

Análisis de rendimiento de modelos gratuitos de *Machine Learning* (aprendizaje automático)  
utilizados en infraestructura de la nube en la predicción de la diabetes

Diego Fernando Alfaro Quintero

Diana Milena Chaparro Macias

Juan Pablo Lozano Novoa

Juan David Palma Roa

Universidad EAN

Especialización en Machine Learning

Ingeniera: Marie José Chéry Leal

Bogotá, Colombia

18 de noviembre de 2024

## **Resumen**

El *Machine Learning* (aprendizaje automático) es una herramienta que, aplicada a la detección de enfermedades como la diabetes, puede optimizar los procesos de prevención, detección y gestión en salud pública, pero la implementación de estas tecnologías comúnmente se ve bloqueada por la brecha en conocimientos técnicos y los altos costos, incluso en etapas piloto. Este estudio presenta un ejercicio de selección de datos, preprocesamiento y utilización de varias herramientas gratuitas de *Machine Learning* en la nube, que ofrecen una configuración sencilla y facilitan el entrenamiento de modelos para minimizar costos y tiempo en la prevención de la enfermedad. Los resultados son evaluados y comparados con diferentes métricas que demuestran la capacidad y facilidad que ofrecen estas herramientas para tareas de clasificación de pacientes sanos y pacientes con diabetes.

*Palabras clave:* Diabetes, machine learning, gestión de salud, algoritmo de entrenamiento, conjunto de datos

## Contenido

Resumen .....	2
Lista de Figuras .....	6
Lista de Tablas .....	7
Introducción .....	8
Problema de Investigación .....	9
Objetivos .....	11
Objetivo general .....	11
Objetivos específicos .....	11
Justificación .....	12
Marco Teórico.....	14
Contexto General Sobre La Diabetes Y Su Diagnóstico .....	14
Los Modelos Fundacionales Y Su Rol En Aplicaciones De Machine Learning Al Diagnóstico Médico .....	15
Servicios Gratuitos En La Nube Para Machine Learning .....	16
Modelos De Machine Learning Aplicados A La Detección De La Diabetes .....	18
Métricas De Rendimiento De Cada Modelo .....	19
Estado Del Arte .....	22
Características Del Conjunto De Datos Y Variables Relevantes Para La Detección De La Diabetes .....	22

Estructura Del Conjunto de Datos Depurado Y Selección De variables Para El	
Entrenamiento Del Modelo.....	23
Hipótesis.....	26
Hipótesis 1.....	26
Hipótesis 2.....	26
Metodología .....	26
Etapa 1. Establecimiento Del Conjunto De Datos .....	26
Actividad 1: Selección de la base de datos .....	26
Actividad 2: Limpieza de datos.....	27
Actividad 3: Transformación de datos .....	28
Etapa 2. Entrenamiento de los modelos de Machine Learning.....	29
Actividad 1. Seleccionar los modelos de Machine Learning.....	29
Actividad 2. Entrenar los modelos de Machine Learning con la base de datos depurada.....	30
Etapa 3. Análisis de los resultados de los modelos.....	32
Actividad 1: Evaluación preliminar mediante tabulación de las métricas .....	32
Actividad 2: Análisis de las métricas seleccionadas .....	32
Definición de Variables.....	33
Definición conceptual y operacional.....	33
Resultados y Análisis de Resultados.....	37

Análisis de rendimiento de modelos gratuitos de Machine Learning (aprendizaje automático) utilizados en infraestructura de la nube en la predicción de la diabetes	5
Resultado 1. Establecimiento del conjunto de datos.....	37
Actividad 1: Selección de la base de datos .....	37
Actividad 2: Limpieza de datos.....	38
Actividad 3: Transformación de datos .....	40
Resultado 2. Entrenamiento de los modelos de Machine Learning .....	42
Actividad 1. Seleccionar los modelos de Machine Learning .....	42
Actividad 2. Entrenar los modelos de Machine Learning con la base de datos depurada.....	45
Resultado 3. Comparación de los resultados de los modelos.....	54
Actividad 1: Evaluación preliminar mediante tabulación de las métricas .....	54
Discusión.....	57
Conclusiones .....	59
Recomendaciones.....	61
Referencias .....	62

### Lista de Figuras

Figura 1. Importancia de características a través de <i>Random Forest Classifier</i> .....	39
Figura 2. Conteo de Variable Clase .....	41
Figura 3. Vistazo General Sagemaker estudio de AWS .....	46
Figura 4. Resultados del modelo mediante Sagemaker estudio de AWS .....	47
Figura 5. Matriz de confusión con Sagemaker de AWS .....	48
Figura 6. Configuración Google Cloud.....	49
Figura 7. Resultados del modelo en Google Cloud .....	50
Figura 8. Resumen de la configuración de la tarea de aprendizaje de AutoML de Azure	52
Figura 9. Resultados obtenidos con AutoML de Azure .....	53
Figura 10. Matriz de confusión con AutoML de Azure.....	54

### **Lista de Tablas**

Tabla 1. Objetivos con su definición de variables a nivel conceptual y operacional.....	34
Tabla 2. Comparación de bases de datos .....	37
Tabla 3. Comparación de los servicios de la nube de Machine Learning .....	43
Tabla 4. Comparación de resultados .....	55

## Introducción

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo. Según la Organización Mundial de la Salud (OMS), en 2021 había aproximadamente 537 millones de adultos viviendo con diabetes, y se espera que esta cifra aumente a 643 millones para 2030 (Organización Mundial de la Salud, 2024). Esta enfermedad no solo representa un desafío significativo para la salud pública, sino que también impone una carga económica considerable debido a los costos asociados con el tratamiento y las complicaciones derivadas. En este contexto, la predicción y el manejo temprano de la diabetes se han convertido en áreas críticas de investigación y desarrollo.

En la actualidad, las herramientas de *Machine Learning* (aprendizaje automático) en la nube están revolucionando la forma en que se aborda la predicción de enfermedades. Estas plataformas permiten a los investigadores y profesionales de la salud analizar grandes volúmenes de datos de manera eficiente y desarrollar modelos predictivos precisos. Por ejemplo, Azure Machine Learning ofrece capacidades avanzadas para la creación y despliegue de modelos de machine learning, facilitando la integración de datos y la automatización de procesos (Microsoft, 2023). De manera similar, Amazon SageMaker proporciona un entorno completo para el desarrollo de modelos, desde la preparación de datos hasta la implementación y monitoreo (Amazon Web Service, 2023). Estas herramientas ofrecen opciones tecnológicas que se aplican exitosamente a diversidad de campos y abren la posibilidad de mejorar la precisión y la eficiencia de los sistemas de predicción de la diabetes, contribuyendo así a una mejor gestión de la enfermedad a nivel global.

## **Problema de Investigación**

La detección temprana de la diabetes es esencial para un manejo efectivo y la prevención de complicaciones graves (Asociación Latinoamericana de Diabetes, 2019). En los últimos años, las técnicas de *Machine Learning* han surgido como herramientas prometedoras para mejorar la precisión y eficiencia en la detección de la diabetes. Sin embargo, la adopción de estas herramientas enfrenta múltiples dificultades como la diversidad de enfoques y variables utilizadas en estos estudios.

El uso de modelos de *Machine Learning* en la predicción de la diabetes se ha visto limitado por la calidad de los conjuntos de datos y la falta de un enfoque metodológico estandarizado (Alam et al., 2023). Las causas subyacentes incluyen la complejidad en la identificación de las variables más relevantes y la presencia de datos ruidosos. Para abordar estas limitaciones, es necesario identificar características clave mediante técnicas de limpieza y selección de variables, seguidas del entrenamiento de modelos (Hoyos et al., 2023). La falta de calidad en los conjuntos de datos y la presencia de datos ruidosos generan modelos de *Machine Learning* que no son precisos en la detección de la diabetes. Esto podría manifestarse en un elevado número de falsos positivos o falsos negativos, afectando tanto el diagnóstico como el tratamiento adecuado de los pacientes. Además, la ineficiencia en el entrenamiento de los modelos debido a la infraestructura inadecuada puede generar resultados obsoletos o inexactos, afectando la toma de decisiones clínicas.

A pesar de la disponibilidad de algoritmos avanzados como redes neuronales y *Support vector machine* (SVM), estudios recientes publicados por expertos en consultoría estratégica vinculados en *Boston Consulting Group*, señalan que los hospitales y sistemas de salud no siempre aprovechan estas técnicas debido a la falta de estandarización en los enfoques

metodológicos. Como resultado, los modelos utilizados son menos eficientes y no alcanzan su potencial en la predicción temprana de la diabetes. Además, la fragmentación en los sistemas de información de salud, donde las instituciones no comparten datos eficazmente, limita el acceso a información completa del paciente. Esto afecta el desarrollo de modelos de *Machine Learning* que requieren conjuntos de datos amplios y completos para mejorar su capacidad predictiva. (Huddle et al., 2023)

En 2021, entre 500 y 564 millones de personas vivían con diabetes a nivel mundial, con una prevalencia del 6.1%. De no implementarse mejoras en su prevención y tratamiento se pronostica que este número aumentará a una cifra entre 1220 y 1390 millones de personas para 2050. La diabetes tipo 2, que representa la mayoría de los casos, puede prevenirse y revertirse mediante detección y control en etapas tempranas. Comprender cómo los diferentes factores de riesgo afectan la carga de diabetes en cada población es fundamental para plantear estrategias efectivas de control (Ong et al., 2023).

La implementación de herramientas en la nube ofrece múltiples ventajas para el manejo de datos. Actualmente, uno de los casos que se pueden tomar como referencia es el de los hospitales en China, que están migrando sus bases de datos desde servidores locales hacia entornos en la nube (Xu, 2024). Esta transición les permite optimizar costos, aprovechar el procesamiento en paralelo, mejorar la capacidad de análisis y generar predicciones más precisas. Considerando estos beneficios, en las próximas décadas se espera que la historia clínica de los pacientes facilite el diagnóstico temprano de enfermedades, lo que a su vez permitirá tratamientos oportunos y mejorará la calidad de vida. La implementación general de estas herramientas en países en vía de desarrollo como Colombia podría iniciar mediante opciones gratuitas de la nube que se analizan en este trabajo de investigación.

Considerando lo anterior, se plantea la siguiente pregunta de investigación: ¿Cómo es el desempeño de diferentes modelos gratuitos de *Machine Learning* implementados en infraestructura de la nube para la predicción de diabetes al compararlos respecto a sus indicadores de rendimiento?

## **Objetivos**

### **Objetivo general**

Evaluar el desempeño de diversos modelos gratuitos de *Machine Learning* utilizados en infraestructura de la nube para la predicción de diabetes a partir de indicadores de rendimiento.

### **Objetivos específicos**

1. Establecer el conjunto de datos mediante investigación y selección de las variables relevantes para el entrenamiento del modelo.
2. Entrenar los modelos de *Machine Learning* mediante los servicios gratuitos de la nube para su comparación.
3. Analizar los resultados de los modelos mediante comparación de métricas para la identificación del modelo con mejor rendimiento.

## **Justificación**

La diabetes es una afección prolongada que tienen que sobrellevar 420 millones de personas en el mundo y cuya mortalidad ha aumentado a 70%, según cifras del 2022 de la Organización Panamericana de la Salud (Organización Panamericana de la Salud, 2022a). El tratamiento de esta enfermedad es costoso, requiere de medicamentos especiales y, en muchos países, su gestión es deficiente, lo que agrava las complicaciones asociadas a la diabetes y pone una carga significativa sobre los sistemas de salud.

Es posible detener el incremento de esta enfermedad mediante la prevención de factores de riesgo, tamizaje y seguimiento, gestión de medicamentos y tratamientos, y sistemas de monitoreo y vigilancia. El desarrollo y mejora de técnicas de *Machine Learning* (ML) utilizadas en estas tareas puede cambiar el panorama poco alentador del manejo de la diabetes a nivel global (Manzini et al., 2022). La aplicación eficaz de estas tecnologías tiene el potencial no solo de mejorar la precisión en la detección temprana, sino también de optimizar la asignación de recursos en los sistemas de salud, reduciendo costos y mejorando los resultados para los pacientes.

Esta investigación busca encontrar herramientas de ML con capa gratuita, que muestren desempeños destacables en la detección de diabetes basada en variables de salud y factores de riesgo. Al identificar las herramientas más efectivas y los desafíos asociados con su implementación, esta investigación contribuirá a la creación de guías clínicas basadas en evidencia que faciliten la adopción de ML en la práctica médica, promoviendo un enfoque más preventivo y personalizado en el tratamiento de la diabetes. De esta manera, la investigación no solo aborda una necesidad crítica en el ámbito de la salud pública, sino que también impulsa la

integración de tecnologías avanzadas en la medicina, beneficiando tanto a pacientes como a profesionales de la salud.

El tema cubierto en esta investigación se acopla a la línea “Gestión en salud”, del grupo GIS (Grupo de Investigación en Salud), del campo “Ciencia, tecnología e innovación” de la Universidad EAN.

## **Marco Teórico**

### **Contexto General Sobre La Diabetes Y Su Diagnóstico**

La diabetes mellitus engloba un conjunto de patologías que alteran el metabolismo de la glucosa, un monosacárido fundamental para el mantenimiento energético de las células del organismo, particularmente en el tejido muscular y en el sistema nervioso central. La glucosa es la principal fuente de energía de las neuronas del cerebro. Las causas subyacentes de esta enfermedad varían según el tipo de diabetes, pero el resultado común es un nivel alto de glucosa en la sangre que, de no ser controlada, puede desencadenar complicaciones multisistémicas graves (Organización Panamericana de la Salud, 2022) .

Hay dos tipos de diabetes crónica: la diabetes tipo 1, autoinmune, y la diabetes tipo 2, asociada con resistencia a la insulina y factores metabólicos. Además, hay formas potencialmente reversibles de esta condición, como la prediabetes y la diabetes gestacional. En el caso de la prediabetes, los niveles de glucosa en sangre están elevados, pero no alcanzan el umbral diagnóstico para diabetes mellitus. Sin intervención adecuada, este estado puede progresar a diabetes tipo 2. La diabetes gestacional, por otro lado, aparece durante la gestación y suele resolverse tras el parto, aunque incrementa el riesgo de desarrollar diabetes en el futuro (Clínica Mayo, 2024).

Según los lineamientos proporcionados por la American Diabetes Association, el diagnóstico de la diabetes se realiza a través de varios métodos, cada uno diseñado para evaluar de manera precisa el control glucémico del paciente. Entre estos métodos destacan la prueba A1c, que mide los niveles promedio de glucosa durante los últimos 2 a 3 meses, y las pruebas de glucosa en ayunas (FPG) y tolerancia oral a la glucosa (OGTT). Estas pruebas permiten detectar

tanto la diabetes como la prediabetes, siendo esenciales para el manejo temprano de la enfermedad (American Diabetes Association, 2024).

