

Sistema Inteligente para la predicción de descompensaciones Clínicas

**Diabetes Tipo 2, Hipertensión Arterial o
ambas patologías**

Fase de Preparación y Procesamiento de Datos

Presentado por:


**Bolaños Escandón María Fernanda
Montaño Cárdenas Fernando Xavier**

26/09/2025

Maestría en Inteligencia Artificial



Contexto del Problema

- El manejo de enfermedades crónicas como la diabetes tipo 2 y la hipertensión requiere monitoreo constante.
 - La falta de seguimiento oportuno incrementa el riesgo de descompensaciones clínicas.
 - Los sistemas tradicionales dependen de visitas médicas periódicas, dejando vacíos de información.
- 



Objetivo del sistema

Diseñar, implementar y evaluar un sistema inteligente basado en modelos de aprendizaje automático supervisado y análisis de series de tiempo, capaz de predecir descompensaciones clínicas en pacientes con enfermedades crónicas (diabetes tipo 2, hipertensión arterial, o la combinación de ambas), utilizando datos históricos de monitoreo fisiológico. El sistema buscará generar soluciones tempranas diferenciadas según el tipo de patología, con el fin de integrarse en una futura plataforma de monitoreo clínico.

Objetivo de La etapa

El Análisis Exploratorio de Datos (EDA) permitió identificar patrones iniciales, detectar outliers clínicos relevantes (ej. glucosa > 400 mg/dL), revisar correlaciones fuertes entre variables (HbA1c y glucosa > 0.8) y caracterizar la distribución de clases, insumo clave para las estrategias de balanceo.

Metodología

1

Preprocesamiento
de datos clínicos
históricos.

2

Análisis exploratorio
y detección de
outliers.

3

Balanceo de clases
mediante técnicas
de oversampling e
híbridas.

4

Entrenamiento de
modelos
supervisados

5

Evaluación y
validación del
rendimiento.

6

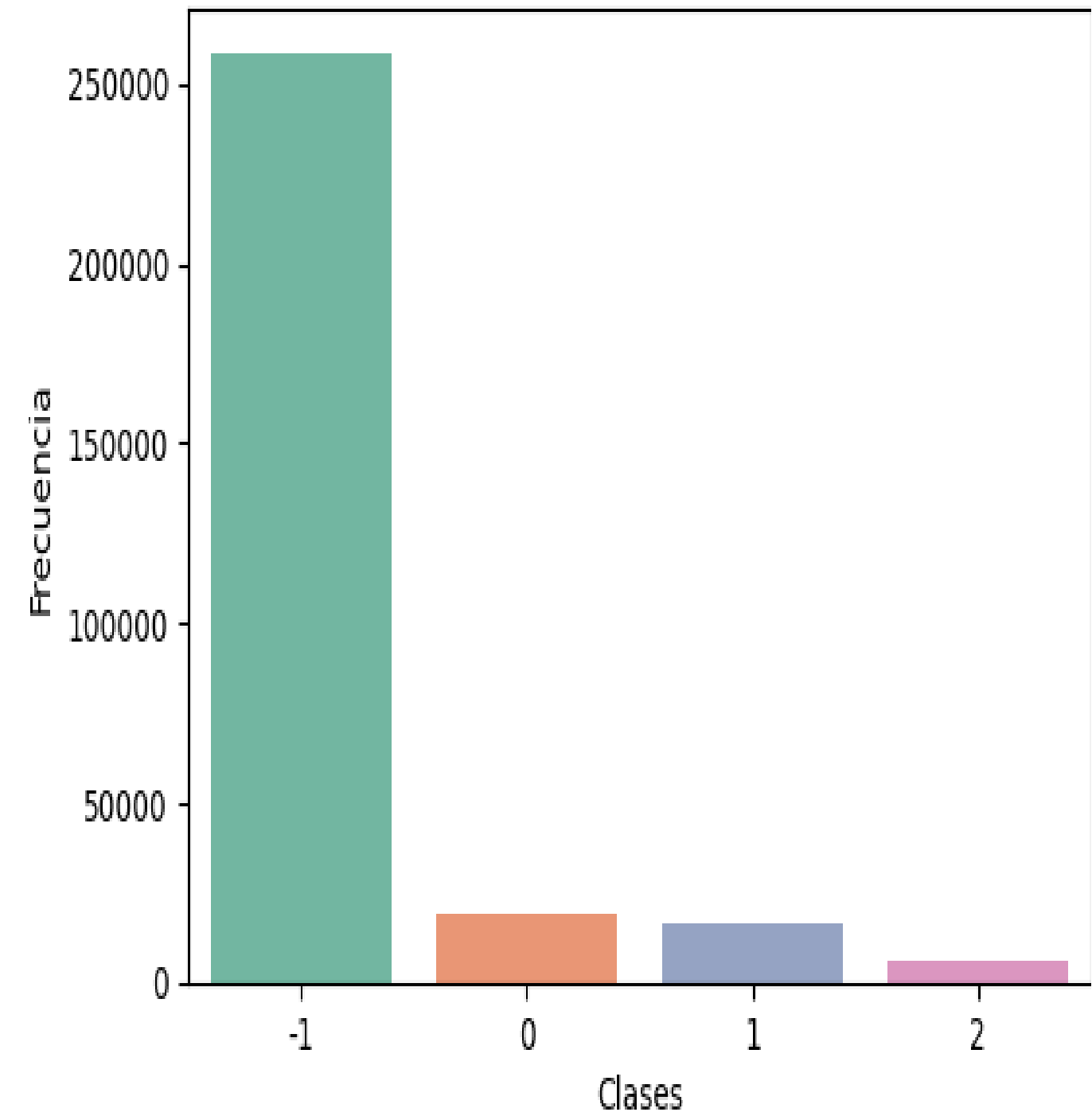
Integración en
plataforma de
monitoreo clínico
en una siguiente
etapa futura

