TRABAJO DE GRADO
Opción Seminario-Diplomado.

ESTRATEGIA COMPUTACIONAL PARA ESTIMAR LA POSIBILIDAD DE UNA PERSONA TENER DIABETES A PARTIR DE DATOS DE HISTORIALES MEDICOS, UTILIZANDO ALGORITMOS DE MACHINE LEARNING

Corporación Universitaria Remington.
Nombre de la facultad: Ingenierías
Nombre del programa académico: Ingeniería Industrial

Estudiantes:
Jose Alejandro Camargo Rios.
Luis Alfonso Oviedo Perez.
Lizeth Valentina Trujillo Carrillo

Tutor: Juan Carlos Briñez de León

Opción de Trabajo de grado Seminario-Diplomado.
2025.

Dedicatoria

A nuestros padres, por ser nuestro pilar más firme en cada etapa de este camino.

Por su amor incondicional, su paciencia infinita y su constante apoyo,

incluso en los momentos más difíciles.

Este logro también es de ustedes.

Gracias por creer en nosotros cuando aún dudábamos de nosotros mismos.

Con todo nuestro cariño y gratitud.

Agradecimientos

Agradecemos a Dios por habernos acompañado a lo largo de este proceso,

por darnos la claridad, la fuerza y la paciencia en cada paso.

Su presencia nos dio serenidad en los momentos difíciles

y esperanza para seguir adelante.

Este logro es también fruto de Su guía y de las puertas que nos ha permitido abrir.

Tabla de Contenidos

Contenido

Figuras

Tablas

Resumen

El presente trabajo de grado tiene como propósito desarrollar una estrategia computacional apoyada en algoritmos de Machine Learning, con el fin de estimar la posibilidad de que una persona tenga un diagnostico de diabetes, a partir del análisis de variables clínicas y antecedentes médicos. Para ello, se utilizará un conjunto de datos reales compuesto por 100.000 registros de historiales médicos, que incluyen información como la edad, género, presencia de hipertensión, problemas cardíacos, historial de tabaquismo, índice de masa corporal, niveles de hemoglobina glicosilada (HbA1c) y glucosa en sangre, así como el diagnóstico correspondiente. La metodología del estudio inicia con un análisis exploratorio de los datos mediante herramientas estadísticas y visuales, como tablas de frecuencia, diagramas de caja y bigotes, gráficos circulares y mapas de correlación. Esta etapa permitirá identificar tendencias, relaciones entre variables y posibles valores atípicos, con el objetivo de preparar adecuadamente los datos para su uso en modelos predictivos.

Posteriormente, se implementarán diferentes modelos de clasificación, con el fin de comparar su precisión, sensibilidad y especificidad. El modelo o los modelos con mejor desempeño será seleccionado como base para el desarrollo de una herramienta computacional que permita ingresar nuevos datos individuales y obtener una predicción binaria: “diabético” o “no diabético”.

Palabras clave

Machine learning, predicción de diabetes, análisis de datos medicos, regresión, detección.

1. Marco conceptual y contextual

1.1 Contexto

En la actualidad, el análisis de datos médicos mediante técnicas de Machine Learning (ML) se ha convertido en una estrategia clave para mejorar los procesos de salud pública y prevención de enfermedades. Estos métodos permiten analizar grandes volúmenes de registros clínicos para desarrollar modelos predictivos que identifican patrones y riesgos asociados a enfermedades crónicas, como la diabetes. Según una revisión sistemática, los modelos de clasificación en ML alcanzan un AUC promedio de 0,81 al aplicarse en entornos comunitarios, lo que demuestra su solidez para la predicción de diabetes en poblaciones reales (Zhou et al., 2021).

Las variables clínicas tradicionales, como la edad, el índice de masa corporal (IMC), los niveles de glucosa y hemoglobina A1c, la hipertensión, las enfermedades cardíacas y los hábitos como el tabaquismo, han sido ampliamente utilizadas en estudios de clasificación. Por ejemplo, un estudio basado en el conjunto de datos *Pima Indians* mostró que algoritmos de clasificación como Random Forest, SVM y K-Nearest Neighbors superaron otros modelos, logrando AUCs de hasta 0,95 en validación cruzada (Liang et al., 2022).

Un estudio comparativo reciente que aplicó modelos de clasificación como SVM, Random Forest, redes neuronales y árboles de decisión en el mismo conjunto de datos encontró que las redes neuronales alcanzaron una precisión del 78,6 %, seguidas por Random Forest con un 76,3 % (Alzboon et al., 2023). Asimismo, investigaciones en entornos clínicos han demostrado que, para conjuntos de datos grandes (más de 70.000 registros), técnicas de

clasificación avanzada como Adaboost, LightGBM ofrecen mejores resultados, aunque requieren mayor capacidad computacional (Choi et al., 2022).

El presente estudio propone aplicar esta base de evidencia para desarrollar un modelo predictivo robusto, utilizando una base real de 100.000 historiales clínicos con variables clave. El enfoque contempla el preprocesamiento de los datos, selección de variables significativas, entrenamiento con algoritmos de clasificación y comparación de su precisión. La validación se realizará mediante métricas robustas (precisión, sensibilidad, especificidad), con el fin de garantizar la aplicabilidad clínica y confiabilidad del modelo en distintas poblaciones. Este enfoque se encuentra plenamente sustentado en evidencia científica y se alinea con la tendencia hacia sistemas de detección de la diabetes basados en datos, contribuyendo a optimizar recursos médicos y mejorar los resultados en salud pública, especialmente en contextos con limitado acceso a laboratorios o diagnóstico oportuno.

1.1.2 Algoritmos de Machine learning en sistemas de predicción de salud.

Primero que todo vamos a identificar qué es el machine learning y la aplicación que se puede presentar en los sistemas de predicción en la salud

¿Qué es el machine learning y cuál es su aplicación en los sistemas de salud?

El Machine Learning (aprendizaje automático) es una rama de la inteligencia artificial que permite a los sistemas aprender de los datos y realizar predicciones o tomar decisiones sin estar programados de forma explícita. A través del análisis estadístico y computacional,

estos algoritmos identifican patrones complejos y relaciones entre variables, lo cual los convierte en herramientas poderosas para resolver problemas del mundo real.

En el campo de la salud, el Machine Learning se ha posicionado como una tecnología innovadora que permite anticipar enfermedades, optimizar diagnósticos, personalizar tratamientos y mejorar los procesos de atención médica. Uno de los usos más importantes es la predicción de enfermedades crónicas, como la diabetes, mediante el análisis de grandes volúmenes de datos clínicos. Variables como edad, antecedentes médicos, niveles de glucosa, antecedentes de tabaquismo, entre otros, son procesadas por modelos computacionales que identifican el riesgo de que un paciente desarrolle una determinada condición. Los sistemas de predicción basados en Machine Learning pueden integrarse en aplicaciones clínicas o comunitarias, permitiendo una intervención oportuna, especialmente en entornos con recursos limitados. Además, estos modelos tienen la capacidad de adaptarse a nuevos datos, lo que los hace escalables y útiles para escenarios cambiantes. Esta combinación de análisis automatizado, precisión diagnóstica y capacidad de actualización constante convierte al Machine Learning en una herramienta clave para fortalecer los sistemas de salud pública y la medicina preventiva.

Algoritmos de Machine Learning en sistemas de predicción de salud

En los sistemas predictivos de salud, se utilizan diversos algoritmos de Machine Learning que ofrecen distintos niveles de precisión e interpretabilidad. Uno de los más empleados es la clasificación, valorada por su simplicidad y claridad, y ampliamente usada como modelo base en investigación clínica.