## **Los Modelos Fundacionales Y Su Rol En Aplicaciones De Machine Learning AI**

### **Diagnóstico Médico**

Los modelos fundacionales son modelos de inteligencia artificial entrenados con enormes cantidades de datos, lo que les permite adquirir un conocimiento generalizable aplicable a diversas tareas sin necesidad de entrenamiento específico para cada una de ellas. Estos modelos, como Llama 3, son esenciales en el campo del *Machine Learning* por su capacidad para realizar tareas complejas, como el procesamiento del lenguaje natural y el análisis multimodal, utilizando grandes conjuntos de datos (Team & Meta, 2024). Su versatilidad y robustez los hacen valiosos para optimizar procesos clínicos como lo son diagnósticos médicos y tratamientos personalizados.

No obstante, poder apropiarse de estos avances en el campo del análisis computacional en el sector médico requiere esfuerzos específicos. En el análisis de la implementación de *Machine Learning* en la atención médica, David Sontag & Pete Szolovits (2020) destacan que, a pesar de los avances tecnológicos, la naturaleza compleja de los datos clínicos, con características como datos faltantes o etiquetados imprecisos, presenta desafíos únicos para la integración de algoritmos de *Machine Learning*. En suma, el éxito dependerá de la robustez de los modelos y de su capacidad para adaptarse a las necesidades clínicas específicas. Al mismo tiempo, los autores enfatizan que la mejora en los sistemas de registros electrónicos de salud (EHR), conocidos en Colombia como la historia médica electrónica, ha facilitado el potencial de estos modelos en diagnósticos más precisos.

En un análisis reciente, se expuso que los sistemas de *Machine Learning* en el ámbito de la salud han evolucionado significativamente, especialmente en la personalización de diagnósticos y tratamientos para enfermedades complejas como la diabetes. El uso de redes neuronales, modelos predictivos y datos electrónicos de salud ha permitido identificar patrones que pueden optimizar la detección temprana de patologías crónicas. Sin embargo, el desafío persiste en la integración ética y equitativa de estos modelos, asegurando que los algoritmos sean robustos y no perpetúen sesgos clínicos (Nadis, 2023).

### **Servicios Gratuitos En La Nube Para Machine Learning**

La computación en la nube es la disponibilidad bajo demanda de recursos como almacenamiento y procesamiento a través de internet, eliminando la necesidad de gestionar infraestructura física. Los usuarios y empresas solo pagan por lo que utilizan, lo que optimiza los costos y aumenta la eficiencia (Google Cloud, 2024). En el campo médico, permite procesar grandes volúmenes de datos clínicos para entrenar modelos de *Machine Learning* que mejoren los diagnósticos, como en el caso de la diabetes. A continuación, se presentan los principales proveedores de almacenamiento y procesamiento en la nube con servicios funcionales para el objeto de la presente investigación:

- Google Cloud AI, de acuerdo con su documentación oficial elaborada por Google Cloud (2024), este servicio destaca por su herramienta AutoML, que facilita el desarrollo de modelos de ML sin requerir un conocimiento profundo en programación. Es ideal para prototipos rápidos, permitiendo a los investigadores entrenar modelos personalizados con pocos datos etiquetados.

- AWS Free Tier (SageMaker), de acuerdo con la documentación oficial de este servicio elaborada por el equipo de AWS Team (2024), esta multinacional ofrece una gama de servicios, como SageMaker, que permite a los usuarios entrenar y desplegar modelos a gran escala. SageMaker facilita la integración de modelos con diferentes herramientas como TensorFlow, PyTorch y MXNet.
- Microsoft Azure, con su herramienta Azure Machine Learning Studio, permite entrenar modelos con una interfaz intuitiva, ideal para usuarios que no están familiarizados con los lenguajes de programación, ofreciendo una infraestructura escalable para experimentos pruebas. (Microsoft Azure ML Team, 2024)

En la publicación de Treveil & Dataiku Team (n.d.), se resalta que los servicios en la nube proporcionan beneficios clave. Google Cloud AI ofrece una solución eficiente con AutoML, ideal para prototipos rápidos y entrenamientos con pocos datos etiquetados. AWS SageMaker, por su parte, facilita la escalabilidad y compatibilidad con múltiples herramientas como TensorFlow y PyTorch, permitiendo modelos a gran escala. Finalmente, Azure Machine Learning Studio proporciona una interfaz intuitiva y escalable, optimizando experimentos sin requerir experiencia en programación, lo que maximiza la accesibilidad y la eficiencia investigativa.

Por otra parte, de acuerdo con lo previsto por Organización de los Estados Americanos Departamento de Derecho Internacional Secretaría de Asuntos Jurídicos (2022), el uso de datos médicos en la nube plantea importantes desafíos en términos de privacidad y seguridad, especialmente con datos sensibles de pacientes. Esto obliga a los investigadores a cumplir con

estrictas regulaciones de gobiernos nacionales en cuanto a la custodia, el almacenamiento y la utilización responsable de esta información altamente sensible.

### **Modelos De Machine Learning Aplicados A La Detección De La Diabetes**

En este segmento se conceptualizarán dos modelos que pueden usarse con los servicios gratuitos de la nube descritos.

El primer modelo son los árboles de decisión. Estos son utilizados en tareas de clasificación y regresión, y se construyen a partir de una secuencia de preguntas tipo "si/entonces" que dividen el conjunto de datos de manera jerárquica, creando nodos de decisión. Este proceso genera un árbol binario donde cada nodo representa una pregunta sobre una característica del conjunto de datos y las hojas finales representan las predicciones del modelo. La simplicidad en la visualización y su capacidad para manejar tanto características continuas como categóricas son algunas de sus fortalezas, aunque tienden a sobreajustar los datos si no se controla la complejidad mediante técnicas de pre-poda o post-poda (Müller & Guido, 2017).

Los árboles de decisión destacan en la detección de enfermedades por su versatilidad para manejar datos categóricos y numéricos. Según académicos como Müller & Guido (n.d.), esto los convierte en modelos sumamente útiles cuando se trabaja con conjuntos de datos médicos, que suelen incluir una combinación de diferentes tipos de variables. Adicionalmente, en entornos de nubes gratuitas, su implementación resulta más accesible, ya que estos modelos no requieren una capacidad computacional elevada en comparación con otros enfoques, permitiendo su uso de forma eficiente y escalable.

El segundo modelo son las redes neuronales. Según David Sontag & Pete Szolovits (2020), las redes neuronales, en particular los perceptrones multicapa, han resurgido como

modelos de referencia en muchas aplicaciones de aprendizaje automático. Su capacidad para capturar patrones complejos en grandes volúmenes de datos las hace ideales para tareas de clasificación y regresión en problemas médicos. Las redes neuronales requieren mucho tiempo de entrenamiento y ajuste cuidadoso de parámetros, lo que las hace costosas computacionalmente. Estas redes son múltiples capas ocultas que transforman los datos de entrada, permitiendo crear modelos muy complejos y no lineales, particularmente útiles para detectar enfermedades como la diabetes (Müller & Guido, 2017).

### **Métricas De Rendimiento De Cada Modelo**

La elección de métricas adecuadas para evaluar los modelos de *Machine Learning* es esencial en tareas de diagnóstico como la detección de diabetes. Las métricas más utilizadas incluyen precisión, sensibilidad, especificidad y el área bajo la curva (AUC) ROC (*Receiver Operator Characteristic*) (AUC-ROC). La curva ROC y el AUC permiten evaluar de manera efectiva la capacidad de un modelo para distinguir entre clases en diferentes umbrales, proporcionando una visión más detallada sobre su rendimiento general. Estas métricas permiten seleccionar el modelo que mejor se adapta a los requisitos clínicos, maximizando tanto la precisión como la seguridad en los diagnósticos (Ahmed et al., 2021).

La sensibilidad y la especificidad son métricas clave en la evaluación de modelos diseñados para la detección de enfermedades crónicas, como la diabetes. La sensibilidad se centra en la capacidad del modelo para detectar correctamente los casos positivos, mientras que la especificidad evalúa su capacidad para identificar los casos negativos. En aplicaciones médicas, la sensibilidad debe ser prioritaria para evitar que los pacientes con diabetes no sean diagnosticados a tiempo, mientras que la especificidad ayuda a evitar diagnósticos incorrectos y tratamientos innecesarios (Ranjit Reddy et al. 2022).

La precisión y el valor predictivo positivo (PPV) son métricas que complementan la sensibilidad y la especificidad en la evaluación de modelos. La precisión mide la proporción de predicciones correctas, mientras que el PPV indica la proporción de verdaderos positivos entre los positivos predichos. Las investigaciones recientes subrayan la importancia del PPV en la detección de enfermedades crónicas, ya que contribuye a reducir los costos derivados de falsos positivos (Boutilier et al., 2021). Un alto PPV asegura que los recursos clínicos se concentren en los casos más probables de diabetes.

La curva ROC y el área bajo la curva (AUC) siguen siendo herramientas fundamentales para evaluar la eficacia de los modelos de clasificación. En un estudio de Stawarz et al. (2023), se destaca que un AUC alto es indicativo de una alta capacidad de discriminación entre individuos con y sin la enfermedad. En el contexto de la diabetes, un AUC elevado significa que el modelo es capaz de realizar distinciones precisas entre pacientes con diferentes niveles de riesgo, lo cual es crucial para una intervención temprana y un manejo adecuado de la enfermedad.

La evaluación integral del rendimiento de un modelo requiere la comparación de múltiples métricas. Qin et al. (2022) señalan que centrarse exclusivamente en la precisión puede ser engañoso, especialmente en el ámbito clínico, donde los falsos negativos tienen un costo significativo. En la detección de diabetes, un análisis exhaustivo que incluya sensibilidad, especificidad, precisión y AUC permite seleccionar el modelo que ofrezca el mejor rendimiento global, minimizando los riesgos de diagnóstico incorrecto y mejorando los resultados para los pacientes.

La validación cruzada es una técnica crucial para evaluar la capacidad de generalización de un modelo. En estudios recientes, como el de Khanam & Foo, (2021), se demuestra que la

validación cruzada reduce el riesgo de sobreajuste al garantizar que el modelo sea capaz de adaptarse a nuevos datos. Esta técnica es especialmente útil en la comparación de modelos de *Machine Learning* en la detección de diabetes, donde es fundamental que el modelo no solo sea preciso con los datos de entrenamiento, sino también con datos clínicos nuevos y diversos.

La regularización ayuda a reducir el sobreajuste en los modelos de Machine Learning, lo que mejora su rendimiento en datos no vistos. Según Jacobs et al. (2024), la implementación de técnicas de regularización como Lasso o Ridge permite simplificar los modelos al eliminar características irrelevantes o redundantes, mejorando su capacidad de generalización. En la comparación de modelos para la detección de diabetes, el uso de regularización puede ser determinante para seleccionar un modelo que mantenga su eficacia con distintos conjuntos de datos.

La elección del modelo óptimo para la detección de diabetes debe basarse en un análisis equilibrado de métricas clave como la sensibilidad, especificidad, precisión y AUC. En estudios recientes sobre aplicaciones de *Machine Learning* en medicina, se ha subrayado la importancia de priorizar modelos que minimicen los falsos negativos, dado el impacto que un diagnóstico no detectado puede tener en los pacientes (Stawarz et al., 2023). El modelo que logre el mejor equilibrio entre estas métricas será el más adecuado para su implementación en entornos clínicos, optimizando tanto los resultados de salud como los recursos disponibles.

## **Estado Del Arte**

### **Características Del Conjunto De Datos Y Variables Relevantes Para La Detección De La Diabetes**

El conjunto de datos es la materia prima con la que se afinan los algoritmos de ML y, por tanto, la selección de las variables y datos utilizados es uno de los puntos más importantes al entrenar el modelo. Como primer paso, se hizo una consulta generalizada de estudios de ML en detección de diabetes, para identificar conjuntos de datos y variables utilizadas (Wee et al., 2024).

En la mayoría de los estudios, como es el caso de Smith et al. (s/f) y Shaukat et al. (2023), se utilizan conjuntos de datos de acceso público, el más popular es el Pima Indian Diabetes Database PIDD de 1988. Este se compone de registros con las siguientes características: Edad, índice de masa corporal BMI, función del pedigrí de la diabetes, número de embarazos, glucosa plasmática en prueba de tolerancia de glucosa oral, presión arterial, grosor del pliegue cutáneo del tríceps, insulina sérica de 2 horas y presencia de diabetes en el cuerpo de la muestra. Este es un conjunto de datos bien validado que no requiere un preprocesamiento complejo.

El conjunto de datos utilizado por Zou et al. (2018) está compuesto por registros de personas saludables y de personas diabéticas, con 14 características: Edad, frecuencia cardíaca, respiración, presión sistólica izquierda (LSP), presión sistólica derecha (RSP), presión diastólica izquierda (LDP), presión diastólica derecha (RDP), altura, peso, índice físico, glucosa en ayunas, cintura, lipoproteína de baja densidad (LDL) y lipoproteína de alta densidad (HDL).

Vidhya & Shanmugalakshmi (2020) utilizó en su estudio un conjunto de datos construido con información recolectada de pacientes mediante cuestionarios, con características como: Antecedentes familiares, índice de masa corporal (IMC), hábitos alimentarios, ejercicio, edad del paciente, consumo de alcohol, atracones nocturnos o hábito de comer, permanencia prolongada de inactividad (sentado), nivel de glucosa en sangre (BGL), tabaquismo, naturaleza del trabajo, género (masculino o femenino).

Rani & Kiran Kumar (n.d.) abordó el problema con un conjunto de datos compuesto por variables digitales (sí o no, 1 o 0) de diferentes indicadores médicos: Edad, poliuria, polidipsia, pérdida repentina de peso, debilidad, polifagia, candidiasis genital, visión borrosa, picazón, irritabilidad, retraso en la cicatrización, paresia parcial, rigidez muscular, alopecia, obesidad y presencia de diabetes en las muestras.

Cada estudio tiene un enfoque diferente y la elección de sus datos de entrenamiento va acorde a su objetivo, como en Czmil et al. (2019) que acotó sus variables de entrada a medidas de la actividad física mediante un dispositivo móvil.

### **Estructura Del Conjunto de Datos Depurado Y Selección De variables Para El Entrenamiento Del Modelo**

La reducción de dimensionalidad es crucial para simplificar las variables de los conjuntos de datos. Existen diferentes métodos los cuales se evalúan de acuerdo con las necesidades de cada modelo y conjunto de datos trabajado. En el año 2024 un grupo de investigadores del Departamento de Ingeniería y Ciencia de la Computación de la Universidad de Negocios y Tecnología de Bangladesh realizaron un estudio en el cual buscaron crear un modelo que mejorara la predicción de la diabetes donde tomaron un conjunto de datos con 8 variables y

calcularon mediante *Random Forest* el valor de las *Feature importance* en el cual determinaron 4 variables importantes (edad, nivel HbA1c, nivel de glucosa en sangre, índice de masa corporal) (Shahin Alam et al., 2024).

