CRISTHIAN ERNESTO ROMERO PULLA

TRABAJO DE TITULACION FINAL.docx

Última entrega, 100% avance completo (Moodle PP)

TRABAJO DE TITULACIÓN - P2111-TEÓRICO-N0159-11-N01 (Moodle PP)

PUCE QUITO MOODLE

Detalles del documento

Identificador de la entrega
trn:oid:::1:3490533256

Fecha de entrega
24 feb 2026, 8:30 p.m. GMT-5

Fecha de descarga
25 feb 2026, 3:35 p.m. GMT-5

Nombre del archivo
408_CRISTHIAN_ERNESTO_ROMERO__PULLA_TRABAJO_DE_TITULACION_FINAL_122343_523928763.docx

Tamaño del archivo
4.0 MB

121 páginas

27.632 palabras

171.246 caracteres

6% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado

Fuentes principales

6% Fuentes de Internet
3% Publicaciones
2% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Fuentes principales

6% Fuentes de Internet
3% Publicaciones
2% Trabajos entregados (trabajos del estudiante)

Fuentes principales

Las fuentes con el mayor número de coincidencias dentro de la entrega. Las fuentes superpuestas no se mostrarán.

1 Internet
vitela.javerianacali.edu.co <1%

2 Trabajos del estudiante
UNIBA <1%

3 Internet
clabes.uct.cl <1%

4 Internet
docta.ucm.es <1%

5 Internet
repositorio.uteq.edu.ec <1%

6 Internet
repositorio.uam.es <1%

7 Internet
www.epn.edu.ec <1%

8 Internet
sired.udenar.edu.co <1%

9 Internet
seryus.ddns.net <1%

10 Trabajos del estudiante
Pontificia Universidad Catolica del Ecuador - PUCE <1%

11 Publicación
Varón Capera, Álvaro. "Modelo Para la Evaluación de la Calidad Intrínseca y Diná... <1%

12	Internet	
aportecivico.gobiernoelectronico.gob.ec		<1%
13	Internet	
www.risti.xyz		<1%
14	Publicación	
Dexon-Mckensy Sambola, Kerry Kenton Kelly Kandler, Deyvon Kestner Ordoñez C...		<1%
15	Internet	
1library.co		<1%
16	Publicación	
Holgado Apaza, Luis Alberto. "Modelamiento de la satisfacción laboral de docent...		<1%
17	Internet	
e-spacio.uned.es		<1%
18	Internet	
repository.libertadores.edu.co		<1%
19	Internet	
revista.ucom.edu.py		<1%
20	Internet	
www.slideshare.net		<1%
21	Trabajos del estudiante	
Universidad de Guayaquil		<1%
22	Internet	
cienciadigital.org		<1%
23	Internet	
dialnet.unirioja.es		<1%
24	Trabajos del estudiante	
Universidad Autónoma de Bucaramanga,UNAB		<1%
25	Internet	
repositorio.puce.edu.ec		<1%

26 Internet
repositorio.uta.edu.ec <1%

27 Publicación
González, José Luis Arguelles. "Retos y Obstáculos en la (RE)-Utilización de Datos ... <1%

28 Trabajos del estudiante
Universidad Nacional de Piura <1%

29 Internet
www.ute.edu.ec <1%

30 Trabajos del estudiante
Universidad Nacional Santiago Antunez de Mayolo <1%

31 Trabajos del estudiante
Universidad de Salamanca <1%

32 Internet
doi.org <1%

33 Internet
repositorio.unjbg.edu.pe <1%

34 Internet
uvadoc.uva.es <1%

35 Internet
www.jove.com <1%

36 Publicación
Aceituno Rojo, Miguel Romilio. "Modelo predictivo de análisis de riesgo crediticio ... <1%

37 Trabajos del estudiante
Universidad Nacional de Educación a Distancia <1%

38 Internet
apidspace.javeriana.edu.co <1%

39 Internet
cdn.nestjs.wipolex.wji.prd.web1.wipo.int <1%

40 Internet

core.ac.uk <1%

41 Internet

hdl.handle.net <1%

42 Internet

repositorio.esan.edu.pe <1%

43 Internet

www.clubensayos.com <1%

Pontificia Universidad Católica Del Ecuador

Facultad De Hábitat, Infraestructura Y Creatividad

Maestría En Sistemas De Información Mención Data Science

Trabajo previo a la obtención del título de Magister en Sistemas de Información mención Data Science

Tema:

Construir un modelo predictivo de riesgo de deserción estudiantil mediante técnicas de aprendizaje automático, utilizando datos abiertos de educación superior del período 2015-2023, con el fin de apoyar la toma de decisiones institucionales orientadas a mejorar la permanencia estudiantil en las universidades públicas del Ecuador.

Autor:

Cristhian Ernesto Romero Pulla

Tutor:

Msc. Jorge Alejandro Alarcon Mena

Quito - marzo 2026

DEDICATORIA

26

A mis padres, por ser el apoyo más grande en mi vida. Gracias por su amor incondicional, por estar siempre presentes y por creer en mi incluso cuando yo dudé. De manera muy especial, a mi madre, por su fuerza, sus palabras de ánimo y por motivarme cada día a seguir adelante con mis estudios. Tu ejemplo y tu apoyo han sido fundamentales para no rendirme y continuar persiguiendo mis metas

A mi hermano, por acompañarme en este proceso, por su apoyo sincero y por estar ahí en cada momento, dándome fuerzas cuando más lo necesitaba.

Finalmente, a Nicole Analuisa, por haber sido parte de este proceso. Gracias por tu acompañamiento, tu apoyo y compartir este camino conmigo, dejando una huella importante en esta etapa de mi vida.

AGRADECIMIENTO

Quiero agradecer de corazón a todas las personas que han sido parte de mi camino y de mi crecimiento a lo largo de esta etapa tan importante de mi vida.

De manera muy especial, a mi padrino Martín Romero, quien estuvo a mi lado en los momentos más difíciles y decisivos de mi etapa estudiantil. Gracias por tu apoyo incondicional, por tu ayuda cuando más lo necesitaba y por creer en mí incluso en los momentos de mayor incertidumbre. Tu respaldo marcó una diferencia profunda en mi vida.

A mis padres, gracias por todo el amor, el sacrificio y el esfuerzo que han hecho por mí. Ustedes han sido mi mayor fortaleza, mi refugio y mi motivación para seguir adelante cuando el camino se hizo cuesta arriba. Nada de esto habría sido posible sin su apoyo constante.

A todas las personas que confiaron en mí, que me tendieron una mano, que me brindaron palabras de ánimo y que creyeron en mis capacidades. De manera especial, agradezco a mis jefes por sus consejos, su guía y por enseñarme con el ejemplo, aportando de forma significativa a mi crecimiento personal y profesional.

