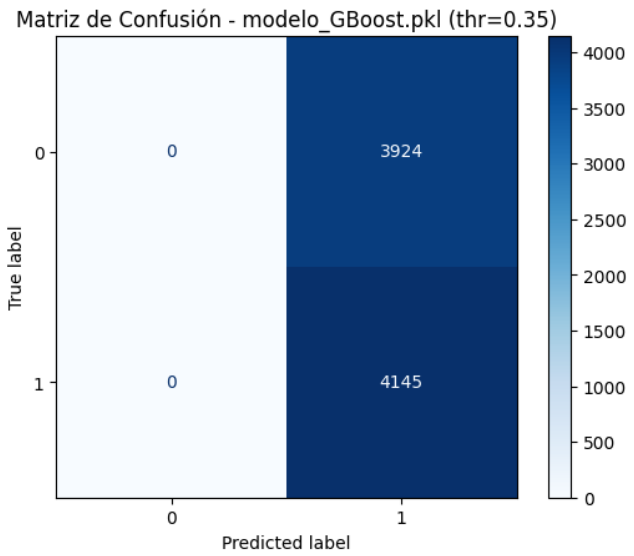


Figura 12. Matriz de confusión de la Etapa 4-2 - Modelo Gradient Boosting con threshold de 0.3



Observando la Figura 12, esta muestra que el modelo clasificó todas las observaciones como positivas (diabetes = 1). Aunque el recall es perfecto, este resultado evidencia que el modelo no tiene capacidad para discriminar entre pacientes con y sin diabetes, ya que asigna la clase positiva a todos los registros. En términos prácticos, esto equivale a asumir que todas las personas evaluadas presentan riesgo de diabetes, generando una gran cantidad de falsos positivos.

Por esta razón, este modelo no constituye la mejor alternativa para un sistema de apoyo a la decisión clínica, pese a alcanzar el máximo valor posible de recall. Si bien en problemas de salud es importante minimizar los falsos negativos, también es necesario mantener un balance adecuado con la precisión y el F1-score, de modo que el modelo sea útil y genere alertas razonables.

En consecuencia, un recall de 1.0000 debe interpretarse con cautela, ya que puede corresponder a un comportamiento trivial del clasificador y no necesariamente a un modelo con buen desempeño global. En este caso, el modelo de la Etapa 3 (Gradient Boosting) ofrece un equilibrio más adecuado, resultando más apropiados para la detección temprana de diabetes (Figura 11).

La selección final del modelo no se basó exclusivamente en la maximización de una métrica, sino también en criterios de interpretabilidad, estabilidad y facilidad de aplicación en contextos reales de apoyo a la decisión clínica. Aplicar una metodología con threshold variado nos llevan a dos tipos de errores posibles:

a) Falso Positivo:

Consecuencias:

- La persona puede ser enviada a exámenes adicionales.
- Se genera preocupación innecesaria.
- Aumentan los costos del proceso de evaluación.

b) Falso Negativo:

Consecuencias:

- La enfermedad puede pasar desapercibida.
- Se retrasa el diagnóstico y tratamiento.
- Puede haber complicaciones de salud.

En el contexto de detección de diabetes, generalmente el error más crítico es el falso negativo, porque implica no identificar a una persona que realmente tiene la enfermedad. En aplicaciones médicas, se suele priorizar la reducción de falsos negativos, ya que sus consecuencias pueden ser más graves para el paciente.

Sin embargo, para el presente proyecto se decidió priorizar la selección de un modelo que ofreciera un desempeño sólido y balanceado, considerando no solo los valores de las métricas de evaluación, sino también la interpretabilidad y la facilidad de implementación en un entorno real de apoyo a la decisión clínica.

Bajo este criterio, el modelo seleccionado corresponde al obtenido en la Etapa 1 (Tabla 16), en la cual se utilizaron todas las variables originales del conjunto de datos. En esta etapa, el mejor desempeño se obtuvo con el modelo de Máquinas de Soporte Vectorial (SVM) con kernel polinómico, alcanzando un F1-score de 0.7656 y un recall de 0.8012. Estos resultados evidencian un equilibrio adecuado entre la capacidad para detectar casos positivos y la precisión general del modelo, sin sacrificar la interpretabilidad de las variables utilizadas.