En el año 2024 tres investigadores de la India crearon un modelo de predicción de diabetes en una etapa temprana donde tomaron un conjunto de datos de un proyecto publicado en el 2020. Para el tratamiento de los datos, primero los normalizaron y aplicaron Análisis de Componentes Principales (PCA) para la reducción de dimensionalidad de los datos (Nagpal et al., 2024).

En el año 2024 dos investigadores buscaron optimizar los modelos de machine Learning para la predicción de Diabetes a través de selección de características y reducción de dimensionalidad en el cual implementaron 2 técnicas de reducción de dimensionalidad t-SNE (*t-distributed Stochastic Neighbor Embedding*) y MDS (*Multidimensional Scaling*), obteniendo una precisión del 92,5% (Aouragh et al., 2024).

En el año 2020, 5 investigadores de diferentes universidades de la India y España decidieron realizar una evaluación de conjuntos de datos de enfermedades crónicas diseñando un Evaluador de Atributos el cual comprendía de la agrupación de *K-Means*, el método *Correlation Feature Selection* (CFS) y el método *Best First Search* (BFS), para generar un subconjunto de atributos optimizado. Este modelo se aplicó a 15 conjuntos de datos evidenciando mejoría en la eficiencia de los modelos (Mishra et al., 2020).

En el año 2021 cuatro investigadores vieron la problemática que se presenta en el entrenamiento de modelos de *Machine Learning* cuando en el conjunto de datos se encuentran datos faltantes o datos atípicos, por lo que analizaron el efecto que esto conlleva en las diferentes

técnicas. Las técnicas de reducción de dimensionalidad utilizadas fueron *missing-values ratio* (relación de valores perdidos), *low variance filter* (filtro de baja varianza), *high correlation filter* (filtro de alta correlación), *random forest*, *principal component analysis* (análisis de componentes principales), *linear discriminant analysis* (análisis discriminante lineal), *backward feature elimination* (eliminación de características hacia atrás), *forward feature construction* (construcción de características hacia adelante), y *rough set theory* (teoría de conjuntos aproximados). El resultado evidenció que aplicar estas técnicas minimizó el sobreajuste (Salam et al., 2021).

En el año 2023 tres investigadores analizaron el rendimiento de modelos de *Machine Learning* para la detección de cáncer de próstata. En el estudio aplicaron tres diferentes técnicas de reducción de dimensionalidad: Análisis de Componentes Principales (PCA), PCA con *Kernel function* y *autoencoder* (tipo de red neuronal utilizada para obtener una representación compacta de los datos de entrada), donde evidenciaron que PCA es una técnica que requiere menores recursos computacionales de *autoencoder*. En este estudio se demostró que aplicar técnicas de reducción de dimensionalidad mejoró el rendimiento de los modelos (Kabir et al., 2023).

En el año 2023 tres investigadores de la India crearon un modelo para la detección y clasificación de la diabetes con datos obtenidos de encuestas de la India mediante *Machine Learning*. En este estudio se aplicó el método de *Kernel entropy* para la reducción de dimensionalidad de los datos, y se aplicaron 10 modelos de *Machine Learning* para la predicción y clasificación. En conclusión, el modelo con mejor resultado fue *Random Forest* (Thotad et al., 2023).

## **Hipótesis**

### **Hipótesis 1**

Todas las herramientas de Machine Learning probadas en este estudio darán resultados satisfactorios. (F1-score mayor a 0.9)

### **Hipótesis 2**

Las herramientas que realizan automáticamente la puesta a punto del modelo, sin dar opciones de personalización al usuario, darán el rendimiento más bajo.

## **Metodología**

La metodología de este proyecto se basó en un enfoque cuantitativo para la detección de diabetes, utilizando una base de datos específica. Inicialmente, se realizó una revisión exhaustiva de la bibliografía para identificar los mejores modelos gratuitos de *Machine Learning* disponibles, implementados en infraestructuras de la nube. Posteriormente, se llevó a cabo el preprocesamiento de los datos, incluyendo la limpieza, transformación y selección de características. Una vez preparados los datos, se entrenaron diversos modelos de *Machine Learning* y se evaluó su rendimiento mediante métricas específicas en cada plataforma, con el objetivo de determinar el modelo más eficaz para la detección de diabetes.

### **Etapas 1. Establecimiento Del Conjunto De Datos**

#### ***Actividad 1: Selección de la base de datos***

El primer paso en la recolección de datos consistió en la consolidación de información. Esto implicó revisar múltiples bases de datos para asegurarse de que contenían todos los datos necesarios y que estos eran relevantes para la detección de diabetes.

Para este proyecto, se llevó a cabo una revisión exhaustiva de fuentes mediante la consulta de bases de datos que contenían las variables más relevantes, según la información obtenida en el estado del arte. Además de la presencia de variables relevantes, se evaluó la calidad de los datos, su completitud y representatividad, asegurando que los conjuntos seleccionados reflejaran adecuadamente la población estudiada, de modo que, se redujeron sesgos y se contribuyó a mejorar la robustez del modelo. Las variables más comunes incluyeron sexo, edad, índice de masa corporal (IMC) y nivel de glucosa en sangre. Además de estas variables básicas, se buscaron otras características específicas que pudieran influir en la detección de diabetes, tales como:

- **Presión arterial:** Indicador clave de la salud cardiovascular, que puede estar relacionado con la diabetes.
- **Niveles de colesterol y triglicéridos:** Factores importantes en la evaluación del riesgo metabólico.
- **Hábitos de vida:** Datos sobre la dieta, actividad física y consumo de tabaco o alcohol, que pueden influir en el desarrollo de la diabetes.

Con base en esta información, se investigó una base de datos que cumple con las características definidas, utilizando tres fuentes principales: Kaggle, UCI Machine Learning Repository y Google Dataset Search.

### ***Actividad 2: Limpieza de datos***

El segundo paso fue realizar una revisión exhaustiva de la base de datos seleccionada. Este proceso se llevó a cabo utilizando el lenguaje de programación Python, aprovechando

librerías especializadas como Pandas, NumPy y Scikit-learn. Estas herramientas permitieron detectar anomalías en el conjunto de datos que puedan generar sesgos en los resultados. La revisión inicial se centró en identificar valores faltantes y evaluar su impacto en el análisis. Además, se implementaron técnicas para manejar posibles desequilibrios en las clases de diabetes/no diabetes, como el submuestreo de la clase mayoritaria, el cual consiste en eliminar aleatoriamente datos de la clase sobrerrepresentada hasta que el conjunto de datos esté equilibrado (Murel, 2024b) y el sobremuestreo de la clase minoritaria, el cual duplica los registros de la clase subrepresentada hasta lograr el equilibrio de clases (Murel, 2024a).

Una vez identificados los valores faltantes, se analizaron sus características estadísticas para determinar el método más adecuado de imputación o eliminación. Se consideraron técnicas como la imputación con la media, mediana o moda, así como la eliminación de registros innecesarios que no aportaran valor al modelo. Para abordar el desequilibrio de clases, se utilizaron métodos como SMOTE (*Synthetic Minority Over-sampling Technique*) y el ajuste de pesos en el modelo. Este enfoque sistemático aseguró que los datos estuvieran limpios y equilibrados, listos para el siguiente paso del análisis (Alarcón Paredes et al., 2024).

### ***Actividad 3: Transformación de datos***

La transformación de la base de datos fue un paso de gran relevancia para preparar los datos para el análisis. Este proceso incluyó la normalización y estandarización de las variables numéricas, asegurando que todas tuvieran una escala comparable. La normalización de los datos fue importante para asegurar que cada variable tuviera el mismo peso, evitando sesgos por la diversidad de los rangos ya que, si se cuenta con una variable “a” con valores en el rango de 0 a 5000 y otra variable “b” con valores entre 0 a 500, la variable “a” tiene más influencia en el

modelo (Blanco, 2023). Además, se realizó la codificación de variables categóricas mediante técnicas como *One-Hot Encoding* o *Label Encoding*, convirtiéndolas en un formato numérico adecuado para los algoritmos de Machine Learning. También se pudieron crear nuevas características a partir de las existentes, si se consideraba que podían mejorar la capacidad predictiva del modelo. Estas transformaciones garantizaron que los datos estuvieran en un formato óptimo para el entrenamiento y evaluación de los modelos, mejorando así la precisión y robustez de los resultados.

Finalmente, la base de datos se dividió en un conjunto de entrenamiento (80%) y un conjunto de validación (20%) para prevenir el sobreajuste. Esta estrategia permite evaluar la capacidad del modelo para generalizar a nuevos datos y es ampliamente utilizada en aplicaciones médicas (Khanam & Foo, 2021). Esta división es clave para identificar posibles debilidades en el modelo antes de su implementación clínica definitiva.

## **Etapas 2. Entrenamiento de los modelos de Machine Learning**

### ***Actividad 1. Seleccionar los modelos de Machine Learning***

El proceso de selección de modelos de *Machine Learning* se llevó a cabo utilizando algoritmos diseñados específicamente para resolver problemas de clasificación binaria, los cuales han sido descritos previamente, dado que la meta es predecir la probabilidad de presencia o ausencia de diabetes. Los modelos se eligieron con base en tres criterios clave: capacidad computacional, interpretabilidad y robustez en entornos médicos. En este sentido, se priorizaron aquellos algoritmos que pueden adaptarse a infraestructuras tecnológicas limitadas, típicas de los sistemas de salud pública en Colombia. Esta evaluación tomó en cuenta la complejidad del contexto operativo y las limitaciones de recursos, garantizando que los modelos seleccionados no

solo proporcionen un desempeño preciso, sino que también sean sostenibles en escenarios con restricciones técnicas.

Posteriormente, el entrenamiento de los modelos se alineó con el propósito general del estudio, que es identificar la combinación óptima entre modelo y plataforma que ofrezca un rendimiento superior en la predicción de diabetes. Esta selección permitió mantener un equilibrio entre precisión y eficiencia operativa, promoviendo la adopción de estos modelos en organizaciones de salud pública. A futuro, se planifica que el despliegue de los modelos no solo maximice su utilidad clínica, sino que también facilite la toma de decisiones en entornos con recursos limitados. En definitiva, este enfoque metodológico tiene como objetivo asegurar que las soluciones desarrolladas puedan ser integradas de manera efectiva en la práctica médica cotidiana, brindando valor tanto a pacientes como a profesionales del sector salud.

### ***Actividad 2. Entrenar los modelos de Machine Learning con la base de datos depurada***

Con los modelos seleccionados y la base de datos depurada, el siguiente paso fue entrenar cada modelo utilizando las plataformas. Por lo cual, es necesario destacar que la consulta permanente de la documentación de cada uno de estos servicios resultó en una actividad indispensable en todo el proceso de entrenamiento del modelo. Estas plataformas no solo brindan la infraestructura necesaria para el entrenamiento, sino que también ofrecen herramientas y configuraciones *low-code* que simplifican el proceso para investigadores con distintos niveles de experiencia, tal como se abordó previamente.

El proceso de entrenamiento sigue una secuencia lógica que asegura la eficiencia y precisión de los modelos, debido a que en el contexto de salud la precisión de la predicción de

diabetes es muy importante. Cada paso se presenta para maximizar la capacidad de generalización y minimizar el riesgo de sobreajuste, garantizando que los modelos se desempeñen adecuadamente en datos no vistos. Los pasos específicos son:

### **Ajuste De Hiperparámetros Mediante Random Search Y Grid Search:**

El ajuste de hiperparámetros es un proceso fundamental para mejorar el rendimiento de los modelos. Se emplearon las técnicas de *random search* y *grid search*, las cuales permiten explorar un espacio de posibles configuraciones y encontrar los parámetros óptimos para cada modelo (Müller & Guido, n.d.). Estas técnicas son compatibles con plataformas como AWS SageMaker, donde se automatiza el ajuste mediante *AutoPilot*.

El proceso de entrenamiento no solo pone en práctica las decisiones tomadas en la fase de selección, sino que también permite validar si esas decisiones fueron acertadas al evaluar el desempeño en un entorno real. La integración entre modelos y plataformas es clave para garantizar que el enfoque metodológico cumpla con el objetivo general de la investigación. Esta fase asegura que cada modelo sea probado bajo condiciones estandarizadas, permitiendo comparaciones objetivas entre algoritmos y plataformas.

En resumen, la sinergia entre las fases de selección y entrenamiento refuerza el enfoque sistemático del estudio, alineando cada paso con los objetivos específicos: seleccionar variables y algoritmos relevantes, entrenar los modelos en servicios gratuitos de la nube, y analizar los resultados para determinar la solución más eficaz. Este enfoque garantiza que los resultados obtenidos no solo tengan validez estadística sino también relevancia clínica, contribuyendo a mejorar la predicción temprana de la diabetes en entornos con recursos limitados.

### **Etapa 3. Análisis de los resultados de los modelos**

#### ***Actividad 1: Evaluación preliminar mediante tabulación de las métricas***

La evaluación de los modelos se realizó utilizando métricas como sensibilidad, precisión y AUC-ROC como métricas principales. Estas fueron registradas en una tabla comparativa entre modelos y métricas con el fin de poder evaluar el rendimiento de cada una.

#### ***Actividad 2: Análisis de las métricas seleccionadas***

De acuerdo con las métricas descritas en la tabla de resultados, se analizó cuál es el modelo que tiene mejor rendimiento comparando los valores resultantes de acuerdo con la definición de cada una de las métricas permitiendo una comparación justa de acuerdo con su sensibilidad, precisión, AUC-ROC y demás.

La sensibilidad es crucial en el contexto clínico, ya que permite detectar de manera temprana casos positivos de diabetes, minimizando falsos negativos (Boutilier et al., 2021). La métrica AUC-ROC proporciona una visión global del rendimiento del modelo al evaluar su capacidad para discriminar entre clases en diferentes umbrales (Hanley & McNeil, 1982).

El AUC-ROC es una métrica importante que mide la capacidad de los modelos para distinguir entre clases. Un AUC alto indica una excelente capacidad de separación entre los casos positivos y negativos, lo que es fundamental en un entorno clínico.

La evaluación de los modelos mediante métricas como sensibilidad, precisión y AUC-ROC, junto con un enfoque en la interpretabilidad y el equilibrio de clases, es crucial para el éxito en la detección de diabetes. Al final de esta etapa de evaluación, se evidenció que los

modelos seleccionados no solo son efectivos en la detección de diabetes, sino que también son respaldados por una comprensión clara de su funcionamiento.

La combinación de un alto rendimiento predictivo y la capacidad de los profesionales de la salud para interpretar los resultados garantizará que los modelos tengan un impacto significativo en la detección y manejo de la diabetes, contribuyendo a una mejor salud pública.

En resumen, la clave para una implementación exitosa radica en un enfoque holístico que considere no solo las métricas de rendimiento, sino también la usabilidad y la confianza en los modelos, lo que a su vez permitirá mejorar los resultados de salud en la población.

## **Definición de Variables**

### **Definición conceptual y operacional**

En la Tabla 1, se muestra la descripción de las variables utilizadas en este estudio, su proceso de medición y su relación con los objetivos específicos.

