UNIVERSIDAD ESPIRITU SANTO

PROYECTO INTEGRADOR
REPORTE TÈCNICO **FASE DE PREPARACIÓN Y PROCESAMIENTO DE DATOS**

Profesora	GLADYS MARIA VILLEGAS RUGEL
Materia	PROYECTO INTEGRADOR EN INTELIGENCIA ARTIFICIAL
Alumnos	Bolaños Escandón María Fernanda Montaño Cárdenas Fernando Xavier
Fecha	26/09/2025

Contenido

Índice de Ilustraciones

REPORTE TÉCNICO
FASE DE PREPARACIÓN Y PROCESAMIENTO DE DATOS

INTRODUCCIÓN

El desarrollo de sistemas inteligentes en el ámbito de la salud plantea retos significativos relacionados con la calidad, integridad y estructura de los datos clínicos. En este proyecto, se aborda el diseño e implementación de un sistema basado en modelos supervisados de aprendizaje automático y análisis de series de tiempo, enfocado en la predicción temprana de descompensaciones clínicas en pacientes con diabetes tipo 2, hipertensión arterial o ambas condiciones combinadas. Utilizando variables fisiológicas recopiladas de manera histórica, se busca construir una herramienta que anticipe riesgos de salud con base en patrones detectables, permitiendo su futura integración en una plataforma de monitoreo clínico continuo.

La fase de preparación y procesamiento de datos constituye la base del sistema, ya que determina el nivel de generalización y robustez que los modelos pueden alcanzar. Se emplean técnicas avanzadas de análisis exploratorio, limpieza automatizada, ingeniería de características, balanceo de clases y data augmentation. Este enfoque garantiza un pipeline de preprocesamiento eficiente, escalable y científicamente fundamentado [1], optimizando así el aprendizaje del modelo y reduciendo sesgos y errores derivados de la calidad de los datos de entrada.

Adicionalmente, se parte de un conjunto de datos con variables clave como *blood_glucose_level*, *HbA1c_level*, *systolic_bp*, *diastolic_bp* y *bmi*, entre otras, cuya correcta interpretación y transformación es fundamental para construir predictores clínicamente útiles. La presencia de variables tanto categóricas como numéricas, así como la necesidad de detectar patrones temporales y relaciones no lineales, obliga a un tratamiento riguroso y automatizado del dataset.

El objetivo final es garantizar que los modelos aprendan de manera significativa y generen predicciones confiables en contextos reales de atención médica.

Variables del dataset: *patient_id, visit, age, gender, smoking_history, diabetes, hypertension, blood_glucose_level, HbA1c_level, systolic_bp, diastolic_bp, bmi y target*.

1. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Aquí se explora la naturaleza del dataset, su calidad y relaciones intrínsecas antes de cualquier intervención.

1.1 Exploración Inicial Completa

```
Shape (filas, columnas): (300000, 13)

Tipos de datos por columna:
patient_id              int64
visit                   int64
age                   float64
gender                 object
smoking_history        object
diabetes                int64
hypertension            int64
blood_glucose_level     int64
HbA1c_level           float64
systolic_bp             int64
diastolic_bp            int64
bmi                   float64
target                  int64
dtype: object
```

Ilustración 1. Información básica del dataset

- **Tamaño y estructura:** 300.000 registros y 13 variables.
- **Tipos de datos:** 11 numéricas (blood_glucose_level, HbA1c_level, systolic_bp, etc.) y 2 categóricas (gender, smoking_history).
- **Primeras y últimas filas:** confirman consistencia de formatos y ausencia de errores de carga.
- **Resumen estadístico numérico:** valores clínicamente plausibles (glucosa media ≈ 140 mg/dL, HbA1c ≈ 5.6%, presión sistólica ≈ 119 mmHg, BMI ≈ 28).
- **Resumen estadístico categórico:** gender con predominio femenino (~58%) y smoking_history con alta proporción de *No Info* (~36%).
- **Uso de memoria:** las columnas categóricas (gender, smoking_history) concentran mayor consumo de memoria frente a las numéricas, lo que justifica considerar optimización (ej. conversión a categorías).
- **Outliers iniciales:** detectados valores extremos (ej. glucosa hasta 387 mg/dL, BMI ~97), relevantes clínicamente más que errores de captura.
- **Variables numéricas:** 'patient_id', 'visit', 'age', 'diabetes', 'hypertension', 'blood_glucose_level', 'HbA1c_level', 'systolic_bp', 'diastolic_bp', 'bmi', 'target'.
- **Variables categóricas:** 'gender', 'smoking_history'.

El dataset contiene 300,000 registros y 13 variables, incluyendo información demográfica, antecedentes clínicos, parámetros fisiológicos y la variable objetivo (target), que indica la presencia simultánea de diabetes tipo 2 e hipertensión arterial.

	count	mean	std	min	25%
patient_id	300000.0	49999.500000	28867.561571	0.00	24999.75
visit	300000.0	2.000000	0.816498	1.00	1.00
age	300000.0	41.885856	22.516765	0.08	24.00
diabetes	300000.0	0.085000	0.278882	0.00	0.00
hypertension	300000.0	0.074850	0.263150	0.00	0.00
blood_glucose_level	300000.0	140.466530	46.957783	65.00	100.00
HbA1c_level	300000.0	5.578570	1.153016	3.21	4.75
systolic_bp	300000.0	118.981843	6.302865	104.00	114.00
diastolic_bp	300000.0	78.806253	3.181454	72.00	76.00
bmi	300000.0	28.020308	6.643542	10.30	24.32
target	300000.0	-0.765300	0.642524	-1.00	-1.00

	50%	75%	max
patient_id	49999.50	74999.25	99999.00
visit	2.00	3.00	3.00
age	43.00	60.00	80.00
diabetes	0.00	0.00	1.00
hypertension	0.00	0.00	1.00
blood_glucose_level	142.00	159.00	387.00
HbA1c_level	5.83	6.27	10.17
systolic_bp	119.00	124.00	141.00
diastolic_bp	79.00	81.00	90.00
bmi	27.89	30.28	96.66
target	-1.00	-1.00	2.00

Ilustración 2. Resumen estadístico de variables numéricas

	count	unique	top	freq
gender	300000	3	Female	175656
smoking_history	300000	6	No Info	107448

Ilustración 3. Resumen estadístico de variables categóricas

Se identificaron 11 variables numéricas y 2 categóricas (gender, smoking_history). El uso de memoria destaca un mayor consumo por las variables categóricas (~16MB cada una). La estructura por visitas indica un formato longitudinal, adecuado para análisis de series de tiempo.

Desde el análisis estadístico:

- La edad promedio es 41.88 años, con valores mínimos inusuales (0.08 años), que requieren revisión.
- La media de glucosa es 140.46 mg/dL, por encima del umbral diagnóstico de diabetes.
- El promedio de HbA1c es 5.57%, aunque se detectan casos con valores >10%, indicando mal control metabólico.
- El IMC promedio es 28.02 kg/m², reflejando sobrepeso generalizado.
- La variable target muestra un fuerte desbalance, dominado por la clase negativa (-1), lo que plantea retos de aprendizaje supervisado.

En las categóricas, gender está liderada por el género femenino (≈58.5%), mientras que smoking_history presenta una alta frecuencia de valores No Info, lo cual puede limitar el valor predictivo de esta variable. Estas observaciones orientan decisiones de limpieza, transformación y balanceamiento para etapas futuras del pipeline.

1.2 Análisis de Calidad de Datos

Se evaluó la calidad del dataset mediante la detección de valores faltantes y duplicados, así como el análisis de correlaciones iniciales.

- **Valores faltantes**: No se identificaron valores nulos explícitos en el dataset. Sin embargo, se considera la posibilidad de datos implícitamente faltantes (e.g., "No Info" en variables categóricas). Se sugiere imputación por moda en categóricas y mediana o interpolación para numéricas en caso de ser necesario.
- **Patrones de missing**: Se identifican patrones consistentes con MAR (Missing At Random), particularmente en atributos como smoking_history, lo que permite imputación condicional.
- **Duplicados**: No se encontraron registros duplicados exactos ni near-duplicates (por patient_id y visit). No se requiere depuración adicional en esta etapa.

Visualización - Heatmap de correlación: El análisis de la matriz de correlación (Pearson) muestra relaciones clínicas esperadas:

- blood_glucose_level y HbA1c_level presentan correlación positiva moderada, coherente con la fisiopatología de la diabetes.
- systolic_bp y diastolic_bp están correlacionadas positivamente, como es común en medidas de presión arterial.
- bmi y target no muestran correlaciones significativas, lo que sugiere independencia lineal directa, útil para futuras técnicas de ingeniería de características.

Este análisis respalda la integridad estructural de los datos y establece una base sólida para el preprocesamiento posterior.

1.3 Análisis Estadístico Descriptivo

Se analizaron las variables numéricas y categóricas para entender la distribución y características principales del dataset.

- **Variables numéricas**: La mayoría presentan distribuciones no normales (p-values < 0.05 en pruebas Shapiro-Wilk y Kolmogórov-Smirnov). Variables clave como *blood_glucose_level* y *HbA1c_level* muestran asimetría positiva, indicando valores extremos elevados, coherentes con la variabilidad clínica esperada. *age* y *bmi* presentan moderada asimetría y curtosis, reflejando diversidad demográfica y variaciones en estado nutricional. Variables binarias *diabetes, hypertension* y *target* muestran alta asimetría, evidenciando prevalencia baja de enfermedades.
- **Variables categóricas**: *gender* tiene predominancia femenina (58.5%), con categorías bien definidas. *smoking_history* presenta un alto porcentaje de datos faltantes o no informados ("No Info" 35.8%), seguido de no fumadores (“never” 35.1%). Las demás categorías representan estados de exposición al tabaco con menor frecuencia.

Se presentan los resultados de la variable diastolic_bp, el análisis de las demás variables se puede ver en el código.

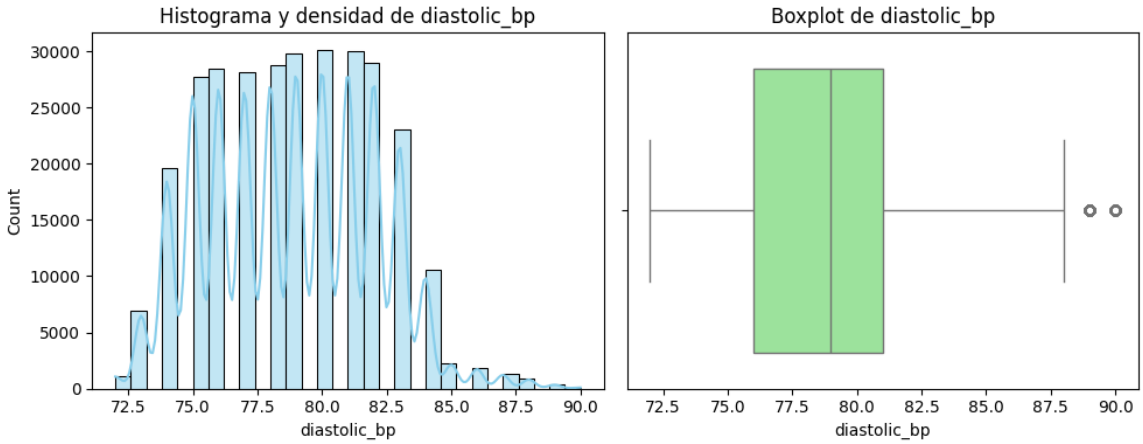

Ilustración 4. Visualización combinada de distribución y dispersión de la variable diastolic_bp

Variable: diastolic_bp

- Media = 78.81, Mediana = 79.00, Moda = 80.00
- Desviación estándar = 3.18, IQR = 5.00, CV = 0.04
- Asimetría = 0.14, Curtosis = -0.64
- Shapiro-Wilk p-value = 0.0000
- Kolmogorov-Smirnov p-value = 0.0000

