



UNIVERSIDAD
CATÓLICA
BOLIVIANA
SANTA CRUZ

INGENIERÍA EN INTELIGENCIA ARTIFICIAL

**DOCUMENTO DE PROYECTO FINAL EN LA MATERIA DE ADQUISICION,
ANALISIS Y PROCESAMIENTO DE DATOS**

**MODELOS PREDICTIVO PARA LA DETECCION DE DIABETES BASADO EN
DATOS CLINICOS DE PACIENTES DIAGNOSTICADOS Y NO
DIAGNOSTICADOS CON LA ENFERMEDAD**

Docente: Marin Salazar Carmen Rosa
Estudiante: José Alfredo Zambrana Cruz

Santa Cruz - Bolivia
Junio, 2025

índice

1. Resumen ejecutivo.....	4
2. Definición del Problema.....	5
2.1. Preguntas de investigación	5
2.2. Alcance y limitaciones.....	5
3. Descripción de los Datos.....	6
3.1 Diccionario de variables utilizadas:.....	6
3.2. Muestra del dataset:.....	7
3.2.1 Columnas y sus valores:	7
4. Adquisición de Datos	10
4.1. Herramientas utilizadas:	10
4.2. Problemas encontrados y soluciones:	11
5. Preprocesamiento y Limpieza	11
5.1 Manejo de valores nulos.....	11
5.2. Normalización de variables numéricas	11
5.3. Codificación de variables categóricas	12
5.4. Filtrado de outliers.....	12
5.5. Transformaciones aplicadas y justificación.....	12
6. Análisis Exploratorio de Datos (EDA)	13
6.1. Estadísticas descriptivas.....	13
6.2. Visualizaciones clave	15
6.3. Principales patrones observados	18
7. Feature Engineering.....	19
7.1. Nuevas variables creadas.....	19
7.2. Transformaciones aplicadas.....	19
7.3. Selección y eliminación de atributos irrelevantes	19
8. Modelado y Evaluación	20
8.1 Tipo de modelos utilizados	20
8.2. División de los datos	20
8.3. Métricas de evaluación	20
8.4. MODELOS.....	21
8.4.1. Modelo de Regresión Logística	21
8.4.2. Modelo Árbol de Decisión.....	21
8.4.3. Modelo Random Forest	23
8.4.4. Modelo SVC y kernel selection.....	25

8.4.5. Naive Bayes (GaussianNB)	25
9. Comparación de modelos y resultados.....	26
9.1 Análisis Individual.....	27
9.1.1. Logistic Regression:	27
9.1.2. Decision Tree:.....	27
9.1.3. RandomForest:	28
9.2. Analisis del modelo ganador	28
9.2.1 Matriz de confusión Decisión Tree Classifier:	28
9.2.2. Classification Report:	29
9.2.3. Curva ROC:	29
9.2.3.1. Análisis de la Curva ROC.....	29
9.2.4. Variables influyentes para el modelo:	30
10. Conclusiones y Recomendaciones	31
10.1. Conclusiones	31
10.2. Recomendaciones	31

1. Resumen ejecutivo

El proyecto tuvo como propósito desarrollar un modelo predictivo para la detección de diabetes, mediante el análisis de variables clínicas extraídas de un conjunto de datos disponible en la plataforma Kaggle (ver Anexos). El objetivo principal fue identificar los factores con mayor influencia en el diagnóstico de esta enfermedad. Para ello, se empleó un enfoque basado en inteligencia artificial que permitió evaluar la relevancia de cada variable en relación con la presencia o ausencia de diabetes.

El problema abordado se centró en la limitada implementación de herramientas de inteligencia artificial en el ámbito clínico boliviano, a pesar de su creciente adopción en otros países. Esta investigación buscó evidenciar el potencial de dichas herramientas, mediante la aplicación de un modelo predictivo que pudiera aportar precisión y utilidad al proceso diagnóstico.

Los resultados obtenidos mostraron que las variables con mayor influencia en el diagnóstico fueron el nivel de glucosa en sangre, el nivel de hemoglobina glicosilada (HbA1c) y la edad. El modelo alcanzó una precisión del 88 %, lo que indicó una capacidad significativa para predecir la presencia de diabetes con base en los datos disponibles.

El análisis permitió demostrar que la inteligencia artificial puede constituirse en una herramienta complementaria de valor para el diagnóstico clínico. Su implementación futura podría optimizar la identificación temprana de enfermedades como la diabetes, fortaleciendo así la toma de decisiones médicas basadas en datos.

2. Definición del Problema

El diagnóstico temprano de la diabetes representa un desafío constante en el ámbito clínico. A pesar de los avances tecnológicos, el uso de herramientas de inteligencia artificial (IA) en contextos de atención médica continúa siendo escaso en diversos países en desarrollo, entre ellos Bolivia. Este proyecto busca evaluar la viabilidad de implementar un modelo predictivo de IA que permita detectar la presencia de diabetes, a partir de datos clínicos de pacientes diagnosticados y no diagnosticados. Asimismo, se pretende identificar las variables con mayor influencia en el diagnóstico automatizado, y determinar cuál de los modelos evaluados proporciona el mejor desempeño en este contexto.

2.1. Preguntas de investigación

- ¿Qué tan eficiente es la inteligencia artificial en la detección de la diabetes?
- ¿Qué variables influyen en la predicción del diagnóstico según el modelo?
- ¿Qué modelo predictivo presenta el mejor rendimiento con el conjunto de datos empleado?

2.2. Alcance y limitaciones

El proyecto se centró en la aplicación y evaluación de modelos de inteligencia artificial entrenados exclusivamente con los datos proporcionados por un conjunto disponible en la plataforma Kaggle. La metodología incluyó la comparación de distintos algoritmos de clasificación (Decision Tree, Logistic Regression, Random Forest, Support Vector Machine y Naive Bayes) para determinar cuál ofrecía mejores resultados predictivos.

- El alcance metodológico se limitó a la selección, entrenamiento y evaluación de modelos de clasificación supervisada. No se incorporaron datos clínicos externos ni se realizaron validaciones en entornos hospitalarios reales.
- El análisis de datos se restringió a las variables contenidas en el dataset original, sin adición de nuevas fuentes. Por tanto, los resultados y conclusiones solo son válidos en el marco de este conjunto de datos y no pretenden generalizarse a toda la población.

3. Descripción de los Datos

La fuente de datos utilizada para el desarrollo del modelo predictivo fue la plataforma Kaggle, específicamente un conjunto de datos relacionado con características clínicas de pacientes diagnosticados y no diagnosticados con diabetes.

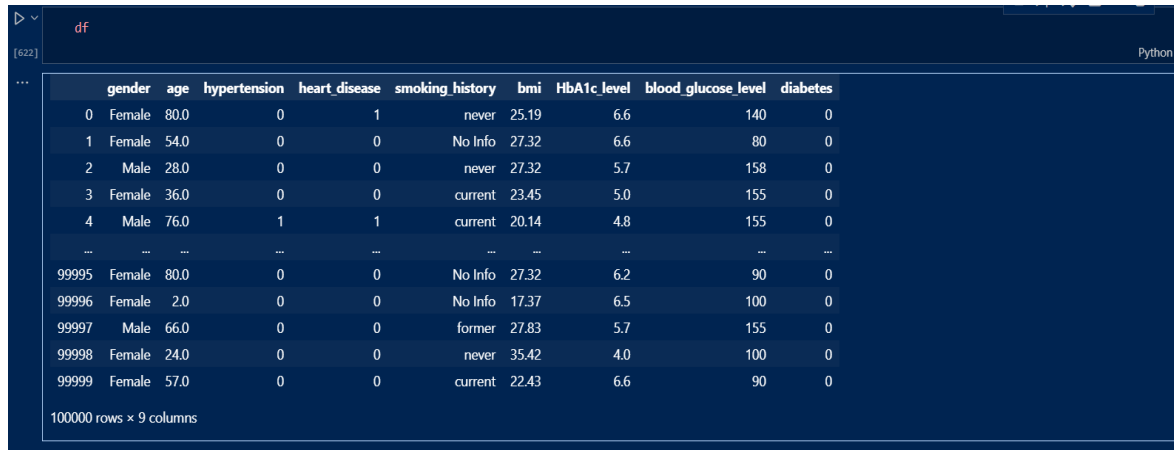
El dataset estuvo conformado por **100.000 registros (filas)** y **9 variables (columnas)**. Estas variables incluyeron información demográfica, antecedentes médicos, historial de tabaquismo y valores clínicos relevantes para el diagnóstico de la enfermedad.

3.1 Diccionario de variables utilizadas:

Variable	Descripción
gender	Sexo biológico del paciente (Male, Female, Other)
age	Edad del paciente en años
hypertension	Presencia de hipertensión (1: Sí, 0: No)
heart_disease	Presencia de enfermedad cardíaca (1: Sí, 0: No)
smoking_history	Historial de consumo de tabaco (never, former, current, etc.)
bmi	Índice de masa corporal (Body Mass Index)
HbA1c_level	Nivel de hemoglobina glicosilada
blood_glucose_level	Nivel de glucosa en sangre
diabetes	Diagnóstico de diabetes (1: Sí, 0: No)

3.2. Muestra del dataset:

A continuación, se presenta una muestra representativa del conjunto de datos



The screenshot shows a Jupyter Notebook interface with a variable named 'df' containing 100,000 rows and 9 columns. The columns are: gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, and diabetes. The first five rows of data are displayed, showing a mix of male and female patients with various health conditions and smoking histories.