Asimismo, quiero expresar un agradecimiento muy especial al coordinador Damián Nicolalde, por su tiempo, su disposición y por atender cada una de mis inquietudes con prontitud y compromiso a lo largo de este proceso. De igual manera, agradezco a mi tutor Jorge Alarcón, por sus sugerencias, orientación y acompañamiento constante durante el desarrollo de mi trabajo de titulación.

Finalmente, quiero agradecer a Nicole Analuisa, por su compañía, apoyo y comprensión durante este proceso. Tu presencia, tus palabras y tu respaldo emocional fueron un impulso importante para no rendirme y continuar avanzando.

Este logro también les pertenece a ustedes.

RESUMEN

El presente trabajo tiene como objetivo construir un modelo predictivo de riesgo de deserción estudiantil mediante técnicas de aprendizaje automático, utilizando datos abiertos de educación superior correspondientes al período 2015–2023, con el fin de apoyar la toma de decisiones institucionales orientadas a mejorar la permanencia estudiantil en las universidades públicas del Ecuador. La investigación se desarrolla bajo un enfoque cuantitativo, con un diseño analítico y predictivo, a partir de registros oficiales del sistema de educación superior en formato agregado.

Para el desarrollo del modelo se aplican técnicas de preprocesamiento, selección de características y entrenamiento de algoritmos de aprendizaje automático. Dado que los datos disponibles corresponden a conteos agregados de matrícula, el riesgo de deserción se operacionaliza a nivel de segmento/cohorte mediante patrones y variaciones temporales de matrícula según características académico-programáticas, demográficas e institucionales. El análisis permite identificar segmentos con mayor nivel de riesgo y evaluar el desempeño del modelo mediante métricas de clasificación.

Se concluye que la aplicación de modelos predictivos basados en aprendizaje automático constituye una herramienta útil para generar señales tempranas de riesgo a nivel agregado y facilitar la priorización de estrategias institucionales orientadas a fortalecer la permanencia estudiantil. Este estudio aporta evidencia y un enfoque reproducible basado en datos abiertos para apoyar decisiones en el sistema universitario público ecuatoriano.

Palabras claves: Deserción Estudiantil, Aprendizaje Automático, Modelos Predictivos, Datos Abiertos, Educación Superior.

ABSTRACT

This study aims to build a predictive model of student dropout risk using machine learning techniques and open higher education data from the 2015–2023 period, in order to support institutional decision-making aimed at improving student retention in Ecuador's public universities. The research follows a quantitative approach, with an analytical and predictive design, based on official higher education system records in aggregated format.

To develop the model, data preprocessing, feature selection, and the training of machine learning algorithms are applied. Since the available data correspond to aggregated enrollment counts, dropout risk is operationalized at the segment/cohort level through patterns and temporal variations in enrollment according to academic-programmatic, demographic, and institutional characteristics. The analysis makes it possible to identify segments with higher risk levels and to evaluate model performance using classification metrics.

It is concluded that the application of machine learning–based predictive models is a useful tool for generating early risk signals at the aggregated level and for facilitating the prioritization of institutional strategies aimed at strengthening student retention. This study provides evidence and a reproducible open-data–based approach to support decision-making in Ecuador's public university system.

Keywords: Student Dropout, Machine Learning, Predictive Models, Open Data, Higher Education.

ÍNDICE

turnitin Página 13 de 127 - Engrega de integridad Identificador de la entrega trn:oid:::1:3490533256

turnitin Página 15 de 127 - Engrega de integridad Identificador de la entrega trn:oid:::1:3490533256

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABLAS

CAPÍTULO I: INTRODUCCIÓN

1.1. Antecedentes

El presente trabajo de titulación tiene como objetivo desarrollar un modelo predictivo basado en técnicas de aprendizaje automático para estimar el riesgo de deserción estudiantil en universidades públicas del Ecuador, empleando para ello datos abiertos de educación superior correspondientes al período 2015 – 2023. La predicción temprana del abandono académico se ha convertido en una herramienta fundamental para las instituciones de educación superior, ya que permite identificar oportunamente segmentos/cohortes con mayor probabilidad de interrupción de estudios y, en consecuencia, diseñar estrategias preventivas que fortalezcan la permanencia y el éxito académico. En este contexto, la aplicación de modelos de aprendizaje automático contribuye a analizar patrones, reconocer variables significativas y generar sistemas de alerta oportuna que apoyen la toma de decisiones institucionales orientadas a mejorar los índices de retención estudiantil.

En los últimos años, la necesidad de identificar tempranamente grupos de mayor riesgo ha impulsado el uso de herramientas tecnológicas y analíticas dentro del ámbito educativo. La predicción del riesgo de deserción se ha convertido en una estrategia esencial para fortalecer la permanencia estudiantil, permitiendo que las instituciones implementen acciones preventivas y focalizadas. Sin embargo, las metodologías tradicionales no siempre logran capturar la complejidad del fenómeno, especialmente cuando la interacción entre variables es no lineal o altamente dependiente del contexto.

El aprendizaje automático se presenta como una herramienta eficaz para analizar grandes volúmenes de información y detectar patrones asociados al riesgo de deserción estudiantil. Algoritmos como Random Forest, Adaptive Boosting y Gradient Boosting han demostrado un buen desempeño en la clasificación de perfiles de riesgo, al manejar datos heterogéneos y modelar relaciones complejas entre variables.

A ello se suma la disponibilidad de datos abiertos proporcionados por la SENESCYT, particularmente los registros de matrícula del período 2015 – 2023, que constituyen una fuente valiosa para entrenar modelos predictivos basados en evidencia. En este contexto, la presente investigación desarrolla un modelo de aprendizaje automático para estimar el riesgo de

deserción a nivel agregado a partir de patrones de matrícula por segmento/cohorte, con un enfoque cuantitativo propio de la Ciencia de Datos y orientado a fortalecer la toma de decisiones institucionales y las estrategias de retención.

Los datos utilizados corresponden al registro estadístico de matrícula y se encuentran estructurados como conteos agregados por año, institución y características de la oferta académica y del estudiantado. En consecuencia, el riesgo de deserción se abordará a nivel agregado, construyendo segmentos/cohortes analíticas (combinaciones de variables) y aproximando el riesgo mediante patrones temporales de matrícula (variaciones o caídas interanuales del total) en el período 2015–2023.

1.2. Objetivos

1.2.1. Objetivos General

Construir un modelo predictivo de riesgo de deserción estudiantil mediante técnicas de aprendizaje automático, utilizando datos abiertos de educación superior del período 2015-2023, con el fin de apoyar la toma de decisiones institucionales orientadas a mejorar la permanencia estudiantil en las universidades públicas del Ecuador.