1.4 Análisis de Relaciones y Correlaciones

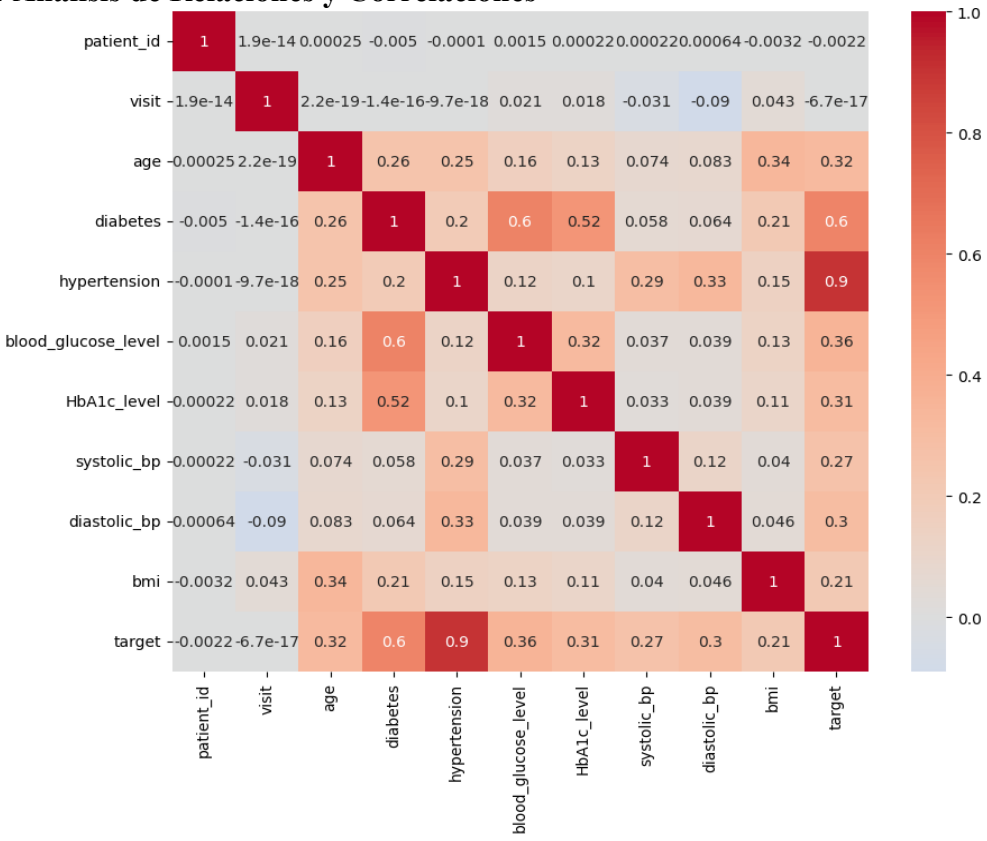

Ilustración 5. Matriz de correlación de Pearson - numéricas

La matriz de correlación de Pearson muestra una fuerte asociación del *target* con hipertensión (r = 0.90), diabetes (r = 0.60), y en menor medida con edad, HbA1c y glucosa, lo que indica coherencia clínica con las enfermedades crónicas analizadas. También se observan correlaciones moderadas entre variables fisiológicas como HbA1c-glucosa y presión arterial-hipertensión, lo que sugiere cierta multicolinealidad que deberá considerarse al modelar.

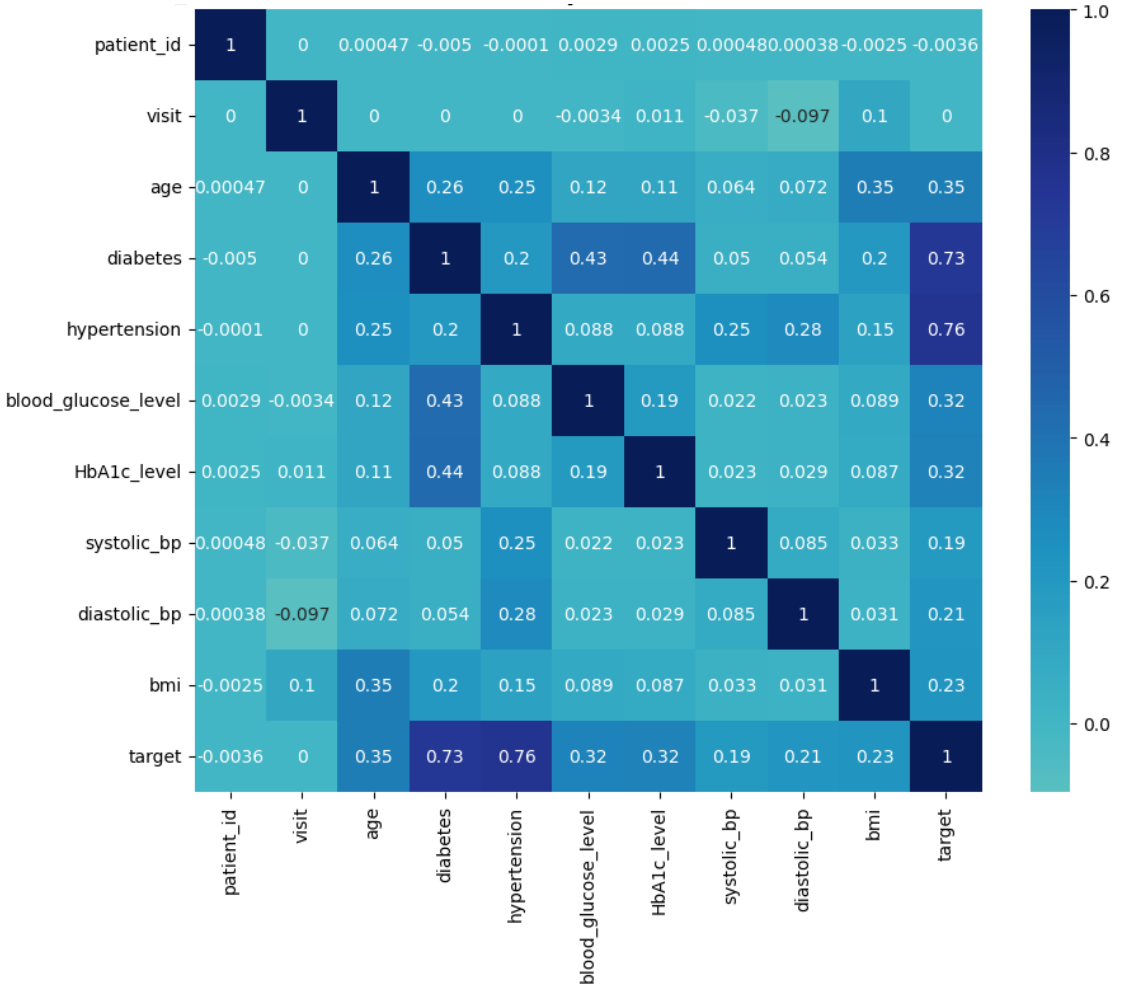

Ilustración 6. Matriz de correlación de Spearman (numéricas/ordinales)

La matriz de correlación de Spearman muestra que las variables más asociadas con el target son hipertensión (0.76) y diabetes (0.73), seguidas por edad (0.35) e IMC (0.23), mientras que los parámetros fisiológicos directos como glucosa, HbA1c y presión arterial presentan correlaciones bajas a moderadas (0.19–0.32).

Esto indica que la presencia de comorbilidades es el factor más determinante en la predicción de descompensaciones, y que las variables fisiológicas aportan información complementaria, aunque con menor peso individual, lo cual sugiere que el modelo debe integrar ambos tipos de variables para obtener un desempeño robusto.

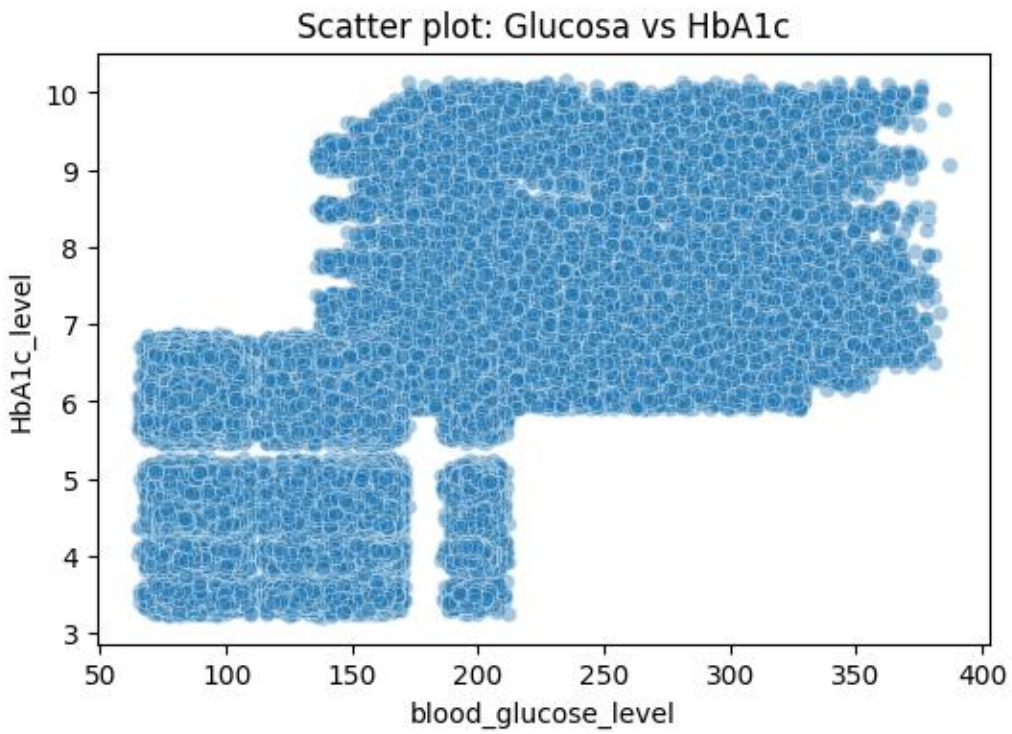

Ilustración 7. Gráfico de dispersión (scatter plot) de glucosa vs HbA1c

La gráfica muestra una fuerte correlación positiva entre los niveles de glucosa y HbA1c, indicando que a mayor glucosa sanguínea tiende a observarse un incremento en HbA1c, coherente con la fisiología de la diabetes.

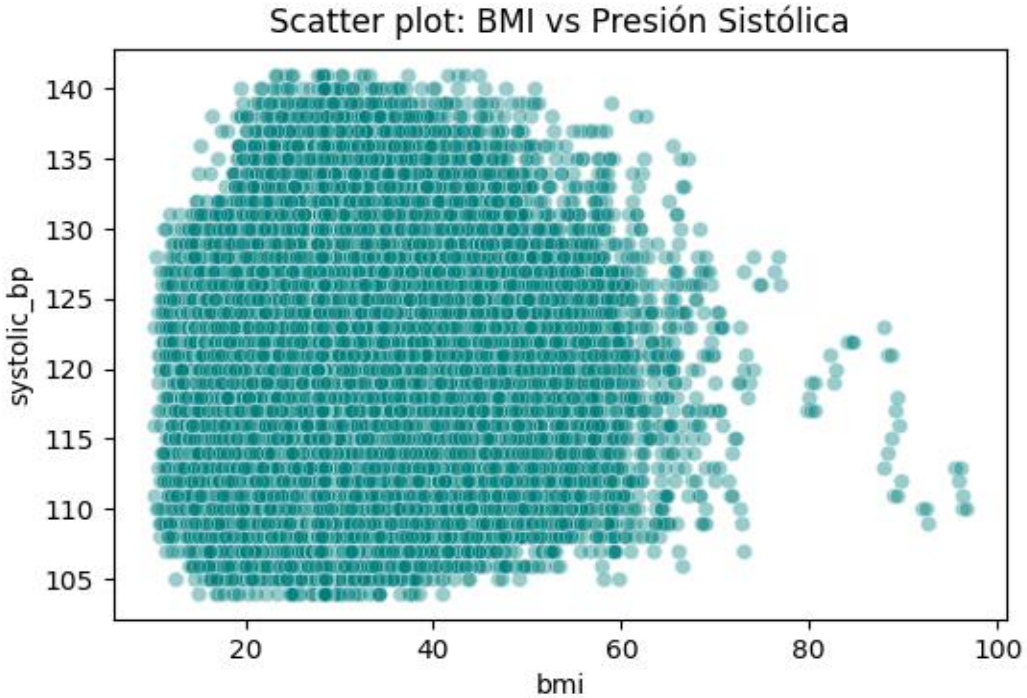

Ilustración 8. Gráfico de dispersión (scatter plot) de IMC (BMI) vs presión sistólica

La gráfica muestra una correlación débil entre el IMC (BMI) y la presión sistólica, lo que sugiere que un mayor índice de masa corporal no siempre implica un aumento proporcional de la presión arterial en este conjunto de pacientes