Los árboles de decisión y Random Forest son modelos más sofisticados que permiten capturar relaciones no lineales y manejar variables heterogéneas. Un estudio basado en datos sanguíneos de rutina mostró que Random Forest alcanzó una precisión del 79,2 %, mientras que XGBoost logró hasta un 80,5 % (Budi et al., 2024). En ese mismo análisis, XGBoost demostró un rendimiento superior (AUC ≈ 0.99 en entrenamiento y ≈ 0.985 en pruebas), resaltando su capacidad para modelar interacciones complejas entre variables.

El algoritmo Support Vector Machine (SVM) también ha mostrado un desempeño destacado en escenarios clínicos, alcanzando una sensibilidad del 88,9 % y un AUC de 0.855 en muestras comunitarias con validación cruzada (Choi et al., 2022). A su vez, estudios comparativos recientes que incluyen clasificación, SVM y reportan que este último ofrece la mejor generalización y estabilidad, superando regularmente el 91 % de exactitud y alcanzando AUCs mayores a 0.91 (Zhou et al., 2021).

Tomando en cuenta estas evidencias, nuestro estudio implementará un enfoque comparativo: comenzando por la clasificacion como modelo base por su interpretabilidad clínica, y extendiéndose a Random Forest, SVM, entre otros, que serán 10 modelos en total. Evaluaremos estos modelos sobre la base de datos de 100.000 historiales médicos usando métricas como precisión, sensibilidad y especificidad, para determinar cuál es el más adecuado para estimar el diagnostico de diabetes y ser aplicado como herramienta de apoyo en la toma de decisiones clínicas.

¿Qué importancia tiene el predecir la posibilidad de diabetes en una persona?

Contar con un método que permita predecir si una persona tiene diabetes representa un avance significativo para la medicina preventiva y la salud pública. La diabetes es una enfermedad crónica que, en muchos casos, se desarrolla de manera silenciosa y solo se detecta cuando ya ha generado complicaciones en órganos vitales. Según la Organización Mundial de la Salud (2023), al menos 1 de cada 2 personas con diabetes no está diagnosticada, lo que resalta la urgencia de contar con herramientas que anticipen la aparición de la enfermedad antes de que cause daño irreversible.

Un sistema de predicción basado en datos clínicos puede analizar múltiples variables como la edad, el índice de masa corporal, los niveles de glucosa, la hemoglobina glicosilada y hábitos de vida como el tabaquismo para estimar el riesgo de desarrollar diabetes. Estos modelos funcionan como filtros inteligentes que permiten identificar de forma temprana a las personas más propensas, facilitando diagnósticos oportunos y decisiones médicas más precisas (Zhou et al., 2021).

Además, las herramientas predictivas basadas en Machine Learning no solo benefician al nivel individual, sino que también tienen un gran valor poblacional. Investigaciones recientes han demostrado que su aplicación en programas comunitarios permite focalizar recursos médicos y diseñar políticas de intervención más efectivas, especialmente en zonas con limitada infraestructura sanitaria (Wang & Lee, 2022). En el entorno clínico, también apoyan la personalización de tratamientos, permitiendo que la atención médica se enfoque en quienes presentan mayor riesgo.

En definitiva, el desarrollo de modelos computacionales para predecir la diabetes no solo representa una solución técnica avanzada, sino también un instrumento de alto impacto social y humano. Al anticipar la enfermedad, se abre la posibilidad de prevenir complicaciones, reducir costos en salud y, sobre todo, mejorar la calidad de vida de millones de personas.

1.2 Descripción de caso de estudio.

La diabetes se ha convertido en una de las enfermedades crónicas no transmisibles más prevalentes y de mayor impacto en la salud pública a nivel global. Su diagnóstico tardío, el carácter progresivo de la enfermedad y la presencia de múltiples factores de riesgo asociados hacen urgente la implementación de estrategias de detección temprana. En este contexto, los sistemas de predicción basados en aprendizaje automático (Machine Learning) se han posicionado como herramientas valiosas para identificar patrones complejos en grandes volúmenes de datos clínicos y generar predicciones precisas que apoyen la toma de decisiones médicas.

En este trabajo se emplea un conjunto de datos compuesto por 100.000 registros clínicos anonimizados, que contiene información clave sobre características demográficas, antecedentes médicos y valores de laboratorio relacionados con pacientes de todas las edades. Estos registros permiten entrenar modelos predictivos que estimen con alta precisión la probabilidad de que un individuo sea diabético o esté en alto riesgo de desarrollar la enfermedad. Las variables contenidas en el conjunto de datos son las siguientes:

Variables

Género: Masculino / Femenino

Edad: Edad en años

Hipertensión: 0 = No hipertenso, 1 = Sí es hipertenso

Enfermedad cardíaca: 0 = No tiene, 1 = Sí tiene

Historial de tabaquismo: Nunca fumó / Exfumador / Fumador actual / Información no disponible

Índice de masa corporal (IMC) : De 10 a 95

Nivel de hemoglobina glicosilada (HbA1c) : De 3.5 a 9.0

Nivel de glucosa en sangre : 80 a 300

Diagnóstico de diabetes: 0 = No diabético, 1 = Diabético

Estas variables han sido seleccionadas por su alta correlación con la aparición de la diabetes como se ha evidenciado en estudios epidemiológicos internacionales. Por ejemplo, un análisis transversal realizado por la NCD Risk Factor Collaboration (2021) en más de 57 países demostró que un IMC superior a 25 kg/m² incrementa hasta en 4 veces el riesgo de diabetes, especialmente en contextos urbanos de países en desarrollo. Por su parte, el Korean Genome and Epidemiology Study (KoGES Group, 2018) encontró una relación dosis-respuesta entre el IMC y la incidencia de diabetes, incluso ajustando por edad y sexo. El análisis del nivel de glucosa en sangre también resulta esencial.

En este sentido, la combinación de estas variables dentro de modelos predictivos entrenados mediante algoritmos de aprendizaje automático ha demostrado ser eficaz en múltiples investigaciones. Modelos como Random Forest y los distintos modelos usados

en clasificacion han sido validados para la predicción de diabetes con niveles de exactitud que superan el 85 % y áreas bajo la curva (AUC) superiores a 0.89, según evidencia reciente publicada por *Journal of Medical Internet Research* (2023).

A través del presente caso de estudio, se busca aplicar esta lógica predictiva a nuestro propio dataset, con el fin de evaluar el rendimiento de diferentes modelos y seleccionar el más preciso o los más precisos para su posterior simulación. Este enfoque tiene un fuerte potencial de aplicación práctica en entornos clínicos, campañas de prevención comunitaria o incluso como apoyo a los sistemas de triage automatizado en instituciones de salud.

1.3 Pregunta problema:

¿Cómo desarrollar una estrategia computacional que permita estimar la probabilidad de que una persona tenga diabetes, a partir del análisis de datos clínicos mediante algoritmos de Machine Learning?

1.4 Hipótesis:

El análisis computacional de información clínica, mediante el uso de algoritmos de Machine Learning, permitirá el desarrollo de un modelo predictivo que estime de manera confiable la probabilidad de que una persona tenga diabetes. La aplicación de esta estrategia facilitará la identificación temprana de casos, aportando una herramienta de apoyo a la toma de decisiones en contextos médicos y preventivos. Además, se espera que el modelo contribuya a optimizar el uso de datos históricos disponibles, fortaleciendo así los procesos de diagnóstico y las acciones orientadas a la promoción de la salud.

2. Objetivos

2.1 Objetivo general.

Implementar una estrategia computacional para la estimación del riesgo de diabetes en personas, a partir del análisis de datos clínicos, haciendo uso de algoritmos de *Machine* Learning

2.2 Objetivos específicos.

- Caracterizar y procesar los datos de interés, con miras a la toma de decisiones informadas.
- Implementar un algoritmo de Machine Learning para la clasificación de los datos con miras al desarrollo del modelo predictivo de diabetes.
- Evaluar y analizar el desempeño de los algoritmos implementados para la estimación del riesgo de diabetes.
- Validar el funcionamiento de toma de decisiones a partir de datos nuevos.