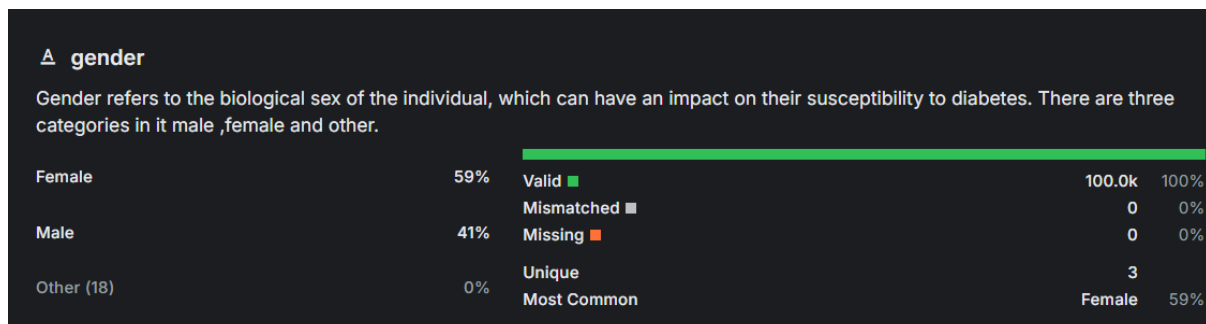
	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
...
99995	Female	80.0	0	0	No Info	27.32	6.2	90	0
99996	Female	2.0	0	0	No Info	17.37	6.5	100	0
99997	Male	66.0	0	0	former	27.83	5.7	155	0
99998	Female	24.0	0	0	never	35.42	4.0	100	0
99999	Female	57.0	0	0	current	22.43	6.6	90	0

100000 rows x 9 columns

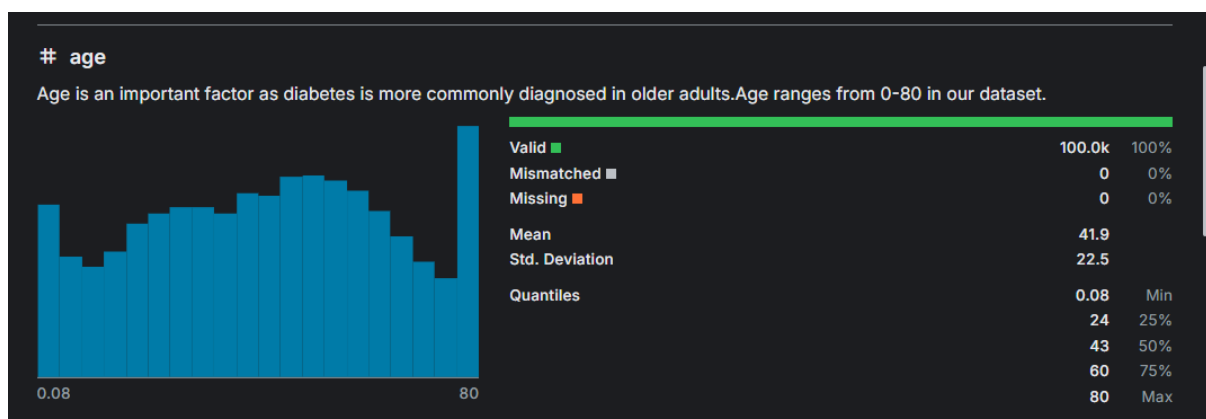
Esta muestra ilustra la estructura de los datos disponibles y los valores esperados para cada una de las variables consideradas en el análisis.

3.2.1 Columnas y sus valores:

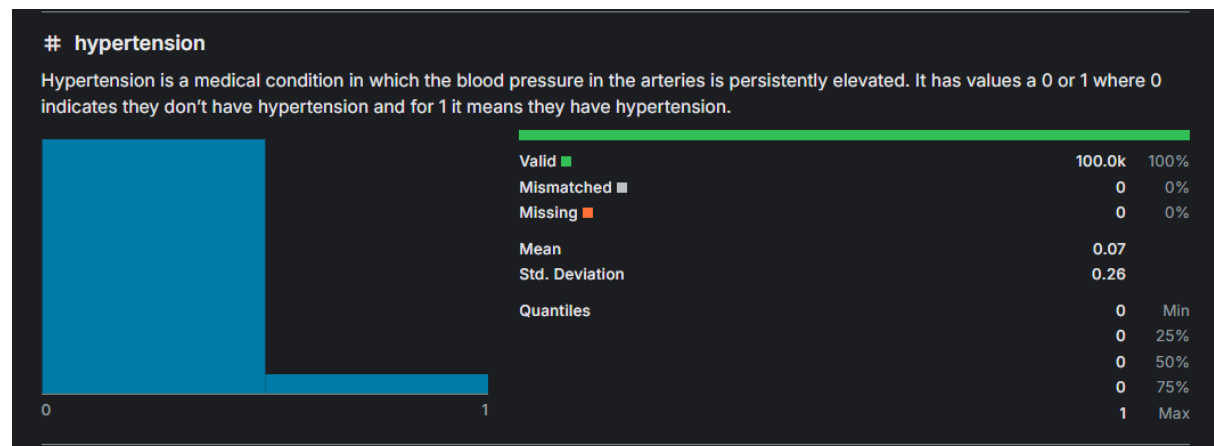
Gender:



Age:



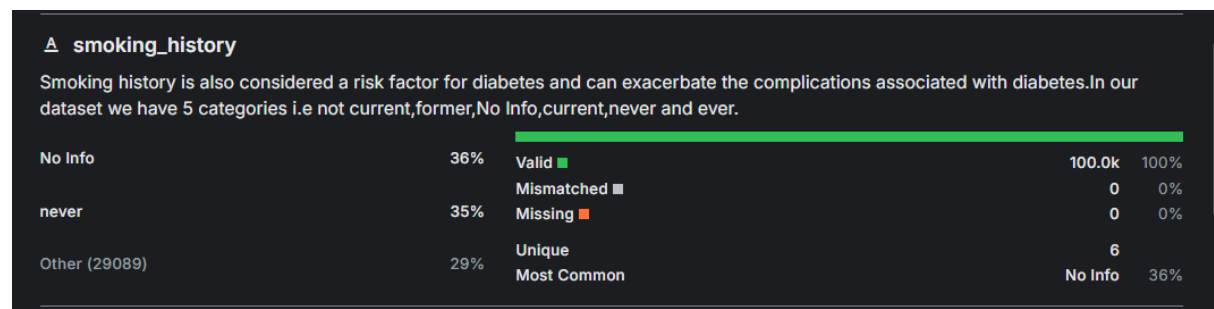
Hypertension:



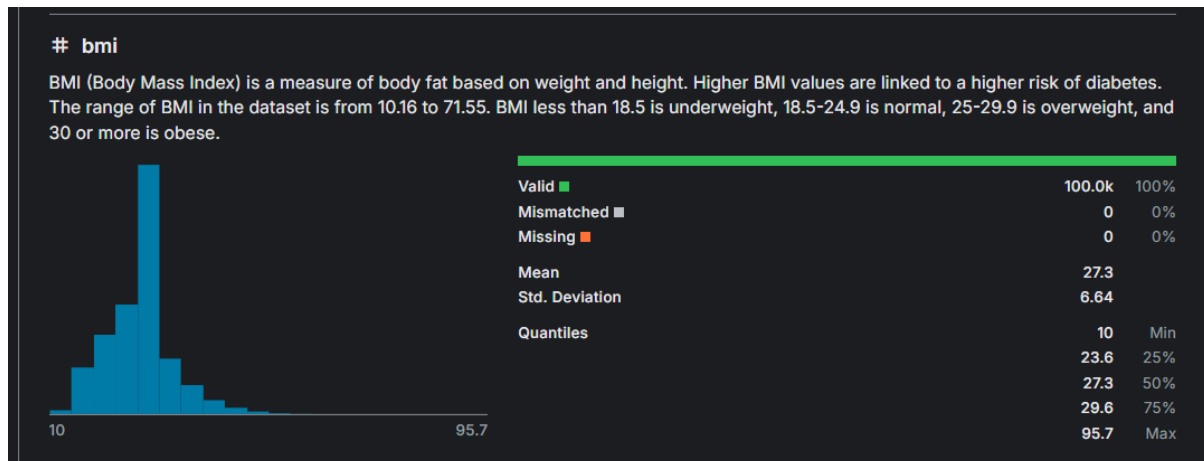
heart_disease:



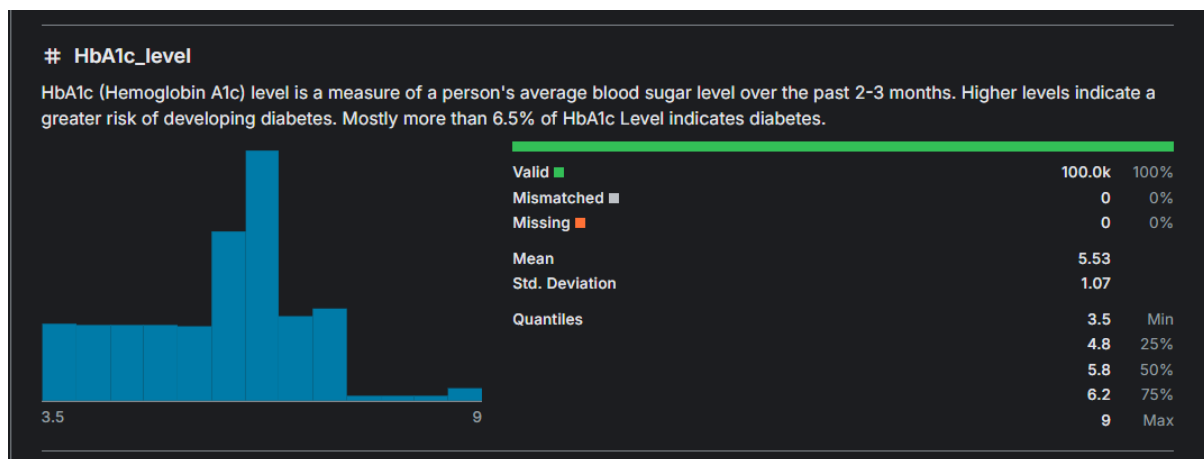
smoking_history:



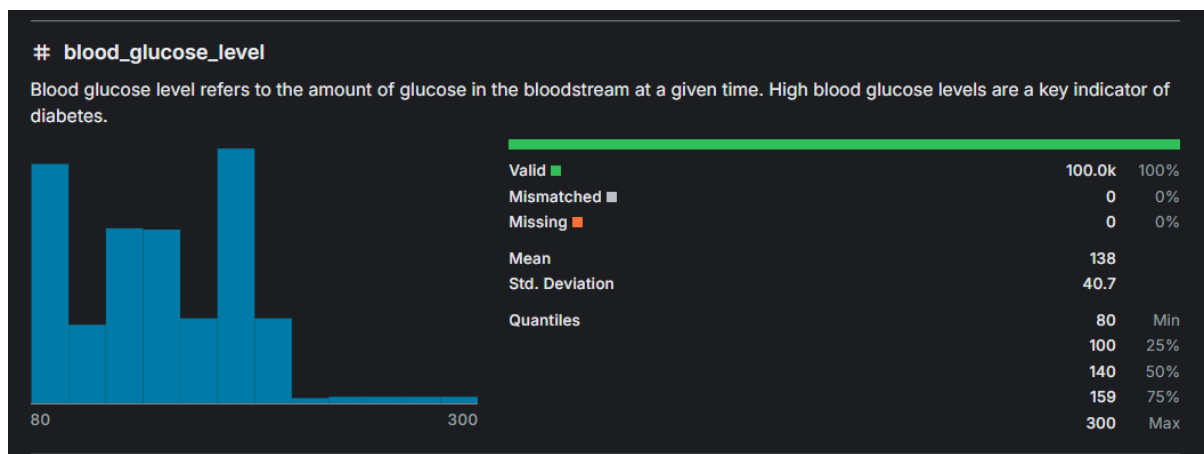
Bmi:



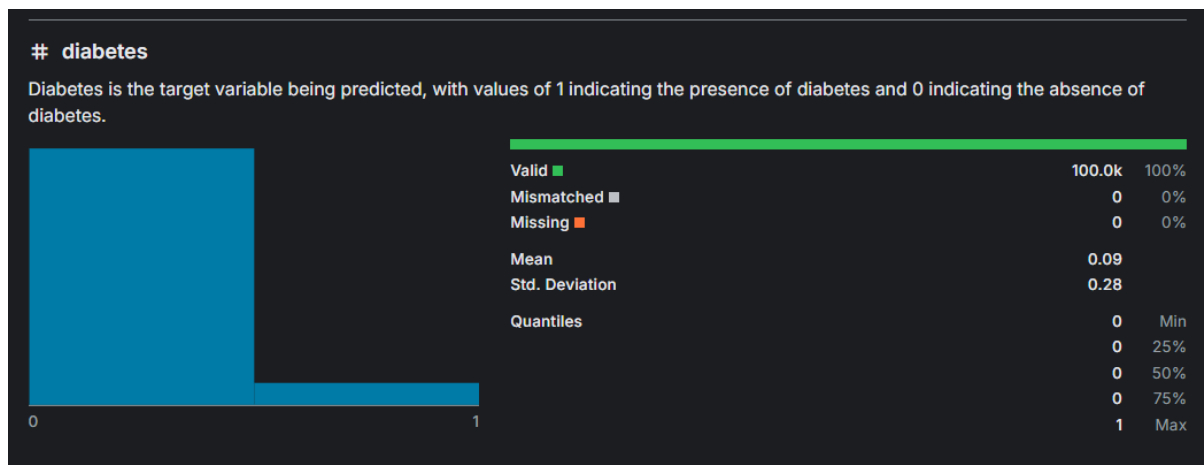
HbA1c_level:



blood_glucose_level:



diabetes:



4. Adquisición de Datos

Los datos fueron obtenidos mediante carga local, a partir de un archivo en formato CSV descargado desde la plataforma Kaggle. Esta fuente proporcionó un conjunto de datos estructurado, listo para su procesamiento y análisis.

4.1. Herramientas utilizadas:

- **Pandas:** Utilizado para la lectura del archivo CSV, así como para la limpieza, transformación y preparación de los datos antes del entrenamiento de los modelos.
- **Matplotlib:** Empleado en la elaboración de gráficos exploratorios como diagramas de barras, diagramas de cajas y curvas ROC, con el fin de facilitar el análisis visual de las variables.
- **Seaborn:** Aplicado en la representación gráfica de matrices de correlación, lo que permitió identificar relaciones entre variables.
- **Scikit-learn (sklearn):** Herramienta principal para el entrenamiento de los modelos predictivos. También se utilizó para el cálculo de métricas de evaluación (precisión, sensibilidad, exactitud, puntuación F1), visualización de árboles de decisión, generación de matrices de confusión y curvas ROC. Adicionalmente, fue empleada en el proceso de *feature engineering*, que incluyó la normalización de variables numéricas, la codificación de variables categóricas y la selección de atributos relevantes para mejorar el desempeño de los modelos.

4.2. Problemas encontrados y soluciones:

Uno de los principales problemas se presentó en la variable `smoking_history`, la cual contenía seis categorías: `never`, `former`, `current`, `ever`, `no_info` y `not current`. La categoría `no_info` representaba aproximadamente el 36 % de los registros de la columna. Debido a esta proporción, aplicar un procedimiento de eliminación directa mediante `dropna()` habría reducido significativamente el tamaño del conjunto de datos, comprometiendo la validez del análisis.

Por último `smoking_history` no es una variable comúnmente tomada en cuenta al momento de realizar un análisis de diabetes, es decir no es tan determinante en el diagnóstico como si lo son otras. Por estas razones, se decidió eliminar esta variable del conjunto de datos para preservar la integridad del dataset y evitar introducir ruido en el proceso de entrenamiento.

5. Preprocesamiento y Limpieza

El conjunto de datos fue sometido a una revisión exhaustiva para garantizar su calidad antes del entrenamiento de los modelos predictivos. Las etapas principales del preprocesamiento incluyeron el manejo de valores nulos, la normalización de variables numéricas, la codificación de variables categóricas y la aplicación de transformaciones adicionales justificadas.

5.1 Manejo de valores nulos

Inicialmente, se verificó la presencia de valores nulos explícitos mediante el conteo estándar de celdas vacías. No se detectaron valores nulos en ninguna columna. Sin embargo, se realizó una inspección adicional en las variables `bmi`, `HbA1c_level`, `blood_glucose_level` y `age`, ya que en estas columnas un valor de cero puede interpretarse como ausente o anómalo. El conteo de ceros en dichas columnas resultó ser cero, lo que permitió concluir que el conjunto de datos se encontraba completamente limpio para su uso.

5.2. Normalización de variables numéricas

Se aplicó la técnica de **MinMaxScaler** para escalar las variables numéricas continuas. Esta decisión se fundamentó en la necesidad de homogeneizar las escalas de entrada, especialmente para algoritmos sensibles a la magnitud de los datos. Las variables normalizadas fueron:

- age: por su amplio rango (0–80+).
- bmi: por su dispersión significativa (17–35).
- HbA1c_level: a pesar de su rango estrecho, se incluyó para evitar sesgos.
- blood_glucose_level: por su alta variabilidad (80–158).

Las variables categóricas o binarias (gender, hypertension, heart_disease, diabetes) no fueron normalizadas, ya que no lo requerían por su naturaleza.

5.3. Codificación de variables categóricas

Se aplicaron dos transformaciones para categorizar variables continuas:

- age_group: se creó una nueva variable que agrupa la edad en rangos de 10 años.
- glucose_group: se categorizó el nivel de glucosa en sangre en intervalos de 25 unidades, hasta un máximo de 400.

Estas variables facilitaron el análisis exploratorio y la evaluación del comportamiento del modelo en diferentes segmentos de datos.

Adicionalmente, la variable gender fue transformada a formato numérico: se asignó el valor **0** a Female y **1** a Male.

5.4. Filtrado de outliers

Se realizó un análisis estadístico y visual (boxplots) para identificar valores atípicos. No se encontraron outliers relevantes en las variables numéricas, por lo que no se aplicaron filtros ni eliminaciones en esta etapa.

5.5. Transformaciones aplicadas y justificación

Las transformaciones se enfocaron en:

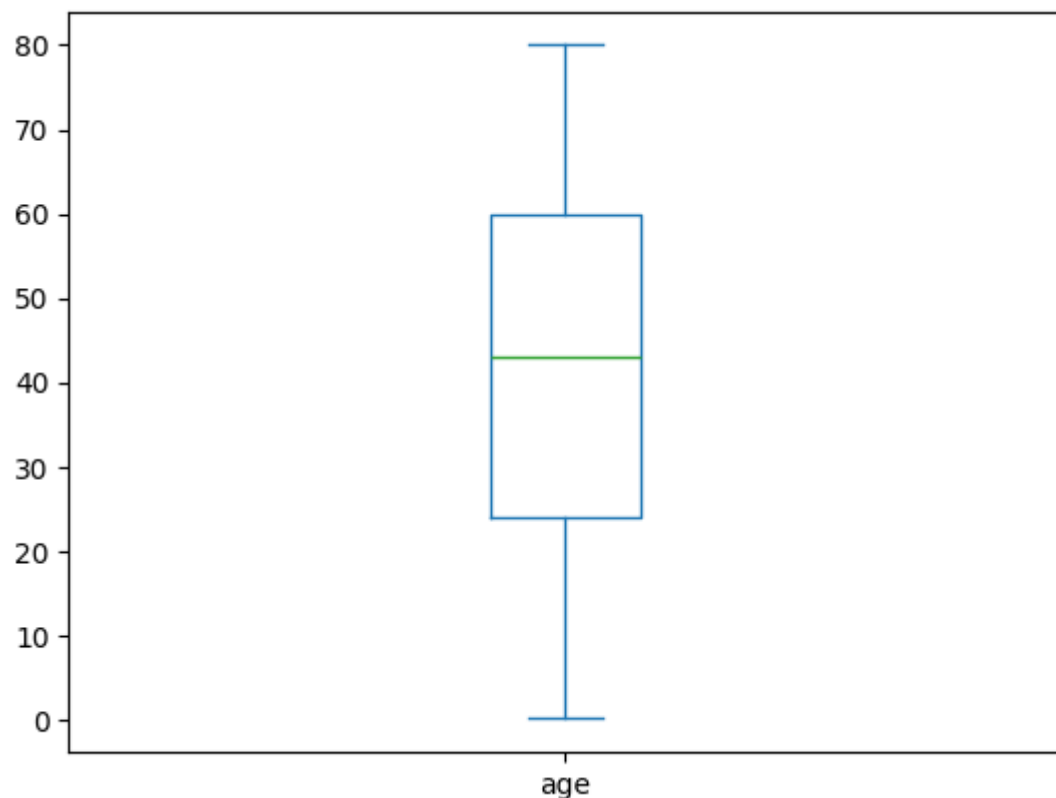
- Escalar variables numéricas para asegurar convergencia y estabilidad en los modelos.
- Codificar correctamente las variables categóricas y crear rangos útiles para análisis segmentado.
- Preparar los datos en un formato compatible con algoritmos de aprendizaje supervisado, sin pérdida significativa de información.