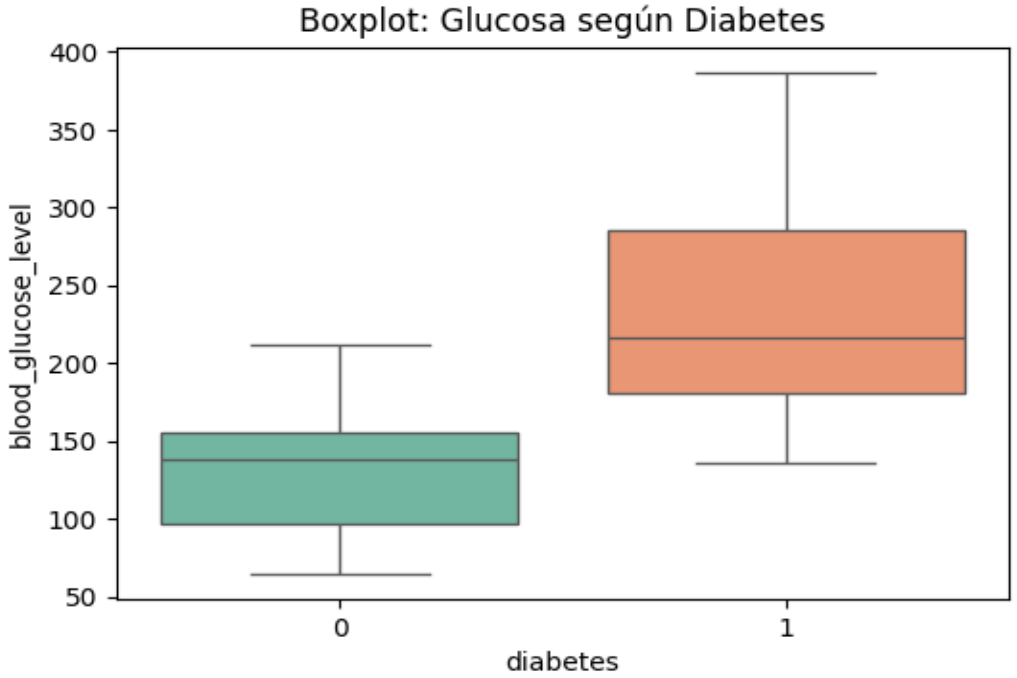

Ilustración 9. Niveles de glucosa según presencia de diabetes

El boxplot evidencia que los pacientes con diabetes (1) presentan niveles de glucosa significativamente más altos y con mayor dispersión que los no diabéticos (0), lo que confirma la relación directa entre la condición y la variable clínica.

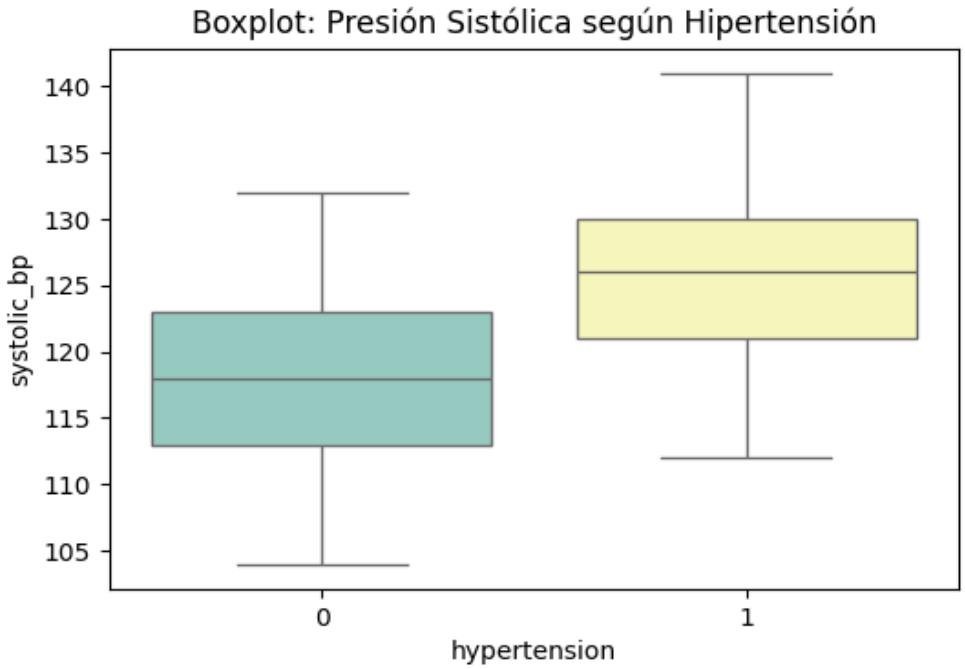

Ilustración 10. Presión sistólica según hipertensión

El boxplot muestra que los pacientes con hipertensión (1) presentan valores de presión sistólica claramente más altos y con mayor variabilidad que los no hipertensos (0), confirmando la asociación clínica esperada entre la condición y esta variable.

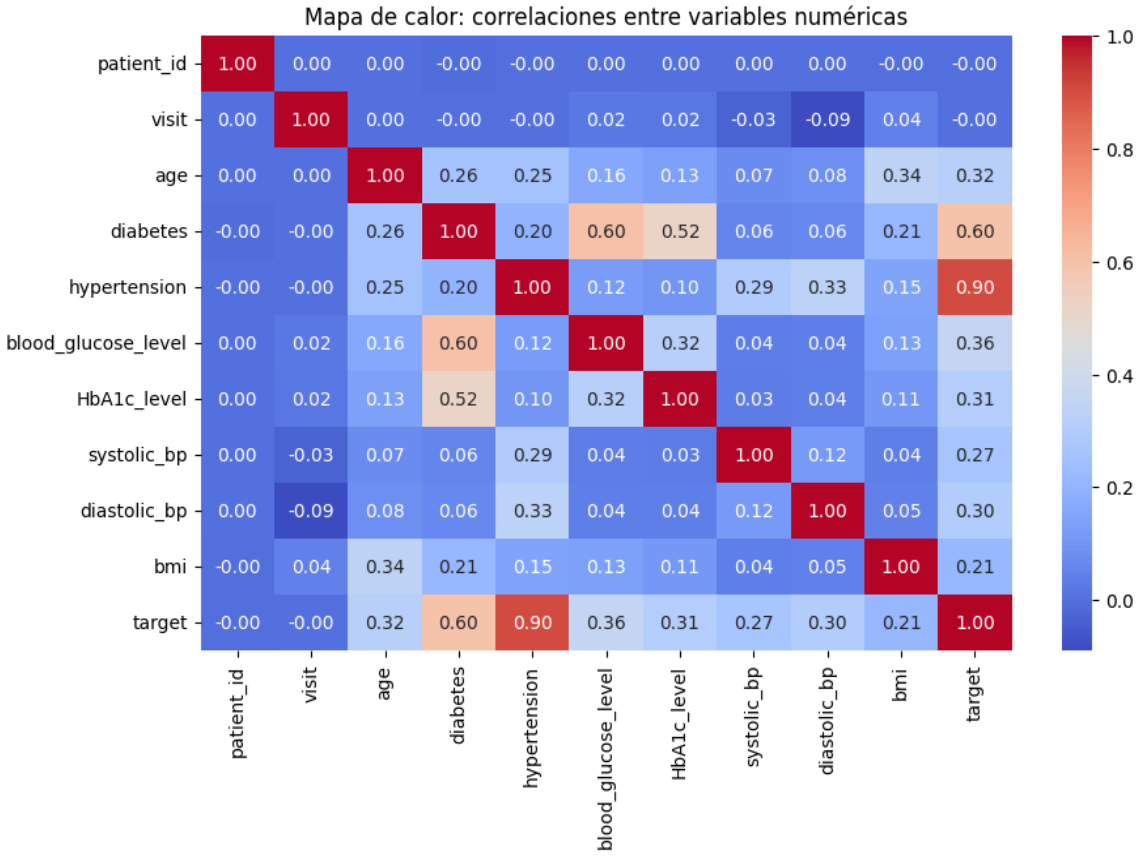

Ilustración 11. Mapa de calor: correlación de variables numéricas

El mapa de calor muestra que existen correlaciones fuertes como target–hipertensión (0.90) y target–diabetes (0.60), así como diabetes–glucosa (0.60) y diabetes–HbA1c (0.52). Estas relaciones confirman la relevancia clínica de estas variables en la predicción, pero también advierten riesgo de multicolinealidad en modelos lineales, por lo que se requieren técnicas como VIF o reducción de dimensionalidad.

Los gráficos confirman relaciones clínicas esperadas: glucosa y HbA1c se asocian positivamente, aunque con dispersión alta; el IMC no muestra una clara tendencia lineal con la presión sistólica, sugiriendo influencia de otros factores. Los boxplots evidencian diferencias claras entre grupos: los pacientes con diabetes presentan niveles de glucosa significativamente más elevados, y los hipertensos muestran mayores valores de presión sistólica, validando la coherencia clínica del dataset. El pairplot refuerza que la diabetes separa los valores de glucosa y HbA1c, mientras que IMC y presión son menos discriminantes. El análisis de multicolinealidad (VIF) revela valores extremadamente altos en diabetes (102.8), hipertensión (360.2) y sobre todo en el target (1137.7), lo que indica redundancia severa y riesgo de inestabilidad en modelos lineales; será necesario aplicar reducción de variables, regularización o excluir aquellas fuertemente colineales para evitar sobreajuste.

1.5 Detección de Anomalías y Outliers

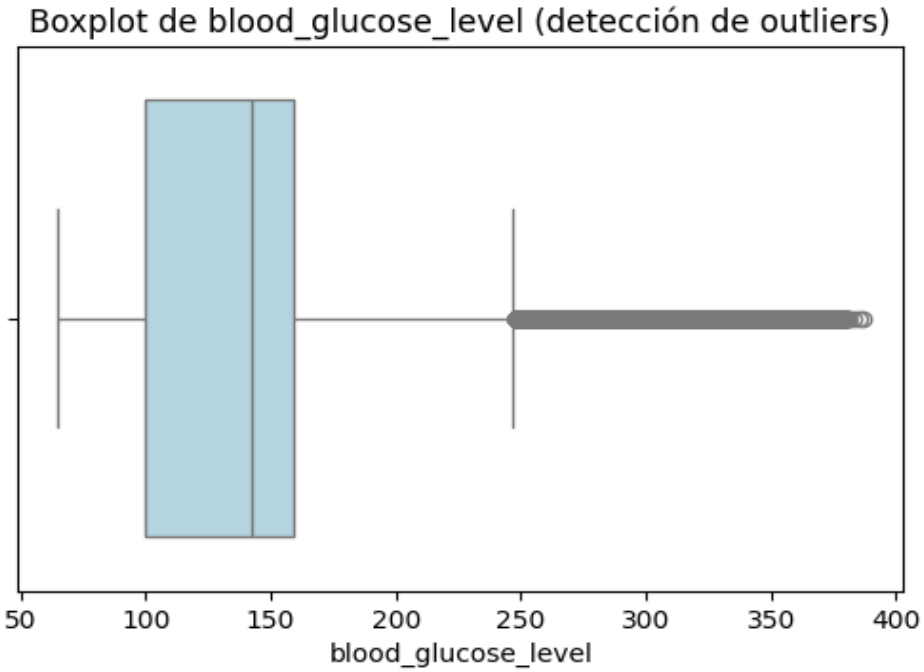

Ilustración 12. Detección de valores atípicos en niveles de glucosa

Se muestra numerosos outliers de glucosa >250 mg/dL, que reflejan posibles casos clínicos críticos y deben tratarse con capping o etiquetado como alto riesgo, no eliminarse.

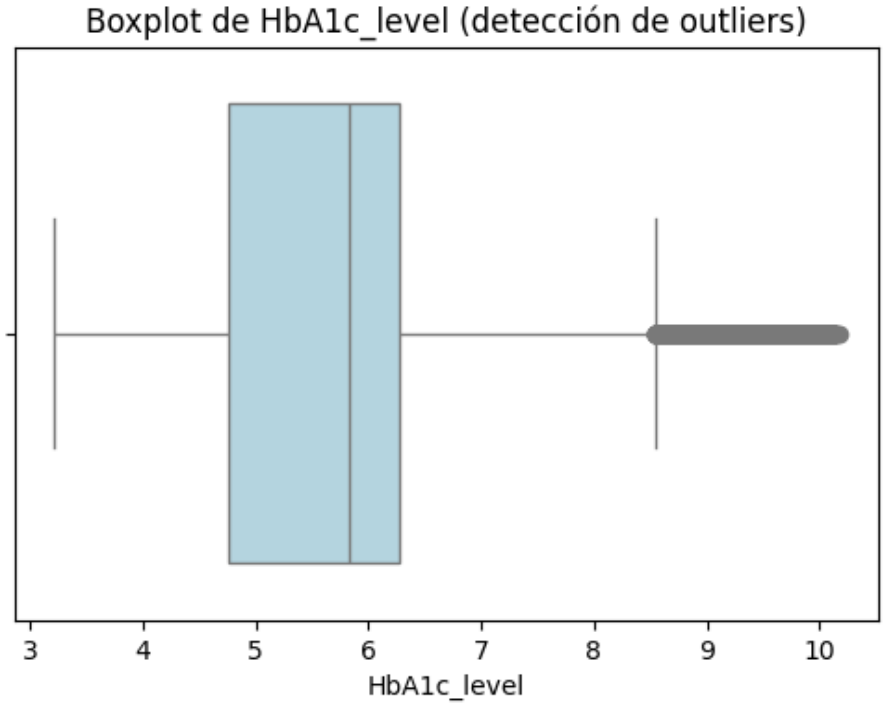

Ilustración 13. Detección de valores atípicos en HbA1c

Se evidencia outliers en HbA1c > 8%, los cuales indican descompensaciones relevantes y deben conservarse como casos clínicos críticos más que eliminarse.