**Tabla 1.**

Objetivos con su definición de variables a nivel conceptual y operacional

Objetivo	Variable	Definición conceptual	Definición operacional
Establecimiento de la base de datos	Número de registros	Cada fila que compone el conjunto de datos correspondiente a una persona estudiada que resultó positiva o negativa para diabetes.	Al tener acceso al conjunto de datos se revisa la cantidad de filas que lo componen.
	Porcentaje de registros con datos nulos	Proporción de registros que no tienen datos completos en todas las columnas del conjunto de datos.	Se detecta la presencia de datos nulos en los registros mediante un análisis hecho con herramientas como la librería Pandas de Python, o incluso con las herramientas de la nube que se van a probar (si esta opción está disponible).
	Número de variables de entrada	Cantidad de variables independientes que se utilizarán para entrenar el modelo. También llamadas características o atributos. Corresponden a las columnas del conjunto de datos diferentes al diagnóstico de diabetes.	En base a la investigación previa se identifican variables relevantes y se contrastan con las registradas en cada conjunto de datos.
	Número de variables cuantitativas	Cantidad de variables que miden un atributo con un valor numérico.	Se identifica el número de columnas del conjunto de datos, con variables tipo <i>Integer</i> , <i>Float</i> y <i>Double</i> .
	Número de variables cualitativas	Cantidad de variables que no pueden ser medidas numéricamente, solo expresadas con palabras.	Se identifica el número de columnas del conjunto de datos, con variables tipo <i>String</i> .
	Desbalanceo de clases	Es una dificultad que presentan algunos conjuntos de datos usados en <i>Machine Learning</i> . Consiste en que una o varias categorías del conjunto de datos tienen muchos menos registros que otras. Esto genera sesgos de clasificación y afecta el rendimiento de los modelos. (Alarcón Paredes et al., 2024)	Se identifica el porcentaje de registros del conjunto de datos con diagnóstico positivo para diabetes. Se hace lo mismo para los registros con diagnóstico negativo, y se comparan ambos porcentajes. Idealmente los porcentajes deberían ser 50% en ambas categorías o lo más cercano posible.

automático) utilizados en infraestructura de la nube en la predicción de la diabetes

	Reputación	Utilización del conjunto de datos en los estudios consultados en el Estado del arte.	Número de estudios consultados en el marco teórico y estado del arte que han utilizado cada conjunto de datos.
	Accesibilidad	Facilidad para acceder a la descarga del conjunto de datos (registro, pagos, acceso parcial).	Costo en dólares de acceder a la totalidad del conjunto de datos.
Entrenamiento de los modelos de <i>machine learning</i>	Cuota de mercado	Porcentaje que una empresa de servicios en la nube representa en el total de ventas de este mercado durante un periodo de tiempo específico. (INEAF Business School, n.d.)	Se investiga en datos públicos las estadísticas de dominio de mercado de las plataformas de soluciones en la nube.
	Oferta de servicios de <i>Machine Learning</i>	Lista de servicios relacionados al aprendizaje automático ofrecidos por la plataforma de soluciones en la nube.	Se investiga en el catálogo de servicios de cada plataforma, cuántos corresponden directamente a <i>Machine Learning</i>
	Alcance de la capa gratuita	Oferta de servicios completos o de servicios con limitaciones que no suponen un cobro por parte de la plataforma.	Se investiga la cantidad y límites de uso, tanto en capacidad como en tiempo, de los servicios de <i>Machine Learning</i> de la capa gratuita de cada plataforma.
	Opciones <i>low-code/no-code</i>	Métodos que facilitan el diseño y desarrollo mediante herramientas intuitivas que eliminan la necesidad de escribir código. (SAP Colombia, n.d.)	Se investiga el grado de conocimiento en codificación que requiere el uso de las herramientas de capa gratuita de cada plataforma, y se clasifica en grados principiante, intermedio y avanzado.
Análisis de los resultados de los modelos	Verdaderos positivos VP	El modelo predice correctamente el resultado como positivo. (Wei-Meng Lee, 2019)	Se realiza la prueba de predicción con una muestra diferente a la utilizada en el entrenamiento del modelo. Se compara el resultado esperado con el obtenido y se forma la matriz de confusión compuesta por VP, VN, FP y FN.
	Verdaderos negativos VN	El modelo predice correctamente el resultado como negativo. (Wei-Meng Lee, 2019)	
	Falsos positivos FP	El modelo predice erróneamente el resultado como positivo. (Wei-Meng Lee, 2019)	

automático) utilizados en infraestructura de la nube en la predicción de la diabetes

Falsos negativos FN	El modelo predice erróneamente el resultado como negativo. (Wei-Meng Lee, 2019)	
Exactitud	La suma de todas las predicciones correctas dividida por el número total de predicciones. (Wei-Meng Lee, 2019)	Con los datos de la matriz de confusión, se calcula con la fórmula: $(VP+VN) / (VP+VN+FP+FN)$
Precisión	Esta métrica se relaciona con la cantidad de predicciones positivas correctas. (Wei-Meng Lee, 2019)	Con los datos de la matriz de confusión, se calcula con la fórmula: $VP/(VP+FP)$
Recall	También conocido como tasa de verdaderos positivos (TPR). Esta métrica se relaciona con la cantidad de predicciones positivas correctas. (Wei-Meng Lee, 2019)	Con los datos de la matriz de confusión, se calcula con la fórmula: $VP/(VP+FN)$
F1-score	También conocido como la media armónica de precisión y recall. Resume la evaluación del modelo en un solo número. (Wei-Meng Lee, 2019)	Con los datos de la matriz de confusión, se calcula con la fórmula: $2 * (precisión * recall) / (precisión + recall)$
AUC-ROC (Opcional)	Aplicado al ámbito estadístico en diagnósticos médicos, es la probabilidad de que un paciente seleccionado al azar con una afección tenga un resultado de prueba que indique una mayor sospecha que un paciente elegido al azar sin la afección. (Hanley & McNeil, 1982)	Si la herramienta da la opción se generará esta gráfica, con el objetivo de establecer el mejor umbral para clasificar pacientes positivos y negativos para diabetes, y optimizar los resultados obtenidos.

*Nota:* Descripción de las variables que se utilizaran en el estudio, con su definición y proceso de medición. Autoría propia.

## Resultados y Análisis de Resultados

A continuación se presentan los resultados de las etapas desarrolladas para alcanzar los objetivos del proyecto.

### Resultado 1. Establecimiento del conjunto de datos

#### *Actividad 1: Selección de la base de datos*

La selección de la base de datos se realizó mediante la comparación de las bases de datos utilizadas en los artículos revisados en el estado del arte. Se consolidó una tabla comparativa en la cual se analizaron las variables de accesibilidad, que describe si la base de datos es pública o confidencial, el número de registros, el número de variables, la cantidad de variables cualitativas y cuantitativas, porcentaje de registros con datos, desbalanceo de clases. Esto se describe en la Tabla 2.

**Tabla 2.**  
Comparación de bases de datos

Bases de datos	Accesibilidad	Número de registros	Variables de entrada	Número de variables cuantitativas	Número de variables cualitativas	Porcentaje de registros con datos nulos	Desbalanceo de clases Diabetes / Saludables
Pima Indian Diabetes Database PIDD de 1988	Acceso público	786	8	8	0	50,12%	268 / 500
Hospital de Luzhou, China 2018	Acceso confidencial	178131	14	14	0	No mencionado	No mencionado
CDC Diabetes Health Indicators UCI	Acceso público	253680	21	21 (Alguna categóricas ya codificadas)	Ya codificadas	0%	39977 / 213703
Hospital Estatal de Rzeszow en Polonia	Acceso confidencial	330	9	8	1	No mencionado	215 / 115

*Nota:* Comparación de bases de datos evidenciadas en el estado del arte. Autoría propia.

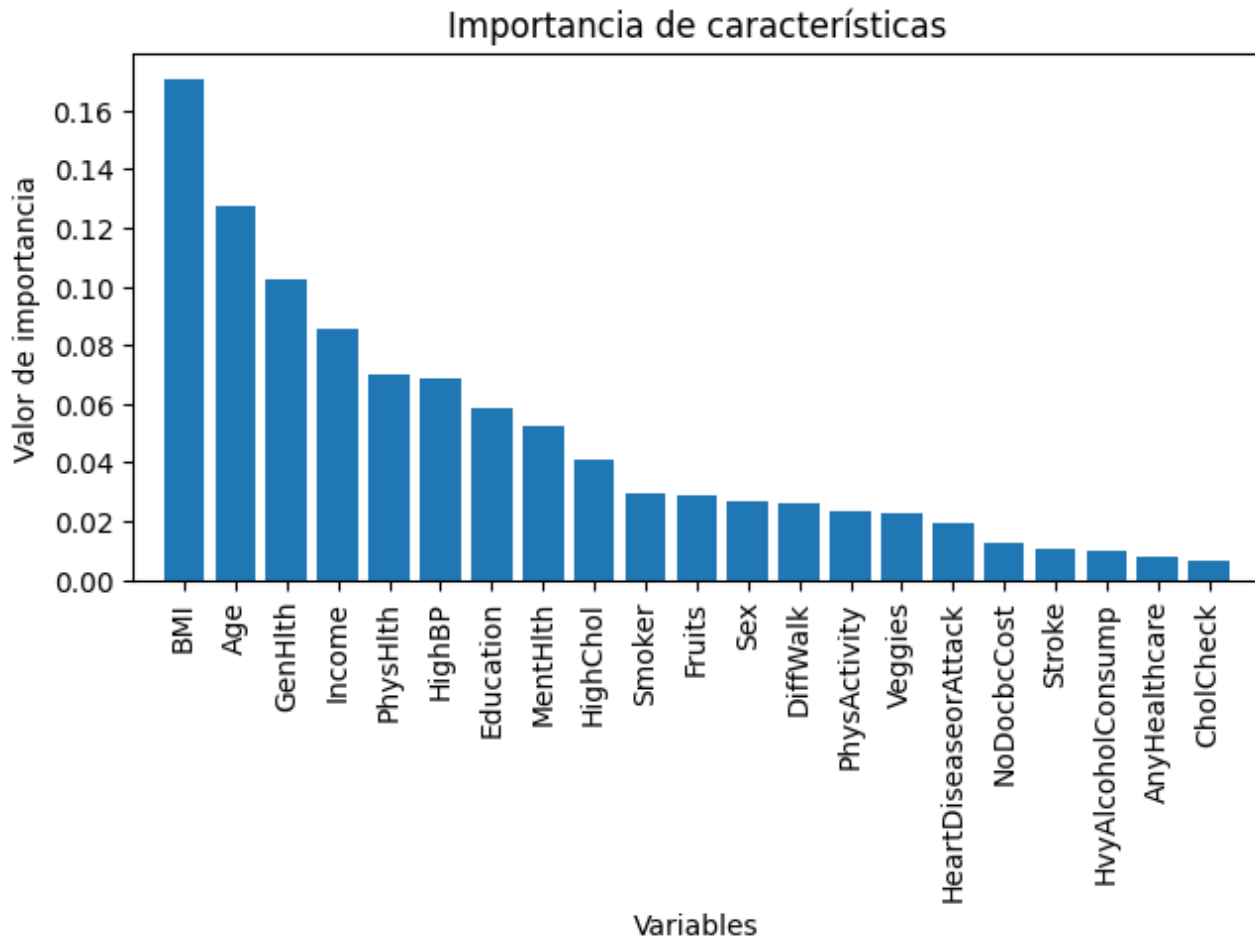
Inicialmente se analizó la variable accesibilidad en la cual dos de las cuatro bases de datos son de acceso confidencial, lo que no permite el fácil uso de ellas. Se continuó el análisis entre Pima Indian Diabetes Database PIDD de 1988 y CDC Diabetes Health Indicators UCI, donde se analizaron los diferentes atributos en los cuales se evidenció que la base de datos PIDD contiene un gran porcentaje de datos nulos en comparación con la CDC, lo que afecta el rendimiento de los modelos, y genera sesgos en las predicciones. Se analizaron las variables y se encuentran que coinciden en las variables BMI (índice de masa corporal) y Age (edad), pero se evidenció que la base de datos de CDC tiene variables relacionadas con el estilo de vida, siendo esto un factor importante para la predicción de la diabetes, por lo tanto, se seleccionó la base de datos CDC Diabetes Health Indicators UCI.

### ***Actividad 2: Limpieza de datos***

En la limpieza de datos inicialmente se verificaron los tipos de datos, donde se obtuvieron 21 variables en formato decimal con 70602 registros (incluida la variable clase de tipo binario). Se realizó conteo de valores nulos y no se evidenció ninguno. Se calculó la cantidad de registros duplicados y se encontraron 1635 que posteriormente fueron eliminados.

Para mejorar el rendimiento del modelo es importante simplificar la base de datos, por lo que mediante el uso de *Random Forest Classifier* se obtuvo el listado de las variables más importantes como se observa en la Figura 1. De acuerdo con la cantidad de variables, se toman las variables con valor de importancia mayor a 0.05.

**Figura 1.**  
Importancia de características a través de *Random Forest Classifier*



*Nota.* Figura que muestra las variables en el orden de importancia de acuerdo con la clasificación a través de Random Forest Classifier. Autoría Propia.

De acuerdo con las variables más significativas evidenciadas en el estado del arte, se encontraron similitudes en la importancia de características establecidas por el procesamiento, lo que indica que los datos son acordes a la literatura.