1.2.2. Objetivos Específicos

- ✓ Analizar y depurar los datos de matrícula 2015 – 2023 para identificar variables relevantes y patrones por segmento/cohorte que permitan operacionalizar el riesgo de deserción a nivel agregado.
- ✓ Construir un pipeline de preprocesamiento con limpieza, transformación/codificación y selección de características, incorporando balanceo cuando corresponda según la clase objetivo.
- ✓ Desarrollar al menos tres modelos de aprendizaje automático para estimar el riesgo de deserción a nivel agregado (segmento/cohorte) con base en los patrones de matrícula del período analizado.
- ✓ Evaluar el desempeño de los modelos utilizando métricas como accuracy, precisión, recall, F1-score y AUC, para seleccionar el algoritmo más eficiente.

1.3. Justificación

La deserción estudiantil en las universidades públicas del Ecuador constituye un fenómeno complejo que afecta la continuidad formativa, limita el cumplimiento de los objetivos institucionales y repercute negativamente en los indicadores de eficiencia académica. Identificar tempranamente segmentos o cohortes con riesgo de abandono resulta esencial para implementar estrategias preventivas que fortalezcan la permanencia y reduzcan los costos sociales y económicos asociados a la deserción.

El desarrollo de esta investigación resulta especialmente relevante porque incorpora técnicas de aprendizaje automático para la construcción de un modelo predictivo fundamentado en datos reales del sistema de educación superior del Ecuador. Aunque existen esfuerzos institucionales orientados a reducir la deserción estudiantil, aún no se dispone de herramientas analíticas basadas en evidencia que permitan anticipar el riesgo de abandono con criterios objetivos. En este sentido, el uso de modelos de aprendizaje automático constituye una alternativa metodológica innovadora, capaz de identificar patrones complejos, analizar interacciones no lineales entre variables y generar sistemas de alerta temprana de carácter estratégico, orientados a la toma de decisiones institucionales para fortalecer la permanencia estudiantil.

El empleo de los datos abiertos de la SENESCYT fortalece el carácter transparente, reproducible y verificable del estudio y, además, proporciona una base robusta para el análisis longitudinal del comportamiento de la matrícula entre 2015 y 2023, considerando la disponibilidad y consistencia de los registros para el período analizado. No obstante, esta información no ha sido aprovechada de forma sistemática para aproximaciones predictivas del riesgo de abandono a nivel agregado, lo que evidencia una brecha que esta investigación busca cubrir.

El modelo resultante permitirá a las instituciones de educación superior mejorar la gestión académica, optimizar recursos, focalizar apoyos y diseñar políticas de intervención basadas en métricas objetivas. Asimismo, contribuirá al campo de la analítica educativa en Ecuador, demostrando el potencial del aprendizaje automático y los datos abiertos para abordar problemas estructurales del sistema universitario mediante la identificación de

segmentos/cohortes analíticas con mayor nivel de riesgo, a partir de patrones temporales de matrícula

1.4. Preguntas de investigación

Pregunta Principal

¿Cómo se puede desarrollar un modelo predictivo basado en técnicas de aprendizaje automático que permita estimar el riesgo de deserción estudiantil a nivel agregado (segmento/cohorte) en universidades públicas del Ecuador, utilizando datos abiertos de educación superior del período 2015 – 2023?

Preguntas Secundarias

- ✓ ¿Qué variables académico-programáticos , demográficas e institucionales presentan mayor importancia predictiva en la estimación del riesgo de deserción a nivel agregado, según los datos abiertos analizados?
- ✓ ¿Qué algoritmo de aprendizaje automático (por ejemplo, Regresión Logística, Random Forest, Gradient Boosting, AdaBoost, entre otros) obtiene el mejor desempeño en la estimación del riesgo de deserción a nivel agregado?
- ✓ ¿Cómo afecta el desbalance de clases al desempeño del modelo y qué técnicas de balanceo permiten mejorar sus resultados en la identificación de segmentos/cohortes con mayor nivel de riesgo?

CAPÍTULO II: DESARROLLO

2. Marco Teórico

2.1. La deserción estudiantil: una aproximación conceptual

La deserción estudiantil constituye un fenómeno de alta relevancia social, económica y académica, con profundas repercusiones para el desarrollo de un país. Según (Granda et al., 2024), el abandono de los estudios superiores intensifica brechas de desigualdad y exclusión social, incrementa la vulnerabilidad frente al desempleo y repercute en la salud pública y la participación ciudadana; además, implica pérdida de talento humano y limita el potencial de innovación al afectar la producción de conocimiento y la competitividad.

En este sentido, (Moreira & Caicedo, 2024) señalaron que la deserción estudiantil es un fenómeno multifactorial, asociado a la interacción de variables académicas, demográficas, personales e institucionales, lo que hace indispensable su análisis integral para diseñar estrategias de prevención y mitigación.

Adicionalmente, investigaciones recientes evidencian que el análisis de patrones, tendencias y perfiles de riesgo permite fortalecer la formulación de políticas educativas. En particular, el aprendizaje automático ha demostrado alta capacidad para identificar perfiles de riesgo con niveles elevados de precisión y apoyar la predicción temprana del abandono académico (Bouihi et al., 2024). En esta investigación, estas dimensiones se utilizan como marco conceptual; sin embargo, dado que la fuente corresponde a registros agregados de matrícula, el modelo predictivo se construye con variables observables en dichos datos, enfocándose en factores académico-programáticos, demográficos, institucionales y territoriales, mientras que variables personales o psicológicas se consideran como contexto teórico.

2.2. Tipos de deserción estudiantil

El estudio contemporáneo de la deserción universitaria no se limita a sus causas, sino que incorpora el análisis del momento en que ocurre el abandono, lo cual permite comprender mejor su dinámica y orientar políticas institucionales más efectivas. (Gutiérrez et al., 2024) clasifican la deserción estudiantil, desde una perspectiva temporal, en tres momentos analíticos: deserción inicial o precoz, deserción temprana y deserción tardía. Esta clasificación resulta

relevante para el enfoque predictivo, ya que permite diferenciar patrones de riesgo según etapas del trayecto académico.

2.2.1. Deserción inicial o precoz

La deserción inicial o precoz se presenta cuando el estudiante ha sido admitido en una institución de educación superior, pero no llega a concretar su matrícula ni a iniciar formalmente el primer período académico. Este tipo de abandono ocurre antes del ingreso efectivo a la vida universitaria y suele asociarse a factores económicos, falta de orientación vocacional, cambios en condiciones personales o dificultades administrativas (Behr et al., 2020).