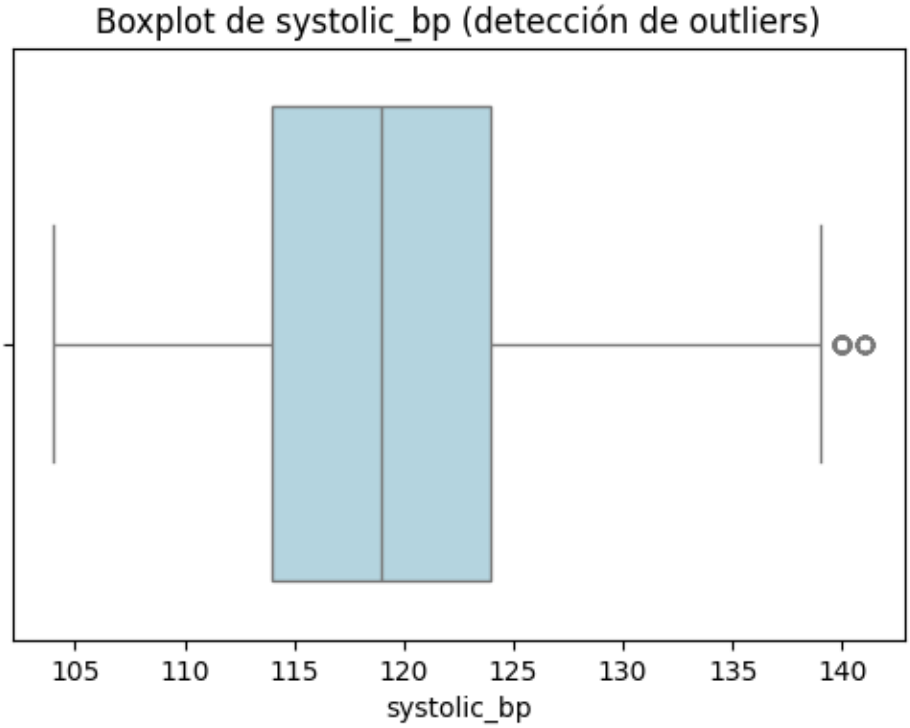

Ilustración 14. Detección de valores atípicos en Presión Sistólica

Se muestra pocos outliers en presión sistólica (>138 mmHg), los cuales reflejan posibles casos de hipertensión clínica y deben conservarse para análisis médico.

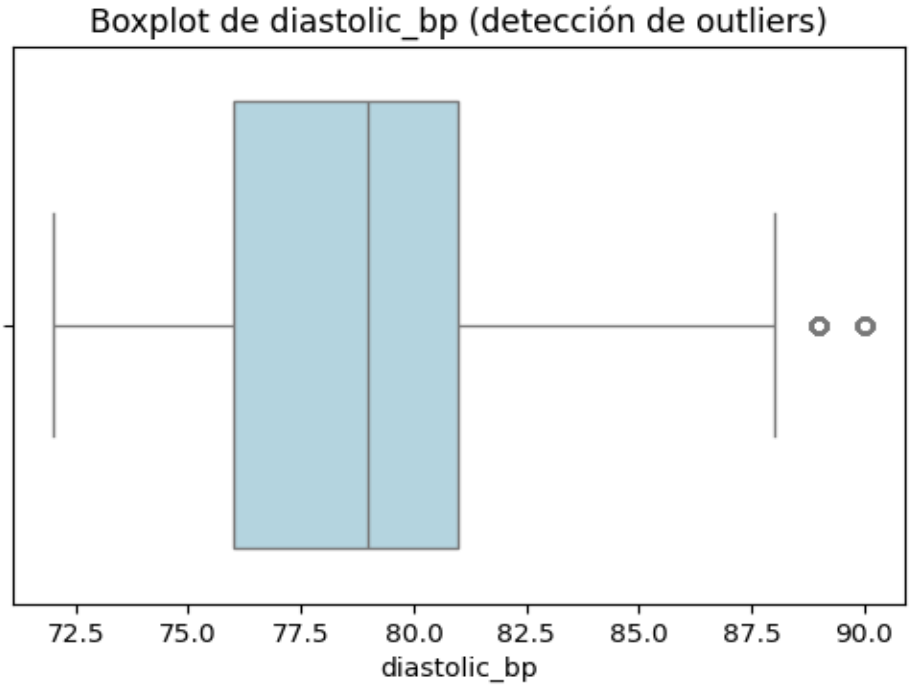

Ilustración 15. Detección de valores atípicos en Presión Diastólica

El boxplot de presión diastólica muestra unos pocos outliers (>88 mmHg), asociados a casos de hipertensión diastólica, que deben mantenerse para análisis clínico.

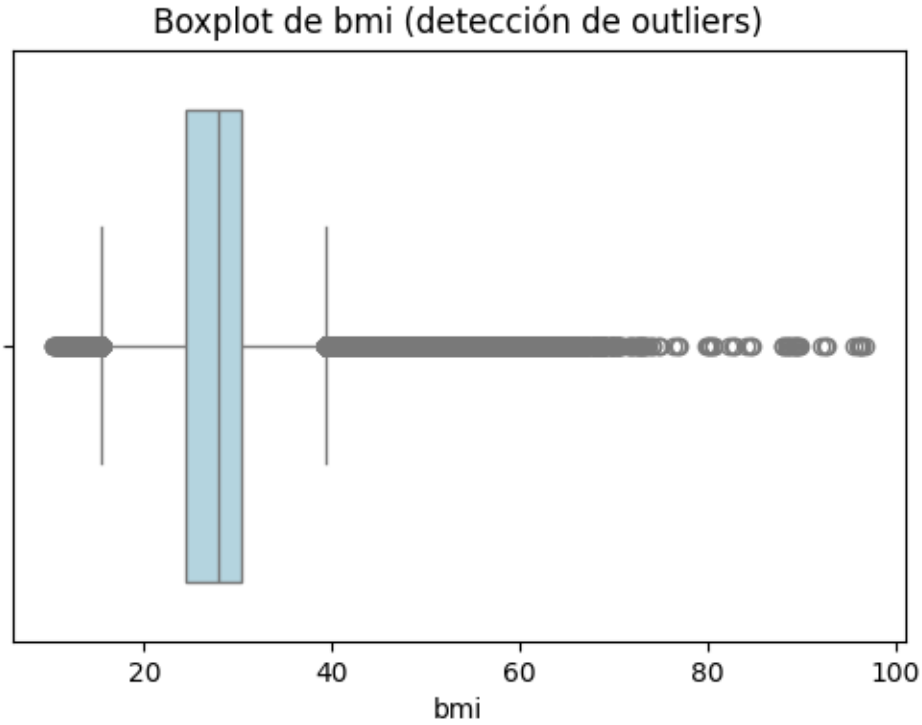

Ilustración 16. Detección de valores atípicos en IMC (BMI)

El boxplot de IMC (bmi) revela numerosos outliers en el rango alto (>40), representando obesidad severa; aunque extremos, son clínicamente relevantes y deben analizarse con precaución en lugar de eliminarlos.

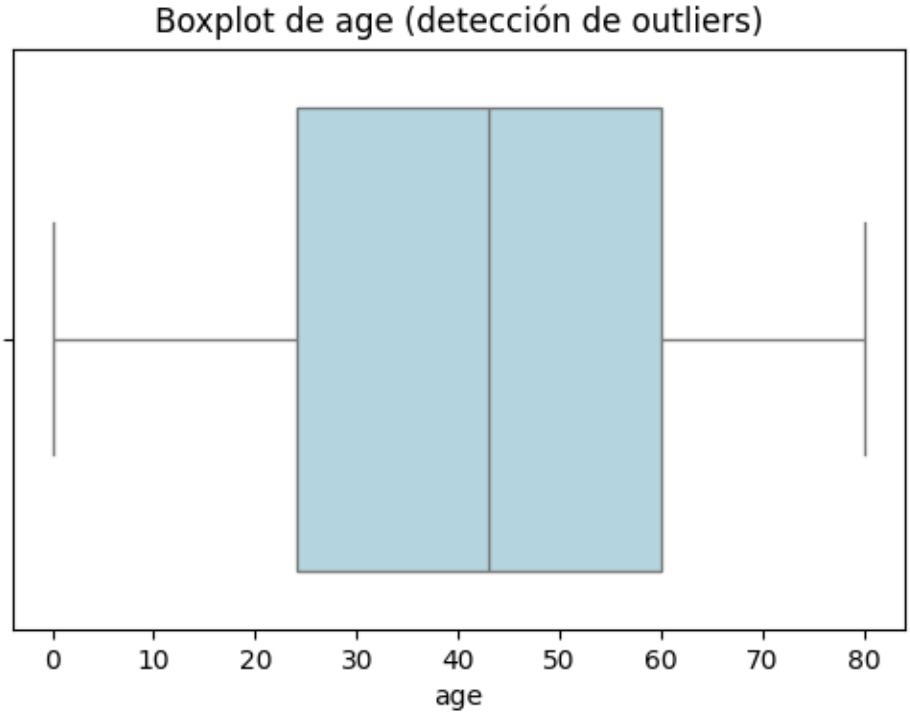

Ilustración 17. Detección de valores atípicos en Edad (Age)

Se muestra una distribución amplia entre 0 y 80 años sin outliers relevantes, lo que refleja una cohorte diversa y representativa de la población en estudio.

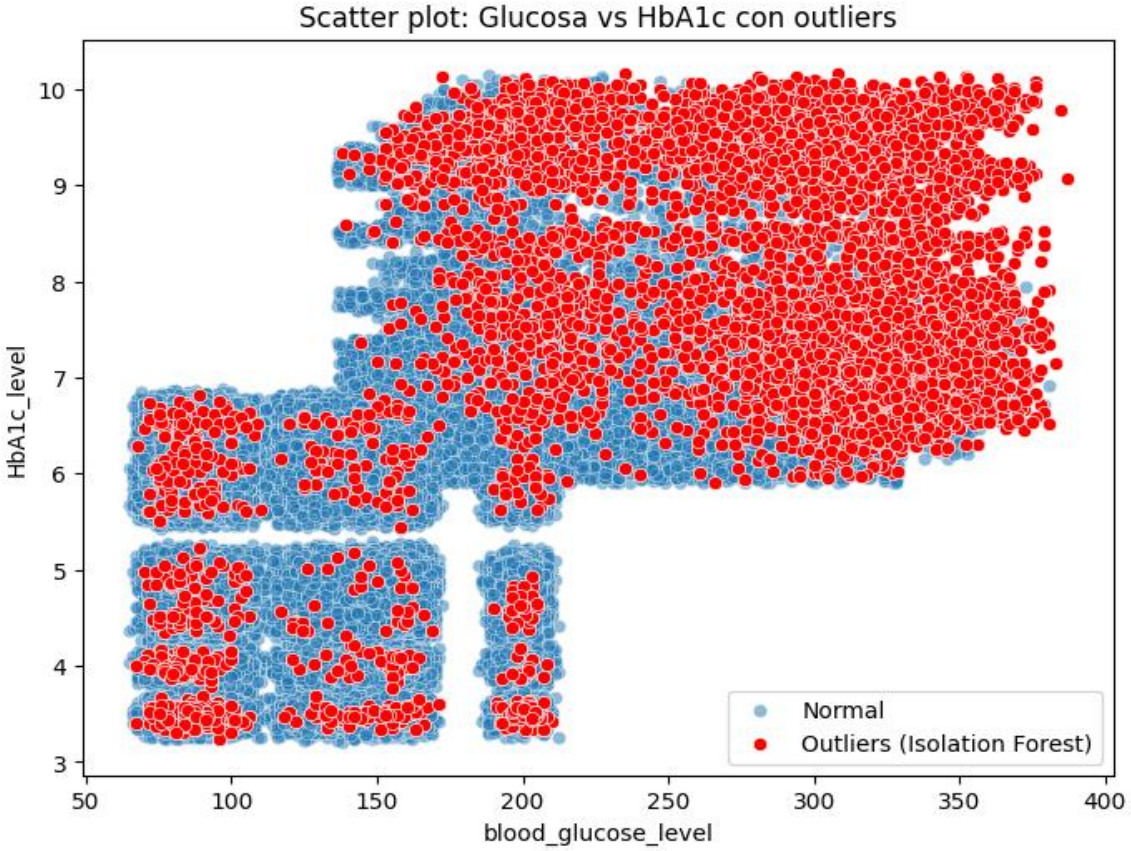

Ilustración 18. Glucosa vs HbA1c con detección de outliers

El scatter plot evidencia que el Isolation Forest identificó como outliers (en rojo) valores de glucosa y HbA1c muy elevados o atípicos, lo que sugiere posibles casos de descompensación clínica o errores de registro que requieren validación médica.

