



# RiskAlert: Diabetes

**Manuel Gutiérrez Tangarife, Estiven Ospina  
Jaramillo, Juan Pablo Ramírez Betancur**

Proyecto Final Bootcamp IA nivel Explorador

# ¿Qué es la diabetes?

- La diabetes es una enfermedad crónica que ocurre cuando el cuerpo no puede regular bien el nivel de azúcar en la sangre.
- Esto sucede porque no produce suficiente insulina o no la usa correctamente.
- Con el tiempo, puede causar daños graves al corazón, los riñones, la vista y otros órganos, incluso si no hay síntomas al principio.

# La motivación detrás del proyecto

- +537 millones de personas viven con diabetes en el mundo.
- 1 de cada 2 no sabe que la tiene.
- Cada 5 segundos alguien es diagnosticado.
- Cada 8 segundos, alguien muere por sus complicaciones.
- La diabetes no da señales claras al inicio, pero puede causar daños graves.
- ¿Y si pudiéramos predecir el riesgo antes de que sea tarde?

# Problema: Diagnóstico tardío de la diabetes

- Muchas personas viven con diabetes sin saberlo.
- La enfermedad avanza sin síntomas claros al inicio.
- El diagnóstico suele llegar cuando ya hay complicaciones.
- Esto dificulta el tratamiento y aumenta los riesgos.
- La detección temprana puede salvar vidas.

# Fuente de datos

<https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-data-set>

## Features

Independiente(Target): 'tipo' 0:no diabetico. 1:pre diabetico 2: diabetico

Dependiente: 'presion\_alterial\_alta', 'colesterol\_alto', 'colesterol\_chequeado', 'imc', etc. Que son datos de estilo de vida y descripción de las personas.

De las cuáles todas las variables son categóricas exceptuando la variable de índice de masa corporal (imc).