2.2.2. Deserción temprana

La deserción temprana ocurre durante los primeros niveles o ciclos del programa académico y se caracteriza porque el estudiante mantiene la intención de continuar sus estudios superiores, aunque no necesariamente en la misma carrera o institución. Puede manifestarse como:

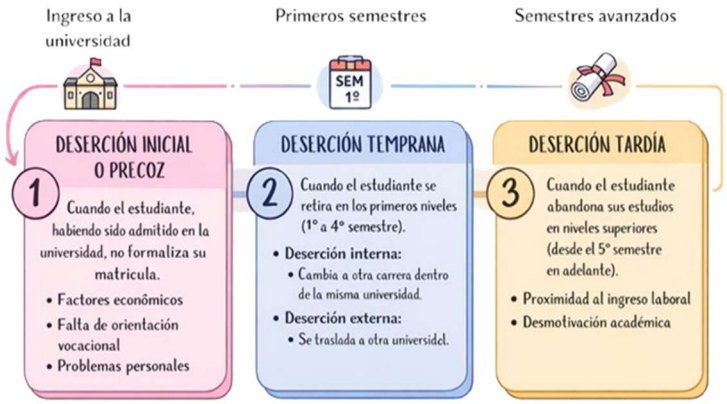
- ***Deserción interna:*** Cuando el estudiante se retira de la carrera actual para trasladarse a otra oferta académica dentro de la misma institución.
- ***Deserción externa:*** Cuando abandona el programa y se matricula en una carrera ofrecida por otra institución de educación superior.

La literatura reporta que este tipo de abandono es frecuente durante los primeros semestres de formación (Gutierrez-Pachas et al., 2023), lo que lo convierte en un foco prioritario para estrategias de alerta temprana.

2.2.3. Deserción tardía

La deserción tardía corresponde al abandono que ocurre después de haber superado la mitad del plan curricular, generalmente a partir del quinto semestre en adelante. En esta etapa, el estudiante ya ha avanzado significativamente, pero interrumpe su trayectoria en niveles superiores o finales. Según (Segura et al., 2022) indican que suele asociarse a factores acumulativos, como desgaste académico, dificultades económicas persistentes o cambios en expectativas profesionales.

Ilustración 1. *Tipos de deserción estudiantil*



Nota. *Fuente: Elaboración propia a partir de (Gutierrez-Pachas et al., 2023).*

En esta investigación, estas categorías se emplean como referencia conceptual; sin embargo, debido al carácter agregado de los datos de matrícula, la estimación del riesgo se realiza a nivel de segmento/cohorte mediante variaciones temporales del total (conteo de matrícula), sin seguimiento individual de trayectorias académicas.

2.3. Factores que influyen en la deserción estudiantil

La deserción en la educación superior ha sido ampliamente reconocida como un fenómeno de carácter multifactorial, resultado de la interacción de diversas dimensiones que inciden en la continuidad académica del estudiante. Según (Quiñónez et al., 2025), estos factores no actúan de manera aislada, sino que se combinan y refuerzan entre sí, incrementando el riesgo de abandono universitario. En el contexto de esta investigación, la identificación de estos factores resulta fundamental no solo para comprender el fenómeno de la deserción, sino también para estructurar variables analíticas que posteriormente serán utilizadas en el desarrollo de modelos predictivos basados en aprendizaje automático, considerando las variables observables en los registros agregados de matrícula (principalmente dimensiones académicas, demográficas, institucionales y territoriales).

Tabla 1. Factores asociados a la deserción estudiantil

Factores	Descripción

Factores académicos	Características de la oferta (carrera, nivel, modalidad, campo) y condiciones de integración institucional pueden asociarse con diferencias en permanencia, especialmente en etapas iniciales de la trayectoria.(Castrillón-Gómez et al., 2020 ; Rivas et al., 2023).
Factores demográficos	Variables como sexo, etnia, discapacidad y residencia pueden influir en la permanencia y aportar valor predictivo(Zerpa & Rodríguez-Montoya, 2024).
Factores personales y psicológicos	La dinámica familiar, el apoyo emocional y la motivación condicionan la continuidad académica, pudiendo convertirse en factores de riesgo ante contextos adversos(Matute et al., 2023; Romualdo Rosario, 2020).
Factores institucionales	La calidad académica, las políticas internas y los programas de acompañamiento influyen directamente en la integración y el compromiso del estudiante con su formación profesional (Saeteros, 2024).

Nota. Fuente: Elaboración propia Romero Cristhian 2026

Desde una perspectiva analítica, estas dimensiones permiten operacionalizar variables predictoras y explorar patrones asociados al abandono académico. Integrarlas en modelos de aprendizaje automático puede apoyar la construcción de alertas tempranas y la toma de decisiones institucionales para fortalecer la permanencia estudiantil en universidades públicas del Ecuador.

2.4. Datos abiertos en la educación superior

Los datos abiertos constituyen un insumo fundamental para el análisis de fenómenos educativos complejos, al permitir el acceso libre, reutilizable y verificable a información

académica, demográfica e institucional, y en algunos casos a indicadores de desempeño estudiantil. Según (López-Pernas et al., 2024), la disponibilidad de datos abiertos en educación superior facilita el desarrollo de procesos analíticos avanzados, tales como el análisis descriptivo, la minería de procesos, el modelado predictivo y las técnicas de agrupamiento, siempre que la información se encuentre adecuadamente anonimizada y documentada.

En el contexto de esta investigación, los datos abiertos representan la fuente principal para la construcción del modelo predictivo de riesgo de deserción, dado que permiten analizar el comportamiento histórico de la matrícula y extraer patrones asociados a posibles situaciones de abandono. Considerando que la información disponible se presenta en formato agregado, el análisis se orienta a identificar tendencias y variaciones temporales por segmento/cohorte, según combinaciones de variables institucionales, académico-programáticas, demográficas y territoriales.

Ilustración 2. *Datos abiertos en la educación superior en Ecuador*

Nota. Fuente: Elaboración propia Romero Cristhian 2026

La ilustración 2 sintetiza el rol de los datos abiertos como base para el análisis educativo seguro cuando existen procesos de anonimización y documentación. En este estudio, dicha condición permite trabajar con registros agregados de matrícula (2015–2023) para estimar riesgo a nivel de segmento/cohorte mediante variaciones temporales del total de matrícula (TOTAL), sin requerir información individual.

2.4.1. Principios internacionales de apertura de datos

Los principios internacionales de apertura de datos constituyen un marco normativo esencial para garantizar la disponibilidad, accesibilidad y reutilización de la información pública empleada en investigaciones científicas y en la formulación de políticas basadas en evidencia. Estos principios se consolidaron en la Carta Internacional de Datos Abiertos, la cual establece que los datos deben ser abiertos por defecto, accesibles, reutilizables, comparables e interoperables, y orientados a fortalecer la gobernanza y el desarrollo sostenible(Open Data Charter, 2015).