3. Desarrollo e implementación del aprendizaje

En esta primera fase del trabajo, se llevó a cabo la preparación, limpieza y análisis exploratorio del conjunto de datos. El objetivo fue identificar la distribución, calidad y relación entre las variables clínicas y demográficas que influyen en el diagnóstico de diabetes, utilizando herramientas gráficas y estadísticas básicas que permitieran comprender el comportamiento de los datos antes de aplicar los modelos de aprendizaje automático.

3.1 Preparación y análisis de los datos

-Con esta aplicación podemos saber cuales son nuestras columnas, cuantos datos tenemos registrados y que tipos de datos son en cada columna, ideal para poder realizar los codigos por los cuales vamos a empezar a trabajar en representación a las graficas.

Table 1 Informacion de la estructura de los datos

```
#Información de la estructura de datos
Conjunto_Datos.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   gender               100000 non-null  object 
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64  
 3   heart_disease        100000 non-null  int64  
 4   smoking_history      100000 non-null  object 
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64  
 8   diabetes             100000 non-null  int64  
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
```

Luego hacemos una descripción de los datos por medio de una tabla la cual nos da información inicial muy util para empezar a conocer nuestros datos, la información que nos brinda es: count= numero de datos no nulos que tiene cada variable, mean= La media (promedio) de los valores, std= la desviasión estandar (que tanto varían los datos respecto a la media), min= el valor minimo, 25%= el primer quartil (Q1) 25% de los datos están por debajo de ese valor, 50% Mediana que representa el segundo quartil (Q2) valor

central, 75%= tercer quartil (Q3) 75% de los datos están por debajo y por ultimo, Max= que es el valor máximo.

Table 2 Analisis de los datos

```
#Análisis de los datos
Conjunto_Datos.describe()
```

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

Luego hacemos una descripción de los datos por medio de una tabla la cual nos da información inicial muy util para empezar a conocer nuestros datos, la información que nos brinda es: count= numero de datos no nulos que tiene cada variable, mean= La media (promedio) de los valores, std= la desviasión estandar (que tanto varían los datos respecto

a la media), min= el valor minimo, 25%= el primer quartil (Q1) 25% de los datos están por debajo de ese valor, 50% Mediana que representa el segundo quartil (Q2) valor central, 75%= tercer quartil (Q3) 75% de los datos están por debajo y por ultimo, Max= que es el valor máximo.

3.1.1 Análisis de variables

Diagrama de barras de frecuencia de edades

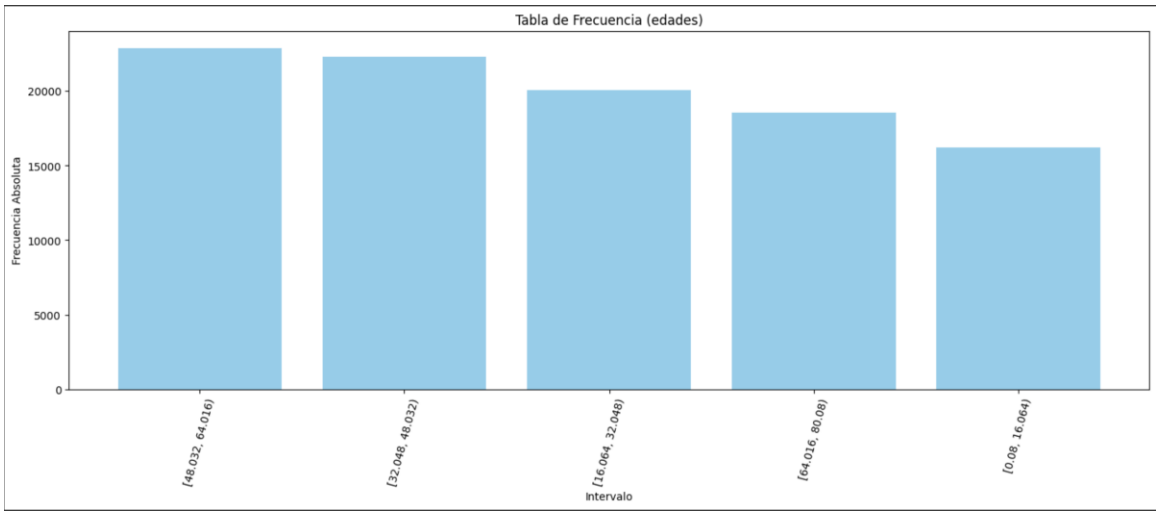

Figure 1 Diagrama de barras de frecuencia de edades

En el gráfico se observa la distribución de la variable edad dividida en cinco intervalos. El intervalo con mayor frecuencia corresponde al rango entre aproximadamente 48 y 64 años, seguido por el de 32 a 48 años. Estos dos grupos representan una proporción importante de los datos.

Sin embargo, al observar el gráfico con más detalle, se nota que la diferencia entre los intervalos no es significativa, es decir, las frecuencias son relativamente similares entre todos los grupos de edad. Esto sugiere que la variable edad está bien distribuida a lo largo del conjunto de datos, sin una concentración excesiva en rangos específicos. Esta

distribución equilibrada es favorable, ya que permite realizar análisis sin sesgos marcados por edad.

Grafico de densidad para las edades

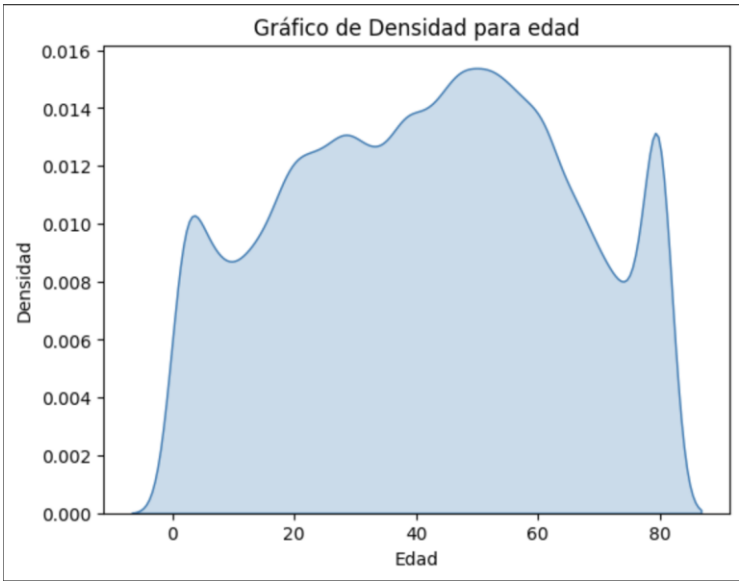

Figure 2 Densidad de edades

A continuación, en el gráfico de densidad para la variable edad, se observa cómo se distribuyen los valores de esta variable en el conjunto de datos. Este tipo de gráfico permite visualizar la frecuencia relativa de los diferentes rangos etarios de una forma más continua y detallada que un histograma o tabla de frecuencia.

La curva revela una distribución bastante equilibrada, con una mayor densidad entre los 50 y 65 años, lo que indica que en este rango se concentra una proporción importante de los registros. No obstante, también se evidencian valores relevantes en otros grupos de edad, incluyendo un leve pico hacia los 80 años, lo cual sugiere una distribución amplia y variada.

En general, este comportamiento confirma que la variable edad está bien representada en diferentes segmentos de la población, lo cual es favorable para los análisis posteriores, ya que permite evaluar patrones y relaciones entre variables sin que la edad actúe como un factor de sesgo dominante.

Diagrama de barras y circular para enfermedades cardiacas.