Aunque las técnicas de reducción de dimensionalidad, como Análisis de Componentes Principales (PCA) y Análisis Discriminante Lineal (LDA), permitieron obtener mejoras marginales en algunas métricas, estas transforman las variables originales en componentes o discriminantes cuya interpretación clínica es limitada. Por esta razón, se consideró más conveniente seleccionar un modelo que opere directamente sobre las variables originales del paciente.

No obstante, los resultados obtenidos permiten identificar modelos alternativos que podrían resultar apropiados en otros contextos:

- Cuando la interpretabilidad no sea un requisito prioritario, los modelos basados en LDA constituyen una alternativa válida, ya que alcanzaron los mejores resultados globales en términos de F1-score y recall. Ejemplo: Etapa 4-1 con el modelo de Máquinas de Soporte Vectorial (SVM) con kernel polinómico.
- Cuando el objetivo principal sea maximizar la detección de casos positivos, puede considerarse el modelo de la Etapa 3, basado en Gradient Boosting con threshold ajustado, el cual alcanzó un recall de 0.9793, dejando sin detectar únicamente 86 casos de diabetes.

Si bien estos modelos presentan un mayor número de falsos positivos, lo que implicaría la realización de evaluaciones complementarias en pacientes que no necesariamente padecen la enfermedad, este comportamiento puede ser aceptable en el ámbito médico, donde generalmente es preferible generar algunas alertas adicionales antes que omitir pacientes con diabetes no diagnosticada.

3.3. Interpretación

Los resultados obtenidos demuestran que es posible predecir de manera razonablemente precisa la presencia de diabetes a partir de variables demográficas, antecedentes clínicos y hábitos de vida reportados por los pacientes. En términos generales, los modelos evaluados alcanzaron valores de F1-score entre 0.75 y 0.78, lo que evidencia una adecuada capacidad para distinguir entre personas con y sin diabetes y confirma que la información disponible en el conjunto de datos contiene patrones relevantes para la identificación del riesgo.

El mejor modelo seleccionado para el proyecto fue el de Máquinas de Soporte Vectorial (SVM) con kernel polinómico, entrenado con todas las variables originales. Este modelo obtuvo un F1-score de 0.7656 y un recall de 0.8012, lo que significa que logra identificar correctamente aproximadamente el 80% de los pacientes con diabetes, manteniendo un equilibrio adecuado entre sensibilidad y precisión. En el contexto del problema, este resultado es altamente valioso, ya que permite detectar de manera temprana una proporción importante de personas con riesgo de padecer la enfermedad, facilitando su remisión a exámenes complementarios, seguimiento médico y estrategias de prevención.

En conjunto, los resultados indican que los modelos de aprendizaje automático pueden constituir una herramienta efectiva de apoyo para la detección temprana de diabetes. Su implementación podría contribuir a focalizar recursos médicos, priorizar pacientes para evaluación clínica y fortalecer las estrategias de prevención y seguimiento de esta enfermedad.

Los resultados obtenidos en el presente estudio son consistentes con lo reportado en la literatura reciente sobre la predicción de diabetes mediante técnicas de aprendizaje automático. En particular, el artículo publicado en la revista *Elsevier* titulado *"Machine Learning Techniques for Diabetes Prediction: A Comparative Study"* evaluó diversos algoritmos supervisados utilizando variables clínicas y demográficas similares a las consideradas en esta investigación, entre ellas edad, índice de masa corporal (BMI), antecedentes médicos y otros factores relacionados con el estado de salud del paciente (Alzahrani & Alzahrani, 2022).

En dicho trabajo los modelos con mejor comportamiento fueron Máquinas de Soporte Vectorial (SVM), XGBoost, Random Forest y Gradient Boosting, lo cual confirma la capacidad de estos algoritmos para capturar relaciones no lineales entre las variables clínicas y el diagnóstico de diabetes.

En dicho estudio, el modelo con mejor desempeño fue Random Forest, alcanzando un F1-score de 0.8226 y recall de 0.8045, lo que evidencia la alta capacidad predictiva de los métodos de ensamble para este tipo de problema. En comparación con nuestro modelo seleccionado, SVM con kernel polinómico, alcanzó un F1-score de 0.7656 y un recall de 0.8012, resultados comparables con lo reportado por el estudio en mención, donde las métricas de desempeño suelen ubicarse en rangos entre 0.75 y 0.85 dependiendo del conjunto de datos y del preprocesamiento aplicado.