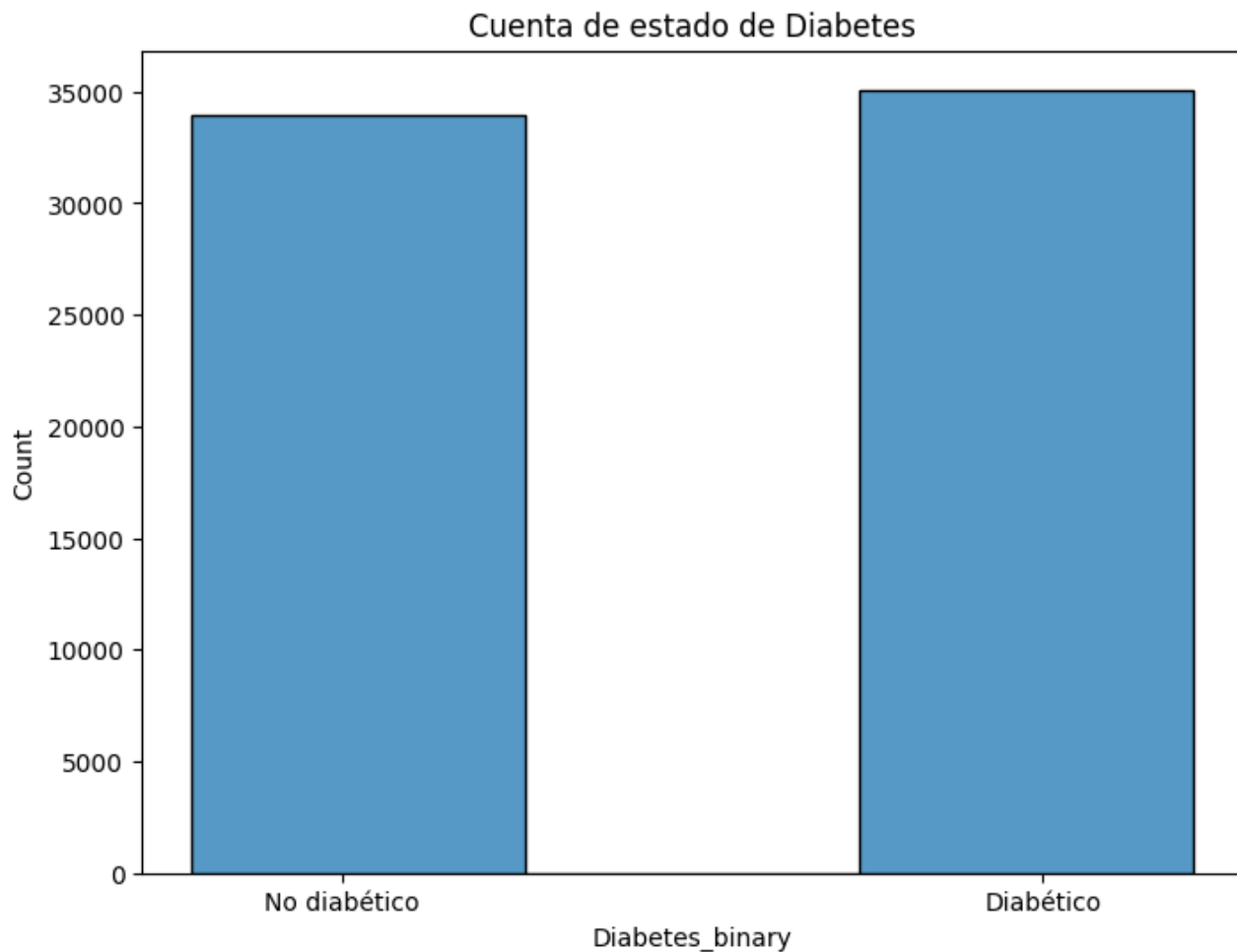
### ***Actividad 3: Transformación de datos***

Con el fin de garantizar que todas las variables contribuyan de manera equitativa al modelo, se verificaron los rangos mínimos y máximos de cada variable, donde se evidenció que estas no se encontraban normalizadas, por lo que se aplicó la técnica de normalizado entre el rango de 0 a 1. Esto se logró mediante el uso de la función *MinMaxScaler* de la librería *scikit-learn*. La fórmula matemática de esta función es:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \text{ (ecuación 1)}$$

Para evitar sesgos en la clasificación de los datos, la variable clase debe estar balanceada. Para esto se realizó conteo de clases donde se evidenció que 35097 registros corresponden a pacientes con diabetes y 33960 corresponden a pacientes sin diabetes como se evidencia en la Figura 2.

**Figura 2.**  
Conteo de Variable Clase



*Nota.* Figura que muestra las clases y la cantidad de registros en cada una de estas. Autoría Propia.

La base de datos cuenta con una gran cantidad de datos, por lo que para el balanceo de clases se utilizó la técnica de submuestreo mediante la función *NearMiss* de la librería *imblearn.under\_sampling*, donde cada clase tiene 33960 registros, por lo tanto la base de datos cuenta con 67920 registros. Finalmente, se dividió la base de datos en un conjunto de entrenamiento (80%) y un conjunto de validación (20%) para iniciar con el entrenamiento de los modelos.

## **Resultado 2. Entrenamiento de los modelos de Machine Learning**

### ***Actividad 1. Seleccionar los modelos de Machine Learning***

Para seleccionar los mejores modelos de entrenamiento, se realiza una tabla comparativa que relaciona las características de Cuota de mercado Q3 2024, Oferta de servicios de Machine Learning, Alcance de la capa gratuita, Opciones low-code/no-code como se evidencia en la Tabla 3.

**Tabla 3.**  
Comparación de los servicios de la nube de Machine Learning

Ofertas\Modelos	AWS	Microsoft Azure	Google Cloud	IBM Watson
Cuota de mercado Q3 2024	31% (Felix Richter, 2024)	20% (Felix Richter, 2024)	12% (Felix Richter, 2024)	2% (Felix Richter, 2024)
Oferta de servicios de Machine Learning	<p>Amazon Polly</p> <p>Amazon Transcribe</p> <p><b>Amazon SageMaker</b></p> <p>Amazon Textract</p> <p>Amazon Kendra</p> <p>Amazon Personalize</p> <p>Amazon Rekognition</p> <p>Amazon Lex</p> <p>Amazon Comprehend</p> <p>Amazon Translate</p> <p>(Amazon Web Services, n.d.-c)</p>	<p><b>Azure Machine Learning</b></p> <p>Azure Cognitive Services</p> <p>Azure SQL Managed Instance Machine Learning Services</p> <p>Machine learning in Azure</p> <p>Synapse Analytics</p> <p>Machine learning and AI with ONNX in Azure SQL Edge</p> <p>Azure Databricks.</p> <p>(Microsoft Learn, 2024a)</p>	<p>Vertex AI Studio</p> <p>Generative AI Document Summarization</p> <p>Vertex AI Platform</p> <p>Vertex AI Notebooks</p> <p><b>AutoML</b></p> <p>Natural Language AI</p> <p>Otras soluciones de video, imagen y conversación.</p> <p>(Google Cloud, n.d.-a)</p>	<p>Watson Studio</p> <p>watsonx Assistant</p> <p>Natural Language Understanding</p> <p>Watson Discovery</p> <p><b>Watson Machine Learning</b></p> <p>OpenPages</p> <p>Otras soluciones de conversación y asistencia. (IBM Cloud, n.d.-a)</p>
Alcance de la capa gratuita	<p>Para Amazon SageMaker, prueba gratuita de 2 meses.</p> <p>250 horas al mes de uso de la notebook t2.medium durante los primeros dos meses.</p> <p>50 horas al mes de m4.xlarge para capacitación durante los primeros dos meses.</p>	<p>El nivel gratuito da acceso limitado a los servicios de Azure ML. Algunos productos tienen período gratuito de 12 meses y \$200 en créditos de Azure por los primeros 30 días. (Azure, n.d.)</p>	<p>El límite de la capa gratuita no caduca, pero puede ser modificado. Para AutoML es de 6 horas de nodo para entrenamiento y predicción. (Google Cloud, n.d.-c)</p>	<p>En la capa gratuita, IBM Watson se puede manejar con 1 usuario autorizado, sin restricción de tiempo. (IBM Cloud, n.d.-c)</p>

## la nube en la predicción de la diabetes

Opciones low-code/no-code	<p>125 horas al mes de m4.xlarge para hospedaje durante los primeros dos meses. (Amazon Web Services, n.d.-a)</p> <p>Cumple. Tiene IDE para científicos de datos e interfaz no-code para analistas de negocios. (Amazon Web Services, n.d.-b)</p>	<p>Cumple. Automated Machine Learning (AutoML) permite crear modelos de Machine Learning sin código. (Microsoft Learn, 2024b)</p>	<p>Cumple. AutoML permite crear modelos con poco esfuerzo y experiencia en Machine learning. (Google Cloud, n.d.-b)</p>	<p>Cumple. Con AutoML se automatizan tareas manuales de construcción y entrenamiento de modelos. (IBM Cloud, n.d.-b)</p>
---------------------------	---	---	---	--

---

*Nota:* Comparación de plataformas Cloud para entrenamiento de modelos Machine Learning en su capa gratuita. Autoría propia.

## ***Actividad 2. Entrenar los modelos de Machine Learning con la base de datos depurada***

Se realizó el entrenamiento de los modelos mediante las herramientas seleccionadas que se describen a continuación.

### ***Amazon Web Services***

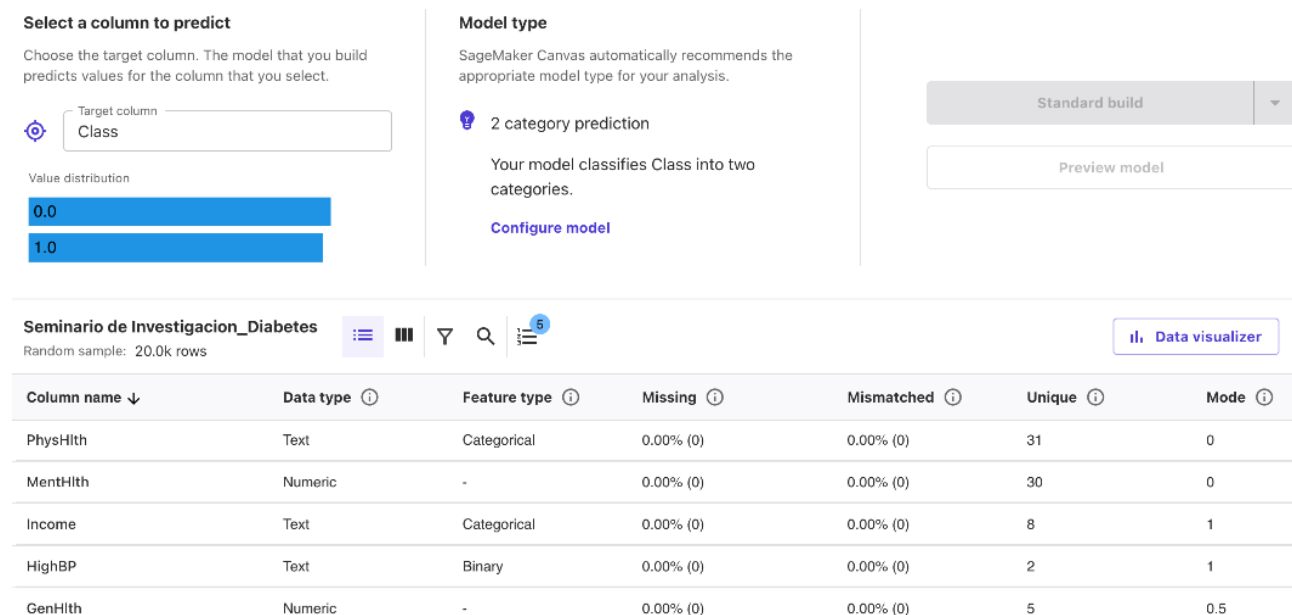
Se realizó el registro en Amazon Web Services (AWS) para acceder a sus servicios gratuitos. En el catálogo de servicios de AWS, se seleccionó Amazon SageMaker Autopilot, la opción de aprendizaje automático automatizado de AWS, que facilita la creación y ajuste de modelos de Machine Learning sin requerir una programación extensa.

Dentro de *SageMaker* Autopilot, se creó un experimento utilizando la base de datos previamente depurada. En la interfaz, se configuraron las opciones paso a paso, incluyendo la selección de la variable objetivo, la definición del tipo de cada variable de entrada, la elección de la métrica de evaluación principal, el establecimiento de límites de tiempo para el entrenamiento, la división en conjuntos de datos de validación y prueba, la selección del tipo de instancia para el procesamiento.

Cabe aclarar que, aunque *SageMaker Autopilot* ofrece algunas opciones gratuitas, el procesamiento puede generar costos si se usan funcionalidades avanzadas o instancias de mayor capacidad. En la Figura 3 se evidencia la interfaz de la herramienta en la cual previamente se ha realizado el cargue de los datos, aquí se muestran la clase predictora y algunas de las variables de la base de datos.

la nube en la predicción de la diabetes

**Figura 3.**  
Vistazo General Sagemaker estudio de AWS



**Select a column to predict**  
Choose the target column. The model that you build predicts values for the column that you select.

Target column  
Class

Value distribution

0.0	1.0
-----	-----

**Model type**  
SageMaker Canvas automatically recommends the appropriate model type for your analysis.

2 category prediction

Your model classifies Class into two categories.

[Configure model](#)

Standard build

[Preview model](#)

**Seminario de Investigacion\_Diabetes**  
Random sample: 20.0k rows

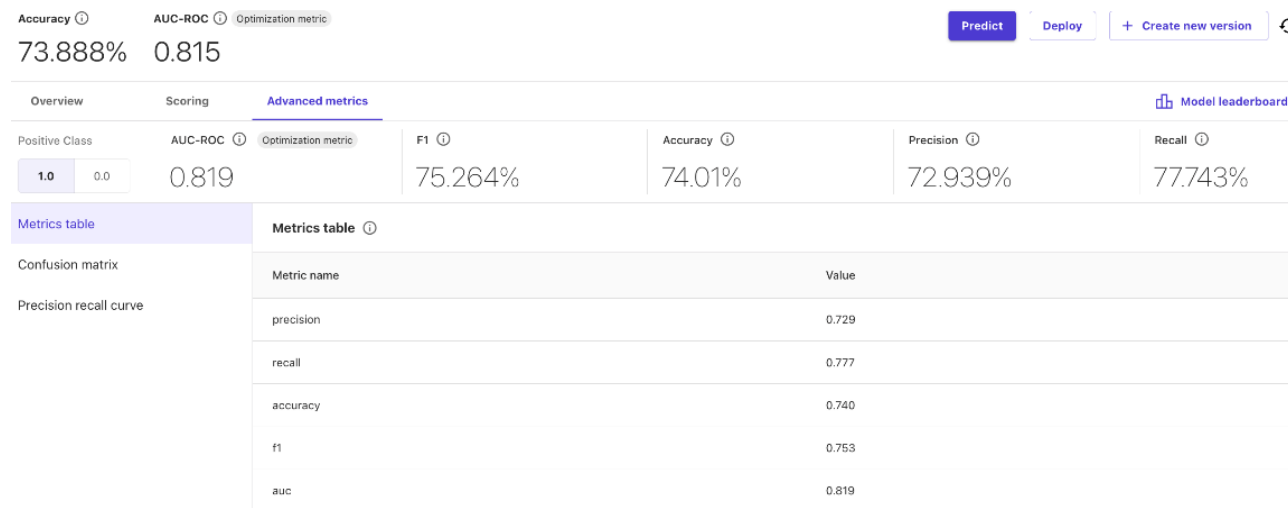
Column name ↓	Data type ⓘ	Feature type ⓘ	Missing ⓘ	Mismatched ⓘ	Unique ⓘ	Mode ⓘ
PhysHlth	Text	Categorical	0.00% (0)	0.00% (0)	31	0
MentHlth	Numeric	-	0.00% (0)	0.00% (0)	30	0
Income	Text	Categorical	0.00% (0)	0.00% (0)	8	1
HighBP	Text	Binary	0.00% (0)	0.00% (0)	2	1
GenHlth	Numeric	-	0.00% (0)	0.00% (0)	5	0.5

*Nota.* Figura que muestra la configuración de columnas y tipo de datos en AWS SageMaker para iniciar el entrenamiento del modelo. Tomado de la interfaz de AWS SageMaker.

En la consola de *Amazon SageMaker*, se seleccionó la opción de *Autopilot*, que permite la creación de modelos de Machine Learning de manera automatizada. SageMaker Autopilot guía al usuario a través de una serie de configuraciones clave para definir el experimento, incluyendo la variable objetivo, el tipo de problema y la métrica de evaluación.