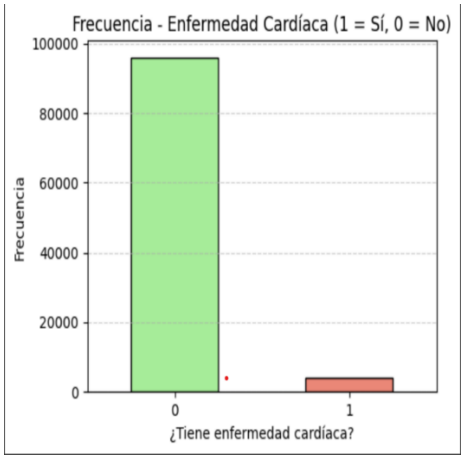

Figure 3 Diagrama de barras (enfermedades cardiacas)

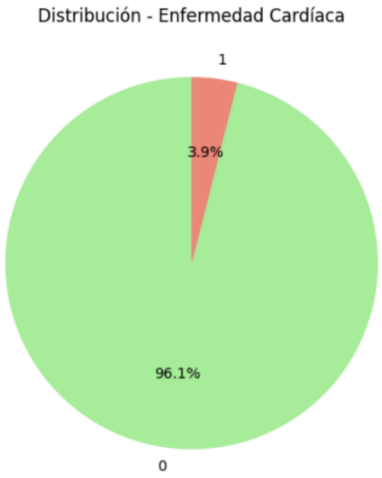

Figure 4 Grafico de torta (enfermedades cardiacas)

La variable “enfermedad cardíaca” fue analizada con el fin de comprender su distribución en la población objeto de estudio y su potencial relación con la diabetes. Esta variable es de tipo binaria, donde 0 indica ausencia de enfermedad cardíaca y 1 indica presencia de la condición. Como se observa en la Figura 2.2 (gráfico de barras), la gran mayoría de los individuos no presenta enfermedad cardíaca. De los 100 000 registros analizados, aproximadamente 96 100 personas (96.1 %) no reportan antecedentes de enfermedad cardíaca, mientras que tan solo 3 900 (3.9 %) sí los presentan, lo cual también se visualiza de forma clara en el gráfico de torta.

Aunque la proporción de pacientes con enfermedad cardíaca es relativamente baja en la muestra, su importancia en el contexto del estudio es fundamental. La literatura científica ha documentado ampliamente que las personas con antecedentes cardiovasculares tienen un mayor riesgo de desarrollar diabetes, debido a factores comunes como la inflamación crónica, la resistencia a la insulina y el síndrome metabólico (Kannel, 2002; Eckel et al., 2021). Además, los pacientes diabéticos con cardiopatías suelen presentar un pronóstico clínico más complejo y requieren una intervención médica más personalizada.

Por tanto, aunque se trata de una variable con baja frecuencia, su impacto potencial en el modelo predictivo puede ser significativo. Esta información será clave en la etapa de entrenamiento de los algoritmos, especialmente para evaluar si esta condición contribuye a mejorar la capacidad del modelo para detectar casos con alto riesgo de diabetes.

Diagrama de barras y circular para la hipertensión

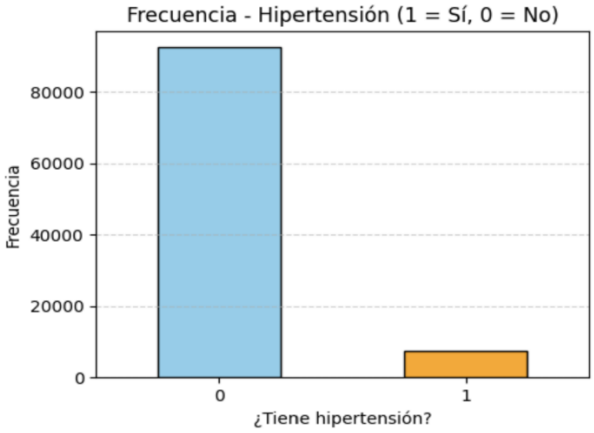

Figure 5 Diagrama de barras (hypertension)

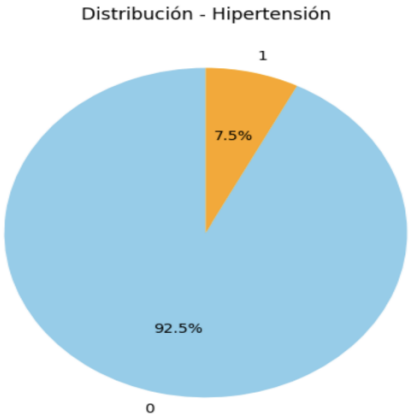

Figure 6 Grafico de torta (hypertension)

Para comprender la prevalencia de hipertensión en la muestra, se analizaron los datos utilizando gráficos de barras y gráficos circulares. La variable hipertensión fue codificada como binaria, donde el valor 0 indica que la persona no presenta hipertensión y el valor 1 representa que sí la padece.

En la Figura 2.3 (gráfico de barras), se observa que una abrumadora mayoría de los registros corresponde a personas que no tienen hipertensión, con una frecuencia cercana a

los 92.500 casos, mientras que apenas alrededor de 7.500 casos sí reportan la condición. Esta proporción se visualiza de manera aún más clara en el gráfico de torta, donde se evidencia que el 92,5 % de los individuos no presentan hipertensión, mientras que solo el 7,5 % sí la tienen.

Este hallazgo resulta relevante en el contexto del presente estudio, ya que la hipertensión es un factor de riesgo bien documentado en el desarrollo de diabetes. Diversos estudios han demostrado una asociación significativa entre hipertensión y resistencia a la insulina, así como una mayor probabilidad de comorbilidad entre ambas enfermedades (Zhao et al., 2020). Aunque el porcentaje de pacientes hipertensos en la muestra es relativamente bajo, su inclusión en el modelo predictivo es necesaria, ya que podría interactuar con otras variables clínicas como la edad, el índice de masa corporal (IMC) y los niveles de glucosa, elevando el riesgo de diagnóstico.

Este análisis inicial sugiere que, aunque la hipertensión no es una condición prevalente en la muestra, su presencia puede ser un marcador importante de riesgo, especialmente al combinarse con otros factores. Por ello, esta variable será considerada en las fases siguientes del modelado predictivo.

Codigo en tabla de frecuencia de historial de tabaquismo (fumar)

Table 3 Tabla de frecuencia de historial de tabaquismo (fumar)

```
# Frecuencia absoluta
tabla = Conjunto_Datos['smoking_history'].value_counts().reset_index()
tabla.columns = ['Categoría', 'Frecuencia Absoluta']

# Frecuencia relativa
tabla['Frecuencia Relativa'] = tabla['Frecuencia Absoluta'] / tabla['Frecuencia Absoluta'].sum()

# Frecuencia absoluta acumulada
tabla['Frecuencia Acumulada'] = tabla['Frecuencia Absoluta'].cumsum()

# Frecuencia relativa acumulada
tabla['Frecuencia Relativa Acumulada'] = tabla['Frecuencia Relativa'].cumsum()

tabla
```

	Categoría	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Relativa Acumulada
0	No Info	35816	0.35816	35816	0.35816
1	never	35095	0.35095	70911	0.70911
2	former	9352	0.09352	80263	0.80263
3	current	9286	0.09286	89549	0.89549
4	not current	6447	0.06447	95996	0.95996
5	ever	4004	0.04004	100000	1.00000