Asimismo, el estudio de referencia destaca que técnicas de reducción de dimensionalidad pueden generar mejoras moderadas en el desempeño de los modelos, aunque no siempre justifican la pérdida de interpretabilidad. Esta observación coincide con los resultados obtenidos en el presente trabajo, donde las estrategias basadas en Análisis Discriminante Lineal (LDA) lograron F1-score ligeramente superiores (hasta 0.7773), pero a costa de transformar las variables originales en discriminantes con menor significado clínico.

Finalmente, tanto el estudio citado como el presente trabajo resaltan la importancia de evaluar múltiples métricas y no limitarse únicamente a la exactitud. En particular, el recall adquiere un papel fundamental cuando el objetivo es la detección temprana de diabetes, ya que permite minimizar la cantidad de pacientes enfermos que no son identificados por el modelo. En este sentido, los resultados obtenidos refuerzan la evidencia de que los modelos de aprendizaje automático constituyen herramientas eficaces para apoyar el tamizaje y la toma de decisiones preventivas en el ámbito de la salud.

El desempeño de los modelos evaluados estuvo influenciado por diversos factores relacionados con la calidad de los datos, las técnicas de preprocesamiento aplicadas y la adecuada configuración de los hiperparámetros.

- En primer lugar, la calidad del conjunto de datos desempeñó un papel fundamental. El dataset utilizado no presentó valores nulos y estuvo conformado por variables clínicas, demográficas y de hábitos de vida ampliamente reconocidas en la literatura como factores asociados al desarrollo de diabetes, tales como hipertensión, colesterol alto, estado general de salud, días de mala salud física y antecedentes médicos. La ausencia de datos faltantes redujo la necesidad de imputaciones y disminuyó el riesgo de introducir sesgos adicionales en el proceso de modelado.
- En segundo lugar, la detección y eliminación de valores atípicos mediante el algoritmo Local Outlier Factor (LOF) contribuyó a depurar el conjunto de entrenamiento, eliminando observaciones con patrones inusuales que podían afectar negativamente el ajuste de algunos modelos, especialmente aquellos sensibles a distancias, como k-Nearest Neighbors y Máquinas de Soporte Vectorial (SVM).
- Otro factor importante fue el tratamiento del desbalance de clases. Aunque el dataset presentaba un desbalance ligero, se aplicó submuestreo aleatorio de la clase mayoritaria mediante imbalanced-learn RandomUnderSampler, lo que permitió entrenar los modelos con una distribución equilibrada y mejorar la capacidad para detectar casos positivos de diabetes.
- En cuanto al preprocesamiento, el escalado de las variables numéricas mediante scikit-learn MinMaxScaler fue determinante para modelos sensibles a la magnitud de los datos, como Regresión Logística, KNN y SVM. Asimismo, la decisión de no aplicar codificación One-Hot-Encoding a variables ordinales como edad y estado general de salud preservó su estructura ordenada y evitó aumentar innecesariamente la dimensionalidad del problema.
- La optimización de hiperparámetros también tuvo un impacto significativo en el desempeño. Parámetros como el valor de C y el tipo de kernel en SVM, el número de vecinos en KNN, la profundidad de los árboles, el número de estimadores en los modelos de ensamble y las tasas de aprendizaje en los algoritmos de boosting permitieron adaptar cada modelo a las características del conjunto de datos y maximizar su capacidad predictiva.
- Finalmente, el ajuste del umbral de decisión (threshold) demostró ser un factor clave para mejorar la sensibilidad de los modelos. Al reducir el threshold por debajo del valor estándar de 0.5, fue posible incrementar significativamente el recall, priorizando la detección de casos positivos, aunque en algunos casos esto implicó un aumento en el número de falsos positivos.

En conjunto, estos factores permitieron construir modelos robustos y competitivos, y explican las diferencias observadas en el desempeño entre los distintos algoritmos evaluados.