Distribución de Clases (Target)

- La clase objetivo presenta un fuerte desbalance: la mayoría de pacientes no presenta condición (-1), mientras que los casos de diabetes (0), hipertensión (1) y ambas condiciones (2) son minoritarios.

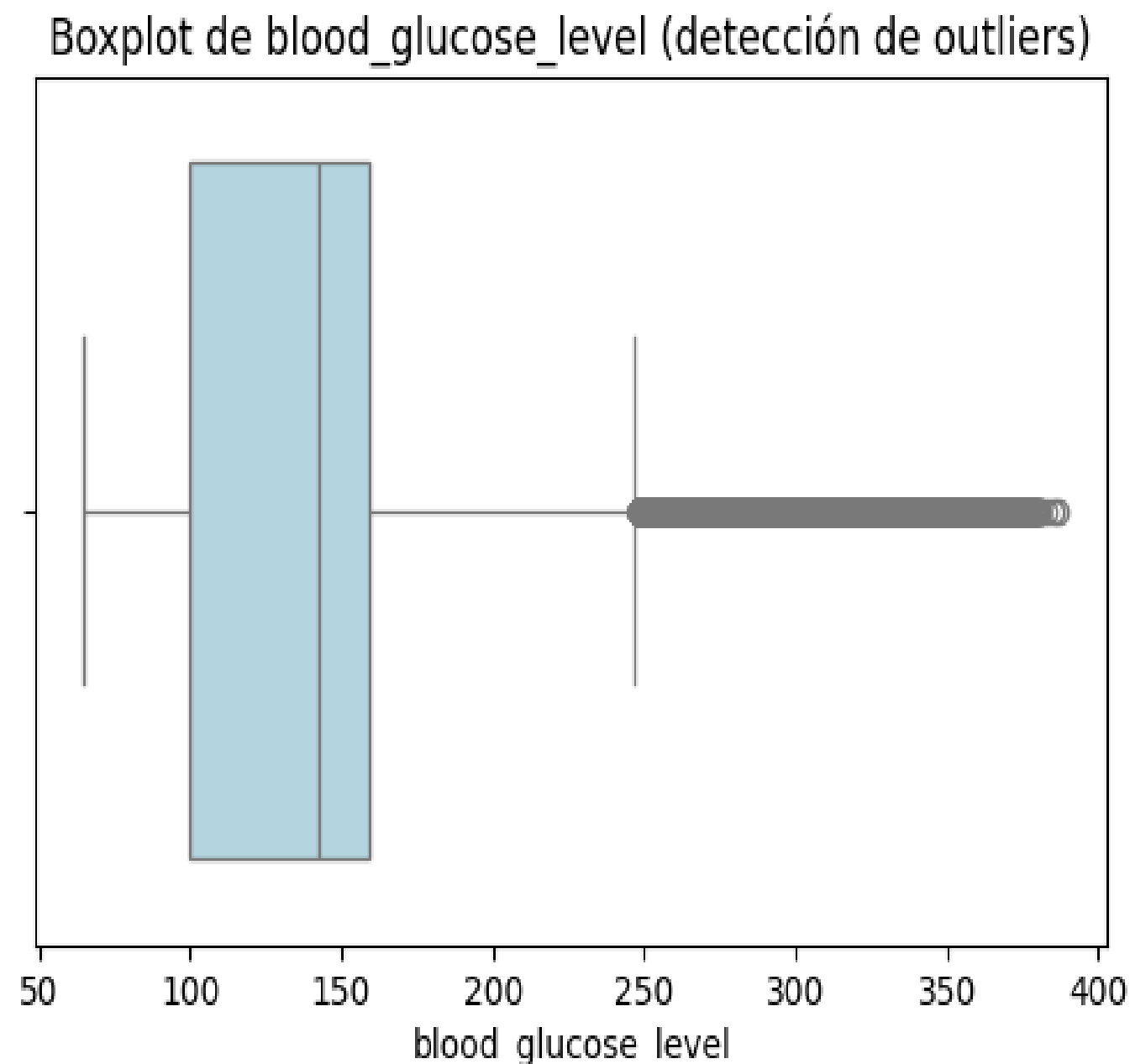
Hallazgo