En el contexto de esta investigación, dichos principios permiten comprender la importancia de contar con datos públicos de educación superior estructurados y estandarizados, los cuales constituyen un insumo clave para el desarrollo de modelos predictivos orientados a estimar el riesgo de deserción estudiantil en universidades públicas del Ecuador. La disponibilidad de datos abiertos de calidad facilita la aplicación de técnicas de aprendizaje automático; sin embargo, su utilidad analítica depende del cumplimiento de estándares internacionales de interoperabilidad, documentación e integridad.

De acuerdo con (Kunigami & Palomino, 2019), el Banco Interamericano de Desarrollo destacó que la apertura de datos debe sustentarse en condiciones técnicas y jurídicas que aseguren su reutilización efectiva, tales como licencias abiertas y formatos procesables por máquina. Este enfoque resulta especialmente relevante en investigaciones educativas, ya que contribuye a fortalecer la transparencia institucional, impulsar el desarrollo de modelos predictivos y sistemas de alerta temprana, y mejorar la toma de decisiones basada en evidencia.

Desde esta perspectiva, la apertura de datos contribuye a:

- *Fortalece la transparencia en los sistemas educativos.*
- *Impulsar el desarrollo de modelos predictivos y sistemas de alerta temprana.*
- *Mejorar la toma de decisiones institucionales basada en evidencia.*
- *Facilitar el análisis comparativo y multivariado a partir de variables institucionales, territoriales y demográficas disponibles en los registros agregados.*

2.4.2. Datos abiertos de educación superior en Ecuador (SENESCYT - SNIESE)

En Ecuador, la información relacionada con la educación superior se encontró centralizada principalmente en la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) y en el Sistema Nacional de Información de la Educación Superior del Ecuador (SNIESE). Estas entidades administraron y publicaron datos oficiales sobre instituciones, carreras, oferta académica y registros de matrícula, constituyéndose en fuentes fundamentales para el análisis del sistema universitario ecuatoriano.

La SENESCYT publicó reportes institucionales y bases de datos relacionadas con universidades, carreras y procesos académicos, mientras que el SNIESE ofreció indicadores agregados sobre el comportamiento del sistema de educación superior. Adicionalmente, parte de esta información estuvo disponible en el Portal Nacional de Datos Abiertos, desde donde fue posible acceder a conjuntos de datos en formatos descargables (Datos Abiertos, 2019).

Aunque estos conjuntos de datos no cumplieron plenamente con todos los estándares internacionales de apertura, tales como interoperabilidad, actualización continua o disponibilidad de metadatos completos, constituyeron una fuente válida para la construcción de indicadores asociados al comportamiento histórico de la matrícula. En el contexto de esta investigación, los datos públicos disponibles permitieron elaborar variables relevantes para el desarrollo del modelo predictivo de riesgo de deserción estudiantil mediante técnicas de aprendizaje automático.

En el contexto ecuatoriano, la deserción estudiantil en universidades públicas ha sido identificada como un desafío estructural del sistema de educación superior. Informes institucionales de la SENESCYT y registros del SNIESE evidencian variaciones en el comportamiento agregado de la matrícula durante el período 2015–2023, asociadas a características académico-programáticas, demográficas e institucionales. Esta situación refuerza la necesidad de desarrollar modelos predictivos basados en datos abiertos que permitan anticipar el riesgo de abandono a nivel agregado y apoyar la formulación de estrategias institucionales orientadas a mejorar la permanencia estudiantil.

2.4.3. Calidad, estructura y ciclo de vida de los datos abiertos

Los datos abiertos utilizados en esta investigación correspondieron a conjuntos de información de libre acceso y reutilización, los cuales resultaron fundamentales para garantizar la transparencia y la reproducibilidad del análisis (World Bank, 2020). Para su uso efectivo en contextos analíticos y predictivos, fue necesario que cumplieran con estándares relacionados con su calidad, estructura y ciclo de vida.

La calidad de los datos se asoció con su exactitud, integridad y fiabilidad, lo que permitió asegurar la consistencia de los resultados obtenidos (Creswell & Clark, 2017). Asimismo, la organización de la información en formatos procesables por máquina facilitó su tratamiento automatizado y su integración en las fases de preprocesamiento y modelado (W3C, 2017). En particular, al tratarse de registros agregados de matrícula por año y características del segmento, fue indispensable mantener consistencia en campos clave y asegurar la comparabilidad temporal para el análisis longitudinal.

Por su parte, el ciclo de vida de los datos comprendió las etapas de generación, publicación, uso, actualización y archivado o despublicación, influyendo en su trazabilidad y reutilización durante el desarrollo del estudio (Data.gov, 2021). La consideración conjunta de estos elementos fortaleció la consistencia metodológica del trabajo y contribuyó a la confiabilidad de los modelos predictivos desarrollados.

Ilustración 3. Calidad, estructura y ciclo de vida de los datos abiertos

Nota. Fuente: Elaboración propia Romero Cristhian 2026

2.4.4. Limitaciones y retos de los datos abiertos en educación superior

En el desarrollo de esta investigación, el uso de datos abiertos en el contexto de la educación superior presentó diversas limitaciones y retos que condicionaron su aprovechamiento analítico. Aunque la disponibilidad de datos abiertos representó una oportunidad para fortalecer la transparencia, la participación ciudadana y la innovación, su implementación efectiva enfrentó desafíos de carácter técnico, legal, institucional y cultural, los cuales afectaron tanto la calidad de la información como su gestión (World Bank, 2020).

Tabla 2. Limitaciones y retos de datos abiertos

Limitaciones y retos	Descripción
Estandarización y accesibilidad	Los datos abiertos en la educación superior enfrentaron limitaciones relacionadas con la falta de estandarización y con deficiencias en accesibilidad e infraestructura tecnológica (Janssen et al., 2012; Open Data Charter, 2015).
Privacidad y resistencia institucional	La protección de la privacidad de los datos y la resistencia institucional constituyeron barreras relevantes para la implementación de datos abiertos en educación superior (Fischer et al., 2022; Dei, 2024).
Identificación de vacíos de conocimiento	A pesar de los estudios existentes, persistieron vacíos en la investigación sobre el impacto de los datos abiertos en aspectos como resultados académicos, políticas educativas y sostenibilidad de portales de datos abiertos (Alvesson & Sandberg, 2011).

Nota. Fuente: Elaboración propia Romero Cristhian 2026

Estas limitaciones influyen en el alcance de la generalización de resultados (validez externa), ya que la calidad y consistencia de los registros pueden variar entre instituciones y períodos. Por ello, el estudio adopta controles de limpieza, validación y consistencia temporal,

además de una estrategia de evaluación con enfoque temporal, con el fin de reducir riesgos de sesgo y fortalecer la robustez del modelo predictivo a nivel agregado.