Una vez configurado el experimento, se inició el proceso de entrenamiento. *SageMaker* exploró automáticamente diferentes algoritmos y configuraciones, ajustando hiperparámetros y seleccionando los modelos más adecuados de forma iterativa. Durante un período de aproximadamente 25 minutos. En la Figura 4 se evidencian los resultados obtenidos en el entrenamiento del modelo.

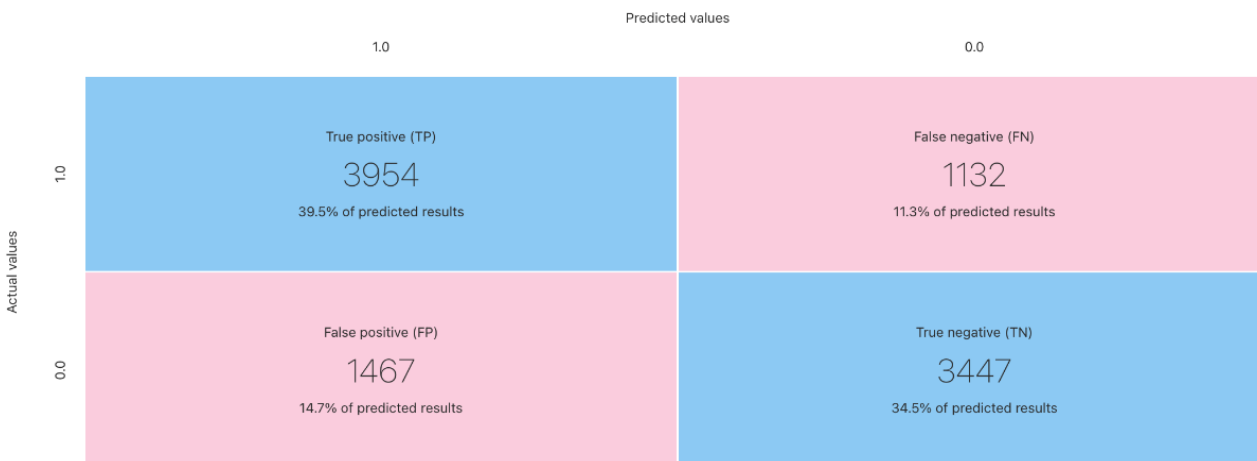
**Figura 4.**  
Resultados del modelo mediante Sagemaker estudio de AWS



*Nota.* Figura que muestra los resultados obtenidos con AWS SageMaker. Tomado de la interfaz de AWS SageMaker.

Al finalizar el entrenamiento, *Autopilot* ordenó los modelos generados según el valor de AUC ponderado. El modelo de mejor rendimiento fue el preprocesamiento automático de escalado, que alcanzó un AUC ponderado de 0.83. Este modelo fue seleccionado para la evaluación final y se consideró para el despliegue. En la Figura 5 se muestra la matriz de confusión obtenida.

**Figura 5.**  
Matriz de confusión con Sagemaker de AWS



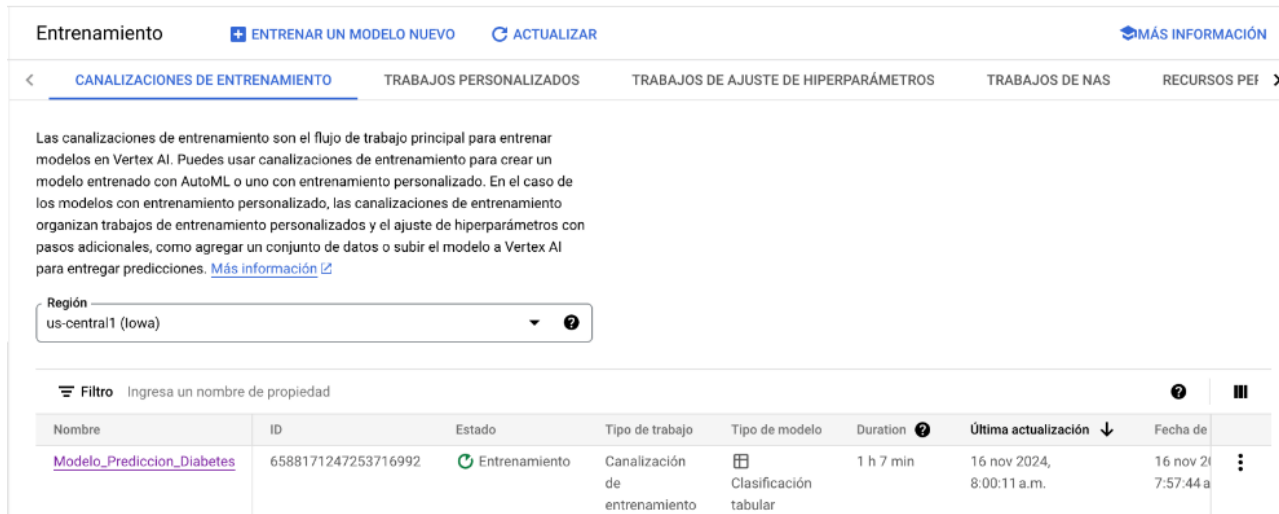
*Nota.* Figura que muestra los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos obtenidos con AWS SageMaker. Tomado de la interfaz de AWS SageMaker.

Los resultados obtenidos con AWS destacan su capacidad para generar modelos con un buen rendimiento general en la tarea de clasificación. El modelo entrenado en Amazon SageMaker alcanzó una exactitud de 0,7401 y un AUC de 0,819, Esto indica que AWS tiene un buen desempeño en la discriminación entre clases, lo que es fundamental en problemas relacionados con la salud.

### ***Google Cloud Platform***

El proceso de implementación inició con un flujo completo que incluyó la configuración del entorno, el entrenamiento de un modelo de clasificación mediante AutoML, y la evaluación detallada de los resultados obtenidos. En la Figura 6 se muestra la interfaz de la plataforma donde se seleccionó la región del servidor y la creación del modelo.

**Figura 6.**  
Configuración Google Cloud



*Nota.* Figura que muestra la configuración de GCP Vertex AI para iniciar el entrenamiento del modelo. Tomado de la interfaz de GCP Vertex AI.

El proyecto se inició habilitando los servicios necesarios en la plataforma, siguiendo las directrices oficiales de la documentación de Vertex AI . Las acciones específicas incluyeron:

- Creación del proyecto y configuración del entorno, en la que se estableció un proyecto en GCP con la región us-central1 y se habilitaron las APIs requeridas, como Vertex AI, Cloud Storage, y IAM.
- Gestión de permisos y roles, en la que se asignaron los roles de Administrador de Vertex AI y Visualizador de Almacenamiento al usuario encargado del entrenamiento. Esto aseguró acceso completo a las funcionalidades necesarias.
- Carga del dataset en Cloud Storage, en la que el dataset fue normalizado y escalado previamente, y posteriormente subido a un bucket de Cloud Storage para ser referenciado directamente desde Vertex AI.

la nube en la predicción de la diabetes

- Selección del método de entrenamiento, en la que se optó por AutoML Tabular, una herramienta de Vertex AI que automatiza la selección de algoritmos y el ajuste de hiperparámetros para optimizar el rendimiento del modelo.

El entrenamiento fue configurado para resolver un problema de clasificación binaria, donde la columna objetivo indica si un individuo presenta o no diabetes. Los datos fueron divididos automáticamente en tres conjuntos: 80% para entrenamiento, 10% para validación y 10% para prueba.

El objetivo de optimización seleccionado fue la Pérdida Logarítmica (Log Loss), debido a su relevancia para tareas de clasificación. El entrenamiento se llevó a cabo durante aproximadamente 2 horas y produjo un modelo con las métricas mostradas en la Figura 7.

**Figura 7.**  
Resultados del modelo en Google Cloud



AUC de PR	AUC de PROC	Puntuación F1 en 0.5	F1-Macro a 0.5	F1-Micro a 0.5	Pérdida logística	Precisión en 0.5	Recuperación en 0.5
0.813	0.816	0.739	0.739	0.739	0.523	73.26%	74.62%

*Nota.* Figura que muestra los resultados obtenidos con GCP Vertex AI. Tomado de la interfaz de GCP Vertex AI.

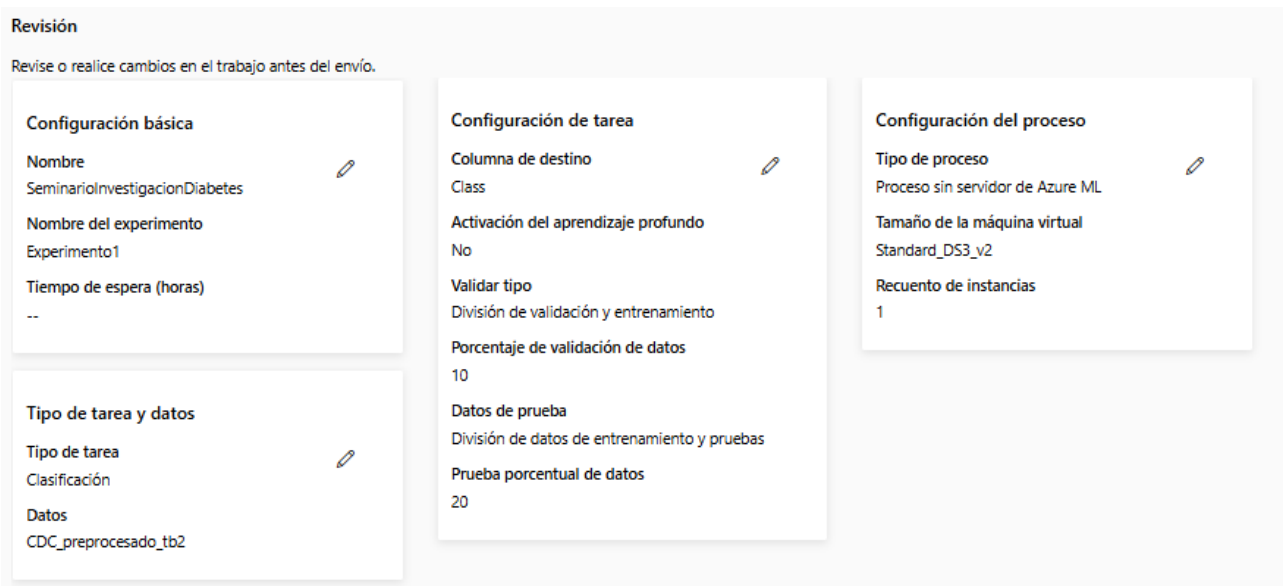
El AUC-ROC de 0.816 refleja una alta capacidad del modelo para distinguir entre individuos con y sin diabetes. Al ser superior a 0.8 es considerado excelente, indicando un desempeño robusto. El F1-Score de 0.739 confirma que el modelo mantiene un buen equilibrio entre precisión y recall. Sobre el Log Loss de 0.523, si bien el valor es aceptable, reducir esta métrica podría mejorar la capacidad del modelo para estimar probabilidades correctas. La precisión de 73.26% y recall de 74.62% muestran que el modelo identifica correctamente una proporción considerable de casos positivos, aunque no logra un nivel perfecto de predicción.

El modelo demostró un sólido desempeño general, alcanzando métricas destacadas como el AUC-ROC y el F1-Score, lo que resalta su capacidad para clasificar correctamente los datos. Además, el uso de AutoML simplificó de manera eficiente el ajuste de hiperparámetros, optimizando el proceso de entrenamiento para un dataset tabular. Por último, la integración con Vertex AI garantizó una escalabilidad sencilla, habilitando tanto predicciones por lotes como en tiempo real, lo que refuerza su aplicabilidad en distintos escenarios empresariales.

### *AutoML de Microsoft Azure*

Se hizo el registro en la plataforma de Microsoft Azure para la prueba gratuita de sus servicios. Dentro del catálogo de servicios de Azure Machine Learning se escogió el AutoML (Machine Learning automatizado), por ser la opción no-code de esta plataforma. Dentro de esta herramienta, se creó un trabajo de aprendizaje tomando el dataset depurado que se generó con anterioridad y asignando las configuraciones que solicita la interfaz en su paso a paso, como la selección de la variable objetivo, el tipo de dato de cada variable de entrada, la métrica primaria, opciones de aprendizaje avanzado, límites de entrenamiento, datos de validación y de prueba y características de la máquina virtual a utilizar (aunque el uso del servicio de AutoML es gratuito, hay costos por procesamiento). En la Figura 8 se evidencia la configuración del modelo de aprendizaje en la herramienta.

**Figura 8.**  
Resumen de la configuración de la tarea de aprendizaje de AutoML de Azure



*Nota.* Figura que muestra la configuración de Azure AutoML para iniciar el entrenamiento del modelo. Tomado de la interfaz de Azure AutoML.

Con esta configuración realizada, se inició el trabajo de aprendizaje. Este se llevó a cabo durante 1 hora y 2 minutos aproximadamente, en los que se crearon 58 modelos con diferentes técnicas y algoritmos y fueron ordenados finalmente de acuerdo con la métrica primaria escogida, que en este caso fue el valor ponderado de AUC. Esta métrica, promedia el AUC de positivos y negativos para diabetes, ponderando el número de instancias de cada clase. En la Figura 9 se evidencian los resultados obtenidos en cada uno de los modelos entrenados por *AutoML* de Azure.