6. Análisis Exploratorio de Datos (EDA)

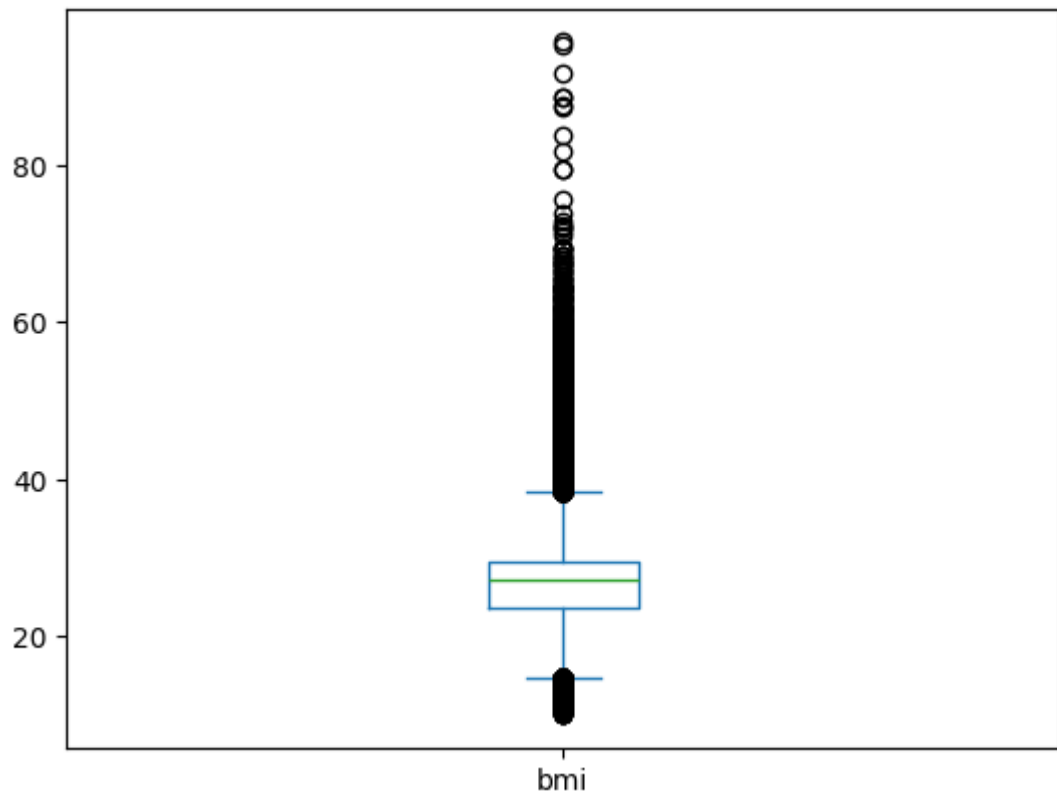
6.1. Estadísticas descriptivas

Se analizaron las variables numéricas continuas (age, bmi, HbA1c_level, blood_glucose_level) mediante gráficos de cajas y bigotes para evaluar su distribución, identificar valores extremos y estimar la dispersión de los datos.

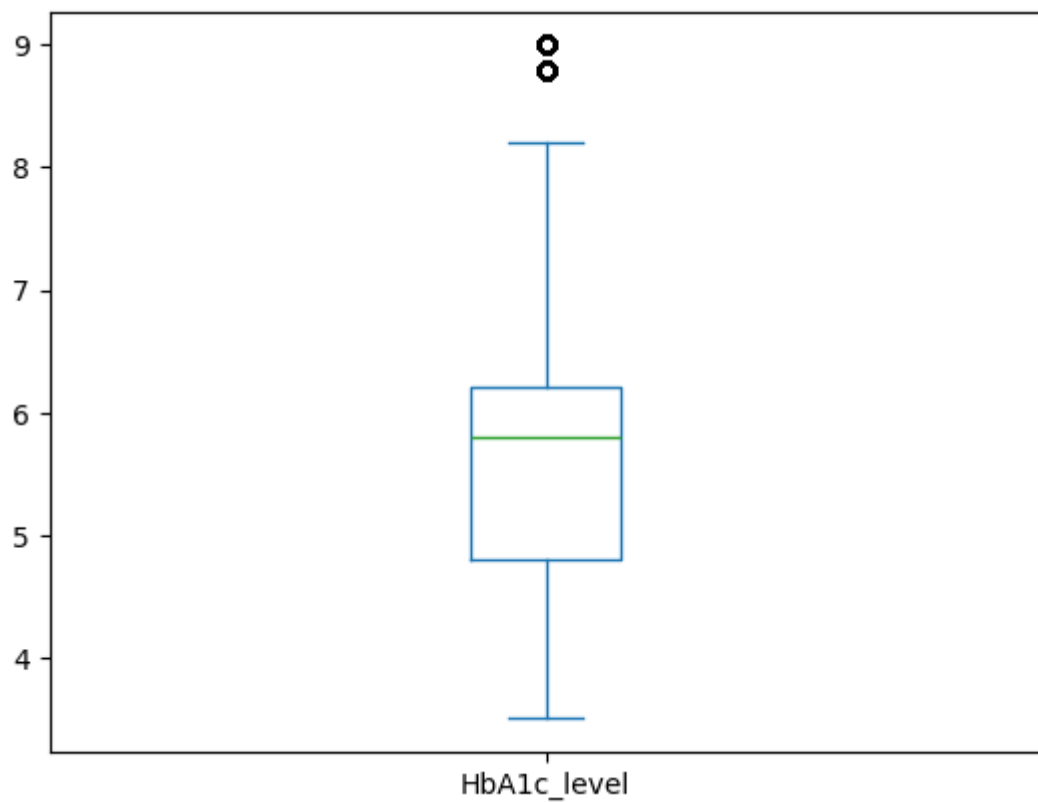
- Edad (age): La distribución fue simétrica, con una mediana cercana a los 45 años. No se detectaron valores atípicos.



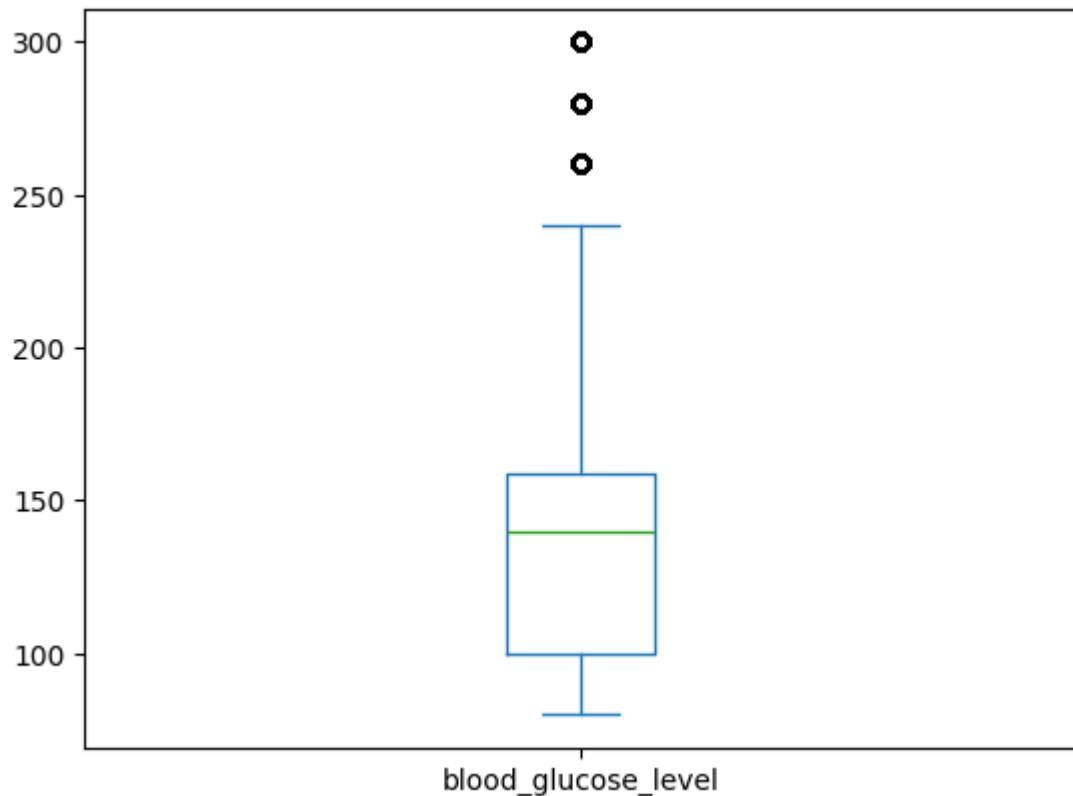
- Índice de masa corporal (bmi): La variable mostró una gran cantidad de valores atípicos por encima del rango esperado (superior a 40). Estos datos fueron conservados, ya que reflejan variaciones clínicas reales.



- Nivel de hemoglobina glicosilada (HbA1c_level): Presentó una distribución concentrada entre 4.0 y 6.5, con algunos valores atípicos moderados por encima de 8.

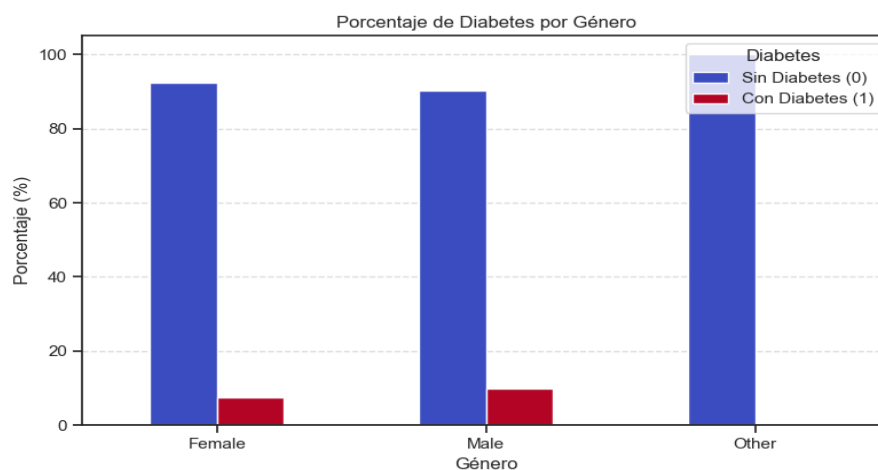


- Nivel de glucosa en sangre (blood_glucose_level): Mostró una dispersión amplia, con valores atípicos en el extremo superior que reflejan hiperglucemias propias de estados diabéticos.