3.4. Evaluación crítica

A pesar de los resultados satisfactorios obtenidos, el presente estudio presenta algunas limitaciones que deben ser consideradas al interpretar los hallazgos:

- En primer lugar, el análisis se desarrolló a partir de un conjunto de datos secundario, construido a partir de información autodeclarada por los participantes. Este tipo de datos puede contener errores de medición, sesgos de recuerdo o respuestas inexactas, lo que podría afectar la calidad de las variables utilizadas para el entrenamiento de los modelos.
- En segundo lugar, aunque el conjunto de datos incluye variables relevantes asociadas al riesgo de diabetes, no incorpora información clínica más específica, como resultados de laboratorio (glucosa en ayunas, hemoglobina glicosilada HbA1c), antecedentes familiares detallados o información genética, que podrían mejorar significativamente la capacidad predictiva de los modelos.
- Otra limitación importante es el uso de detección y eliminación de valores atípicos mediante el algoritmo Local Outlier Factor (LOF) y submuestreo aleatorio (undersampling) para balancear las clases. Si bien estas técnicas permitieron entrenar los modelos con una distribución equilibrada, también implicó descartar una parte de los registros, lo que pudo ocasionar pérdida de información potencialmente útil.
- Asimismo, los resultados obtenidos dependen del conjunto de datos y de la partición utilizada para entrenamiento, validación y prueba. Aunque se aplicó validación cruzada y se reservó un conjunto independiente de prueba, el desempeño del modelo podría variar al aplicarse en poblaciones con características demográficas o epidemiológicas diferentes.
- Desde el punto de vista metodológico, las técnicas de reducción de dimensionalidad como PCA y Análisis Discriminante Lineal mejoraron ligeramente algunas métricas, pero redujeron la interpretabilidad clínica de los resultados. Esta situación evidencia la necesidad de equilibrar el rendimiento predictivo con la facilidad de interpretación y uso por parte de los profesionales de la salud.
- Finalmente, el ajuste del threshold permitió incrementar la sensibilidad de algunos modelos, pero en ciertos casos produjo un aumento considerable de falsos positivos. Esto implica que, aunque se reducen los casos de diabetes no detectados, también se incrementa el número de pacientes que requerirían evaluaciones adicionales sin presentar realmente la enfermedad.

Existen varias oportunidades para fortalecer y ampliar el presente estudio:

- En primer lugar, sería conveniente incorporar variables clínicas adicionales, especialmente resultados de laboratorio y antecedentes familiares más detallados, con el fin de enriquecer la información disponible y mejorar el poder predictivo de los modelos.
- En segundo lugar, podrían evaluarse técnicas alternativas para el manejo del desbalance de clases, como SMOTE, ADASYN o enfoques basados en aprendizaje sensible al costo, que permitan aprovechar toda la información disponible sin descartar observaciones de la clase mayoritaria.
- También sería útil explorar métodos avanzados de interpretabilidad, como SHAP o LIME, para explicar las predicciones individuales y cuantificar la contribución de cada variable al resultado del modelo, incluso en algoritmos complejos como Random Forest o XGBoost.
- Otra línea de mejora consiste en realizar una validación externa con datos provenientes de otras poblaciones o instituciones de salud, lo que permitiría evaluar la robustez y capacidad de generalización del modelo en contextos diferentes.
- Adicionalmente, podrían incorporarse técnicas de calibración de probabilidades y estrategias sistemáticas de optimización del threshold, ajustando el modelo a objetivos específicos de negocio o criterios clínicos.
- Finalmente, una posible extensión del trabajo sería implementar el modelo seleccionado en una aplicación web o herramienta de apoyo clínico que permita ingresar la información de un paciente y obtener automáticamente una estimación de su riesgo de diabetes, facilitando su utilización práctica por parte de profesionales de la salud.

4. Conclusiones

En el presente trabajo se diseñaron e implementaron diversos modelos de aprendizaje automático supervisado con el propósito de predecir tempranamente la presencia de diabetes a partir de variables demográficas, clínicas y de hábitos de vida. Los resultados obtenidos demostraron que este tipo de técnicas permite identificar patrones relevantes en los datos y generar predicciones con un nivel de desempeño adecuado para apoyar procesos de tamizaje y toma de decisiones preventivas en salud.

Durante el desarrollo del estudio se realizó un análisis exploratorio de las variables, el cual permitió identificar la distribución de cada variable, detectar valores atípicos y evaluar su relación con la presencia de diabetes. Este análisis evidenció que factores como la hipertensión, el colesterol alto, el estado general de salud, la edad y los días de mala salud física presentan una asociación importante con la presencia de diabetes, en concordancia con lo reportado en la literatura científica.