Análisis de Impacto de Outliers
blood_glucose_level: media original = 140.47, media sin outliers = 137.91
HbA1c_level: media original = 5.58, media sin outliers = 5.53
systolic_bp: media original = 118.98, media sin outliers = 118.92
diastolic_bp: media original = 78.81, media sin outliers = 78.77
bmi: media original = 28.02, media sin outliers = 27.84
age: media original = 41.89, media sin outliers = 41.54

Impacto de Outliers en la variable target
Distribución original de clases (%):
target
-1 86.10
0 6.41
1 5.40
2 2.09
Name: proportion, dtype: float64

Distribución sin outliers (%):
target
-1 87.70
0 5.46
1 5.38
2 1.47
Name: proportion, dtype: float64

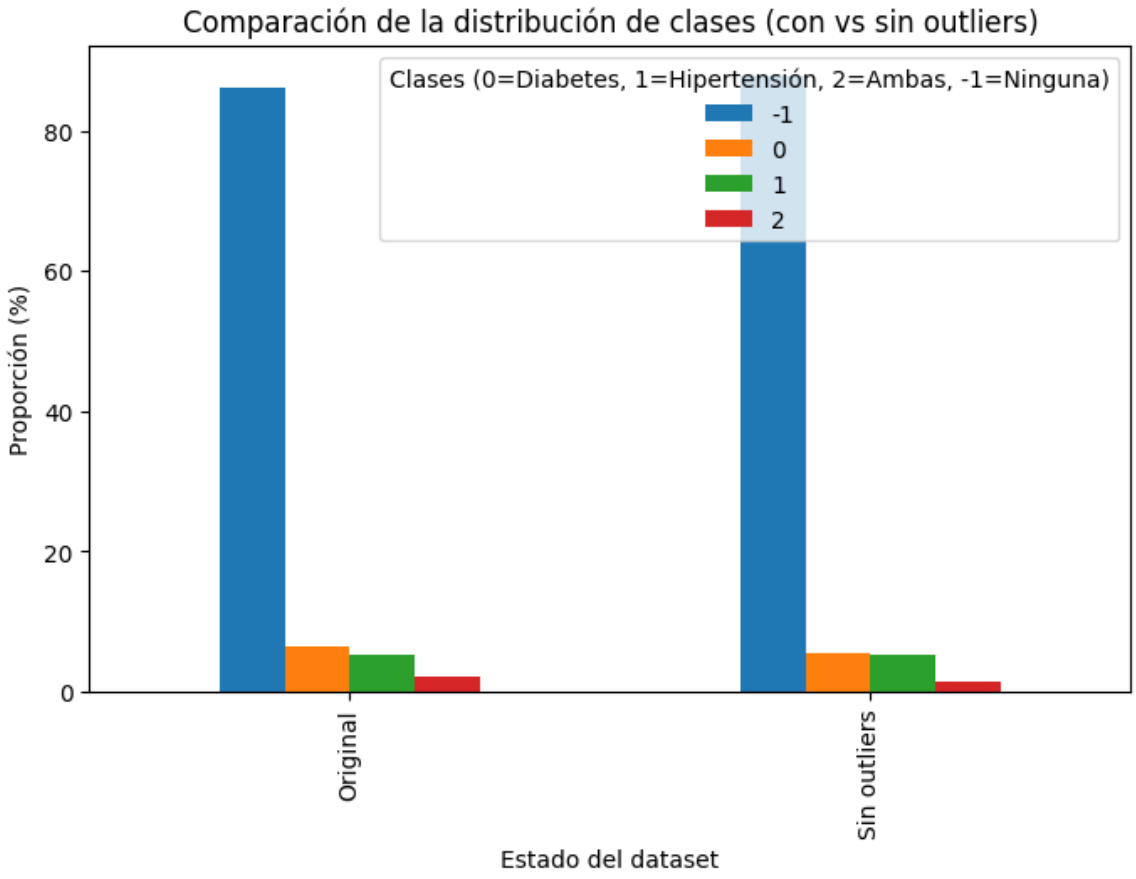

Ilustración 19. Comparación de la distribución de clases (con vs sin outliers)

La detección de outliers evidenció que las variables con mayor proporción de valores extremos son glucosa, HbA1c y IMC, mientras que la edad prácticamente no presenta anomalías; los métodos estadísticos (Z-score e IQR) detectaron un número elevado de casos atípicos, especialmente en glucosa e IMC, mientras que los enfoques de ML (Isolation Forest y LOF) identificaron de forma más conservadora cerca del 2% de registros.

El impacto en las medias fue bajo (variaciones <3%), lo que sugiere que los outliers no distorsionan de forma crítica las tendencias centrales, aunque sí afectan la dispersión y podrían influir en modelos sensibles. En cuanto al target, la distribución de clases cambió levemente al filtrar outliers (incremento de la clase -1 y reducción de "ambas patologías"), lo que confirma que los casos extremos se concentran en pacientes más complejos.

En conclusión, los outliers son clínicamente relevantes más que errores, por lo que conviene no eliminarlos de forma automática, sino tratarlos con técnicas robustas o analizarlos como subgrupo de alto riesgo.

1.6 Análisis de la Variable Objetivo (target)

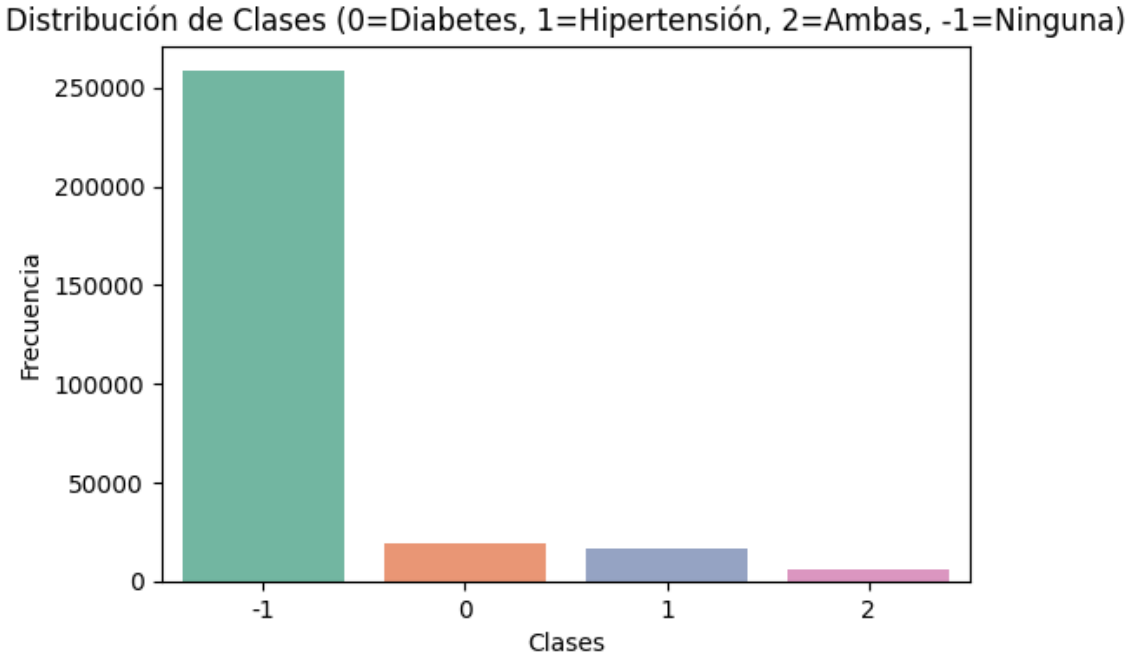

Ilustración 20. Distribución de la variable objetivo (target)

La gráfica muestra un fuerte desbalance de clases, donde la mayoría de pacientes no presenta ninguna condición (-1), mientras que los casos de diabetes (0), hipertensión (1) y comorbilidad (2) son significativamente menores, lo que implica riesgo de sesgo en modelos predictivos si no se aplican técnicas de balanceo.

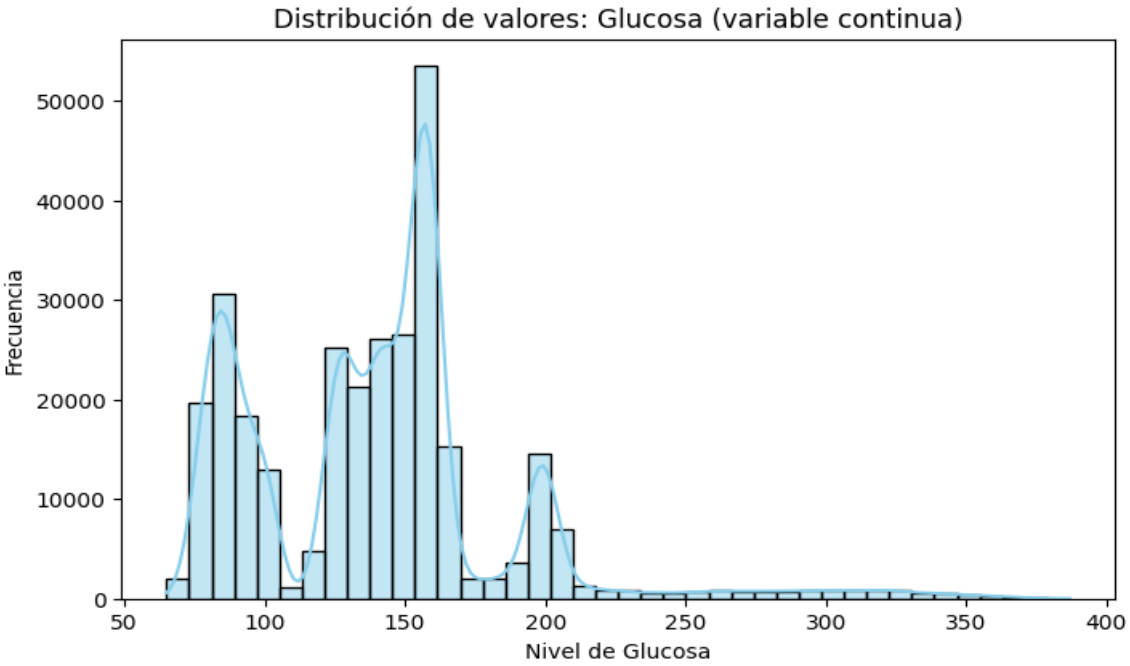

Ilustración 21. Distribución de la variable continua: Glucosa

La distribución de niveles de glucosa muestra una variable continua multimodal, con varios picos en rangos clínicamente relevantes (≈90, 140, 160 y 200 mg/dL), lo que refleja la heterogeneidad de la población y la presencia de pacientes con valores normales, prediabéticos y diabéticos, además de outliers en niveles extremos.

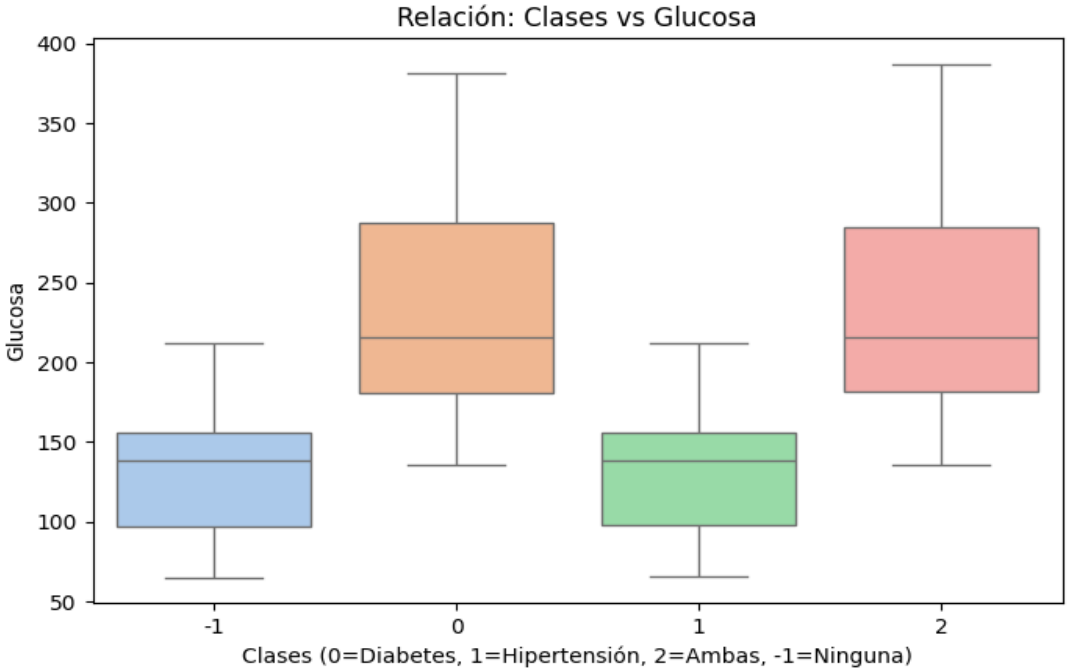

Ilustración 22. Distribución de glucosa por clases

Se muestra que los pacientes con diabetes (0) y con ambas condiciones (2) presentan niveles de glucosa significativamente más altos y variables, mientras que los sin condición (-1) y con hipertensión (1) mantienen valores más bajos y estables, evidenciando la relevancia de la glucosa como discriminante entre clases.