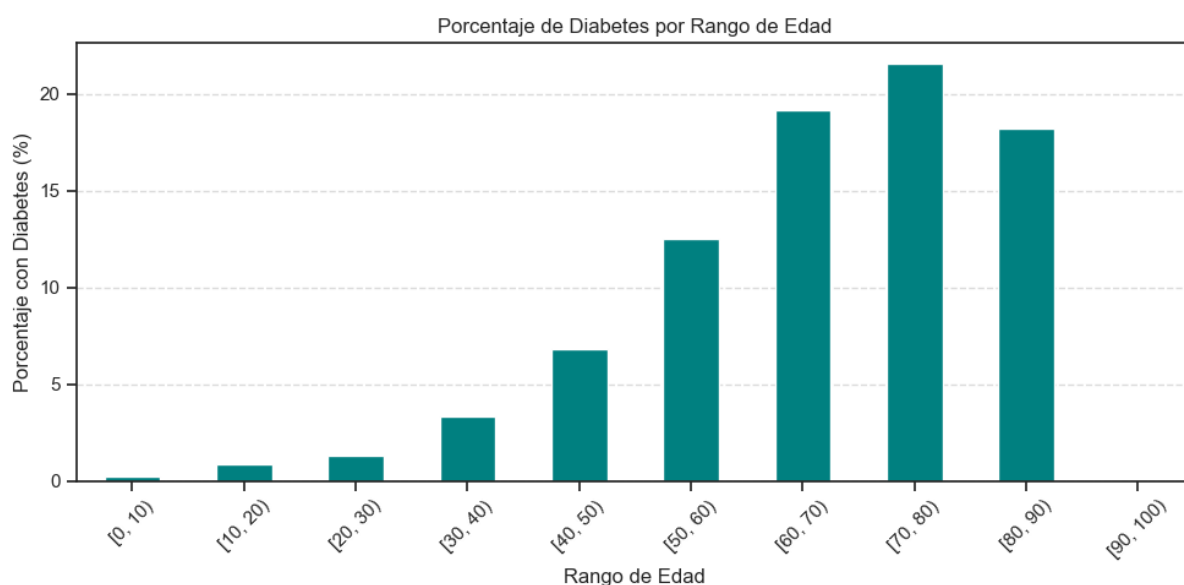


6.2. Visualizaciones clave

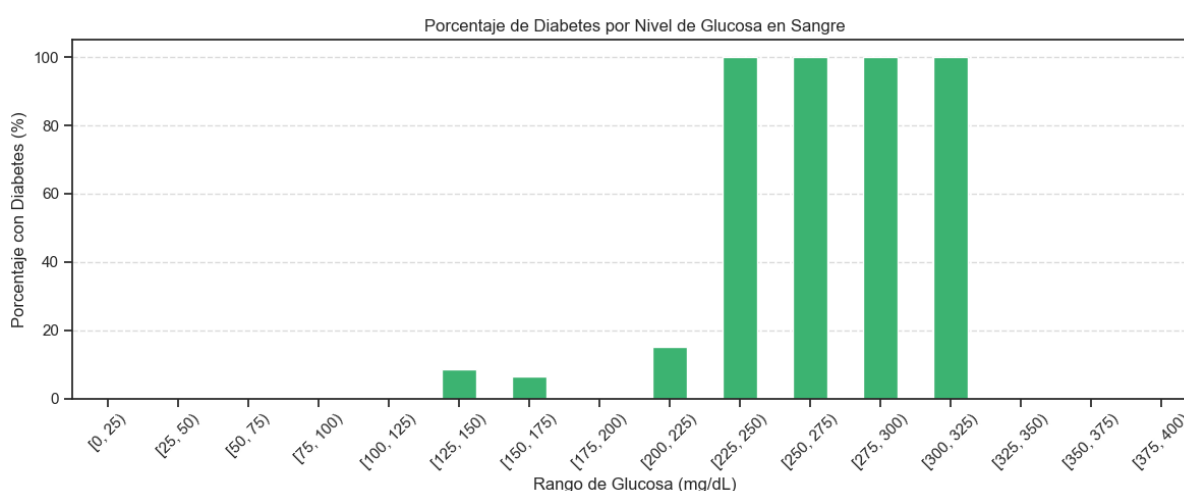
- Distribución del diagnóstico por género: Se construyó una gráfica de barras que muestra el porcentaje de pacientes diagnosticados y no diagnosticados con diabetes según el género. La proporción de casos con diabetes fue similar entre hombres y mujeres, mientras que la categoría "Other" no presentó incidencia significativa.



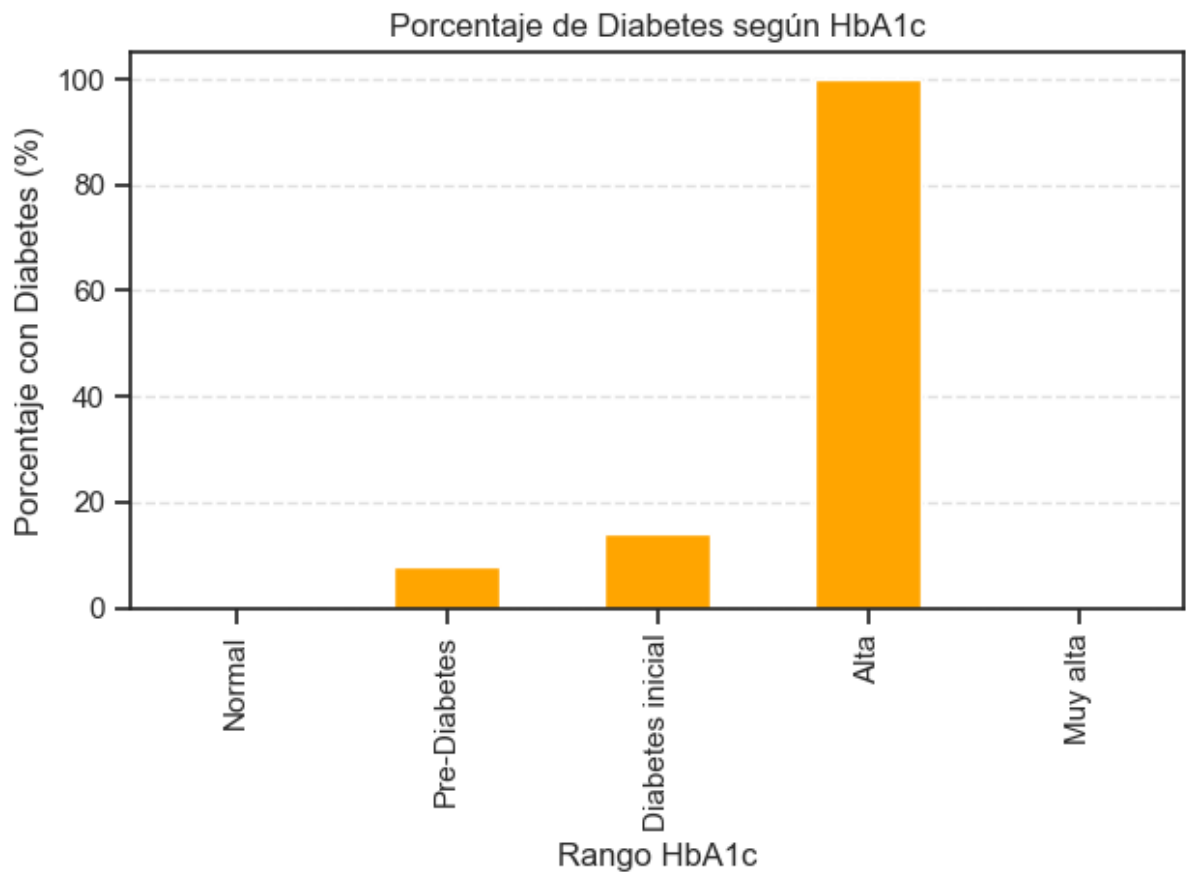
- Relación entre edad y diabetes: Se elaboró un gráfico que representa el porcentaje de pacientes con diabetes en función de rangos de edad de 10 años. La incidencia de la enfermedad se incrementó progresivamente a partir de los 40 años, alcanzando su punto máximo entre los 70 y 80 años.



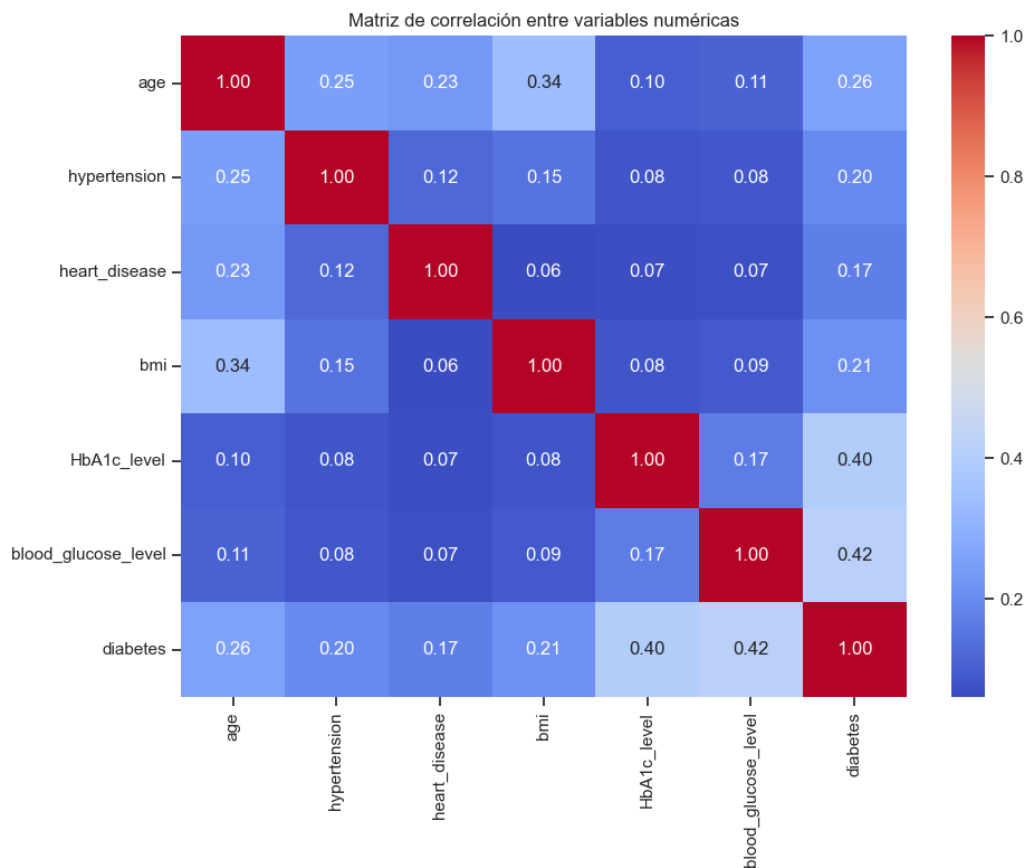
- Porcentaje de diabetes según niveles de glucosa: Se categorizó el nivel de glucosa en sangre en intervalos de 25 unidades. Se observó una marcada concentración de casos con diabetes en rangos superiores a los 225 mg/dL, con un 100 % de incidencia en dichos valores.