- Esto justifica el uso de técnicas de balanceo.



Análisis de Glucosa

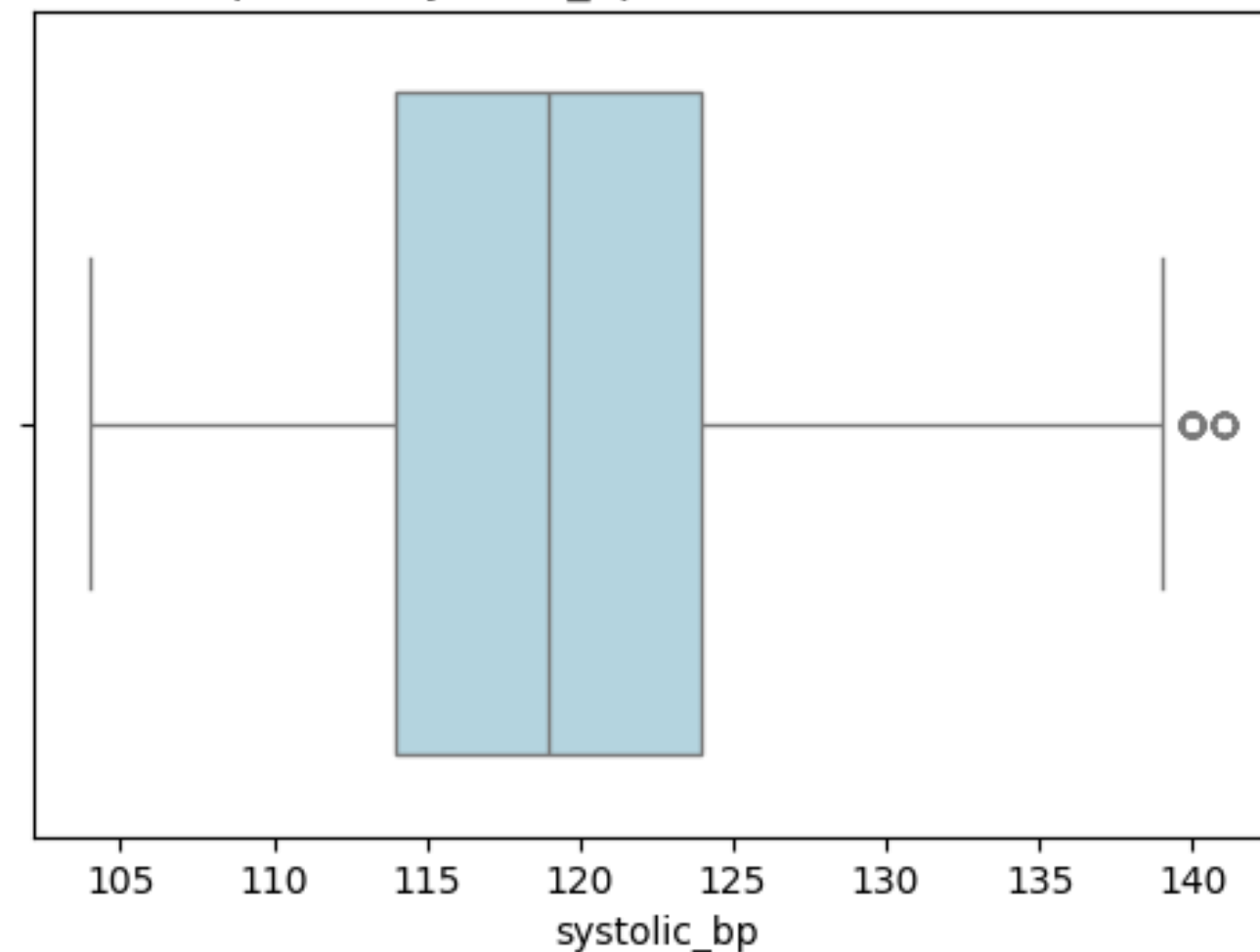
La glucosa en sangre muestra una distribución multimodal con presencia de outliers. Se observa diferencia significativa en niveles de glucosa según presencia de diabetes.