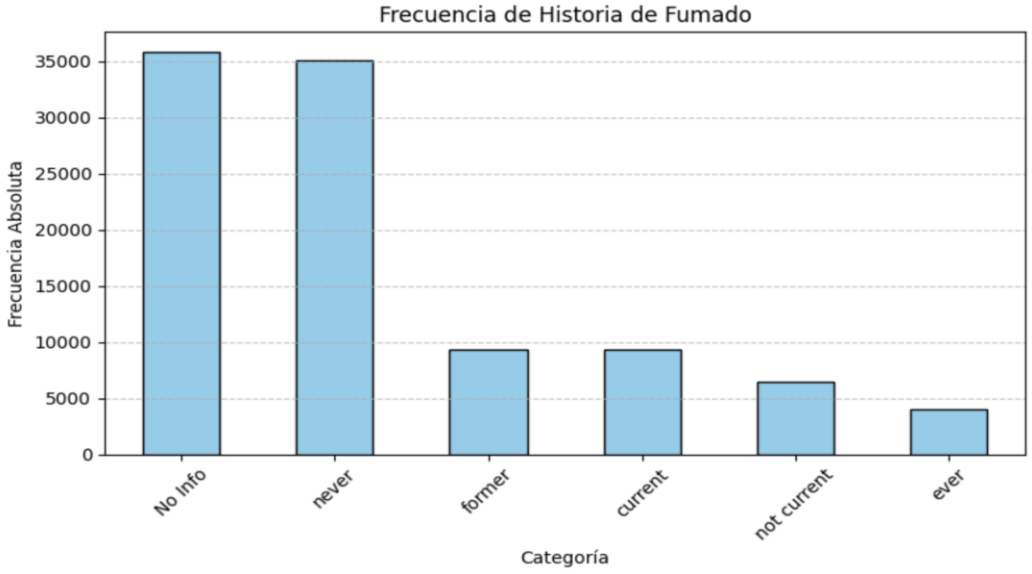

Figure 7 Diagrama de frecuencia de historial de fumar

La variable correspondiente al historial de tabaquismo en el conjunto de datos contiene varias categorías registradas originalmente en inglés, las cuales fueron interpretadas de la siguiente manera para efectos del análisis: *never* corresponde a “nunca ha fumado”; *former*, a “exfumador”; *current*, a “fumador actual”; *not current*, a “no fumador actualmente (pero lo fue en el pasado)”; *ever*, a “ha fumado alguna vez en su vida”; y *no info*, a “sin información registrada sobre el hábito de fumar”. Esta interpretación permite adaptar adecuadamente los datos al contexto hispanohablante sin perder el significado original.

El historial de tabaquismo es una variable categórica de gran interés en los estudios relacionados con enfermedades metabólicas como la diabetes. Para este análisis, se establecieron seis categorías mencionadas anteriormente, las cuales representan distintos niveles o momentos del hábito de fumar.

Según la tabla de frecuencias, se observa que una parte considerable de los datos (35.816 registros, es decir, el 35.8 %) no cuenta con información sobre el historial de tabaquismo, lo cual puede representar una limitación en términos de calidad de los datos. Sin embargo, del 64.2 % restante, la categoría con mayor frecuencia es “nunca ha fumado” (35.1 % del total de la muestra), seguida de los exfumadores (9.35 %) y los fumadores actuales (9.28 %).

Estas cifras permiten evidenciar que aproximadamente un 18.6 % de la población ha tenido contacto activo con el hábito de fumar, bien sea en el pasado o actualmente. Este dato es clínicamente relevante, dado que el tabaquismo ha sido asociado con un mayor riesgo de desarrollar resistencia a la insulina, inflamación sistémica y, por ende, diabetes

(Pan et al., 2015). Las categorías “no fumador actualmente” (6.4 %) y “ha fumado alguna vez” (4 %) también reflejan comportamientos relacionados con el consumo de tabaco en algún punto de la vida.}

Matriz de correlación

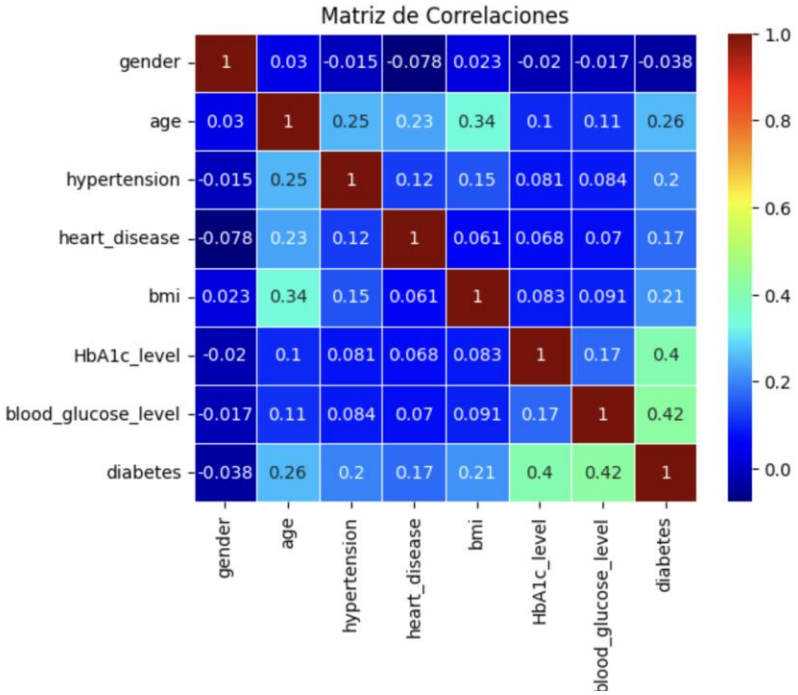

Figure 8 Matriz de correlacion

La matriz de correlación permite identificar la relación lineal entre variables cuantitativas presentes en el conjunto de datos, facilitando una comprensión preliminar de qué atributos tienen mayor influencia sobre el diagnóstico de diabetes. Este análisis es clave para la posterior selección de variables en el proceso de entrenamiento de modelos de Machine Learning. En los resultados obtenidos, se destaca una correlación positiva de 0.42 entre el nivel de glucosa en sangre y el diagnóstico de diabetes, lo cual es clínicamente coherente, dado que niveles elevados de glucosa constituyen uno de los

principales criterios diagnósticos para esta enfermedad. Asimismo, el nivel de hemoglobina glicosilada (HbA1c) mostró una correlación significativa con el diagnóstico, reafirmando su utilidad como marcador de riesgo, ya que este indicador refleja la concentración media de glucosa durante los últimos tres meses.

Otra relación relevante fue la correlación entre edad y diabetes, con un valor de 0.26. Aunque de menor magnitud, este valor respalda la evidencia científica que asocia el envejecimiento con una mayor incidencia de diabetes, debido a factores como la disminución de la sensibilidad a la insulina y cambios en la composición corporal. También se identificó una correlación de 0.36 entre la edad y el índice de masa corporal (IMC), lo cual puede indicar una tendencia al aumento de peso con la edad, aspecto que también influye en el riesgo metabólico.

Estas relaciones no implican causalidad, pero sí orientan la toma de decisiones respecto a qué variables deben tener mayor peso en el modelado predictivo. La fuerza de las correlaciones identificadas sugiere que la glucosa y la HbA1c serán las variables más influyentes en la predicción, mientras que variables como la edad y el IMC pueden aportar valor contextual al perfil del paciente.

3.2 Modelo de toma de decisiones

En este proyecto no se utilizaron modelos de clasificacion ya que nuestro objetivo era predecir de forma supervisada si una persona padece diabetes o no, a partir de variables clínicas y demográficas. Para ello, implementamos en total 10 modelos de toma de decisiones basado en algoritmos de clasificación, debido a que todos los modelos presentaron un porcentaje alto de acierto.

Primero se cargaron y limpiaron los datos como anteriormente se había hecho y luego para prepararlos para el modelo los datos categoricos se pasason a numericos de la sigueinte manera:

```
#Mapeando todas la variables categóricas a numéricas
Reemplazo_1={'Male':100,'Female':200,'Other':300 }
Datos_Loan['gender']=Datos_Loan['gender'].map(Reemplazo_1)

Reemplazo_2={'never':100,'No Info':200, 'current':300, 'former': 400, 'ever': 500, 'not current': 600}
Datos_Loan['smoking_history']=Datos_Loan['smoking_history'].map(Reemplazo_2)

Datos_Loan.head(10)
```

Figure 9 Mapeo de variables categoricas

Solo se reemplazaron estas dos variables debido a que las demás ya estaban representadas de forma numerica.