2.5. Analítica educativa y ciencia de datos.

En este estudio, la analítica educativa se operacionaliza mediante un pipeline de ciencia de datos (preprocesamiento, ingeniería de variables, modelado y evaluación temporal) aplicado a registros agregados de matrícula. La analítica educativa se entiende como la aplicación de métodos analíticos orientados al análisis de los procesos de enseñanza y aprendizaje, permitiendo el uso de datos para apoyar la toma de decisiones y la identificación de patrones asociados al comportamiento estudiantil (Siemens, 2013).

Dentro de este marco analítico, la analítica predictiva adquiere especial relevancia al permitir anticipar comportamientos futuros a partir de datos históricos. En particular, los modelos de aprendizaje automático, y especialmente aquellos basados en enfoques de *ensemble*, se presentan como herramientas adecuadas para la estimación del riesgo de deserción estudiantil.

En esta investigación, la estimación del riesgo se aborda a nivel de segmento/cohorte, dado que los datos disponibles corresponden a conteos agregados de matrícula.

2.5.1. La analítica educativa (Learning Analytics)

Desde el enfoque del *Learning Analytics*, la analítica educativa se centró en el análisis de datos relacionados con el desempeño de los estudiantes y su interacción con los entornos educativos, permitiendo optimizar los procesos de enseñanza y aprendizaje mediante la recopilación, análisis y visualización de información relevante (Siemens, 2013). En este estudio, la analítica educativa se aplica principalmente a partir de datos administrativos agregados de matrícula, más que sobre registros individuales de desempeño o interacción.

2.5.2. Componentes de la analítica educativa.

La implementación de la analítica educativa se estructuró en fases secuenciales que permitieron transformar los datos educativos en información útil para la toma de decisiones académicas (Siemens, 2013).

Tabla 3. Fases esenciales de la analítica educativa

Fases	Descripción
Recopilación de datos	Recolección de datos provenientes de portales de datos abiertos y plataformas digitales oficiales, para acceder a información de educación superior.
Análisis de datos	Procesamiento y análisis de los datos mediante herramientas estadísticas y técnicas de aprendizaje automático para identificar patrones y tendencias.
Validación de resultados	Presentación de los resultados a través de visualizaciones que facilitaron la interpretación y la toma de decisiones académicas.
Intervención educativa	Uso de los resultados para orientar estrategias institucionales de permanencia y retención, focalizando apoyos en segmentos/cohortes de mayor riesgo.

Nota. Fuente: Elaboración propia Romero Cristhian 2026

2.5.3. Aplicaciones de la analítica educativa.

La analítica educativa se aplica en el contexto académico para el monitoreo del comportamiento estudiantil, la detección temprana de riesgo y la optimización de decisiones académicas e institucionales, contribuyendo a mejorar la efectividad de los procesos formativos(Siemens, 2013). En esta investigación, estas aplicaciones se abordan desde una perspectiva agregada, centrada en la identificación de segmentos/cohortes en riesgo a partir de patrones históricos de matrícula, más que en la personalización individual del aprendizaje.

2.5.4. Desafíos de la analítica educativa

Durante la implementación de la analítica educativa se identificaron desafíos que condicionan su aplicación efectiva. En primer lugar, está la privacidad y el uso ético de los

datos, por lo que se requiere anonimización, seguridad y trazabilidad, incluso cuando se trabaja con información agregada. En segundo lugar, la calidad de los datos (inconsistencias, faltantes y discontinuidades temporales) puede afectar la estabilidad de los análisis y del modelado.

Además, existe riesgo de sesgo por representaciones desiguales entre categorías y por la forma de construcción de variables. Un reto clave es la calidad de la etiqueta/proxy, ya que en este estudio el riesgo se operacionaliza con la variación interanual de matrícula (TASA_1), lo cual puede capturar también pausas temporales o movilidad académica y no solo abandono definitivo. Finalmente, debe evitarse la fuga de información (data leakage), asegurando consistencia temporal en las variables y validación con enfoque temporal.

2.5.5. Tipos de Analítica

La analítica de datos se clasificó en distintos tipos según el objetivo del análisis y el tipo de decisiones que permitió apoyar. Estos enfoques resultaron aplicables al ámbito educativo, al facilitar la comprensión de comportamientos estudiantiles, la identificación de causas subyacentes y la estimación de escenarios futuros relevantes para la toma de decisiones académicas (Shmueli et al., 2017).

Tabla 4. Tipos de analítica en el ámbito educativo

Tipos	Descripción
Analítica descriptiva	Analizó datos históricos para describir lo ocurrido, mediante resúmenes estadísticos que permitieron identificar patrones y comportamientos previos.
Analítica diagnóstica	Explicó las causas de los resultados observados mediante técnicas como correlación y regresión, respondiendo a la pregunta "¿por qué ocurrió?" (Fawcett & Provost, 2013).
Analítica predictiva	Permitió anticipar eventos futuros a partir de patrones históricos, utilizando modelos estadísticos y de aprendizaje automático (Bertsimas & Kallus, 2018).

Analítica prescriptiva	Propuso recomendaciones orientadas a optimizar la toma de decisiones mediante modelos matemáticos y algoritmos avanzados (Davenport & Harris, 2007).

Nota. Fuente: Elaboración propia Romero Cristhian 2026

2.5.6. Aplicaciones de la analítica predictiva en educación

La analítica predictiva adquirió relevancia en el ámbito educativo al permitir anticipar comportamientos y resultados académicos a partir del análisis de patrones históricos de datos. Su aplicación facilitó la identificación temprana de segmentos en riesgo y apoyó la toma de decisiones informadas orientadas a mejorar los procesos formativos.

Una de sus principales aplicaciones se centró en la estimación del riesgo de deserción estudiantil. Mediante el análisis de variables relacionadas con el rendimiento académico, la asistencia y factores demográficos, los modelos predictivos permitieron identificar perfiles de riesgo y apoyar la implementación de estrategias preventivas orientadas a fortalecer la retención estudiantil (Lee et al., 2020).

Asimismo, la analítica predictiva fue utilizada para la personalización del aprendizaje y la evaluación del rendimiento académico, al posibilitar el ajuste de estrategias pedagógicas y la provisión de retroalimentación oportuna en función del comportamiento y desempeño previo de los estudiantes, contribuyendo a una educación más inclusiva y eficaz (Gonzalez & Chiappe, 2024). En esta investigación, la estimación se realiza con datos agregados de matrícula; por ello, el riesgo se aproxima mediante variaciones temporales por segmento/cohorte, sin variables individuales como asistencia o calificaciones

2.5.7. Consideraciones éticas en el uso de datos educativos

El uso de datos educativos en la analítica del aprendizaje y en la construcción de modelos predictivos implicó importantes consideraciones éticas durante el desarrollo de la investigación. La protección de la privacidad de los estudiantes constituyó uno de los aspectos más relevantes, por lo que fue necesario garantizar que los datos fueran recolectados, almacenados y utilizados bajo principios de anonimización y seguridad, asegurando el manejo responsable de la información sensible (Fu & Weng, 2024).