## Instancias

253680 Entrenamiento/prueba (70/30%) para un total de 100%--- evaluación datos

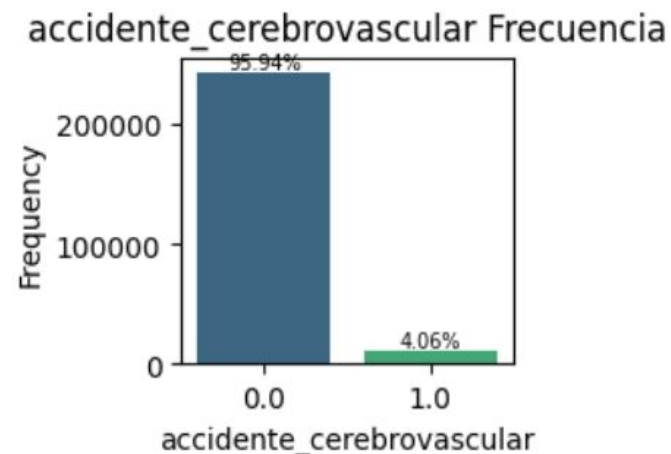
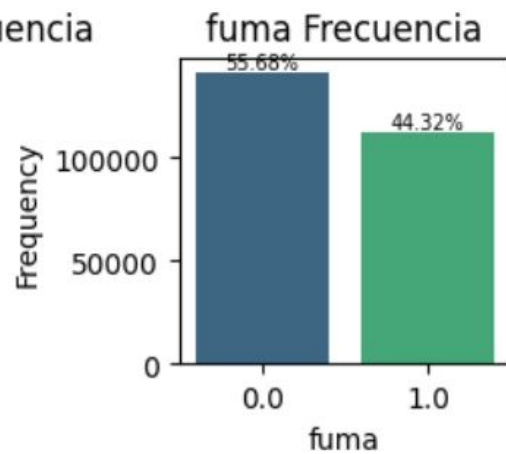
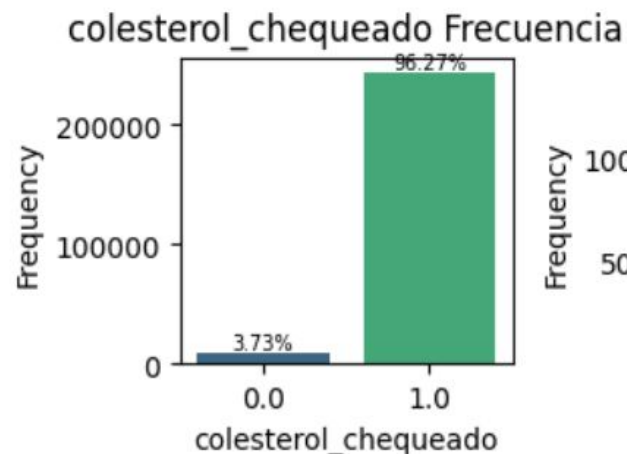
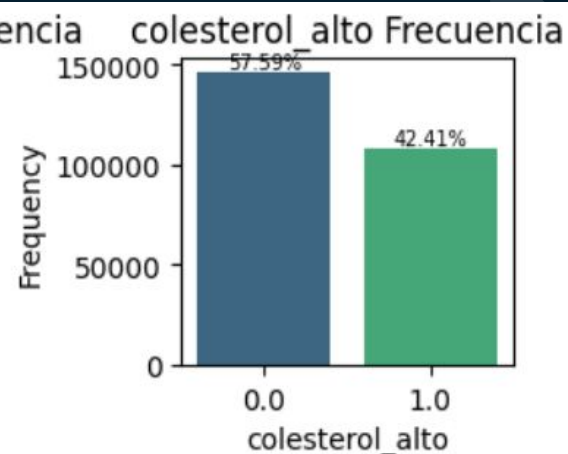
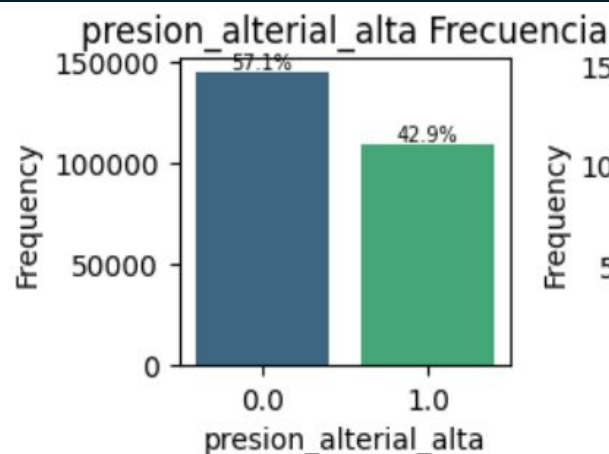
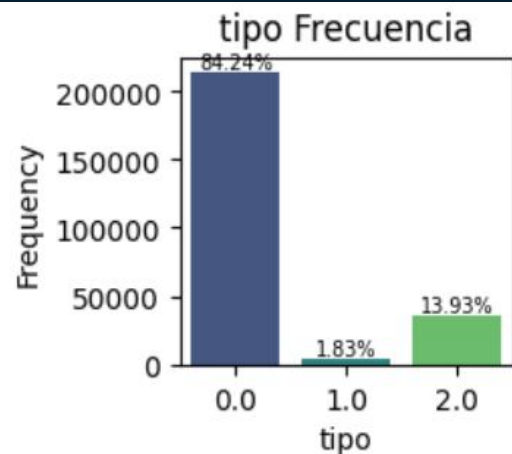
# Análisis Descriptivo y Exploratorio