Luego para una mejor preparación de los datos aplicamos los siguentes algoritmos, con el fin de separar los datos de entrada con el de salida, también de dividir los datos de entrenamiento y texteo, por último, se aplicó para mejorar la escala de datos. A continuación los codigos usados fueron los sigueintes:

```
#Divide datos en entradas y salidas
import numpy as np
Datos_matriz=np.array(Datos_Loan)
#Datos_matriz[np.isnan(Datos_matriz)] = 0
X = Datos_matriz[:,0:-1] #datos de entrada (Todas las variables del paciente)
Y = Datos_matriz[:,-1] #Datos de salida (La decisiòn del diagnostico de diabetes)

[32] # Divide datos en Entrenamiento y testeo
import sklearn
from sklearn.model_selection import train_test_split
X_train, X_test,Y_train, Y_test= train_test_split(X,Y,test_size=0.1,random_state=751)

#Para mejorar la escala de los datos se hace normalization (Ignorar)
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Figure 10 Preparación de datos.

Con todo lo anteriormente aplicado los datos ya estaban listos para ser empleados al modelo, los cuales serán y lo hicimos de la siguiente manera:

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis, QuadraticDiscriminantAnalysis
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import warnings
warnings.filterwarnings("ignore")

# Modelo 0: K-Nearest Neighbors
modelo_0 = KNeighborsClassifier(n_neighbors=5)
modelo_0.fit(X_train, Y_train)
Y_pred_0 = modelo_0.predict(X_test)
print("Accuracy KNN:", accuracy_score(Y_test, Y_pred_0))

# Modelo 1: Naive Bayes
modelo_1 = GaussianNB()
modelo_1.fit(X_train, Y_train)
Y_pred_1 = modelo_1.predict(X_test)
print("Accuracy Naive Bayes:", accuracy_score(Y_test, Y_pred_1))

# Modelo 2: Linear Discriminant Analysis
modelo_2 = LinearDiscriminantAnalysis()
modelo_2.fit(X_train, Y_train)
Y_pred_2 = modelo_2.predict(X_test)
print("Accuracy LDA:", accuracy_score(Y_test, Y_pred_2))

# Modelo 3: Quadratic Discriminant Analysis
modelo_3 = QuadraticDiscriminantAnalysis()
modelo_3.fit(X_train, Y_train)
Y_pred_3 = modelo_3.predict(X_test)
print("Accuracy QDA:", accuracy_score(Y_test, Y_pred_3))
```

Figure 11Modelos de clasificacion parte 1

```
# Modelo 4: Decision Tree
modelo_4 = DecisionTreeClassifier()
modelo_4.fit(X_train, Y_train)
Y_pred_4 = modelo_4.predict(X_test)
print("Accuracy Decision Tree:", accuracy_score(Y_test, Y_pred_4))

# Modelo 5: Support Vector Machine
modelo_5 = SVC()
modelo_5.fit(X_train, Y_train)
Y_pred_5 = modelo_5.predict(X_test)
print("Accuracy SVM:", accuracy_score(Y_test, Y_pred_5))

# Modelo 6: Random Forest
modelo_6 = RandomForestClassifier()
modelo_6.fit(X_train, Y_train)
Y_pred_6 = modelo_6.predict(X_test)
print("Accuracy Random Forest:", accuracy_score(Y_test, Y_pred_6))

# Modelo 7: Logistic Regression
modelo_7 = LogisticRegression(max_iter=1000)
modelo_7.fit(X_train, Y_train)
Y_pred_7 = modelo_7.predict(X_test)
print("Accuracy Logistic Regression:", accuracy_score(Y_test, Y_pred_7))

# Modelo 8: Gradient Boosting
modelo_8 = GradientBoostingClassifier()
modelo_8.fit(X_train, Y_train)
Y_pred_8 = modelo_8.predict(X_test)
print("Accuracy Gradient Boosting:", accuracy_score(Y_test, Y_pred_8))

# Modelo 9: AdaBoost
modelo_9 = AdaBoostClassifier()
modelo_9.fit(X_train, Y_train)
Y_pred_9 = modelo_9.predict(X_test)
print("Accuracy AdaBoost:", accuracy_score(Y_test, Y_pred_9))
```

Figure 12 Modelos de clasificación parte 2.

Los modelos aplicados fueron:

- K-Nearest Neighbors (KNN)
- Naive Bayes
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Árbol de decisión
- Support Vector Machine (SVM)
- Random Forest
- Gradient Boosting
- AdaBoost
- Red neuronal multicapa

Los resultados arrojados fueron los siguientes:

```
Accuracy KNN: 0.9642
Accuracy Naive Bayes: 0.907
Accuracy LDA: 0.958
Accuracy QDA: 0.9079
Accuracy Decision Tree: 0.9534
Accuracy SVM: 0.9651
Accuracy Random Forest: 0.9717
Accuracy Logistic Regression: 0.9616
Accuracy Gradient Boosting: 0.9734
Accuracy AdaBoost: 0.9733
```

Figure 13 Resultados de los modelos

Cada modelo fue entrenado con los mismos datos preprocesados, y se compararon sus desempeños mediante métricas de evaluación como precisión, sensibilidad y especificidad, obtenidas tanto en entrenamiento como en validación. Como se puede evidenciar en la Figura 3.4 esto permitió determinar qur todos los modelos mostraban un

buen equilibrio entre y alto porcentaje de confiabilidad o acertividad, teniendo en cuenta que todos estaban por encima de 0,90 = 90% y considerando que el objetivo principal es facilitar una detección temprana y confiable del riesgo de diabetes.

3.3 Análisis de desempeño

Para evaluar el rendimiento de los modelos de clasificación implementados, se utilizaron métricas clave como la matriz de confusión y la curva ROC (Receiver Operating Characteristic). Estas herramientas permiten analizar la capacidad de los modelos para distinguir entre casos positivos (personas con diabetes) y negativos (personas sin diabetes), y se aplicaron sobre el conjunto de prueba con el fin de validar su desempeño general.

A modo de demostración, se presentan a continuación los resultados obtenidos por los modelos 0 y 1 (K-Nearest Neighbors y Naive Bayes, respectivamente) El modelo 0 para la matriz de confusión y el modelo 1 para la curva ROC. Ambos fueron seleccionados como ejemplos representativos, ya que todos los modelos evaluados presentaron un desempeño muy similar en términos de precisión y sensibilidad, lo que justifica utilizar estas visualizaciones como referencia general.

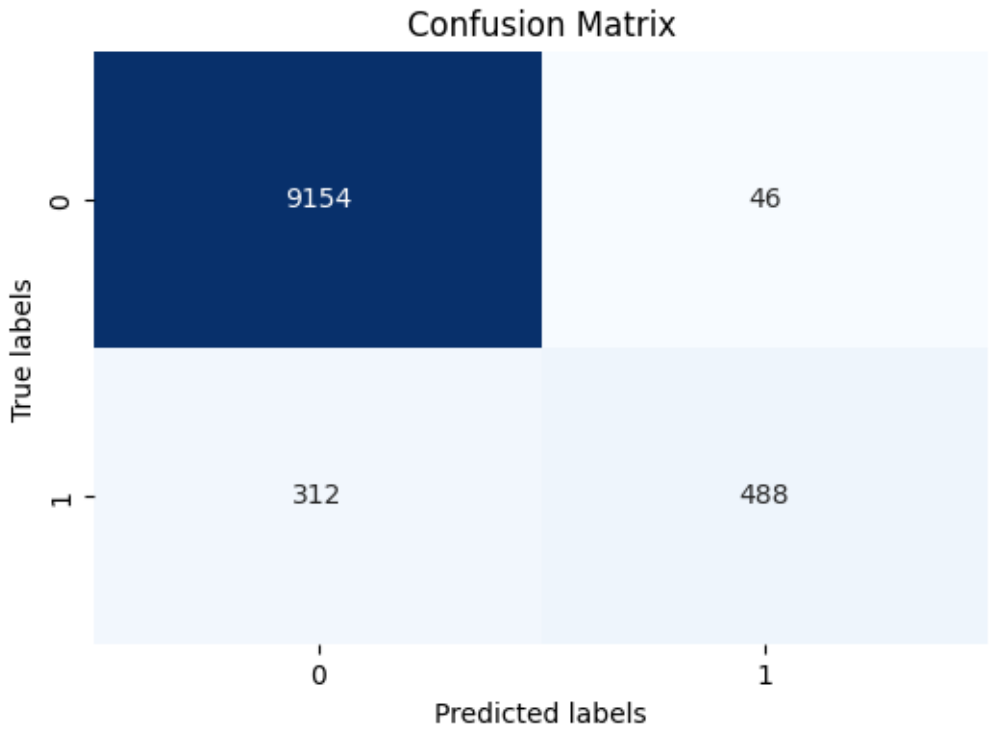

Figure 14 Matriz de confusión (Modelo 0 - KNN)