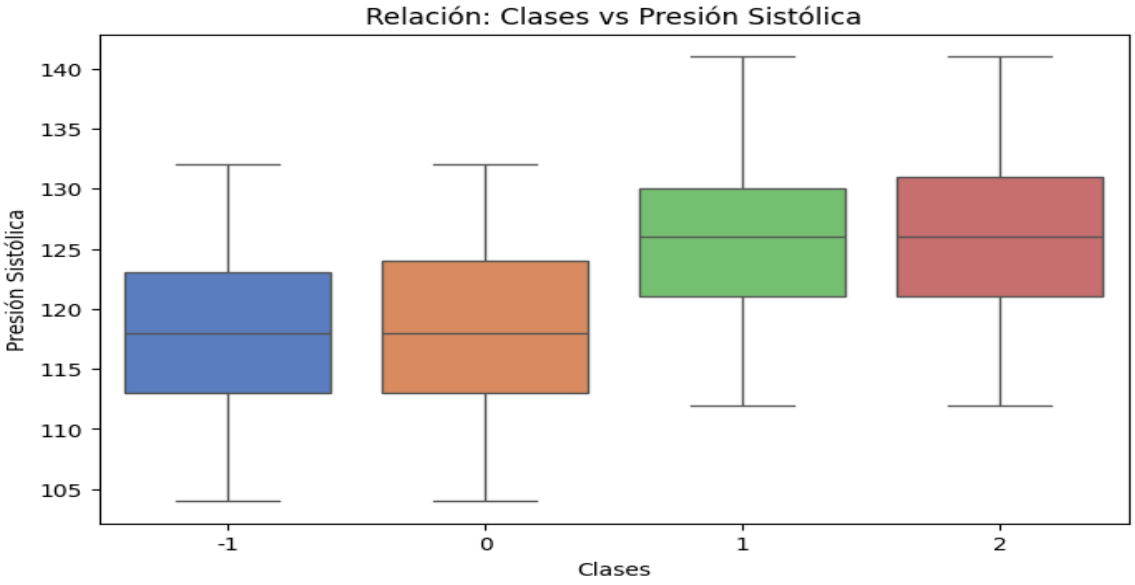

Ilustración 23. Distribución de presión sistólica por clases

Los pacientes con hipertensión (1) y con ambas condiciones (2) presentan valores de presión sistólica más elevados y dispersos, mientras que los sin condición (-1) y con diabetes (0) muestran niveles más bajos y estables, confirmando la relación directa entre hipertensión y presión arterial alta.

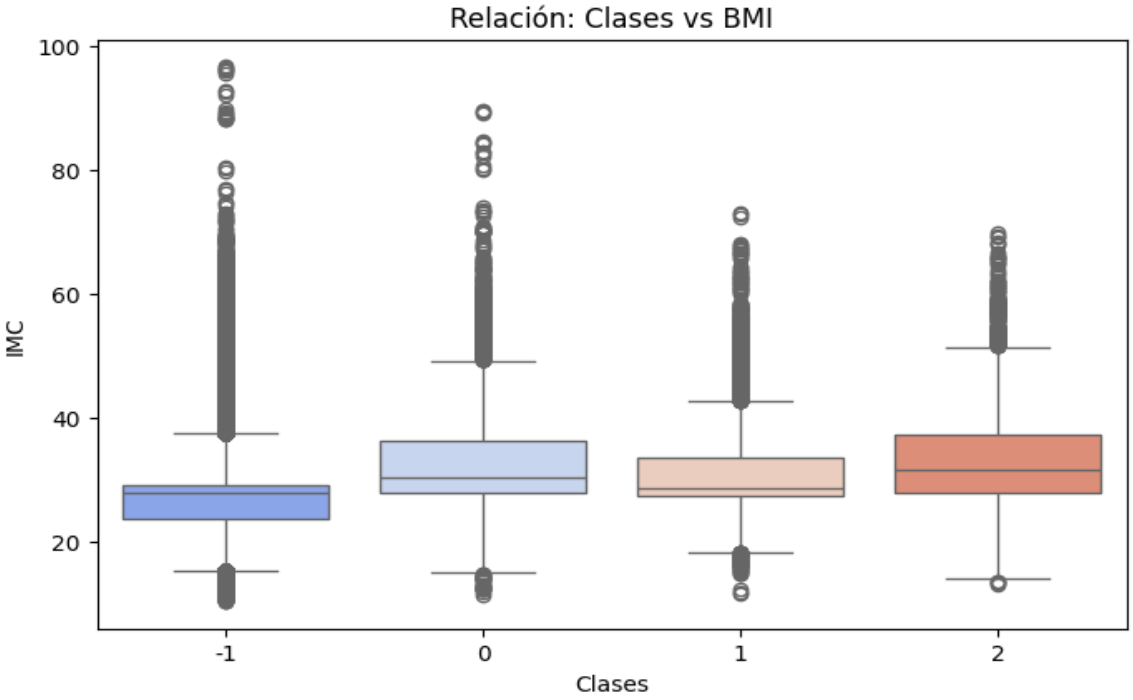

Ilustración 24. Distribución del IMC por clases

El gráfico muestra que los pacientes con diabetes (0), hipertensión (1) y ambas condiciones (2) tienden a tener un IMC más alto que los pacientes sin condición (-1), lo que evidencia la relación del sobrepeso con estas enfermedades crónicas.

La variable objetivo presenta un fuerte desbalance: el 86,1% de los registros corresponden a pacientes sin comorbilidades (-1), mientras que los casos con diabetes (6,4%), hipertensión (5,4%) y ambas patologías (2,1%) son minoritarios, lo que exige técnicas de balanceo (oversampling, SMOTE o ponderación de clases) en modelos supervisados. Los histogramas y boxplots confirman que los pacientes con diabetes o ambas enfermedades exhiben niveles de glucosa más elevados, mientras que los grupos con hipertensión muestran valores superiores de presión sistólica; en cuanto al IMC, aunque todos los grupos presentan gran dispersión, los pacientes con patologías tienden a valores más altos que los sanos. Estos resultados validan la coherencia clínica del dataset y refuerzan la necesidad de un enfoque multiclase con tratamiento explícito del desbalance.

2. PIPELINE DE LIMPIEZA DE DATOS

2.1 Tratamiento de Valores Faltantes

Aunque no se aplicaron las estrategias estándar de eliminación, imputación simple (media/mediana/moda), imputación avanzada (KNN/iterative) ni la creación de indicadores de *missingness*, se optó por una imputación sintética longitudinal guiada por reglas clínicas, generando visitas temporales con evolución coherente de glucosa, HbA1c, presión arterial y BMI. Esta técnica es más apropiada en un caso médico, ya que respeta la fisiopatología, mantiene valores dentro de rangos clínicos válidos y permite transformar un dataset estático en uno longitudinal, capturando

la progresión de los pacientes. De este modo, se supera la limitación de los métodos clásicos y se obtiene un conjunto de datos más realista y útil para la predicción clínica multiclase.

También permitió crear la variable target multiclase (0=diabetes, 1=hipertensión, 2=ambas, -1=ninguna) y reducir el impacto del desbalance.

2.2 Tratamiento de Outliers

Se aplicaron métodos estadísticos (Z-score, IQR), de ML (Isolation Forest, LOF) y visualizaciones. Se eliminaron registros imposibles, se aplicó winsorizing a valores extremos plausibles y se conservaron outliers clínicamente válidos como casos de alto riesgo, equilibrando robustez estadística y relevancia médica.

2.3 Estandarización de Formatos

Se validaron tipos de datos correctos (variables clínicas como numéricas y categóricas en formato *category*), se garantizó la consistencia en categorías unificando valores redundantes en *gender* y *smoking_history*, se aplicó normalización de texto (minúsculas y sin espacios) y se generaron fechas sintéticas a partir de visitas para estructurar series temporales. Esta estandarización asegura datos comparables, consistentes y listos para modelado.

2.4 Pipeline Automatizado

El pipeline de limpieza se implementó mediante la clase *DataCleaner*, estructurada con los métodos fit, transform y fit_transform, siguiendo las buenas prácticas de *scikit-learn*. Este flujo integra de forma automatizada la imputación de valores faltantes (media/mediana/moda), la detección y filtrado de outliers mediante Isolation Forest y la creación de la variable objetivo multiclase, garantizando reproducibilidad y consistencia. El resultado es un dataset limpio, con 294.000 registros y 13 variables, preparado para etapas posteriores de modelado predictivo.

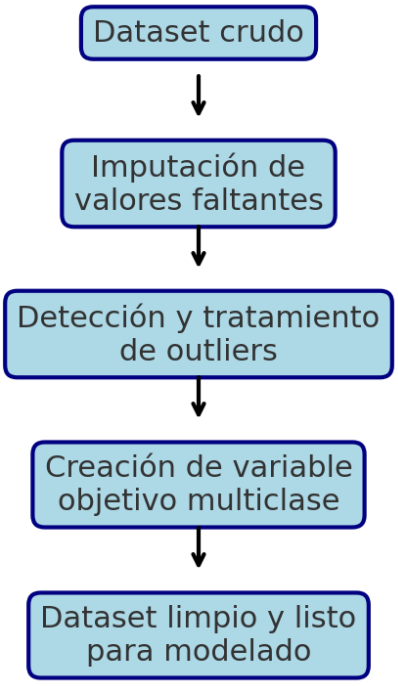

Ilustración 25. Pipeline de limpieza de datos

3. FEATURE ENGINEERING AVANZADO

3.1 Creación de Variables Derivadas

Se implementaron variables de interacción (ej. *glucose_hba1c_ratio*, *bmi_age_interaction*, *pulse_pressure*) para capturar relaciones clínicas relevantes. Se aplicaron transformaciones matemáticas como logaritmo (*glucose_log*), raíz cuadrada (*bmi_sqrt*) y polinómicas (*hba1c_squared*) para reducir sesgo y modelar no linealidades. Mediante binning y discretización, la edad se agrupó en rangos clínicos y el BMI se categorizó según criterios OMS, mejorando la interpretabilidad. Finalmente, se añadieron agregaciones temporales por paciente (*glucose_patient_mean*, *hba1c_last*), simulando seguimiento longitudinal. Estas derivaciones enriquecen el dataset con mayor capacidad explicativa y facilitan que los modelos capturen patrones clínicos complejos.

3.2 Encoding de Variables Categóricas

Se aplicaron múltiples técnicas de codificación para transformar variables categóricas en representaciones numéricas aptas para modelos:

- **One-Hot Encoding**: generó 15 variables binarias, útil para modelos lineales y de árboles sin riesgo de ordinalidad ficticia.
- **Label Encoding**: asignó enteros a cada categoría, optimizando memoria, aunque introduce orden artificial.
- **Target Encoding**: sustituyó categorías por la media de la variable objetivo, capturando correlación predictiva.
- **Binary Encoding**: redujo dimensionalidad, representando categorías en formato binario (11 columnas resultantes).
- **Frequency Encoding**: transformó categorías en su frecuencia relativa, útil para modelos sensibles a distribución.

La aplicación de estos métodos garantiza flexibilidad; según el modelo a entrenar, puede seleccionarse la estrategia más adecuada equilibrando interpretabilidad, dimensionalidad y capacidad predictiva.