## 1. Visualización de datos:

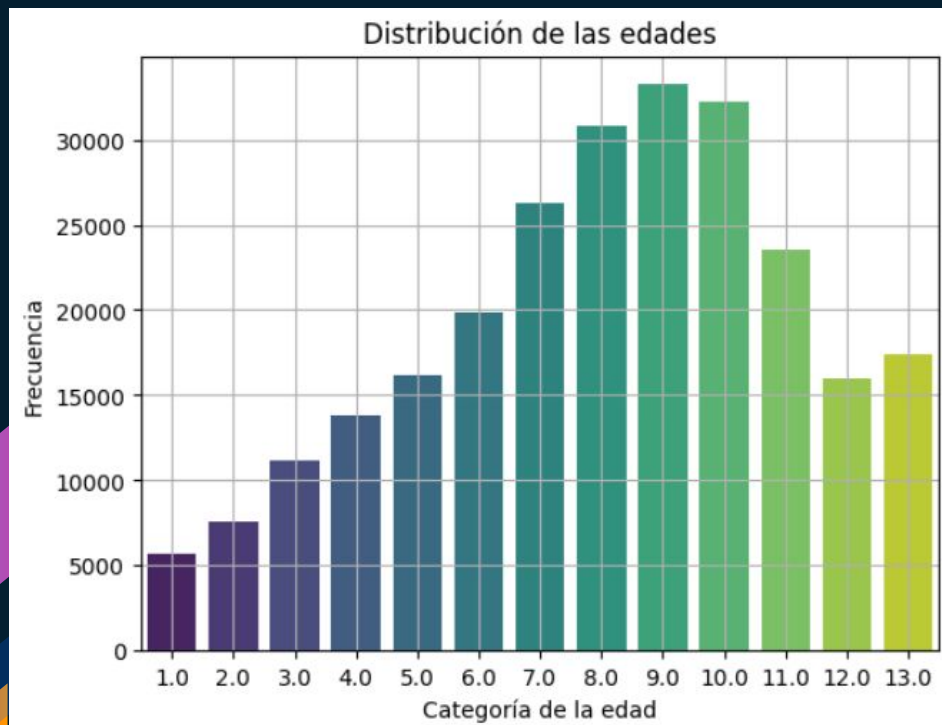
tipo	presion_alterial_alta	colesterol_alto	colesterol_chequeado	imc	fuma	accidente_cerebrovascular
0.0	1.0	1.0	1.0	40.0	1.0	0.0
0.0	0.0	0.0	0.0	25.0	1.0	0.0
0.0	1.0	1.0	1.0	28.0	0.0	0.0
0.0	1.0	0.0	1.0	27.0	0.0	0.0
0.0	1.0	1.0	1.0	24.0	0.0	0.0