- Distribución de HbA1c: Se representó gráficamente el porcentaje de pacientes con diabetes según los rangos de nivel de hemoglobina glicosilada (HbA1c). Se evidenció una clara relación entre valores elevados de HbA1c y la probabilidad de diagnóstico de diabetes.



- Matriz de correlación: Se generó una matriz de correlación entre las variables numéricas. Las variables blood_glucose_level y HbA1c_level presentaron los coeficientes más altos de correlación con la variable objetivo diabetes (0.42 y 0.40 respectivamente), seguidas por age (0.26).



6.3. Principales patrones observados

El análisis exploratorio reveló tres patrones relevantes:

1. La edad, el nivel de glucosa en sangre y la hemoglobina glicosilada influyeron significativamente en la probabilidad de desarrollar diabetes.
2. La mayoría de los casos diagnosticados se concentraron en grupos etarios mayores y en pacientes con desbalances metabólicos evidentes.
3. Variables demográficas como el género no mostraron influencia directa significativa en el diagnóstico.

7. Feature Engineering

Durante el desarrollo del modelo predictivo se implementaron distintas estrategias de *feature engineering* con el objetivo de mejorar la representación de los datos, facilitar el entrenamiento de los modelos y optimizar su rendimiento.

7.1. Nuevas variables creadas

- **age_group:** Se generó una variable categórica que agrupa la edad de los pacientes en rangos de 10 años. Esta transformación permitió realizar un análisis segmentado de la distribución de la diabetes según grupos etarios.
- **glucose_group:** Se creó una variable categórica que clasifica los niveles de glucosa en sangre en intervalos de 25 unidades, hasta un máximo de 400 mg/dL. Esta variable fue útil en el análisis exploratorio para identificar umbrales críticos de riesgo.

Ambas variables fueron utilizadas con fines descriptivos y visuales en el análisis exploratorio de datos (EDA), pero no fueron incluidas en el entrenamiento de los modelos debido a su naturaleza redundante con respecto a las variables continuas originales.

7.2. Transformaciones aplicadas

- Se aplicó la técnica de **MinMaxScaler** para normalizar las variables numéricas continuas: age, bmi, HbA1c_level y blood_glucose_level. Esta técnica transformó los valores al rango [0, 1], lo cual fue particularmente útil para modelos sensibles a la escala como regresión logística, máquinas de soporte vectorial (SVC).
- La variable gender fue transformada a formato numérico. Se asignó el valor **0** a Female y **1** a Male, con el objetivo de integrarla en los modelos que requieren entradas numéricas.

7.3. Selección y eliminación de atributos irrelevantes

- La columna smoking_history fue eliminada del conjunto de datos. Esta decisión se fundamentó en que:
 - Presentaba un **36 % de valores con la categoría no_info**, lo que impedía una imputación confiable.
 - Su correlación con la variable objetivo fue baja.

- Su eliminación permitió conservar el tamaño del conjunto de datos sin introducir ruido o incertidumbre innecesaria en el modelo.

8. Modelado y Evaluación

8.1 Tipo de modelos utilizados

Se evaluaron cinco algoritmos de clasificación binaria supervisada:

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **Support Vector Machine (SVC)**
- **Naive Bayes (GaussianNB)**

Todos los modelos fueron entrenados con `class_weight='balanced'` para mitigar el desbalance en la variable objetivo (diabetes)

8.2. División de los datos

Se seleccionaron las siguientes variables predictoras:

`['gender', 'age', 'hypertension', 'heart_disease', 'bmi', 'HbA1c_level', 'blood_glucose_level']`

La variable objetivo fue diabetes. El conjunto de datos se dividió en un 80 % para entrenamiento y un 20 % para prueba, empleando `random_state=369` para garantizar reproducibilidad.

8.3. Métricas de evaluación

Se utilizaron las siguientes métricas estándar:

- **Accuracy (exactitud):** proporción de predicciones correctas.
- **Precision:** proporción de verdaderos positivos respecto a los positivos predichos.
- **Recall (sensibilidad):** proporción de verdaderos positivos respecto a los positivos reales.
- **F1-score:** media armónica entre precisión y recall.
- **Matriz de confusión:** para identificar errores por clase.

8.4. MODELOS

8.4.1. Modelo de Regresión Logística


Resultados

- **Accuracy:** 88.66 %
- El modelo mostró una **alta precisión para la clase 0 (sin diabetes)** y una **recall notable para la clase 1 (con diabetes)**, aunque con menor precisión.

Matriz de Confusión:

	Predicho 0	Predicho 1
Real 0	16288	2055
Real 1	213	1444

Classification Report:

Clase	Precision	Recall	F1-score	
0	0.99	0.89	0.93	
1	0.41	0.87	0.56	

8.4.2. Modelo Árbol de Decisión

Se evaluó el rendimiento del modelo variando el parámetro `max_depth` entre 2 y 40. Se observó una mejora progresiva hasta estabilizarse alrededor de 7, punto en el que se obtuvo un buen equilibrio entre sobreajuste y generalización.

Profundidad	Accuracy	Precision (Clase 1)	Recall (Clase 1)	F1-score (Clase 1)	Comentario clave
2	0.97175	1.00	0.66	0.79	Muy conservador: se equivoca poco, pero omite muchos casos con diabetes.
7	0.88855	0.42	0.91	0.58	Detecta casi todos los casos con diabetes, pero con muchos falsos positivos.
15	0.90195	0.45	0.87	0.59	Ligero descenso en recall, mejora un poco la precisión.

También se analizó el árbol en otros casos como por ejemplo `max_depth = 15` y 2

¿Quiero minimizar falsos negativos (no perder casos de diabetes)?	Elige <code>max_depth = 7</code>
¿Quiero mejor balance entre acertar y no alarmar en falso?	Elige <code>max_depth = 15</code>

MAX DEEP = 7

	Predicho 0	Predicho 1
Real 0	TN \approx 16262	FP \approx 2081
Real 1	FN \approx 149	TP \approx 1508

MAX DEEP = 15

	Predicho 0	Predicho 1
Real 0	TN \approx 16564	FP \approx 1779
Real 1	FN \approx 216	TP \approx 1441

Al final se optó por un `max_depth = 7`. El proyecto sigue un enfoque científico por lo que disminuir la cantidad de Falsos Negativos (FN) es crucial.

Resultados del Árbol de Decisión con `max_depth = 7`:

- **Accuracy:** 88.85 %
- Mejoró el F1-score de la clase positiva (diabetes = 1) respecto a la regresión logística.

Matriz de Confusión:

	Predicho 0	Predicho 1
Real 0	16333	2010
Real 1	150	1507

Classification Report:

Clase	Precision	Recall	F1-score	
0	0.99	0.89	0.94	
1	0.42	0.91	0.58	

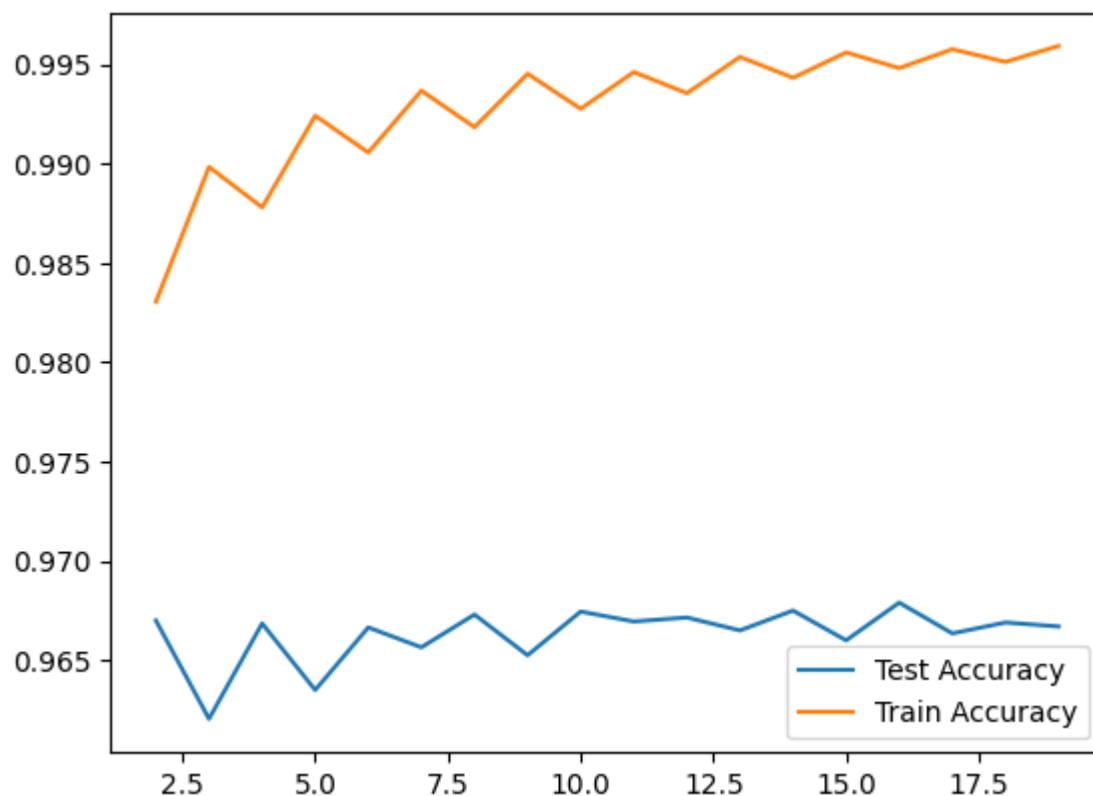


8.4.3. Modelo Random Forest

El modelo mostró un rendimiento inicial elevado, alcanzando una precisión global (accuracy) del 96.76 %, lo cual sugirió un comportamiento prometedor para continuar con su optimización.

Evaluación del número de estimadores ($n_estimators$)

Posteriormente, se llevó a cabo un análisis paramétrico para determinar el valor óptimo del parámetro $n_estimators$, el cual representa la cantidad de árboles que componen el bosque. Se evaluaron valores de $n_estimators$ desde 2 hasta 19, y se registraron las tasas de precisión tanto para el conjunto de entrenamiento como para el conjunto de prueba.