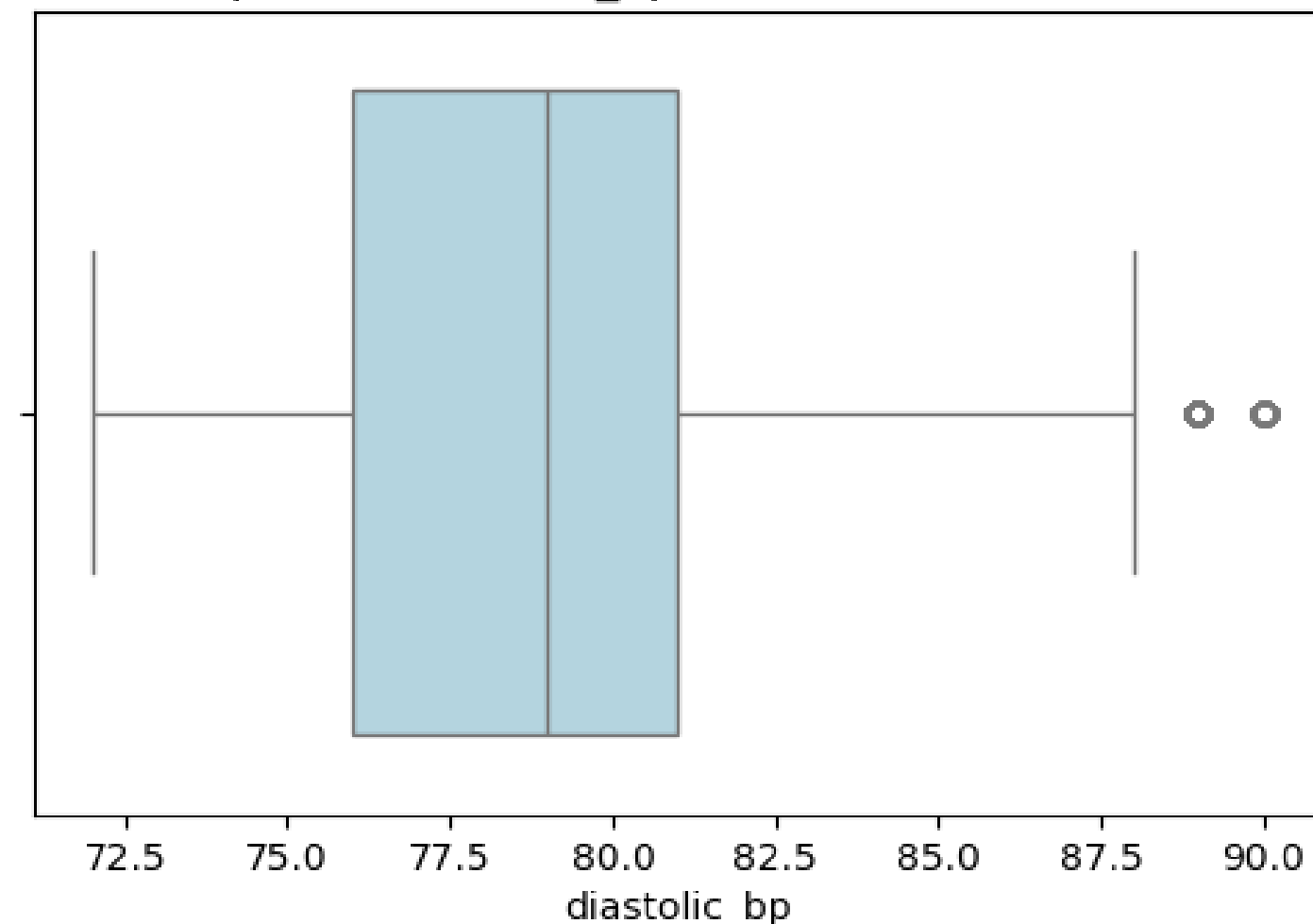
Análisis de Presión Arterial

La presión sistólica y diastólica presentan distribuciones concentradas, con elevación significativa en pacientes con hipertensión.