En esta investigación, los datos provienen de fuentes abiertas en formato agregado, por lo que no contienen identificadores personales; sin embargo, se mantienen criterios de uso responsable, trazabilidad y resguardo de la información.

Por otro lado, se consideró indispensable que los sistemas basados en inteligencia artificial y aprendizaje automático garantizaran la equidad y evitaran la discriminación. El control de sesgos en los datos de entrenamiento permitió reducir el riesgo de reproducir desigualdades existentes y asegurar un tratamiento justo de los estudiantes, independientemente de su contexto socioeconómico (Guan et al., 2023; Nguyen et al., 2023).

Finalmente, la transparencia y la explicabilidad de los modelos predictivos garantizaron la comprensión y auditoría de las decisiones basadas en datos, así como el respeto al derecho al olvido y el control de los estudiantes sobre el uso de su información personal (Misiejuk et al., 2025; Guan et al., 2023).

2.5.8. Aprendizaje automático aplicado a la predicción del abandono académico

El aprendizaje automático se empleó en el ámbito educativo como una herramienta para apoyar la toma de decisiones académicas y la gestión institucional. En particular, su aplicación a la predicción del abandono académico permitió identificar tempranamente segmentos/cohortes con mayor riesgo de deserción, facilitando la implementación de intervenciones preventivas focalizadas orientadas a fortalecer la permanencia estudiantil (Gonzalez & Chiappe, 2024; Olive et al., 2025).

En el marco de esta investigación, los modelos analizan patrones en datos agregados de matrícula a nivel de segmento/cohorte, incorporando variables institucionales, académicas, demográficas y territoriales. Algoritmos como Random Forest, Gradient Boosting y Redes Neuronales son adecuados para clasificar niveles de riesgo al capturar relaciones complejas entre variables (Rabelo & Zárate, 2025).

Tabla 5. Aspectos de aprendizaje automático

Aspecto	Descripción
Modelos Predictivos y Algoritmos Comunes	Aplicación de algoritmos como Random Forest, Gradient Boosting y Redes Neuronales para la clasificación del

	riesgo de deserción estudiantil (Bertsimas & Kallus, 2018; Chen & Guestrin, 2016; Rodrigues et al., 2024).
Factores Predictivos del Abandono Académico	Integración de variables académicas, demográficas y de comportamiento estudiantil para mejorar la estimación del riesgo de abandono (Olive et al., 2025; Rebelo Marcolino et al., 2025).
Desafíos y Oportunidades	Presencia de sesgos en los datos y baja interpretabilidad de algunos modelos, junto con oportunidades para diseñar intervenciones personalizadas basadas en análisis predictivo(Jagačić et al., 2024; Forero-Corba & Bennasar, 2024).

Nota. Fuente: Elaboración propia Romero Cristhian 2026

Adicionalmente, modelos como la Regresión Logística y las Redes Neuronales Artificiales han sido ampliamente utilizados como líneas base o enfoques complementarios. No obstante, en la presente investigación se priorizó el uso de modelos ensemble debido a su mayor robustez frente a datos heterogéneos y desbalanceados.

2.6. Técnicas predictivas basadas en modelos de Ensemble.

Las técnicas predictivas basadas en modelos de *ensemble* se fundamentaron en la combinación de múltiples modelos de aprendizaje automático para mejorar la precisión, estabilidad y capacidad de generalización de las predicciones. Estos enfoques resultaron especialmente adecuados para el análisis de datos complejos y heterogéneos del ámbito educativo, donde interactúan factores académicos, demográficos e institucionales (Prenkaj et al., 2020).

En la predicción del riesgo de deserción estudiantil, los modelos *ensemble* permitieron capturar patrones complejos y reducir los errores asociados a modelos individuales, incrementando la robustez de las estimaciones predictivas. Entre las principales técnicas consideradas se encontraron Random Forest, Adaptive Boosting y Gradient Boosting, las cuales se describen en las subsecciones siguientes.

2.6.1. Random Forest

Random Forest se empleó como un algoritmo supervisado basado en modelos *ensemble*, adecuado para el análisis de datos educativos heterogéneos, al combinar múltiples árboles de decisión entrenados sobre subconjuntos aleatorios y generar predicciones mediante votación o promedio, reduciendo la varianza y mejorando la robustez del modelo (Cutler et al., 2012).

Tabla 6. Técnicas predictivas en modelos ensemble de Random Forest

Teoría y funcionamiento	Algoritmo supervisado que combinó múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de datos y variables, generando predicciones robustas mediante votación o promedio
Ventajas y limitaciones	Alto desempeño predictivo y resistencia al sobreajuste, con capacidad para identificar variables relevantes; menor interpretabilidad y mayor costo computacional(Prenkaj et al., 2020 ; Géron, 2022).

Nota. Fuente: Elaboración propia Romero Cristhian 2026

2.6.2. Adaptive Boosting (AdaBoost)

Adaptive Boosting (AdaBoost) se utilizó como una técnica *ensemble* basada en entrenamiento secuencial, orientada a mejorar el desempeño predictivo mediante la corrección iterativa de errores, lo que resultó adecuado para el análisis de datos educativos complejos y desbalanceados (Wu et al., 2019; Géron, 2022).

Tabla 7. Técnicas predictivas en modelos ensemble de Adaptive Boosting

Principios del *boosting*	**Método *ensemble* que optimizó la predicción corrigiendo errores residuales de forma secuencial.**
Reajuste de pesos	Priorizó observaciones mal clasificadas, siendo útil en la identificación de segmentos en riesgo en contextos educativos desbalanceados (Prenkaj et al., 2020).

19

Aplicaciones en deserción estudiantil	Facilitó la identificación de segmentos de alto riesgo y el diseño de intervenciones focalizadas (Nabil et al., 2022).

Nota. Fuente: Elaboración propia Romero Cristhian 2026

2.6.3. Gradient Boosting

Gradient Boosting se empleó como una técnica *ensemble* basada en el entrenamiento secuencial de modelos, orientada a minimizar una función de pérdida mediante el uso del gradiente. Este enfoque permitió corregir errores residuales de forma progresiva y resultó adecuado para el análisis de datos educativos complejos asociados al riesgo de deserción estudiantil (Chen & Guestrin, 2016).

Tabla 8. Técnicas predictivas en modelos ensemble de Gradient Boosting

Optimización mediante gradiente	**Construyó modelos de forma secuencial optimizando una función de pérdida y reduciendo progresivamente el error total.**
Desarrollo iterativo del modelo	Utilizó árboles de baja profundidad y control de la tasa de aprendizaje para equilibrar precisión y capacidad de generalización(Prenkaj et al., 2020).
Captura de relaciones no lineales	Modeló relaciones no lineales e interacciones complejas entre variables, facilitando la identificación de patrones de riesgo de deserción(Wu et al., 2019).