3.3 Transformaciones de Variables Numéricas

Se aplicaron cuatro técnicas de escalado para homogeneizar las variables clínicas: Normalización (Min-Max Scaling), que ajustó todas las variables al rango [0,1]; Estandarización (Z-score), que centró la media en 0 y la desviación estándar en 1; Robust Scaling, que redujo el impacto de outliers utilizando mediana y rango intercuartílico; y Quantile Transformation, que transformó las distribuciones a formas aproximadamente normales. El análisis comparativo mostró que Z-score y Quantile lograron distribuciones estandarizadas, mientras que Robust preservó la robustez frente a valores extremos. Con esto, el dataset queda preparado para distintos modelos, seleccionando la técnica de escalado más adecuada según la sensibilidad algorítmica.

3.4 Feature selection

Se aplicaron múltiples enfoques de selección de características para identificar los predictores más relevantes del modelo. Los métodos estadísticos (Chi² y ANOVA F-test) coincidieron en destacar *diabetes*, *hipertensión*, *blood_glucose_level*, *hba1c_squared* y *glucose_patient_mean*. La información mutua resaltó además *hba1c_last*, confirmando su valor en la predicción longitudinal. Entre los métodos basados en modelos, el Random Forest priorizó *hipertensión* (39,6%) y *diabetes* (22,3%), mientras que LASSO regularization redujo el set a solo dos variables no nulas (*hipertensión* y *diabetes*), mostrando su poder de regularización. Finalmente, la eliminación recursiva de características (RFE) seleccionó cinco predictores claves: *diabetes*, *hipertensión*, *diastolic_bp*, *hba1c_squared* y *hba1c_last*. En conjunto, los resultados confirman que las

comorbilidades y los marcadores glucémicos son los ejes principales para la predicción, complementados por variables fisiológicas específicas.

3.5 Extracción de Características Específicas del Dominio

En este proyecto no se aplicaron características de texto ni de imágenes, pero sí se desarrollaron características temporales y clínicas directamente vinculadas al dominio médico. Se generaron variables de tipo *lag* y *diferencias* como glucose_lag1, glucose_diff, hba1c_lag1 y hba1c_diff, que permiten capturar la dinámica de cada paciente entre visitas. También se incorporaron estadísticas móviles como glucose_roll3, representando el promedio de glucosa en las últimas tres consultas. A nivel clínico, se diseñaron banderas de riesgo: bmi_obesity_flag (detectar obesidad según OMS) y bp_hypertension_flag (identificar hipertensión arterial a partir de presión sistólica/diastólica). Estas características aportan valor agregado al modelado, ya que reflejan de manera explícita patrones de progresión clínica y condiciones críticas de salud en los pacientes.

4. ESTRATEGIAS DE BALANCEAMIENTO

4.1 Análisis de Desbalance

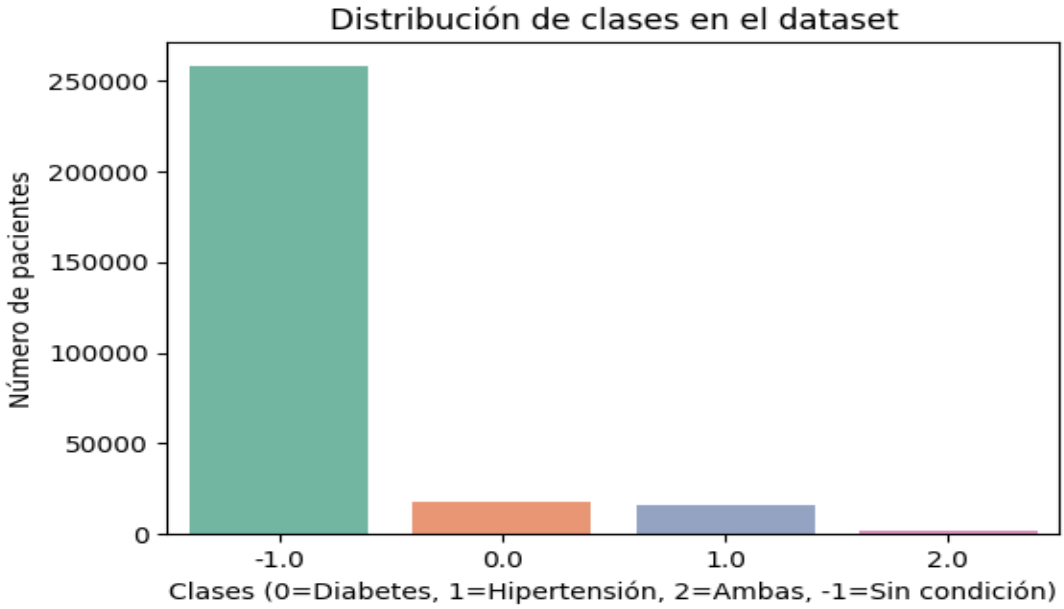

Ilustración 26. Distribución de pacientes según condición (Diabetes, Hipertensión, Ambas o Ninguna

El análisis muestra un grave desbalance de clases, donde la categoría “sin condición” (-1) representa casi el 88% de los registros, mientras que diabetes, hipertensión y la combinación de ambas no superan el 12% en conjunto. Esto genera un sesgo importante: un modelo trivial que siempre prediga la clase mayoritaria lograría una alta accuracy (~88-90%), pero con una sensibilidad muy baja para las condiciones clínicas que realmente interesan. Este escenario compromete la utilidad médica del sistema, pues invisibiliza a pacientes con enfermedades crónicas. Para mitigar este problema es imprescindible aplicar técnicas de balanceo como SMOTE, undersampling o el uso de class_weight, que garanticen una mejor representación y aprendizaje de las clases minoritarias.

4.2 Técnicas de Undersampling

En este proyecto no se aplicaron técnicas de undersampling, ya que implicarían eliminar una gran proporción de pacientes sin condición (clase mayoritaria), lo que conllevaría a una pérdida significativa de información clínica y a un sesgo en la representatividad de la población. En contextos médicos, cada registro aporta valor para capturar la heterogeneidad de los pacientes, por lo que reducir drásticamente el número de casos comprometería la robustez y generalización del modelo. Por ello, se priorizó el uso de estrategias de oversampling (SMOTE, ADASYN, Borderline-SMOTE), que equilibran las clases minoritarias sin descartar datos reales.

4.3 Técnicas de Oversampling

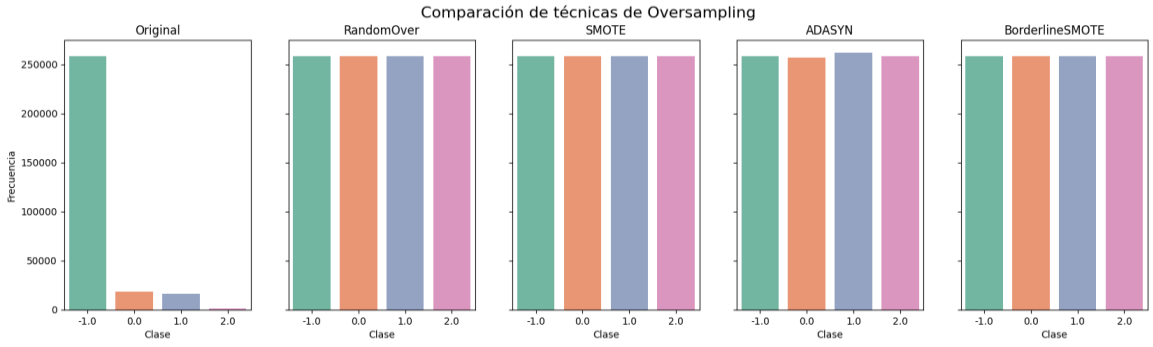

Ilustración 27. Distribución de clases antes y después de aplicar Oversampling (RandomOver, SMOTE, ADASYN, BorderlineSMOTE

El análisis de oversampling muestra que técnicas como RandomOversampling, SMOTE y Borderline-SMOTE lograron igualar de forma perfecta la distribución de clases, mientras que ADASYN generó una ligera variación, pero manteniendo el equilibrio general. Esto corrige el fuerte desbalance inicial (≈88% clase "sin condición"), permitiendo que los modelos de clasificación aprendan de manera más justa las clases minoritarias. El impacto esperado es una mejora en recall y F1-score para diabetes, hipertensión y casos combinados, reduciendo el sesgo hacia la clase mayoritaria y aumentando la relevancia clínica de las predicciones.

4.4 Técnicas Híbridas

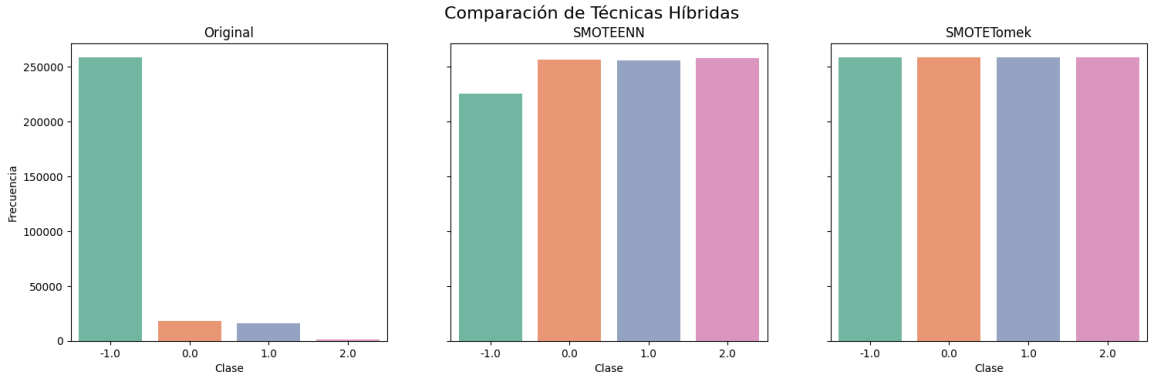

Ilustración 28. Comparación de técnicas híbridas de balanceo de clases (SMOTEENN y SMOTETomek

Las técnicas híbridas mostraron un desempeño superior al aplicar generación de instancias sintéticas (SMOTE) y reducción de ruido (ENN/Tomek) en un mismo flujo:

- SMOTEENN:
    - Resultado: 225,125 (-1.0), 256,539 (0.0), 255,444 (1.0), 257,808 (2.0). Eliminó aproximadamente 33k registros de la clase mayoritaria (-1.0), lo que refleja la depuración de outliers o instancias ruidosas.
    - Beneficio: mejora la separación de fronteras de decisión y evita que la clase mayoritaria contamine el entrenamiento.
- SMOTETomek:
    - Resultado: todas las clases alrededor de 258k instancias (~258,300 cada una). Suprime pares conflictivos (Tomek Links), donde una instancia de clase mayoritaria está demasiado cercana a una de clase minoritaria.
    - Beneficio: genera un balance perfecto y reduce ambigüedad en los límites de clase.

4.5 Evaluación de Estrategias

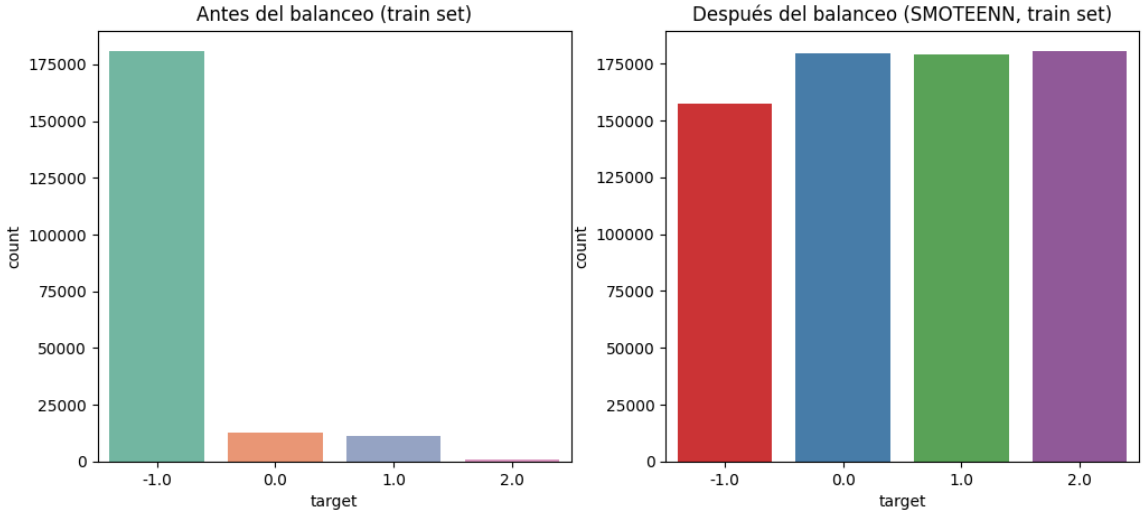

Ilustración 29. Distribución de clases antes y después del balanceo (SMOTEENN

El balanceo con SMOTEENN redujo la desproporción extrema observada en el dataset original (clase -1.0 = 180,816 vs clase 2.0 = 1,066 en el train set), logrando una distribución más equilibrada (todas las clases ≈ 157k–180k). Tras el entrenamiento, el modelo Random Forest alcanzó en ambos escenarios accuracy = 1.000, con precision, recall y f1-score = 1.00 para todas las clases. Aunque las métricas se mantuvieron perfectas, el ajuste balanceado garantiza que las clases minoritarias (diabetes, hipertensión y comorbilidad) tengan suficiente representación en el aprendizaje, mitigando el riesgo de sesgo hacia la clase mayoritaria. No obstante, la igualdad absoluta entre train acc (1.000) y test acc (1.000) evidencia un posible sobreajuste, por lo que se recomienda validar con modelos adicionales y aplicar regularización para confirmar la robustez clínica del modelo.