Boxplot de systolic_bp (detección de outliers)

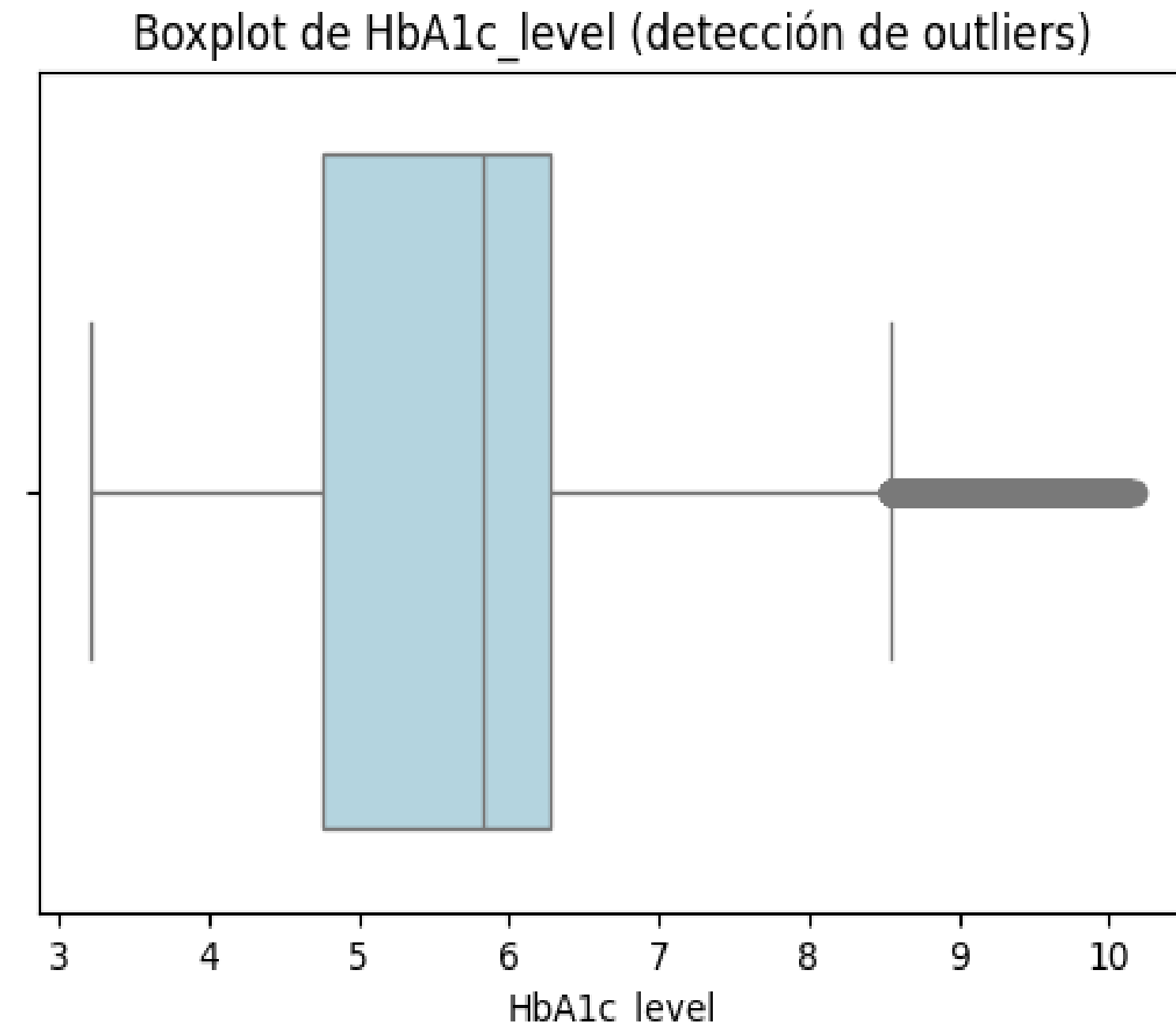


Boxplot de diastolic_bp (detección de outliers)