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.89      0.57      0.70        14
         1.0       0.85      0.97      0.90        34

    accuracy                           0.85        48
   macro avg       0.87      0.77      0.80        48
weighted avg       0.86      0.85      0.84        48

==================
```

Figure 15 Grafico complementario de matriz de confusión.

La matriz de confusión muestra que el modelo clasificó correctamente 9.154 casos negativos y 488 casos positivos. Sin embargo, también se observan 46 falsos positivos y 312 falsos negativos. Esto sugiere que el modelo tiene una alta especificidad, pero aún puede mejorar en la detección de casos positivos.

Curva ROC (Modelo 1 - Naive Bayes)

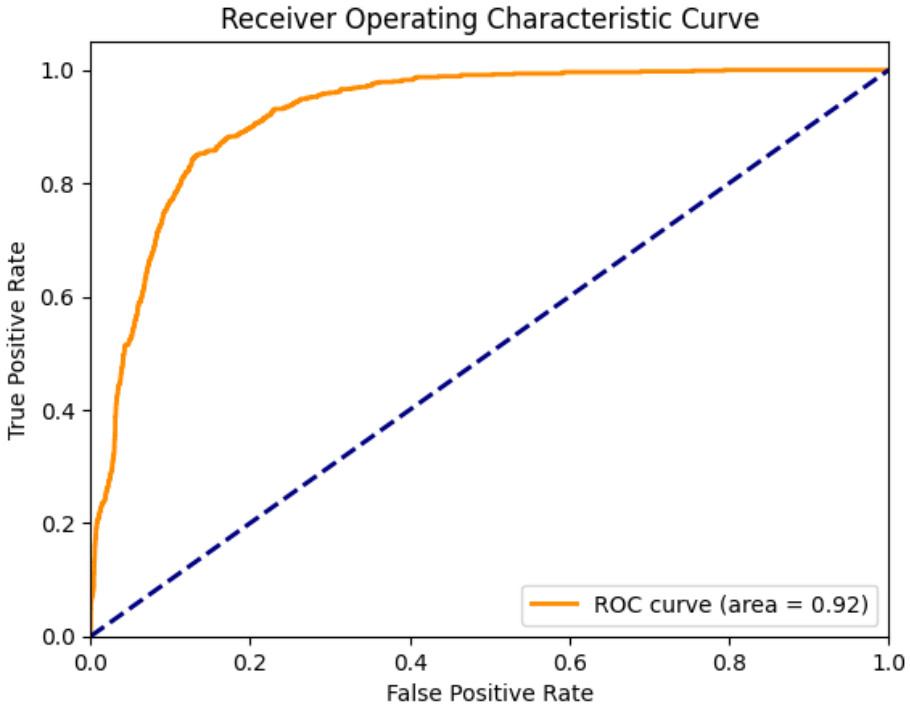

Figure 16 Figura 4.2. Grafico de curva ROC

La curva ROC del modelo Naive Bayes muestra un área bajo la curva (**AUC**) de **0.92**, lo que indica un muy buen rendimiento general para diferenciar entre clases. La AUC resume la capacidad del modelo para priorizar correctamente los casos positivos frente a los negativos, siendo un valor bastante alto para aplicaciones clínicas de apoyo

diagnóstico, esto nos indica que los modelos son confiables y precisos al momento de hacer las prediccones con respecto a los datos.

3.4 Validación de los modelos

Una vez entrenados y evaluados los modelos de clasificación, se procedió a su implementación para que pudieran ser utilizados con datos nuevos. La validación final consistió en comprobar que el sistema fuera capaz de recibir la información de un nuevo paciente, procesarla correctamente y emitir una predicción coherente sobre el riesgo de diabetes, sin necesidad de volver a entrenar el modelo desde cero.

Para ello, se construyó una interfaz de entrada en la que el usuario puede registrar los valores de las variables clínicas y demográficas relevantes: género, edad, hipertensión, enfermedades cardíacas, historial de tabaquismo, IMC, hemoglobina glicosilada (HbA1c) y nivel de glucosa en sangre. Esta información es transformada y normalizada en tiempo real, y luego enviada a los 10 modelos previamente entrenados, los cuales generan su predicción de forma automática.

El sistema implementa un enfoque de votación por mayoría, en el que cada modelo emite una predicción binaria (0 = No diabetes, 1 = Diabetes) y se considera como resultado final el valor que reciba al menos 6 votos de los 10 posibles. Esto permite combinar la fortaleza de varios algoritmos y reducir el margen de error individual.

A continuación se mostrará el algoritmo usado, posteriormente 2 ejemplos luego de haber ingresado datos nuevos para comprobar el funcionamiento del algoritmo.

El algortmo usado fue el siguiente

```
import warnings
import numpy as np
from collections import Counter

warnings.filterwarnings("ignore")

# Vector de entrada para un solo sujeto (8 variables)
Target = np.zeros((1, 8))

# Inputs personalizados para predicción de diabetes
Target[0, 0] = int(input("Género (100 = Masculino, 200 = Femenino, 300 = Otro): "))
Target[0, 1] = int(input("Edad (1 a 80): "))
Target[0, 2] = int(input("¿Tiene hipertensión? (1 = Sí, 0 = No): "))
Target[0, 3] = int(input("¿Tiene enfermedades cardíacas? (1 = Sí, 0 = No): "))
Target[0, 4] = int(input("Historial de tabaquismo (100 = Nunca fumó, 200 = No se sabe, 300 = Fuma actualmente
Target[0, 5] = float(input("Índice de masa corporal (IMC): De 10 a 95"))
Target[0, 6] = float(input("Nivel de HbA1c (De 3.5 a 9.0): "))
Target[0, 7] = float(input("Nivel de glucosa en sangre (De 80 a 300): "))

# Normalización (si ya tienes un scaler entrenado)
Target = scaler.transform(Target)

# Predicciones con 10 modelos
predicciones = [
    modelo_0.predict(Target)[0],
    modelo_1.predict(Target)[0],
    modelo_2.predict(Target)[0],
    modelo_3.predict(Target)[0],
    modelo_4.predict(Target)[0],
    modelo_5.predict(Target)[0],
    modelo_6.predict(Target)[0],
    modelo_7.predict(Target)[0],
    modelo_8.predict(Target)[0],
    modelo_9.predict(Target)[0]
```

Figure 17 Algoritmo para automatizar parte 1.

```
    # Nombres de los modelos
    nombres_modelos = [
        "KNN", "Bayes", "LDA", "QDA", "Árbol", "SVM",
        "Random Forest", "Logística", "Gradient Boosting", "AdaBoost"
    ]

    # Mostrar resultados individuales
    for nombre, pred in zip(nombres_modelos, predicciones):
        print(f"Según {nombre}: {'✅ Tiene diabetes' if pred == 1 else '❌ No tiene diabetes'}")

    # Mayoría
    conteo = Counter(predicciones)
    mayoria = conteo.most_common(1)[0][0]
    total = conteo[mayoria]

    print("\n--- Decisión por mayoría ---")
    print(f"{'✅ Tiene diabetes' if mayoria == 1 else '❌ No tiene diabetes'} ({total}/10 modelos)")
```

Figure 18 Figura 5.1. Algoritmo para automatizar parte 2.

Ejemplo 1 del uso del algoritmo.