No hay na's y los datos ya están numéricos pero ojo, se deben tratar como categóricos

## 1. Análisis univariado:



## 1. Análisis univariado:

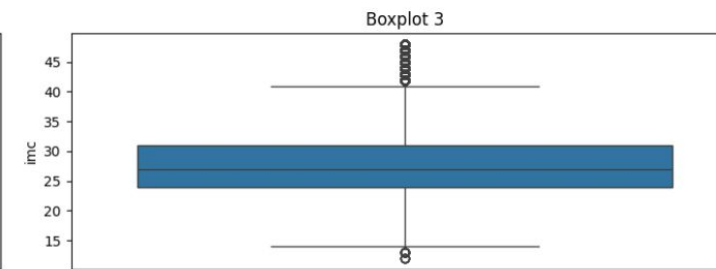
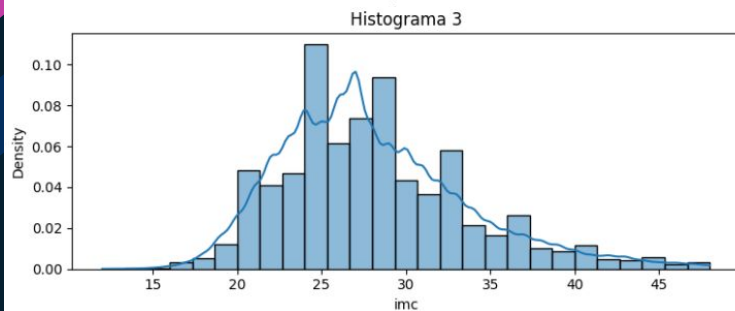
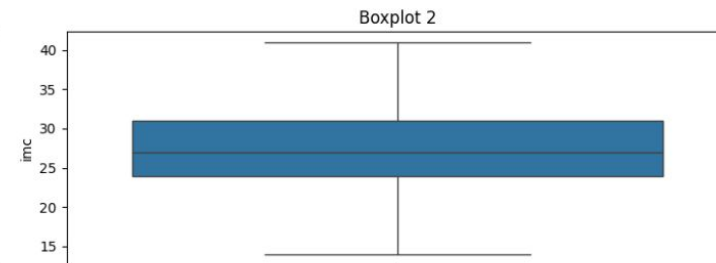
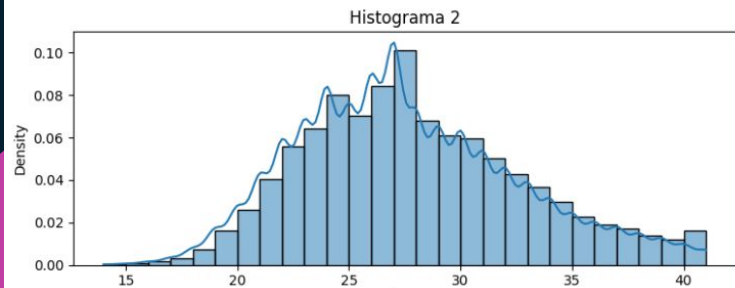
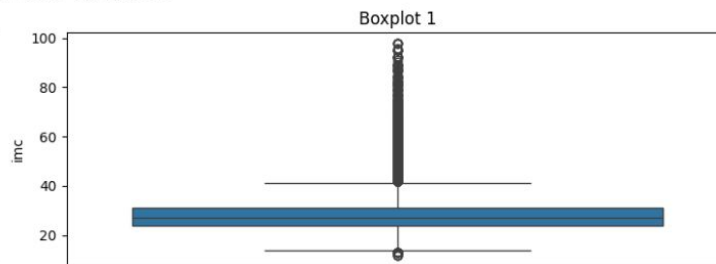
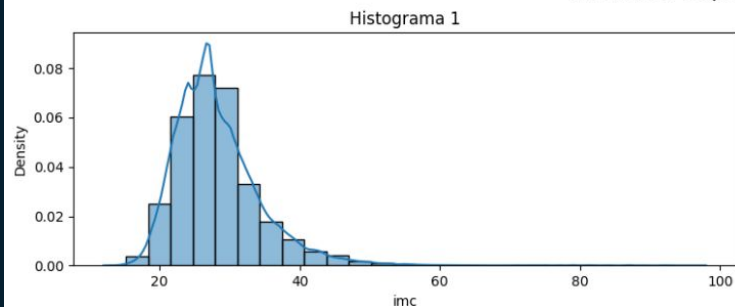