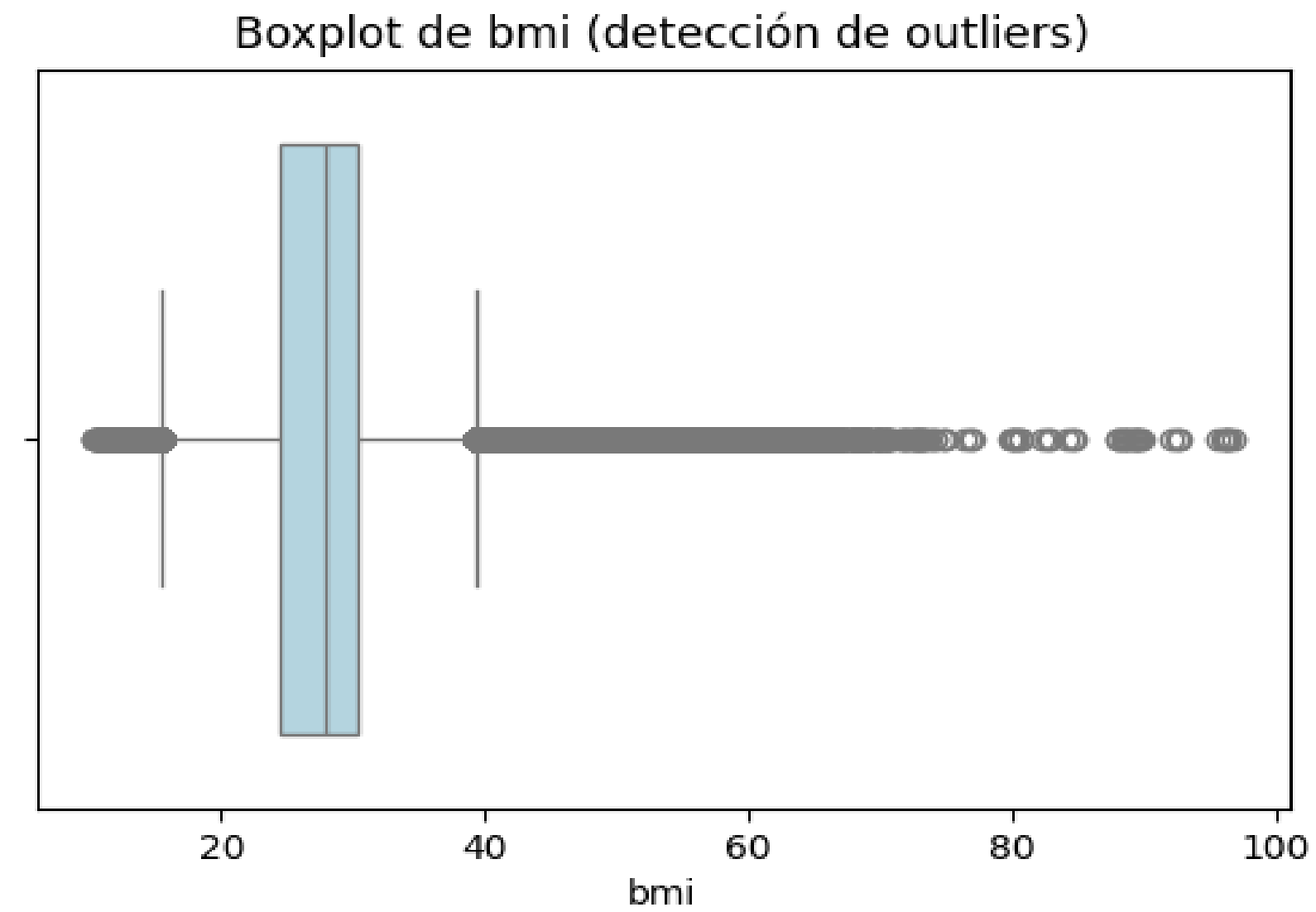
Análisis de Hemoglobina

La HbA1c presenta valores extremos (outliers) que superan el umbral clínico del 6.5%, lo cual indica pacientes con mal control glucémico. Estos valores elevados son relevantes para la detección de diabetes tipo 2 y reflejan el riesgo de complicaciones metabólicas.



Análisis de BMI (Índice de Masa Corporal)

El IMC presenta valores extremos (outliers), asociados a pacientes con obesidad severa. Es un predictor importante en la combinación de patologías.