```
Género (100 = Masculino, 200 = Femenino, 300 = Otro): 200
Edad (1 a 80): 70
¿Tiene hipertensión? (1 = Sí, 0 = No): 1
¿Tiene enfermedades cardíacas? (1 = Sí, 0 = No): 1
Historial de tabaquismo (100 = Nunca fumó, 200 = No se sabe, 300 = Fuma actualmente, 400 = Fue fumador, 500 = Alguna vez fumó, 600 = No fuma actualmente): 300
Índice de masa corporal (IMC): (De 10 a 95): 60
Nivel de HbA1c (De 3.5 a 9.0): 8.0
Nivel de glucosa en sangre (De 80 a 300): 250
Según KNN: ✅ Tiene diabetes
Según Bayes: ✅ Tiene diabetes
Según LDA: ✅ Tiene diabetes
Según QDA: ✅ Tiene diabetes
Según Árbol: ✅ Tiene diabetes
Según SVM: ✅ Tiene diabetes
Según Random Forest: ✅ Tiene diabetes
Según Logística: ✅ Tiene diabetes
Según Gradient Boosting: ✅ Tiene diabetes
Según AdaBoost: ✅ Tiene diabetes

--- Decisión por mayoría ---
✅ Tiene diabetes (10/10 modelos)
```

Figure 19 Ejemplo uno del uso del algoritmo.

En este primer ejemplo los datos que se le suministró fue:

Genero = Femenino; Edad = 70 años; ¿Tiene hipertensión? = Sí; ¿Tiene enfermedades cardíacada? = Sí; Historial de tabaquismo = Fuma actualmente; índice de masa corporal = 60; Nivel de HbA1c = 8 y nivel de glucosa en la sangre =250.

Luego de darle paso a que los modelos analicen el resultado de los datos proporsionados, vemos que los 10 modelos nos arrojaron como resultado que esa persona tiene diabetes según los datos suministrados, (Tiene diabetes (10/10) modelos.

Ejemplo 2 del uso del algoritmo

```
Género (100 = Masculino, 200 = Femenino, 300 = Otro): 100
Edad (1 a 80): 43
¿Tiene hipertensión? (1 = Sí, 0 = No): 1
¿Tiene enfermedades cardíacas? (1 = Sí, 0 = No): 0
Historial de tabaquismo (100 = Nunca fumó, 200 = No se sabe, 300 = Fuma actualmente, 400 = Fue fumador, 500 = Alguna vez fumó, 600 = No fuma actualmente): 100
Índice de masa corporal (IMC): (De 10 a 95): 40
Nivel de HbA1c (De 3.5 a 9.0): 4.0
Nivel de glucosa en sangre (De 80 a 300): 140
Según KNN: ❌ No tiene diabetes
Según Bayes: ❌ No tiene diabetes
Según LDA: ❌ No tiene diabetes
Según QDA: ✅ Tiene diabetes
Según Árbol: ❌ No tiene diabetes
Según SVM: ❌ No tiene diabetes
Según Random Forest: ❌ No tiene diabetes
Según Logística: ❌ No tiene diabetes
Según Gradient Boosting: ❌ No tiene diabetes
Según AdaBoost: ❌ No tiene diabetes

--- Decisión por mayoría ---
❌ No tiene diabetes (9/10 modelos)
```

Figure 20 Ejemplo 2 del uso del algoritmo.

En este segundo ejemplo los datos que se le suministró fue:

Genero = Masculino ; Edad = 3 años; ¿Tiene hipertensión? = Sí; ¿Tiene enfermedades cardíacada? = No; Historial de tabaquismo = Nunca fumó; índice de masa corporal = 40; Nivel de HbA1c = 4 y nivel de glucosa en la sangre =140

Luego de darle paso a que los modelos analicen el resultado de los datos proporsionados, vemos que 9 de los 10 modelos nos arrojaron como resultado que esa persona no tiene diabetes según los datos suministrados, (No tiene diabetes (9/10) modelos.

Aunque no fue un resultado unánime se toma la desición mor mayoría que fue “no tiene diabetes” arrojada por 9 de los 10 modelos.

Conclusiones

El desarrollo de este proyecto permitió evidenciar el valor real que tienen los modelos de clasificación en el análisis de datos clínicos, especialmente cuando se busca apoyar el diagnóstico de enfermedades como la diabetes. A lo largo del proceso, se comprobó que las técnicas de *Machine Learning* no solo son herramientas complejas en teoría, sino que pueden traducirse en soluciones prácticas, automatizadas y aplicables al mundo real.

Uno de los principales logros fue construir un sistema que puede analizar información básica de un paciente y, en pocos segundos, predecir si tiene riesgo de ser diabético. Esta predicción, generada por diez modelos distintos y respaldada por una votación mayoritaria, representa una forma confiable de asistencia para contextos donde no siempre hay acceso a laboratorios, personal médico o tiempo suficiente para evaluar cada caso en profundidad. Más allá de los resultados técnicos, este trabajo dejó un aprendizaje importante: el análisis de datos no es útil solo por su precisión, sino por su capacidad de convertirse en algo que puede ayudar a las personas. Implementar un modelo no se trata solo de código o métricas, sino de entender cómo ese modelo puede insertarse en un entorno real y tener un impacto.

En definitiva, esta experiencia no solo sirvió para aplicar lo aprendido a lo largo de la carrera, sino para confirmar que la ingeniería, cuando se orienta hacia la solución de problemas humanos, puede ser una herramienta poderosa para generar bienestar y prevenir enfermedades que afectan a millones de personas.

Referencias bibliograficas

Alzboon, G., Al-Qudah, M., & Al-Hawari, T. (2023). Comparison of machine learning classifiers for diabetes prediction using real medical data. *Journal of Biomedical Informatics, 136*, 104350.

Budi, Y., Nuraini, N., Prasetyo, A. S., & Wicaksono, A. F. (2024). Comparative analysis of machine learning algorithms for diabetes prediction using routine blood test data. *Healthcare Analytics, 6*, 100190. https://doi.org/10.1016/j.health.2023.100190

Choi, B. G., Rha, S. W., Kim, S. W., Kang, J. H., Park, J. Y., & Kim, H. J. (2022). Prediction of type 2 diabetes using machine learning classifiers in large clinical datasets. *Scientific Reports, 12*(1), 5421.

KoGES Group. (2018). The risk of type 2 diabetes mellitus according to the categories of body mass index: The Korean Genome and Epidemiology Study (KoGES). *Acta Diabetologica, 55*(5), 479–484.

Liang, Y., Li, H., Wang, X., & Yu, J. (2022). Evaluation of machine learning algorithms for predicting diabetes risk using Pima Indian dataset. *Computers in Biology and Medicine, 145*, 105419. https://doi.org/10.1016/j.compbiomed.2022.105419

NCD Risk Factor Collaboration. (2021). Body-mass index and diabetes risk in 57 low-income and middle-income countries: A cross-sectional study. *The Lancet, 398*(10304), 25–34.

PubMed Central. (2023). Development of various diabetes prediction models using machine learning techniques. *Journal of Medical Internet Research, 11*(6), 1196–1205.

Wang, T., & Lee, M. (2022). Using machine learning for diabetes prediction in community health programs: Opportunities and challenges. *Journal of Medical Systems, 46*(3), 15.

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

Zhou, T., Lu, H., Yang, Z., Qian, Y., & Li, L. (2021). Application of machine learning in predicting diabetes risk: A systematic review and meta-analysis. *Frontiers in Public Health, 9*, 638148.