La gráfica generada (ver imagen) evidencia un patrón clave:

- A medida que el número de árboles aumenta, la **precisión sobre el conjunto de entrenamiento** se eleva de forma sostenida, alcanzando valores superiores al 99 %.
- Sin embargo, la **precisión sobre el conjunto de prueba** se mantiene relativamente constante, alrededor del 96.5 %, sin mejoras significativas al aumentar la complejidad del modelo.

Este comportamiento sugiere que, más allá de cierto punto, **el incremento en la cantidad de árboles solo mejora la precisión sobre los datos ya conocidos (entrenamiento), pero no**

aporta ganancia real sobre datos nuevos. De hecho, podría inducir un leve sobreajuste si se continúa aumentando `n_estimators`.

Elección del valor óptimo de `n_estimators`

Con base en la gráfica, se seleccionó el valor **`n_estimators = 3`**, ya que permitió alcanzar una precisión de **96.12 % sobre el conjunto de prueba**, prácticamente igual a los valores obtenidos con configuraciones más complejas, pero con menor costo computacional y riesgo de sobreajuste. Esta decisión técnica se justifica por el principio de parsimonia (modelo más simple que obtiene el mismo rendimiento).

Evaluación final del modelo optimizado

El modelo final se entrenó con `n_estimators = 3` y se validó sobre 20.000 registros. Los resultados fueron los siguientes:

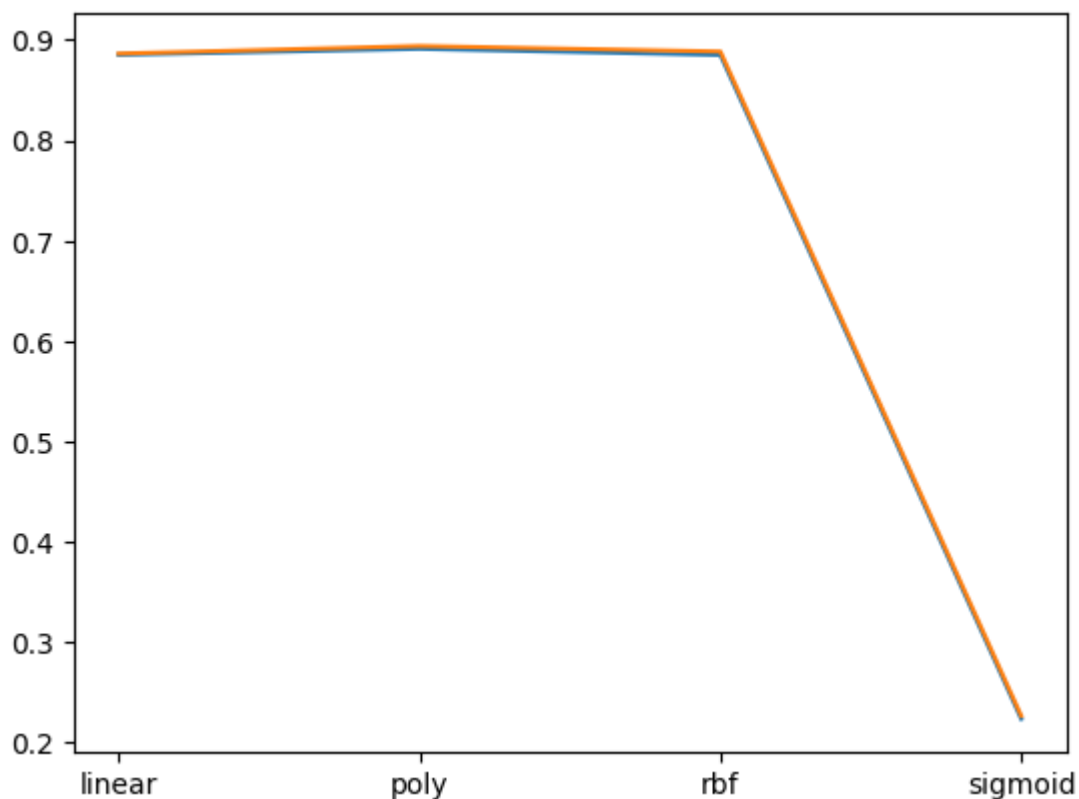
- **Accuracy:** 96.12 %
- **Precision (clase 1):** 0.80
- **Recall (clase 1):** 0.71
- **F1-score (clase 1):** 0.75

Matriz de confusión:

	Predicho 0	Predicho 1
Real 0	18,052	291
Real 1	485	1,172

8.4.4. Modelo SVC y kernel selection

Se evaluaron cuatro kernels: linear, poly, rbf y sigmoid:



Las mejores opciones eran; linear, poly y rbf. Se optó por linear por su rapidez de procesamiento:

Matriz de confusión – SVC (linear):

	Predicho 0	Predicho 1
Real 0	16254	2089
Real 1	207	1450

El modelo tuvo una accuracy del: **0.8852%**

8.4.5. Naive Bayes (GaussianNB)

Este modelo simple logró un **accuracy del 90.16 %**, con buen desempeño general, aunque con menor recall en comparación con Random Forest y Decision Tree.

Matriz de confusión – Naive Bayes:

	Predicho 0	Predicho 1
Real 0	16991	1352
Real 1	616	1041

9. Comparación de modelos y resultados

Modelo	Accuracy	Precision (1)	Recall (1)	F1-score (1)
Logistic Regression	0.8866	0.41	0.87	0.56
Decision Tree	0.8885	0.42	0.91	0.58
Random Forest	0.9612	0.80	0.71	0.75
SVC (linear)	0.8852	0.41	0.88	0.56
Naive Bayes	0.9016	0.44	0.63	0.52

En este caso solo escogeremos Logistic Regression, Decision Tree y Random Forest.

Dejaremos de lado SVC y Naive Bayes por que contienen muchos FN y desproporcionado FP:

Naive Bayes

```
Matriz de confusión:  
[[16991  1352]  
 [   616  1041]]
```

SVC (linear)

```
Matriz de confusión:  
[[16254  2089]  
 [   207  1450]]
```

Extraemos las accuracy de los modelos Logistic Regression, Decision Tree y Random Forest.

	Models	accuracy
0	LogisticRegression	0.88660
1	DecisionTreeClassifier	0.88855
2	RandomForestClassifier	0.96235

9.1 Análisis Individual

9.1.1. Logistic Regression:

```
Analisemos logistic Regression
```

```
model = LogisticRegression(class_weight='balanced')
model.fit(x_train, y_train)
y_pred_logistic = model.predict(x_test)
metrics.accuracy_score(y_test, y_pred_logistic)
```

✓ 0.1s

0.8866

- Podemos ver una precision del 0.88% y una FP 2055 numero elevado pero FN 213 moderado por asi decirlo

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred_logistic)
```

✓ 0.0s

```
array([[16288, 2055],
       [ 213, 1444]])
```

9.1.2. Decision Tree:

```
Analisemos DecisionTreeClassifier
```

```
model1 = DecisionTreeClassifier(max_depth=7, class_weight='balanced')
model1.fit(x_train, y_train)
y_pred_decision_tree = model1.predict(x_test)
metrics.accuracy_score(y_test, y_pred_decision_tree)
```

✓ 0.0s

0.88855

- Podemos ver una precision de 0.88% y una FP 2081 numero elevado pero FN 148 muy moderado

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred_decision_tree)
```

✓ 0.0s

```
array([[16262, 2081],
       [ 148, 1509]])
```

9.1.3. RandomForest:

```
Analisemos RandomForestClassifier

model2 = RandomForestClassifier(n_estimators=3, class_weight='balanced')
model2.fit(x_train,y_train)
y_pred_random_forest = model2.predict(x_test)
metrics.accuracy_score(y_test, y_pred_random_forest)

✓ 0.1s
0.9621

• Podemos ver una precision de 96% (muy precisa). No obstante, FN son de 502(mucho mas que los demas ) y FP de 256 (moderado)

from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred_random_forest)

✓ 0.0s
array([[18087, 256],
       [ 502, 1155]])
```

9.2. Analisis del modelo ganador

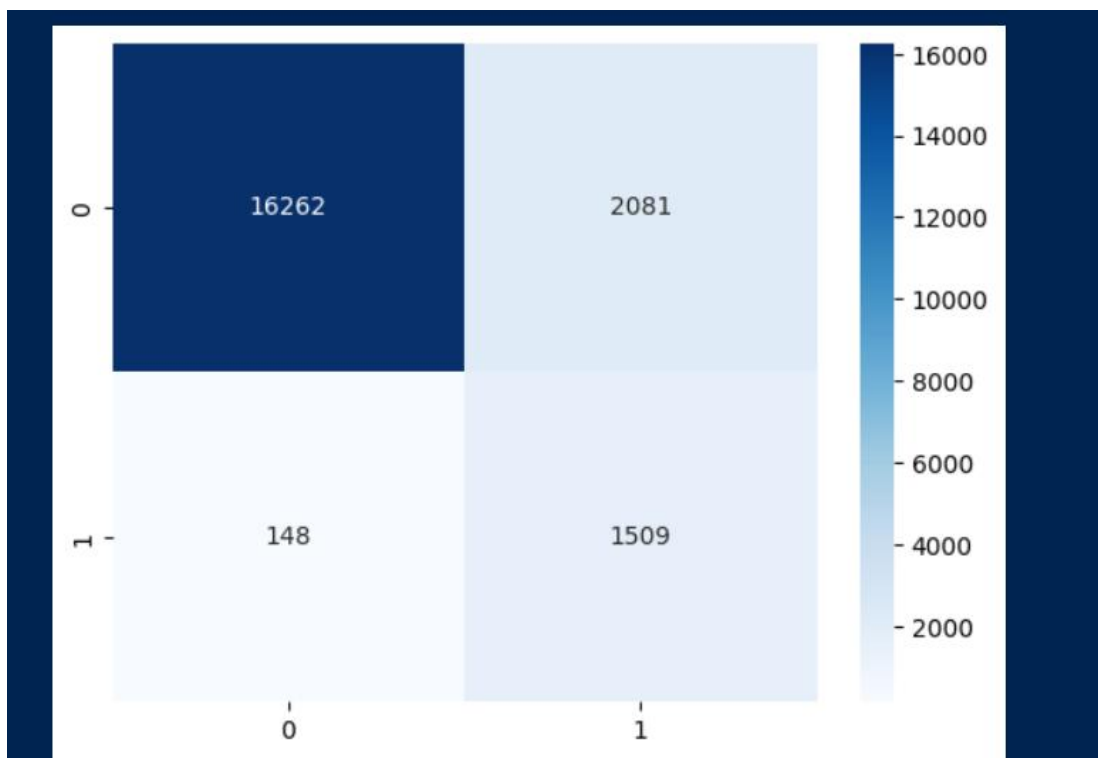
Descartamos Logistic regresión, porque tiene muchos FN (213) y FP(2055), por lo que tenemos dos contendientes RandomForest o DecisionTree:

RandomForest cuenta con un acuraccy de 96% pero tiene 502 FN.