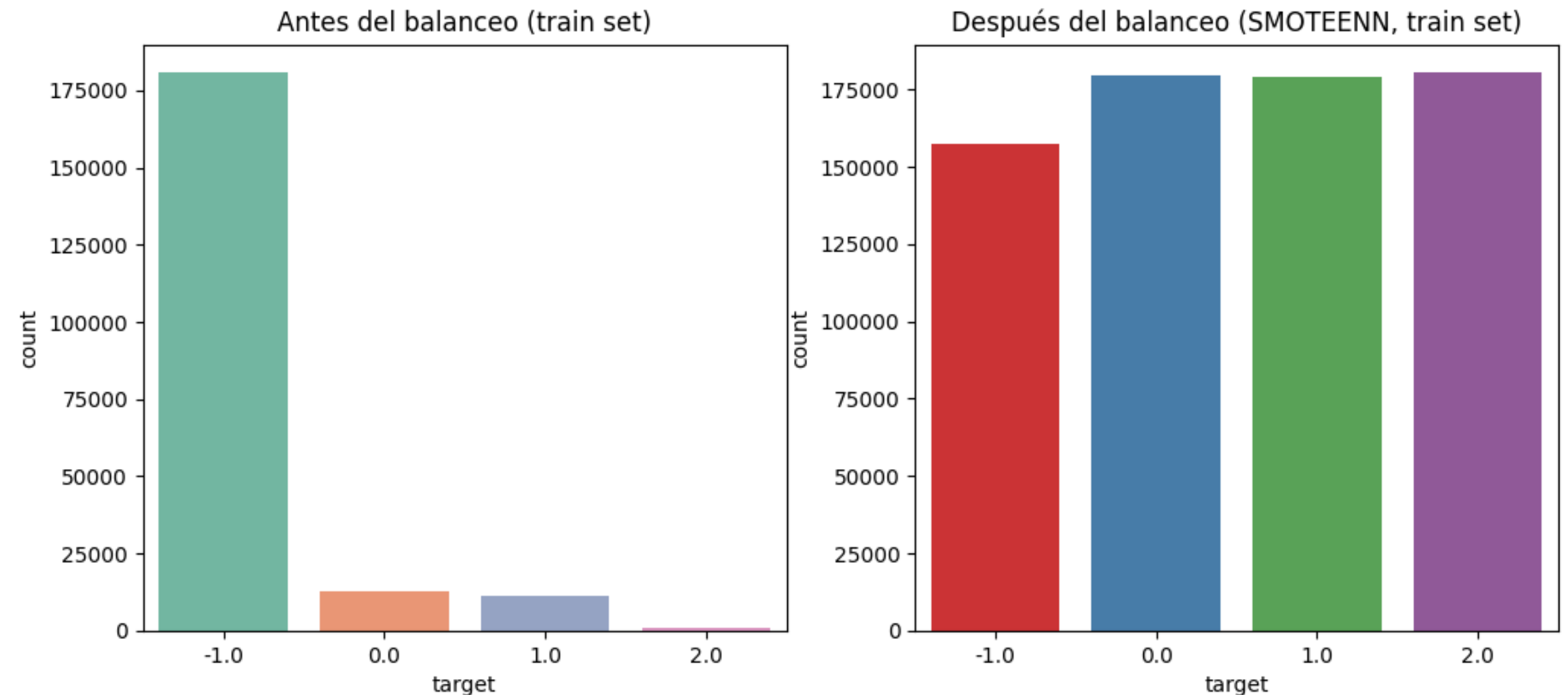
Balanceo de Clases

- Se aplicaron técnicas de oversampling (SMOTE, ADASYN, BorderlineSMOTE) y técnicas híbridas (SMOTEENN, SMOTETomek).
- El objetivo fue equilibrar la distribución de las clases para mejorar el rendimiento de los modelos predictivos.



Resultados del Balanceo

Tras aplicar SMOTEENN, las clases se distribuyen de manera balanceada, reduciendo el sesgo hacia la clase mayoritaria.