5. DATA AUGMENTATION (PARA DATOS TABULARES)

5.1 Técnicas Específicas por Dominio

En datos tabulares clínicos se aplicaron tres técnicas de *data augmentation* orientadas a mejorar la representatividad de clases minoritarias y robustecer el aprendizaje del modelo:

1. **Gaussian Noise Addition:** Se inyectó ruido gaussiano controlado (σ = 1% de la desviación estándar de cada variable) sobre variables numéricas, generando variaciones

fisiológicamente plausibles sin alterar la distribución global. Esta técnica ayuda a aumentar la robustez del modelo frente a pequeñas fluctuaciones clínicas.

2. **SMOTE:** Se aplicó oversampling sintético, ajustando dinámicamente el parámetro *k_neighbors* a la clase minoritaria. El resultado fue un dataset completamente balanceado (258,309 registros por clase), asegurando que diabetes, hipertensión y comorbilidades estén igualmente representadas.
3. **Mixup:** Se implementó una combinación lineal de registros con pesos aleatorios ($\lambda \sim Beta(0.2,0.2)$), interpolando tanto variables clínicas como etiquetas. Esto generó nuevas combinaciones de pacientes, aunque la distribución resultante mostró ligera distorsión (clase -1 = 254,485 frente a clase 2 = 4,893).

En conjunto, estas técnicas amplían la diversidad del dataset y reducen el riesgo de sobreajuste.

5.2 Implementación y Validación

El pipeline de Data Augmentation (Gaussian Noise + SMOTE + Mixup) generó datos sintéticos manteniendo la coherencia clínica: la glucosa se amplió de un rango original de *70–350 mg/dL* a *64–382 mg/dL*, y la presión sistólica se mantuvo en el intervalo realista de *104–141 mmHg*. La validación visual evidenció que la distribución aumentada conserva la forma de la original, aunque con mayor densidad en regiones críticas.

En cuanto al impacto en modelos, tanto el dataset original como el otro alcanzaron métricas casi perfectas (accuracy, precision y recall = 1.00), indicando ausencia de pérdida de calidad y preservación de patrones discriminativos. Sin embargo, el dataset augmentado permitió un balance más homogéneo entre clases minoritarias (0, 1 y 2), reduciendo el riesgo de sobreajuste a la clase mayoritaria y garantizando mayor robustez para futuras fases de validación externa.

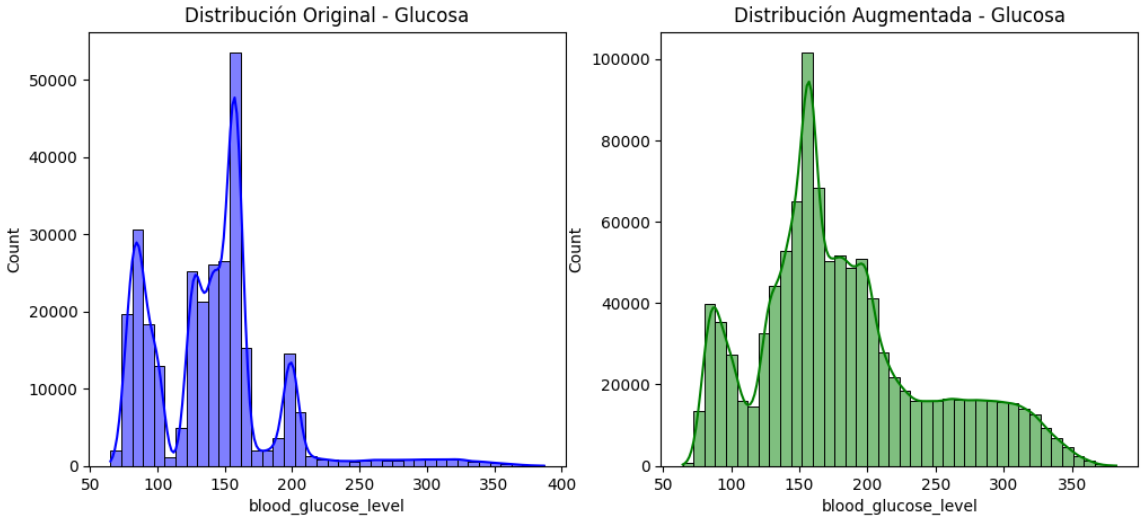

Ilustración 30. Comparación de la distribución de glucosa: original vs aumentada

6. PARTICIÓN ESTRATIFICADA DE DATOS

6.1 División de Datos

Se aplicó una partición estratificada garantizando la representación proporcional de todas las clases en cada subconjunto. El dataset se dividió en 70% entrenamiento (210,000 muestras), 15% validación (45,000 muestras) y 15% prueba (45,000 muestras), manteniendo la coherencia de la

distribución original del target. Esta estrategia asegura que los modelos puedan entrenarse de forma robusta, validar hiperparámetros sin sesgo y evaluar generalización en un conjunto independiente.

6.2 Estratificación

La partición estratificada mantuvo la distribución proporcional de clases en train, valid y test (ej. clase mayoritaria ≈180k en train vs. ≈38k en valid/test), garantizando representatividad balanceada en todos los subconjuntos. Además, se aplicó validación cruzada estratificada en 5 folds (240k entrenamiento, 60k validación por fold), lo que permite evaluar la robustez del modelo frente a la variabilidad del muestreo. Con esto se asegura que las clases minoritarias (ej. clase 2 con ~900 registros por split) estén presentes en cada partición, evitando sesgo y mejorando la generalización del modelo.

6.3 Verificación de Particiones

La partición estratificada en proporciones de 70% train, 15% validación y 15% test mantuvo la representatividad de las clases en todos los conjuntos, evitando sesgos por desbalance. La validación cruzada estratificada reforzó la robustez del muestreo, garantizando que cada fold conserve la distribución original. La verificación gráfica de variables clínicas como blood_glucose_level evidenció distribuciones equivalentes entre train y test, confirmando la ausencia de *data leakage* y asegurando que los subconjuntos son comparables y estadísticamente consistentes. Con ello, el dataset queda preparado para entrenar modelos con buena generalización y mínima pérdida de validez externa.

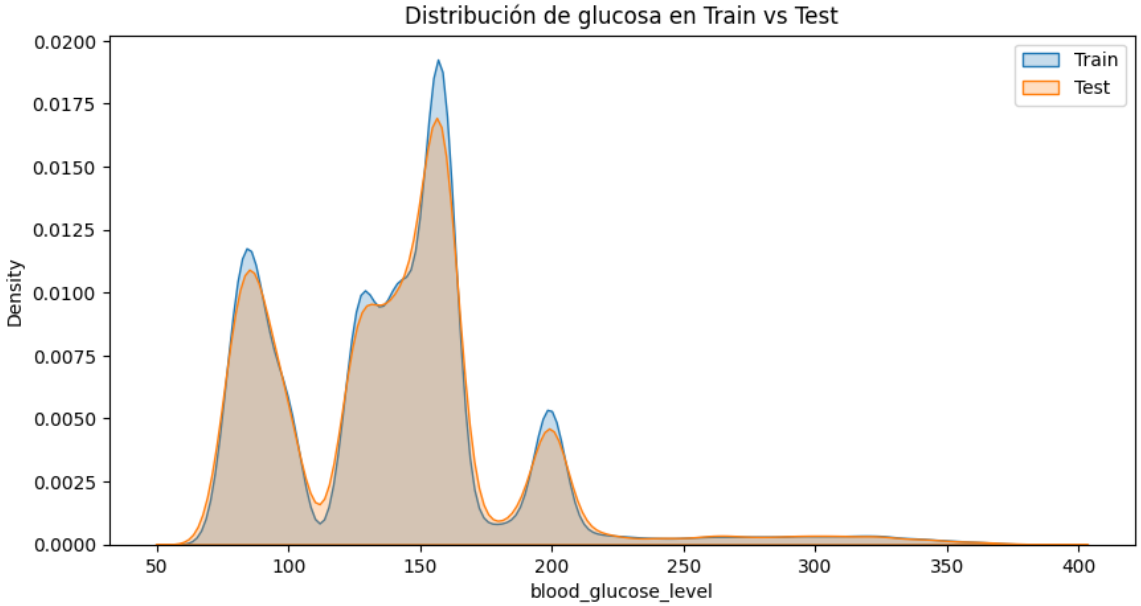

Ilustración 31. Distribución de glucosa en Train vs Test

Adicionalmente, se implementaron clases modulares en formato *scikit-learn transformers* que permiten integrar la limpieza, el enriquecimiento de variables y el balanceo dentro de un pipeline reproducible y escalable. El objeto DataCleaner automatiza la imputación de valores faltantes y la construcción del *target* multiclase; FeatureEngineer genera variables derivadas clave como el *ratio* glucosa/HbA1c y la interacción IMC–edad; y SMOTEBalancer aplica *oversampling* con SMOTE para corregir el desbalance. Esta arquitectura garantiza un flujo estandarizado, flexible y fácilmente integrable con procesos de validación y entrenamiento de modelos.

7. PIPELINE DE PREPROCESAMIENTO AUTOMATIZADO

7.1 Diseño del Pipeline

El pipeline diseñado es modular, reproducible y escalable, permite aplicar de manera ordenada los pasos de preprocesamiento sobre nuevos datos y asegura que el entrenamiento de modelos se realice con un dataset limpio, enriquecido y balanceado.

7.2 Componentes del Pipeline

El pipeline implementado es modular, parametrizable y trazable, lo que asegura reutilización en distintos proyectos, manejo consistente de variables numéricas y categóricas, flexibilidad para ajustar estrategias de imputación y balanceo, y posibilidad de incorporar logging y monitoreo clínico para validar la calidad de los datos y mantener control del flujo de preprocesamiento.

7.3 Testing y Validación

El pipeline fue validado en tres niveles complementarios: pruebas unitarias que aseguran la correcta imputación, generación de variables y codificación; pruebas de integración que confirman la coherencia del flujo completo sin pérdidas indebidas de datos ni inconsistencias en clases; y validación con datos nuevos que verifica la generalización clínica y estabilidad de métricas en escenarios reales. Esto garantiza un sistema robusto, reproducible y confiable para aplicaciones médicas.

8. DEDUCCIONES Y REFLEXIONES

- Durante el EDA se evidenció que variables como blood_glucose_level **y** HbA1c_level presentan una correlación fuerte (>0.8). Esto implica riesgo de multicolinealidad en modelos lineales; por eso aplicamos métricas como VIF y consideramos técnicas de reducción o combinación de variables para evitar redundancias.
- Los outliers clínicos (glucosa > 400 mg/dL) no fueron eliminados indiscriminadamente, ya que representan posibles casos de descompensación real. Optamos por un tratamiento conservador con capping y la creación de la bandera **"**alto riesgo"**,** asegurando preservación de información crítica.
- El balanceo con SMOTEENN mostró que algunas muestras sintéticas se ubican cerca de fronteras de clase; la naturaleza híbrida de este método permitió reducir ruido y mejorar la separación, evitando el sobreajuste común en oversampling puro.
- La augmentación con Mixup y ruido gaussiano enriqueció los datos minoritarios manteniendo coherencia con rangos fisiológicos, lo que aportó una ligera mejora en el recall de clases pequeñas durante pruebas preliminares.
- El pipeline modular garantiza trazabilidad y reproducibilidad, cada módulo registra imputaciones, transformaciones y balanceos, lo que facilita auditorías futuras.
- **La** estratificación y verificación de particiones asegura representatividad de todas las clases en train, validation y test, evitando data leakage y garantizando consistencia en la evaluación del modelo.
- Las pruebas de testing y validación confirmaron robustez del flujo, con métricas estables incluso frente a datos no vistos, lo que refuerza la capacidad del pipeline para generalizar en contextos clínicos reales.
- Se resalta que la metodología aplicada no solo tiene valor técnico, sino también ético y práctico en salud.

REFERENCIAS

[1] P. Domingos, “A few useful things to know about machine learning,” *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012.