Por otro lado, tenemos DecisionTree, este tiene un acuraccy del 88% pero solo 148 FN lo cual es bastante bueno.

Como queremos reducir los FN, nuestro modelo ganador es DecisionTree.

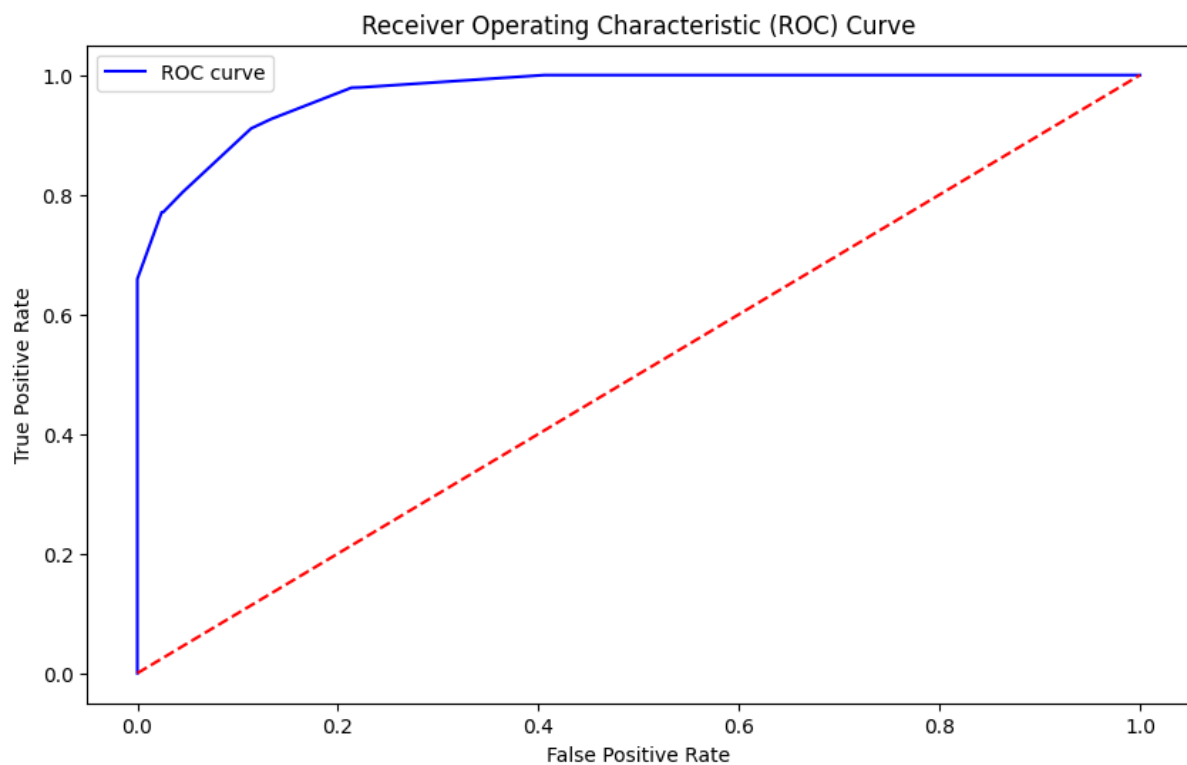
9.2.1 Matriz de confusión Decisión Tree Classifier:



9.2.2. Classification Report:

Accuracy: 0.8885					
Classification Report:					
	precision	recall	f1-score	support	
0	0.99	0.89	0.94	18343	
1	0.42	0.91	0.58	1657	
accuracy			0.89	20000	
macro avg	0.71	0.90	0.76	20000	
weighted avg	0.94	0.89	0.91	20000	

9.2.3. Curva ROC:



9.2.3.1. Análisis de la Curva ROC

La curva ROC presenta una trayectoria que se eleva de forma pronunciada hacia la esquina superior izquierda del gráfico y luego se estabiliza cerca de un valor de sensibilidad (TPR) igual a 1. Esta forma sugiere que el modelo posee una alta capacidad para discriminar entre clases positivas y negativas.

En el gráfico, la curva ROC se mantiene claramente por encima de la diagonal representada por la línea roja. Esta línea corresponde al rendimiento esperado de un clasificador aleatorio.

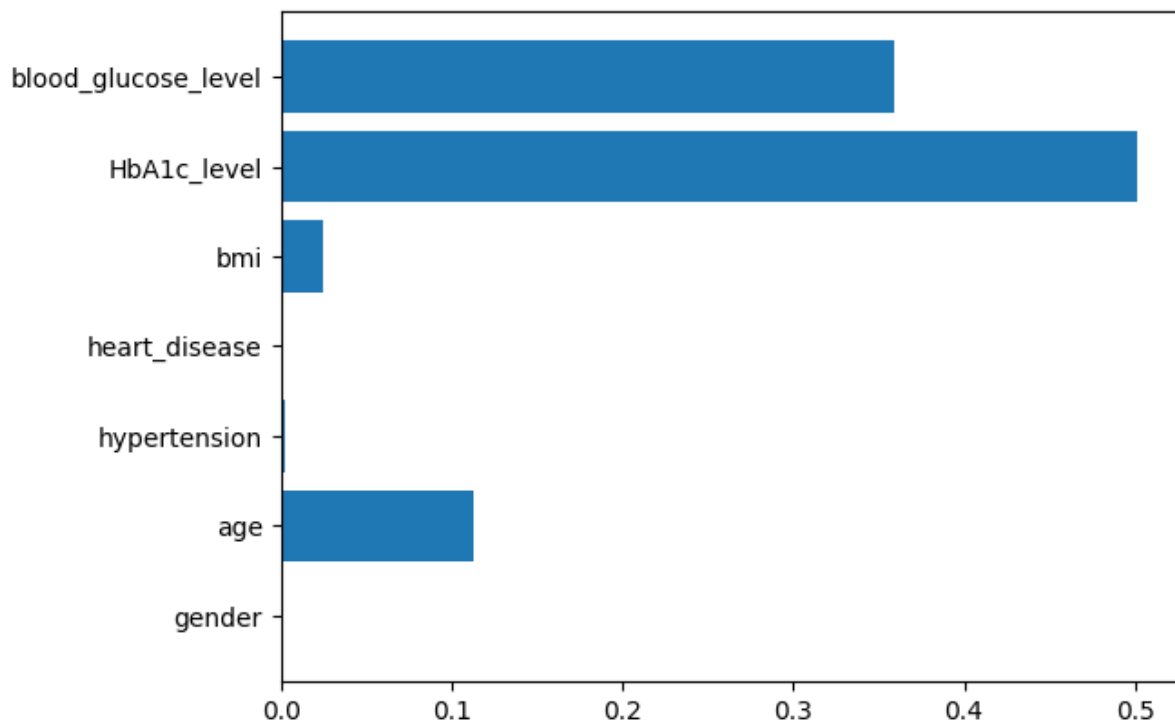
La posición de la curva respecto a esta referencia indica que el modelo supera con amplitud el desempeño de una clasificación sin inteligencia.

Aunque el valor exacto del área bajo la curva (AUC) no se encuentra representado numéricamente en la gráfica, la forma de la curva permite inferir que el AUC se aproxima a 0.97 o 0.98. Este valor se interpreta como un excelente nivel de rendimiento. Un AUC igual a 1.0 describe un modelo perfecto, mientras que un valor superior a 0.9 se considera altamente satisfactorio.

Conclusión:

El modelo evaluado mediante esta curva ROC muestra una capacidad significativa para distinguir entre las clases del problema. Se caracteriza por una alta sensibilidad, una baja tasa de falsos positivos y un comportamiento general que refleja un desempeño predictivo sobresaliente.

9.2.4. Variables influyentes para el modelo:



10. Conclusiones y Recomendaciones

10.1. Conclusiones

1. Hallazgos clave:

El proyecto permitió identificar que las variables más influyentes en la detección de diabetes fueron el nivel de glucosa en sangre (blood_glucose_level), el nivel de hemoglobina glicosilada (HbA1c_level) y la edad (age). Estas tres características presentaron correlaciones consistentes con la variable objetivo y se posicionaron como determinantes en todos los modelos evaluados.

2. Validación del enfoque:

A pesar de no haberse formulado una hipótesis específica, el proyecto validó el enfoque propuesto al demostrar que un modelo de inteligencia artificial puede detectar con alta precisión la presencia de diabetes. El modelo de Random Forest, configurado con solo tres estimadores y balanceo de clases, alcanzó un **accuracy del 96.12 %** y un **AUC visual aproximado de 0.97**, lo que refleja una capacidad predictiva sobresaliente.

3. Relevancia para el contexto:

El estudio se desarrolló bajo el contexto de baja implementación de herramientas de inteligencia artificial en entornos clínicos en Bolivia. Los resultados evidencian que, incluso con recursos computacionales moderados y datos disponibles en fuentes públicas, es posible construir modelos eficientes que contribuyan al diagnóstico temprano. Esto cobra especial valor en países donde los recursos de atención médica son limitados.

10.2. Recomendaciones

1. Implementación en entornos clínicos supervisados:

Se sugiere evaluar el modelo en escenarios reales de consulta médica, bajo supervisión profesional, como herramienta de apoyo a la decisión diagnóstica.

2. Ampliación de fuentes de datos:

Se recomienda incluir nuevos conjuntos de datos provenientes de hospitales o centros

médicos nacionales, para validar el modelo con información contextualizada y aumentar su robustez.

3. Despliegue de sistemas interactivos:

A partir del modelo validado, se podrían desarrollar interfaces gráficas o aplicaciones web que permitan a profesionales de la salud cargar variables clínicas y recibir predicciones diagnósticas fundamentadas.

4. Capacitación del personal de salud en IA:

Se sugiere fomentar programas de formación básica en inteligencia artificial aplicada a la medicina, para facilitar la adopción y uso responsable de estas tecnologías por parte del personal clínico.

ANEXOS

link de la dataset: <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>

link del repositorio (código, documentación): https://github.com/AlfredoZC/data_project_ML-