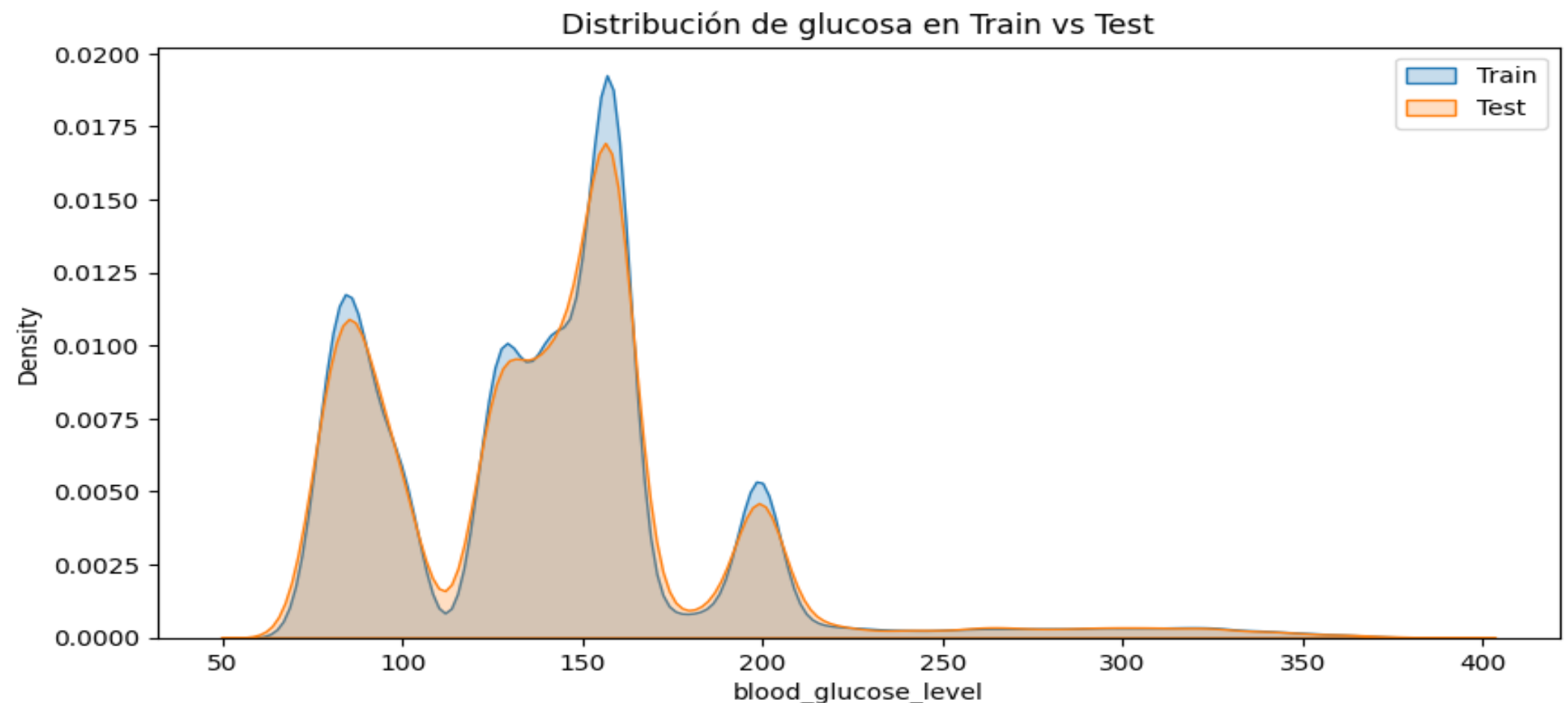


Data Augmentation

- Se aplicaron Gaussian Noise, SMOTE y Mixup para ampliar diversidad y balancear clases.
- Los datos sintéticos mantuvieron coherencia clínica en rangos de glucosa y presión arterial.
- Mejora la representatividad de clases minoritarias y reduce riesgo de sobreajuste.

Distribución en Train vs Test

Se verificó que las distribuciones de variables como glucosa se mantengan consistentes en los conjuntos de entrenamiento y prueba, asegurando validez estadística en la evaluación del modelo.



Pipeline de Preprocesamiento Automatizado

- Pipeline modular, escalable y trazable para limpieza, imputación y balanceo de datos clínicos.
- Validado con pruebas unitarias, de integración y datos nuevos, garantizando robustez y coherencia clínica.
- Flujo reproducible y confiable, con soporte para logging y monitoreo de calidad.



Resultados Preliminares y Conclusiones

- El pipeline de preprocesamiento logró un dataset balanceado ($\approx 258k$ registros por clase) mediante SMOTEENN y Data Augmentation, manteniendo coherencia clínica.
- Los modelos preliminares (Random Forest) alcanzaron métricas perfectas en accuracy, precision y recall (1.00), lo que indica adecuada preservación de patrones discriminativos.
- Las técnicas de augmentación (Gaussian Noise, Mixup) permitieron fortalecer clases minoritarias, reduciendo la dependencia de la clase mayoritaria y mejorando la robustez frente a escenarios clínicos reales.
- Se garantiza un flujo modular, reproducible y trazable, adecuado para futuras fases de validación externa y despliegue en una plataforma de monitoreo clínico
- **Próximos pasos:** Validar el pipeline y explorar modelos avanzados de series de tiempo como LSTM y CatBoost para realizar la clasificación multiclase para predicciones más precisas y tempranas, entre otras.