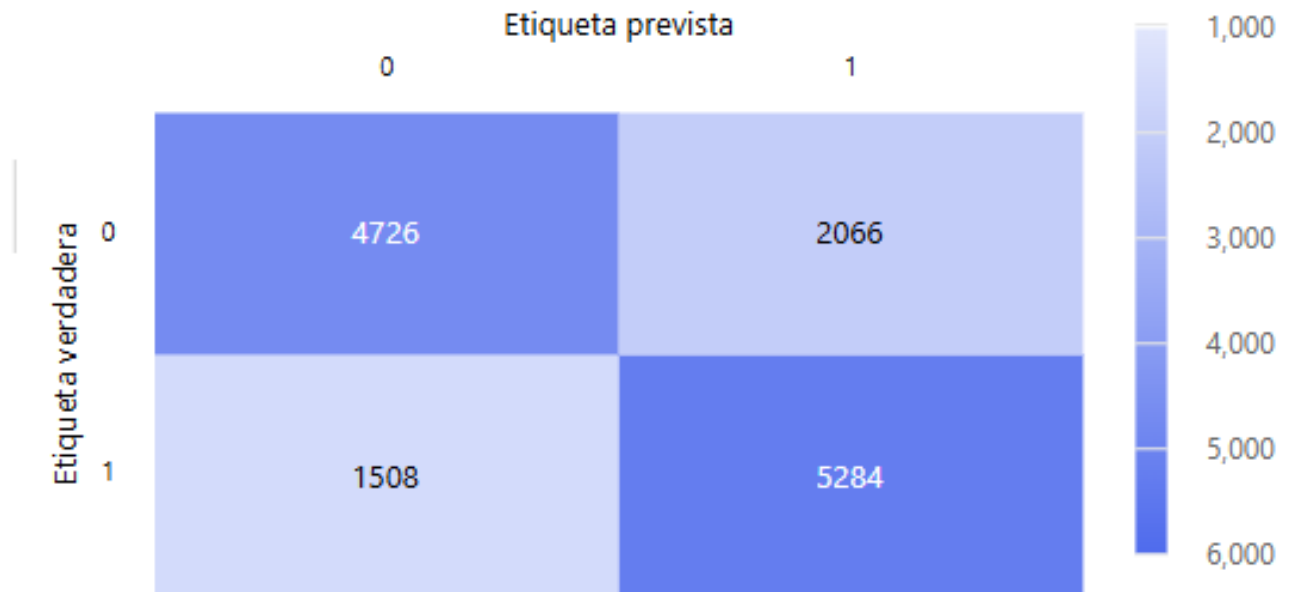
**Figura 9.**  
Resultados obtenidos con AutoML de Azure

Nombre del algoritmo	Explicado	IA responsable	Valor pondera... ↓	Muestreo	Fecha de creación	Duración
VotingEnsemble	<a href="#">Ver explicación</a>		0.81792	100.00 %	Nov 10, 2024 5:40 PM	58 s
StandardScalerWrapper, LightGBM			0.81655	100.00 %	Nov 10, 2024 5:17 PM	41 s
MaxAbsScaler, LightGBM			0.81626	100.00 %	Nov 10, 2024 4:46 PM	25 s
StandardScalerWrapper, XGBoostClassifier			0.81527	100.00 %	Nov 10, 2024 5:04 PM	42 s
MaxAbsScaler, LightGBM			0.81427	100.00 %	Nov 10, 2024 4:46 PM	24 s
StandardScalerWrapper, LightGBM			0.81415	100.00 %	Nov 10, 2024 5:36 PM	42 s
SparseNormalizer, XGBoostClassifier			0.81410	100.00 %	Nov 10, 2024 4:46 PM	23 s
StandardScalerWrapper, LightGBM			0.81405	100.00 %	Nov 10, 2024 5:19 PM	41 s
StandardScalerWrapper, XGBoostClassifier			0.81386	100.00 %	Nov 10, 2024 4:46 PM	19 s

*Nota.* Figura que muestra la lista de modelos generados por Azure AutoML y organizados de mayor a menor de acuerdo con su métrica de AUC ponderada.

El modelo con mejor desempeño fue *StandardScalerWrapper LightGBM*, con un valor ponderado de AUC de 0. 8101990. Este modelo es evaluado con el dataset de prueba que fue configurado con el 20% de registros totales, es decir 13584 registros. La matriz de confusión se evidencia en la Figura 10.

**Figura 10.**  
Matriz de confusión con AutoML de Azure



*Nota.* Figura que muestra los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos obtenidos con Azure AutoML. Tomado de la interfaz de Azure AutoML.

### Resultado 3. Comparación de los resultados de los modelos

#### *Actividad 1: Evaluación preliminar mediante tabulación de las métricas*

Para analizar los resultados de las métricas de aprendizaje entre AWS, GCP y Azure, se evaluó el desempeño de los modelos que fueron entrenados y probados con el mismo conjunto de datos. Esto permite identificar las fortalezas y las limitaciones de cada herramienta en aspectos de clasificación y predicción, como la capacidad de detección en casos positivos, el equilibrio entre errores y aciertos y la eficiencia en la diferenciación de clases. Estos parámetros se evidencian en la Tabla 4.

**Tabla 4.**  
Comparación de resultados

Resultados\Modelos	AWS	GCP	Azure
Verdaderos positivos	3954	2146	5284
Falsos positivos	1467	839	2066
Verdaderos negativos	3447	1888	4726
Falsos negativos	1132	613	1508
Exactitud	0,7401	0,7897	0,7369
Precisión	0,7293	0,7326	0,7189
Recall	0,7774	0,7462	0,7780
F1-score	0,7526	0,7390	0,7473
AUC	0,819	0,816	0,8102

*Nota:* Comparación de métricas obtenidas en las pruebas realizadas con las tres herramientas escogidas AWS SageMaker, GCP Vertex AI y Azure AutoML.

En métricas de Verdaderos positivos, Falsos positivos, Verdaderos negativos, Falsos negativos se evidencian diferencias en los modelos. Esto se presenta debido a las limitaciones que presenta cada una de las herramientas trabajadas, Azure no cuenta con limitaciones en el cargue de los datos, por lo que el entrenamiento se realizó con todos los registros, AWS tiene un límite de cargue de 50000 registros (Amazon SageMaker, n.d.), por lo que la herramienta realizó la selección de los datos a entrenar.

De acuerdo con la hipótesis N1: “Todas las herramientas de Machine Learning probadas en este estudio darán resultados satisfactorios. (F1-score mayor a 0.9)”. Ninguna de las herramientas alcanzó un F1-score mayor a 0.9. Los valores de F1-score obtenidos fueron 0.7526

(AWS), 0.7390 (GCP) y 0.7473 (Azure), indicando que, aunque son buenos resultados, no se consideran “satisfactorios” bajo el umbral establecido en esta hipótesis.

Con respecto a la hipótesis N2: ‘Las herramientas que realizan automáticamente la puesta a punto del modelo, sin dar opciones de personalización al usuario, darán el rendimiento más bajo’, se observa que GCP, conocida por su enfoque completamente automatizado, obtuvo los valores más bajos en métricas clave en comparación con AWS y Azure, que ofrecen un mayor grado de personalización en el entrenamiento y configuración de los modelos. Sin embargo, las diferencias entre los resultados no son significativas. Por lo tanto, no se puede confirmar plenamente la hipótesis planteada.

## Discusión

En este apartado se analizan los resultados obtenidos con las 3 herramientas, haciendo especial énfasis en la exactitud, el F1-score y el AUC, dado que tienen en cuenta de forma general la detección de casos positivos y negativos. Para algunos casos con un enfoque especial en detección de casos positivos puede ser más interesante el análisis de métricas como la precisión.

En la tabla 4 se evidencia que, pese a que la cantidad de datos de prueba fue diferente en cada herramienta por su configuración, las métricas obtenidas fueron similares. Esto muestra que los 3 modelos generados lograron resultados positivos y similares en la implementación.

Las métricas de F1-score sobre 0.73 y AUC sobre 0.8 indican que los tres modelos tienen una capacidad de clasificación correcta muy buena y, aunque no alcanzan el F1-score de 0.9 propuesto inicialmente, son un primer paso positivo que permite la identificación de la mayoría de los casos exitosamente.

Respecto al proceso de entrenamiento, Vertex AI demostró la potencia de Google Cloud Platform para tareas de *Machine Learning*, combinando herramientas avanzadas con facilidad de uso.

Por su parte, AutoML logró este resultado mediante la configuración de parámetros básicos que se establecieron mediante una interfaz amigable que lleva paso a paso al usuario y también ofrece algunas opciones de ajuste avanzado. Este proceso de aprendizaje utilizó recursos de procesamiento con un costo de 0.7 USD, pero al estar dentro de los 200 USD de crédito para usuarios nuevos no genera cobro, esto permite que sea una herramienta muy buena para iniciar la

implementación de modelos de *Machine Learning* en entidades relacionadas a la prevención, detección y tratamiento de la diabetes.

Finalmente, AWS SageMaker permitió llevar a cabo el proceso de entrenamiento con pocos pasos y una interfaz de fácil interacción, sin muchas opciones de personalización, pero con la sencillez ideal para usuarios sin conocimientos técnicos avanzados.

Para la recomendación de una herramienta y dado que las métricas obtenidas en los tres modelos son similares, se tuvieron en cuenta otros factores como las dificultades presentadas en el proceso de entrenamiento de cada modelo, las limitaciones y la capacidad de personalización de los modelos expuestos anteriormente. Como resultado de esta discusión se selecciona Azure AutoML como la mejor herramienta probada en este estudio, aunque se debe tener en cuenta que para esta selección se utilizaron algunos factores de evaluación subjetiva y que para implementaciones de este tipo cualquiera de las 3 herramientas dará buenos resultados.

## Conclusiones

El proceso de elección de una base de datos es un elemento crucial en la ejecución de proyectos asociados a la salud, por lo que es importante tener presente las variables relevantes en estudios previos para obtener mejores métricas de rendimiento. Esto permitió limitar la selección de la base de datos mejorando su calidad y facilidad de entrenamiento, permitiendo que esta cumpla con las características necesarias para el entrenamiento de los modelos.

En el entrenamiento de los modelos de *Machine Learning* en los servicios gratuitos de la nube se evidenciaron dificultades en el preprocesamiento de los datos en la herramienta de GCP Vertex AI, siendo esta una herramienta no muy intuitiva que requirió más tiempo en el procesamiento de los datos, a diferencia de las demás herramientas utilizadas. Por otra parte, Azure AutoML permitió opciones de personalización lo que implicó un mayor conocimiento del procesamiento utilizado en los modelos y finalmente AWS generó limitaciones en la cantidad de datos, por lo que no se recomienda para bases de datos con grandes volúmenes de información.

Los resultados obtenidos con las tres herramientas de *Machine Learning* demuestran un desempeño consistente y satisfactorio, con métricas clave como el AUC y el F1-score superiores a 0.7, lo que refleja la capacidad de todas las plataformas para generar modelos predictivos efectivos en la detección de diabetes, aunque se observen diferencias sutiles en el rendimiento. Estos resultados permitieron definir que la elección de la herramienta depende de las necesidades específicas del usuario, como la personalización del modelo, el presupuesto y la facilidad de implementación.

En general, las herramientas no-code lograron resultados satisfactorios y al no necesitar conocimientos avanzados ni un gran presupuesto, son un excelente primer paso en la

implementación de modelos de clasificación con *Machine Learning*. Esto es aplicable a todo tipo de campos, en los que la incursión en estas tecnologías puede llevar a la optimización de procesos y la obtención de mejores resultados.

### **Recomendaciones**

Para futuros estudios se recomienda que se evalúen diferentes métricas de reducción de dimensionalidad con el fin de obtener una base de datos optimizada para evitar sesgos en los modelos de entrenamiento. De igual forma, podría explorar el rendimiento de las herramientas con la base de datos sin procesar para verificar las métricas obtenidas. También, se recomienda explorar otras herramientas, e incluso retomar las que se probaron en este estudio con algunas opciones de su capa de pago. Asimismo, explorar las opciones de las plataformas con herramientas que permitan codificación, y de esta manera tener más control sobre el procesamiento y entrenamiento de los modelos para realizar ajuste de hiperparámetros y obtener modelos más precisos.

Por ultima recomendación, se recomienda evaluar la posibilidad de aplicar los modelos obtenidos en instituciones de salud mediante un software que sea accesible a diferentes profesionales de la salud para ayudar en la predicción de la diabetes con el fin que los pacientes en ayuda de los profesionales de la salud puedan tomar medidas preventivas para minimizar el riesgo de tener la enfermedad.

### Referencias

- Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., & Paul, B. K. (2021). Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering*, 2. <https://doi.org/10.1016/j.ijcce.2021.12.001>
- Alam, M., Khan, I. R., Alam, M. A., Siddiqui, F., & Tanweer, S. (2023). The diabacare cloud: predicting diabetes using machine learning. *Acta Scientiarum. Technology*, 46(1), e64783. <https://doi.org/10.4025/actascitechnol.v46i1.64783>
- Alarcón Paredes, A., Alegre Ventura, R. J., & Alonso Silverio, G. A. (2024). Mejora en la clasificación de datos con desbalance de clases mediante una redistribución de clases por k-means. *XIKUA Boletín Científico de La Escuela Superior de Tlahuelilpan*, 12(Especial), 111–116. <https://doi.org/10.29057/xikua.v12iespecial.12768>
- Amazon SageMaker. (n.d.). *How custom models work - Amazon SageMaker*. Retrieved November 16, 2024, from <https://docs.aws.amazon.com/sagemaker/latest/dg/canvas-build-model.html>
- Amazon Web Service. (2023). *Machine Learning: Amazon Web Services*. <https://aws.amazon.com/es/sagemaker/>
- Amazon Web Services. (n.d.-a). *Amazon Sagemaker*. <https://aws.amazon.com/>. Retrieved November 11, 2024, from <https://aws.amazon.com/es/sagemaker/?p=ft&c=ml&c=3>
- Amazon Web Services. (n.d.-b). *Servicio de machine learning - Amazon SageMaker*. Retrieved November 11, 2024, from

[https://aws.amazon.com/es/pm/sagemaker/?gclid=EAIaIQobChMItOKIz7vAiQMVzEL\\_AR3BMgSNEAAYASAAEgKnh\\_D\\_BwE&trk=3bdce3a9-f8fc-491e-aef1-3614170a3346&sc\\_channel=ps&ef\\_id=EAIaIQobChMItOKIz7vAiQMVzEL\\_AR3BMgSNEAAYASAAEgKnh\\_D\\_BwE:G:s&s\\_kwcid=AL!4422!3!647999769051!p!!g!!aws%20sagemaker!19677234618!147473973793](https://aws.amazon.com/es/pm/sagemaker/?gclid=EAIaIQobChMItOKIz7vAiQMVzEL_AR3BMgSNEAAYASAAEgKnh_D_BwE&trk=3bdce3a9-f8fc-491e-aef1-3614170a3346&sc_channel=ps&ef_id=EAIaIQobChMItOKIz7vAiQMVzEL_AR3BMgSNEAAYASAAEgKnh_D_BwE:G:s&s_kwcid=AL!4422!3!647999769051!p!!g!!aws%20sagemaker!19677234618!147473973793)

Amazon Web Services. (n.d.-c). *Servicios gratuitos de machine learning en AWS*.

[https://aws.amazon.com/es/free/machine-learning/?gclid=EAIaIQobChMIs6yt27jAiQMVNEP\\_AR1\\_Xgc9EAAYASAAEgJxSPD\\_BwE&trk=a94d9805-fe48-4f0a-b812-0974ceb5f7fe&sc\\_channel=ps&ef\\_id=EAIaIQobChMIs6yt27jAiQMVNEP\\_AR1\\_Xgc9EAAYASAAEgJxSPD\\_BwE:G:s&s\\_kwcid=AL!4422!3!648041757757!e!!g!!amazon%20web%20services%20machine%20learning!19677234450!146461591096](https://aws.amazon.com/es/free/machine-learning/?gclid=EAIaIQobChMIs6yt27jAiQMVNEP_AR1_Xgc9EAAYASAAEgJxSPD_BwE&trk=a94d9805-fe48-4f0a-b812-0974ceb5f7fe&sc_channel=ps&ef_id=EAIaIQobChMIs6yt27jAiQMVNEP_AR1_Xgc9EAAYASAAEgJxSPD_BwE:G:s&s_kwcid=AL!4422!3!648041757757!e!!g!!amazon%20web%20services%20machine%20learning!19677234450!146461591096)

American Diabetes Association. (2024, September 12). *Diabetes Diagnóstico General*.