1. Edad 18 a 24
2. Edad 25 a 29
3. Edad 30 a 34
4. Edad 35 a 39
5. Edad 40 a 44
6. Edad 45 a 49
7. Edad 50 a 54
8. Edad 55 a 59
9. Edad 60 a 64
10. Edad 65 a 69
11. Edad 70 a 74
12. Edad 75 a 79
13. Edad 80 o más

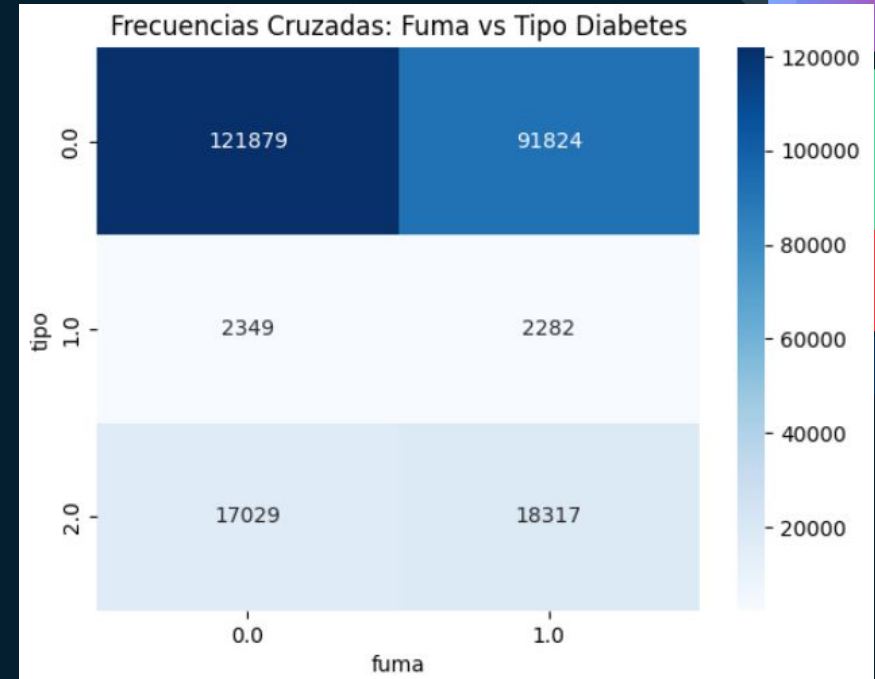
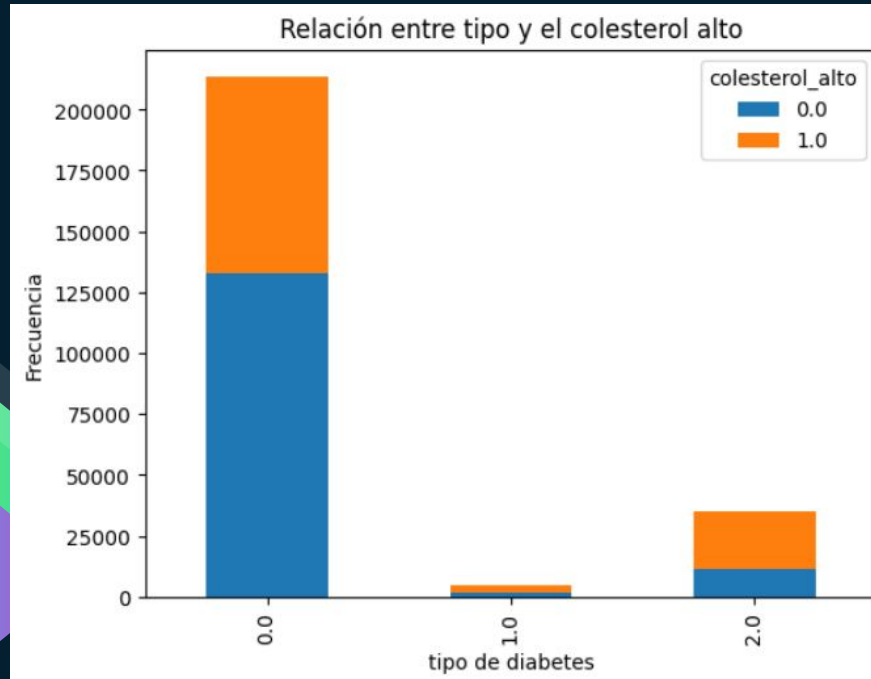


# 1. Análisis univariado (variable numérica):

Análisis IMC por base de datos



## 2. Análisis Bivariado :



# Desarrollo de la solución

Se implementó un modelo de clasificación de machine learning para intentar predecir la posibilidad de que un individuo sea diabético, prediabetico o no diabético.

Se llevaron a cabo varios escenarios por cada data frame seleccionado anteriormente que se separan en dos grupos

## No balanceados

Los algoritmos usados fueron:

1. Regresión logística
2. K vecinos más cercanos.
3. Árbol de decisión.

## Balanceados

Los algoritmos usados fueron:

1. Regresión logística
2. K vecinos más cercanos.
3. Árbol de decisión.
4. XGboost.
5. Light Gradient Boosting Machine

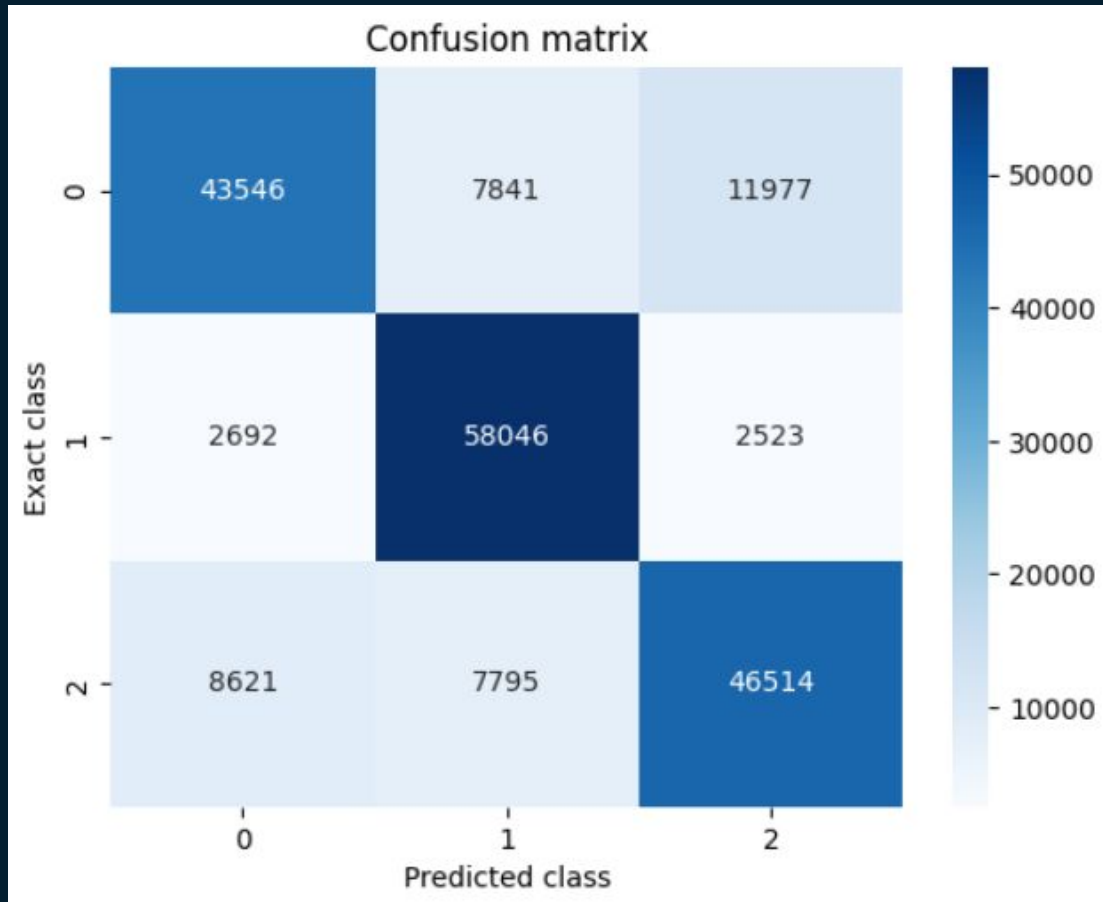
# Algoritmo con mejor resultado “Árbol de decisión” de datos balanceados

Versión del Modelo	Accuracy	F1-score (0)	F1-score (1)	F1-score (2)
Hasta 3 desviaciones estándar	0.7813	0.74	0.85	0.75
Base de datos completa	0.7816	0.74	0.85	0.75
Sin atípicos	0.7768	0.73	0.84	0.74

Versión del Modelo	Specificity (0)	Specificity (1)	Specificity (2)
Hasta 3 desviaciones estándar	0.9104	0.8762	0.8855
Base de datos completa	0.9087	0.8769	0.8868
Sin atípicos	0.9099	0.8721	0.8834

Finalmente se seleccionó el modelo del data frame de hasta 3 desviaciones balanceada con el algoritmo de árbol de decisión.

# Matriz de confusión



# Conclusiones

- **1. Limitaciones de los datos médicos**

La falta de información numérica y la predominancia de variables categóricas dificultan el entrenamiento óptimo del modelo y la selección adecuada de variables.

- **2. Desequilibrio en las clases**



Los datos estaban desbalanceados, predominando personas sin diabetes. Esto sesgó los modelos y afectó la detección de casos reales. Se aplicaron técnicas de balanceo para corregir este problema.

- **3. Resultados y oportunidades de mejora**

El modelo alcanzó un 78% de accuracy, mostrando un buen desempeño general. No obstante, se requieren mejoras en la predicción de prediabetes, optimizando variables y ampliando el dataset.

## Pasos a seguir

Ampliar el dataset con más variables clínicas y diversidad poblacional, probar modelos avanzados y desarrollar una app facilitarían la detección temprana de diabetes y apoyarían a profesionales de la salud.



**“La inteligencia artificial no reemplazará a los médicos, pero los médicos que la usen reemplazarán a los que no lo hagan.”**

— *Eric Topol*, cardiólogo y experto en medicina digital