Se implementaron nueve modelos de aprendizaje supervisado, entre ellos Regresión Logística, Naive Bayes, k-Nearest Neighbors, Máquinas de Soporte Vectorial, Árboles de Decisión, Random Forest, AdaBoost, Gradient Boosting y XGBoost. Durante este proceso se aplicaron técnicas de preprocesamiento, balanceo de clases mediante RandomUnderSampler, escalado con MinMaxScaler, optimización de hiperparámetros y estrategias de reducción de dimensión como SFFS, PCA y Análisis Discriminante Lineal.

Se emplearon métricas como accuracy, precisión, recall, F1-score y AUC, utilizando conjuntos independientes de validación y prueba. El modelo seleccionado fue el de Máquinas de Soporte Vectorial (SVM) con kernel polinómico, entrenado con todas las variables originales del conjunto de datos. Este modelo alcanzó un F1-score de 0.7656 y un recall de 0.8012 lo que significa que el modelo logra identificar correctamente aproximadamente ocho de cada diez personas con diabetes, constituyéndose en una herramienta útil para apoyar la detección oportuna de la enfermedad.

Adicionalmente, se comprobó que el ajuste del umbral de decisión (threshold) permite adaptar el comportamiento del modelo según las prioridades del problema. Cuando el objetivo es maximizar la sensibilidad, se pueden obtener valores de recall cercanos a 0.95, aunque con un incremento en el número de falsos positivos. No obstante, para este estudio se optó por un modelo más conservador y balanceado, que ofrece un adecuado compromiso entre capacidad predictiva, interpretabilidad y aplicabilidad práctica.

En conclusión, los resultados evidencian que los modelos de aprendizaje automático supervisado constituyen una alternativa efectiva para la predicción temprana de la diabetes utilizando información fácilmente disponible del paciente. Asimismo, el estudio demuestra que es posible desarrollar herramientas de apoyo a la decisión clínica que contribuyan a fortalecer las estrategias de prevención y detección oportuna de esta enfermedad.

Referencias bibliográficas

Alzahrani, A. S., & Alzahrani, A. A. (2022). Machine learning techniques for diabetes prediction: A comparative study. *Intelligent Systems with Applications, 16*, 200114. https://doi.org/10.1016/j.iswa.2022.200114

American Diabetes Association. (2024). *Standards of Care in Diabetes—2024*. Diabetes Care, 47(Supplement_1), S1–S350. https://doi.org/10.2337/dc24-Sint

Centers for Disease Control and Prevention. (2023). *La diabetes y la salud mental*. https://www.cdc.gov/diabetes/spanish/living/mental-health.html

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

Kaggle. (2022). *Diabetes Prediction Competition (TFUG CHD Nov 2022)*. https://www.kaggle.com/competitions/diabetes-prediction-competitiontfug-chd-nov-2022

National Institute of Diabetes and Digestive and Kidney Diseases. (s. f.). *Síntomas y causas de la diabetes*. https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/sintomas-causas

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Oguntibeju, O. (2013). *Diabetes Mellitus - Insights and Perspectives*. IntechOpen. https://doi.org/10.5772/3038

Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.

Raschka, S. (s. f.). *SequentialFeatureSelector: The popular forward and backward feature selection approaches (including floating variants)*. https://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/

Raschka, S. (s. f.). *LinearDiscriminantAnalysis: Linear discriminant analysis for dimensionality reduction*. Mlxtend User Guide. https://rasbt.github.io/mlxtend/user_guide/feature_extraction/LinearDiscriminantAnalysis/

World Health Organization. (2023). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes

Anexos

Anexo 1. Notebooks con los experimentos realizados durante el proyecto de monografía:

- Experimento 1: No se aplicaron técnicas de reducción de dimensión.
- Experimento 2: Se aplicaron métodos de Selección Secuencial de Características (SSF - Sequential Feature Selection); en particular, se utilizó la técnica de Selección Secuencial hacia Adelante (SFFS - Sequential Forward Feature Selection).
- Experimento 3: Se empleó el Análisis de Componentes Principales (PCA).
- Experimento 4: Se aplicó el Análisis Discriminante Lineal (LDA), se utilizó la versión de la librería scikit-learn.
- Experimento 5: Se aplicó el Análisis Discriminante Lineal (LDA), se utilizó la implementación de la librería mlxtend.

Estos anexos pueden encontrarse en el siguiente repositorio de GitHub:

https://github.com/BrayanDeSousa/Prediccion_temprana_diabetes_BDS