<https://Diabetes.Org/Espanol/Diagnostico>.

Aouragh, A. A., Bahaj, M., & Toufik, F. (2024). Diabetes Prediction: Optimization of Machine Learning through Feature Selection and Dimensionality Reduction. *International Journal of Online and Biomedical Engineering*, 20(8), 100–114.

<https://doi.org/10.3991/IJOE.V20I08.47765>

Asociación Latinoamericana de Diabetes. (2019). Guías ALAD sobre el Diagnóstico, Control y Tratamiento de la Diabetes Mellitus Tipo 2 con Medicina Basada en Evidencia. *ALAD*.

[www.revistaalad.com](http://www.revistaalad.com)

AWS Team. (2024). *sagemaker-dg-2*.

Azure. (n.d.). *Explore free Azure services*. <https://Azure.Microsoft.Com/>. Retrieved November 11, 2024, from <https://azure.microsoft.com/en-us/pricing/free-services>

Blanco, J. (2023, April 28). “¿Por qué la normalización es clave e importante en Machine Learning y Ciencia de Datos? | by Jorge I. Blanco | Medium.

<https://jorgeiblanco.medium.com/por-qu%C3%A9-la-normalizaci%C3%B3n-es-clave-e-importante-en-machine-learning-y-ciencia-de-datos-4595f15d5be0>

Boutilier, J. J., Chan, T. C. Y., Ranjan, M., & Deo, S. (2021). Risk Stratification for Early Detection of Diabetes and Hypertension in Resource-Limited Settings: Machine Learning Analysis. *Journal of Medical Internet Research*, 23(1). <https://doi.org/10.2196/20123>

Clínica Mayo. (2024, March 27). *Diabetes*. <https://Www.Mayoclinic.Org/Es/Diseases-Conditions/Diabetes/Symptoms-Causes/Syc-20371444>.

Czmlil, A., Czmlil, S., & Mazur, D. (2019). A method to detect type 1 diabetes based on physical activity measurements using a mobile device. *Applied Sciences (Switzerland)*, 9(12). <https://doi.org/10.3390/app9122555>

David Sontag, & Pete Szolovits. (2020a). Machine learning for healthcare: What makes healthcare unique? . In [https://ocw.mit.edu/courses/6-s897-machine-learning-for-healthcare-spring-2019/2020812c6b3be9e2c27873595d1a9aed\\_MIT6\\_S897S19\\_lec1.pdf](https://ocw.mit.edu/courses/6-s897-machine-learning-for-healthcare-spring-2019/2020812c6b3be9e2c27873595d1a9aed_MIT6_S897S19_lec1.pdf). MIT.

David Sontag, & Pete Szolovits. (2020b). Machine learning for healthcare: What makes healthcare unique? . In [https://ocw.mit.edu/courses/6-s897-machine-learning-for-healthcare-spring-2019/2020812c6b3be9e2c27873595d1a9aed\\_MIT6\\_S897S19\\_lec1.pdf](https://ocw.mit.edu/courses/6-s897-machine-learning-for-healthcare-spring-2019/2020812c6b3be9e2c27873595d1a9aed_MIT6_S897S19_lec1.pdf).

*healthcare-spring-2019/2020812c6b3be9e2c27873595d1a9aed\_MIT6\_S897S19\_lec1.pdf.*

MIT.

Felix Richter. (2024, November 1). *Amazon Maintains Cloud Lead as Microsoft Edges Closer.*

<https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/>

Google Cloud. (n.d.-a). *AI and machine learning products.* <https://cloud.google.com/>.

Retrieved November 11, 2024, from <https://cloud.google.com/products/ai?hl=en>

Google Cloud. (n.d.-b). *Datos tabulares de Hello: crea un conjunto de datos y entrena un*

*modelo de clasificación de AutoML .* <https://cloud.google.com/>.

Google Cloud. (n.d.-c). *Get free hands-on experience with Google Cloud.*

<https://cloud.google.com/>. Retrieved November 11, 2024, from

<https://cloud.google.com/free?hl=en>

Google Cloud. (2024a). *AI and machine learning products.*

[https://cloud.google.com/products/ai/?utm\\_source=google&utm\\_medium=...vrlqhqRQi7UoZt0j](https://cloud.google.com/products/ai/?utm_source=google&utm_medium=...vrlqhqRQi7UoZt0j)

[5yOI7TEppstw0D8aAl-9EALw\\_wcB&gclid=aw.ds&hl=en](https://cloud.google.com/products/ai/?utm_source=google&utm_medium=...vrlqhqRQi7UoZt0j5yOI7TEppstw0D8aAl-9EALw_wcB&gclid=aw.ds&hl=en)

Google Cloud. (2024b). *¿Qué es la computación en la nube\_ \_ Google Cloud.*

<https://cloud.google.com/learn/what-is-cloud-computing?hl=es-419>

Hanley, J. A., & McNeil, B. J. (1982). *The Meaning and Use of the Area under a Receiver*

*Operating Characteristic (ROC) Curve1.*

Hoyos, W., Hoyos, K., & Ruiz-Pérez, R. (2023). Modelo de inteligencia artificial para la detección temprana de diabetes. *Biomédica*, 43(Suppl 3), 110.

<https://doi.org/10.7705/BIOMEDICA.7147>

Huddle, M., Kellar, J., Srikumar, K., Deepak, K., & Martines, D. (2023). *Generative AI Will Transform Health Care Sooner Than You Think*.

IBM Cloud. (n.d.-a). *Catálogo de IBM Cloud*. <https://Cloud.Ibm.Com/>. Retrieved November 11, 2024, from

[https://cloud.ibm.com/catalog?category=ai&search=label%3Aibm\\_created#services](https://cloud.ibm.com/catalog?category=ai&search=label%3Aibm_created#services)

IBM Cloud. (n.d.-b). *IBM AutoAI*. <https://Www.Ibm.Com/>. Retrieved November 11, 2024, from <https://www.ibm.com/products/watson-studio/autoai>

IBM Cloud. (n.d.-c). *IBM Cloud free tier*. <https://Www.Ibm.Com/>. Retrieved November 11, 2024, from <https://www.ibm.com/cloud/free>

INEAF Business School. (n.d.). *Glosario Jurídico INEAF Business School*. Glosario Jurídico INEAF Business School. Retrieved October 11, 2024, from <https://www.ineaf.es/glosario-juridico/cuota-de-mercado>

Jacobs, P. G., Herrero, P., Facchinetti, A., Vehi, J., Kovatchev, B., Breton, M. D., Cinar, A., Nikita, K. S., Doyle, F. J., Bondia, J., Battelino, T., Castle, J. R., Zarkogianni, K., Narayan, R., & Mosquera-Lopez, C. (2024). Artificial Intelligence and Machine Learning for Improving Glycemic Control in Diabetes: Best Practices, Pitfalls, and Opportunities. *IEEE Reviews in Biomedical Engineering*, 17. <https://doi.org/10.1109/RBME.2023.3331297>

Kabir, M. F., Chen, T., & Ludwig, S. A. (2023). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*, 3, 100125. <https://doi.org/10.1016/J.HEALTH.2022.100125>

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4). <https://doi.org/10.1016/j.ict.2021.02.004>

Manzini, E., Garrido-Aguirre, J., & Perera-Lluna, A. (2022). *El papel de la inteligencia artificial en el tratamiento de la diabetes: promesas y realidad*. <https://www.revistadiabetes.org/tecnologia/el-papel-de-la-inteligencia-artificial-en-el-tratamiento-de-la-diabetes-promesas-y-realidad/>

Microsoft. (2023). *Azure Machine Learning - ML as a Service | Microsoft Azure*. <https://azure.microsoft.com/en-us/products/machine-learning/>

Microsoft Azure ML Team. (2024, September 11). *Azure AI Studio documentation*. <https://learn.microsoft.com/en-us/azure/ai-studio/>

Microsoft Learn. (2024a, April 15). *Compare Microsoft machine learning products and technologies*. <https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/data-science-and-machine-learning>

Microsoft Learn. (2024b, August 28). *What is automated machine learning (AutoML)?* <https://learn.microsoft.com/en-us/azure/machine-learning/concept-automated-ml?view=azureml-api-2>

Mishra, S., Mallick, P. K., Tripathy, H. K., Bhoi, A. K., & González-Briones, A. (2020). Performance Evaluation of a Proposed Machine Learning Model for Chronic Disease

Datasets Using an Integrated Attribute Evaluator and an Improved Decision Tree Classifier.

*Applied Sciences* 2020, Vol. 10, Page 8137, 10(22), 8137.

<https://doi.org/10.3390/APP10228137>

Müller, A. C., & Guido, S. (n.d.). *Introduction to Machine Learning with Python A GUIDE FOR DATA SCIENTISTS Introduction to Machine Learning with Python.*

Murel, J. (2024a, April 29). *¿Qué es el sobremuestreo? | IBM.* <https://www.ibm.com/es-es/topics/upsampling>

Murel, J. (2024b, June 15). *¿Qué es el submuestreo? | IBM.* <https://www.ibm.com/mx-es/topics/downsampling>

Nadis, S. (2023). *How machine-learning models can amplify inequities in medical diagnosis and treatment MIT researchers investigate the causes of health care disparities among underrepresented groups.* <https://news.mit.edu/2023/how-machine-learning-models-can-amplify-inequities-medical-diagnosis-treatment-0817>

Nagpal, A., Sabharwal, M., & Tripathi, R. (2024). A novel ensemble machine learning framework for early stage diabetes mellitus prediction. *Multidisciplinary Science Journal*, 6(3), 2024031–2024031. <https://doi.org/10.31893/MULTISCIENCE.2024031>

Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., Dalton, B. E., Duprey, J., Cruz, J. A., Hagins, H., Lindstedt, P. A., Aali, A., Abate, Y. H., Abate, M. D., Abbasian, M., Abbasi-Kangevari, Z., Abbasi-Kangevari, M., Abd ElHafeez, S., Abd-Rabu, R., ... Vos, T. (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of

Disease Study 2021. *The Lancet*, 402(10397), 203–234. [https://doi.org/10.1016/S0140-6736\(23\)01301-6](https://doi.org/10.1016/S0140-6736(23)01301-6)

Organización de los Estados Americanos Departamento de Derecho Internacional Secretaría de Asuntos Jurídicos. (2022). *Principios Actualizados sobre la Privacidad y la Protección de Datos Personales* (OEA, Ed.).

Organización Mundial de la Salud. (2024). *Diabetes*. <https://www.who.int/news-room/fact-sheets/detail/diabetes>

Organización Panamericana de la Salud. (2022a, November 14). *Día Mundial de la Diabetes 2022*. Organización Panamericana de La Salud. <https://www.paho.org/es/campanas/dia-mundial-diabetes-2022#:~:text=El%20n%C3%BAmero%20de%20personas%20que,la%20Regi%C3%B3n%20de%20las%20Am%C3%A9ricas>.

Organización Panamericana de la Salud. (2022b, November 14). *Día Mundial de la Diabetes 2022*. Organización Panamericana de La Salud. <https://www.paho.org/es/campanas/dia-mundial-diabetes-2022#:~:text=El%20n%C3%BAmero%20de%20personas%20que,la%20Regi%C3%B3n%20de%20las%20Am%C3%A9ricas>.

Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., Yu, J., Li, C., Yu, F., & Ren, Z. (2022). Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. *International Journal of Environmental Research and Public Health*, 19(22). <https://doi.org/10.3390/ijerph192215027>

Rani, Vv., & Kiran Kumar, K. (n.d.). *SIGNIFICANCE OF MULTILAYER PERCEPTRON MODEL FOR EARLY DETECTION OF DIABETES OVER ML METHODS.*

Ranjit Reddy, M., Lakshmi Sagar, P., & Shaik, N. S. (2022). Diabetes Mellitus Detection and Self Management based on Machine Learning. *Journal of Pharmaceutical Negative Results*, 13(4). <https://doi.org/10.47750/pnr.2022.13.04.138>

Salam, M. A., Azar, A. T., Elgendy, M. S., & Fouad, K. M. (2021). The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem. *International Journal of Advanced Computer Science and Applications*, 12(4), 641–655. <https://doi.org/10.14569/IJACSA.2021.0120480>

SAP Colombia, S. A. S. (n.d.). *¿Qué es el desarrollo de aplicaciones low-code/no-code?* <https://www.sap.com/Latinamerica/Products/Technology-Platform/Build/What-Is-Low-Code-No-Code.html>.

Shahin Alam, M., Jannatul Ferdous, M., & Sarkar Neera, N. (2024). Enhancing Diabetes Prediction: An Improved Boosting Algorithm for Diabetes Prediction. *IJACSA) International Journal of Advanced Computer Science and Applications*, 15(5). [www.ijacsa.thesai.org](http://www.ijacsa.thesai.org)

Shaukat, Z., Zafar, W., Ahmad, W., Haq, I. U., Husnain, G., Al-Adhaileh, M. H., Ghadi, Y. Y., & Algarni, A. (2023). Revolutionizing Diabetes Diagnosis: Machine Learning Techniques Unleashed. *Healthcare (Switzerland)*, 11(21). <https://doi.org/10.3390/healthcare11212864>

Smith, J. W., Everhart, J. E., Dicksont, W. C., Knowler, W. C., & Johannes, R. S. (n.d.). *Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus.*

Stawarz, K., Katz, D., Ayobi, A., Marshall, P., Yamagata, T., Santos-Rodriguez, R., Flach, P., & O’Kane, A. A. (2023). Co-designing opportunities for Human-Centred Machine Learning in supporting Type 1 diabetes decision-making. *International Journal of Human Computer Studies*, 173. <https://doi.org/10.1016/j.ijhcs.2023.103003>

Team, L., & Meta, A. @. (2024). *The Llama 3 Herd of Models*.

Thotad, P. N., Bharamagoudar, G. R., & Anami, B. S. (2023). Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 17(1), 102690. <https://doi.org/10.1016/J.DSX.2022.102690>

Treveil, M., & Dataiku Team, the. (n.d.). *Mastering MLops with Dataiku*. [www.dataiku.com](http://www.dataiku.com)

Vidhya, K., & Shanmugalakshmi, R. (2020). Deep learning based big medical data analytic model for diabetes complication prediction. *Journal of Ambient Intelligence and Humanized Computing*, 11(11), 5691–5702. <https://doi.org/10.1007/s12652-020-01930-2>

Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., & Juwono, F. H. (2024). Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools and Applications*, 83(8), 24153–24185. <https://doi.org/10.1007/s11042-023-16407-5>

Wei-Meng Lee. (2019). *Python Machine Learning* (1st ed.). John Wiley & Sons, Inc.

Xu, J. (2024). The Current Status and Promotional Strategies for Cloud Migration of Hospital Information Systems in China: Strengths, Weaknesses, Opportunities, and Threats Analysis. *JMIR Medical Informatics*, 12. <https://doi.org/10.2196/52080>

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics, 9*.  
<https://doi.org/10.3389/fgene.2018.00515>