Nota. Fuente: Elaboración propia Romero Cristhian 2026

2.6.4. Comparación de enfoques predictivos para la deserción estudiantil

Durante el desarrollo de la investigación, se compararon distintos enfoques predictivos utilizados en el análisis de la deserción estudiantil, considerando sus ventajas, limitaciones y nivel de pertinencia para el contexto de la educación superior. Esta comparación permitió evidenciar que los modelos basados en *ensemble* presentaron un mejor equilibrio entre

precisión, robustez y capacidad de generalización, en relación con enfoques estadísticos tradicionales, reglas heurísticas y modelos de aprendizaje automático individuales.

Tabla 9. Comparación de enfoques predictivos

Enfoque	Ventajas	Limitaciones	Pertinencia
Estadística tradicional	Interpretabilidad	Baja precisión	Media
Reglas heurísticas	Simple	Subjetiva	Baja
ML individual	Buen desempeño	Sensible a ruido	Alta
Ensemble (propuesto)	Alta precisión, robustez	Mayor complejidad	**Muy alta**

Nota. Fuente: Elaboración propia Romero Cristhian 2026

CAPÍTULO III: METODOLOGÍA DE CONSTRUCCIÓN Y EVALUACIÓN DEL MODELO PREDICTIVO

3.1. Enfoque y tipo de investigación

La investigación adoptó un enfoque cuantitativo, de tipo aplicado y con un alcance explicativo-predictivo, al orientarse a la construcción y evaluación de un modelo de aprendizaje automático para estimar el riesgo de deserción estudiantil a nivel agregado (segmento/cohorte) en universidades públicas del Ecuador. El análisis se basó en datos históricos de educación superior correspondientes al período 2015–2023, lo que permitió identificar patrones y variables relevantes asociados al abandono académico, en concordancia con metodologías propias de la Ciencia de Datos y la analítica educativa(Creswell & Clark, 2017).

Además, la metodología de la investigación se estructuró siguiendo el enfoque del modelo CRISP-DM (Cross-Industry Standard Process for Data Mining), el cual permitió organizar de manera sistemática las etapas de comprensión del problema, análisis de los datos, preparación de la información, construcción del modelo predictivo y evaluación de los resultados. Este enfoque resultó adecuado para el desarrollo de modelos de aprendizaje automático aplicados a la estimación del riesgo de deserción estudiantil a partir de datos abiertos de educación superior(Sifuentes et al., 2023).

Ilustración 4. Enfoque y tipo de investigación

Nota. Fuente: Elaboración propia a partir de (Creswell & Clark, 2017).

La Ilustración 4 resume el enfoque cuantitativo y aplicado, con alcance explicativo-predictivo, adoptado en este estudio. La investigación utiliza datos agregados SNIESE/SENESCYT (2015–2023) y el campo TOTAL (conteo de matrícula) por combinaciones de variables institucionales, académico-programáticas, demográficas y territoriales para analizar variaciones temporales a nivel segmento/cohorte. El proceso se estructuró mediante CRISP-DM, organizando la preparación de datos, el modelado y la evaluación de resultados.

3.2. Diseño de la investigación

La investigación se desarrolló bajo un diseño no experimental y longitudinal, debido a que se analizaron datos históricos de educación superior correspondientes al período 2015-2023, sin manipulación directa de las variables. Este diseño permitió examinar el comportamiento del riesgo de deserción a nivel agregado (segmento/cohorte) a lo largo del tiempo y evaluar patrones asociados al abandono académico mediante técnicas de aprendizaje automático. Asimismo, el estudio se estructuró con un enfoque analítico propio de la Ciencia de Datos, orientado a la construcción y evaluación de un modelo predictivo a partir de información real y observacional(Hernández Sampieri & Fernandez-Collado, 2014).

3.3. Fuente y recolección de los datos

La fuente de información utilizada en la presente investigación correspondió a datos abiertos de educación superior publicados por la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) y el Sistema Nacional de Información de la Educación Superior del Ecuador (SNIESE). Los datos empleados incluyeron conteos agregados de

matrícula(campo TOTAL) de universidades públicas del Ecuador, correspondientes al período 2015-2023, los cuales constituyeron el insumo principal para la construcción del modelo predictivo de riesgo de deserción estudiantil a nivel agregado(segmento/cohorte), aproximado mediante variaciones temporales de la matrícula.

La recolección de los datos se realizó mediante la descarga directa de los conjuntos de datos disponibles en los portales oficiales de datos abiertos, en formatos estructurados y procesables por máquina. Posteriormente, la información fue consolidada y organizada para su análisis, garantizando el uso de datos anonimizados y de acceso público, en concordancia con principios de transparencia, reutilización de la información y apoyo a la toma de decisiones institucionales orientadas a fortalecer la permanencia estudiantil mediante la identificación de segmentos/cohortes con mayor nivel de riesgo.

3.4. Descripción del conjunto de datos utilizado (SNIESE, 2015–2023)

El conjunto de datos utilizado en la investigación correspondió a registros abiertos de educación superior publicados por el Sistema Nacional de Información de la Educación Superior del Ecuador (SNIESE), administrado por la SENESCYT, y comprendió información del período 2015–2023 relacionada con las matrículas en universidades públicas del Ecuador en formato agregado.

Los datos incluyeron conteos agregados de matrícula (campo TOTAL) organizados por combinaciones de variables institucionales, académicas, demográficas y territoriales (p. ej., institución, carrera, nivel de formación, modalidad, sexo, etnia y residencia). Dado que la información no contiene identificadores individuales ni variables directas de “abandono” por estudiante, el riesgo de deserción se operacionalizó a nivel de segmento/cohorte mediante patrones y variaciones temporales de matrícula entre períodos.

Previo a su utilización, el conjunto de datos fue revisado para verificar su consistencia temporal, estructura y completitud, con el fin de asegurar su idoneidad para el desarrollo de modelos predictivos basados en técnicas de aprendizaje automático.

Tabla 10. Características generales del conjunto de datos utilizado

Característica	Detalle

Fuente institucional	Sistema Nacional de Información de la Educación Superior del Ecuador (SNIESE)
Entidad administradora	SENESCYT
Tipo de datos	Datos abiertos de educación superior
Cobertura	Universidades públicas del Ecuador
Período de estudio	2015–2023
Unidad de análisis	Segmento/cohorte (combinación de variables)
Dimensiones consideradas	Académica, institucional y demográfica/territorial
Condición analizada	Riesgo de deserción aproximado por variaciones temporales a nivel agregado
Nivel de agregación	Datos agregados, anonimizados y conteos
Formato original	Archivos estructurados
Uso metodológico	Entrenamiento y evaluación de modelos predictivos

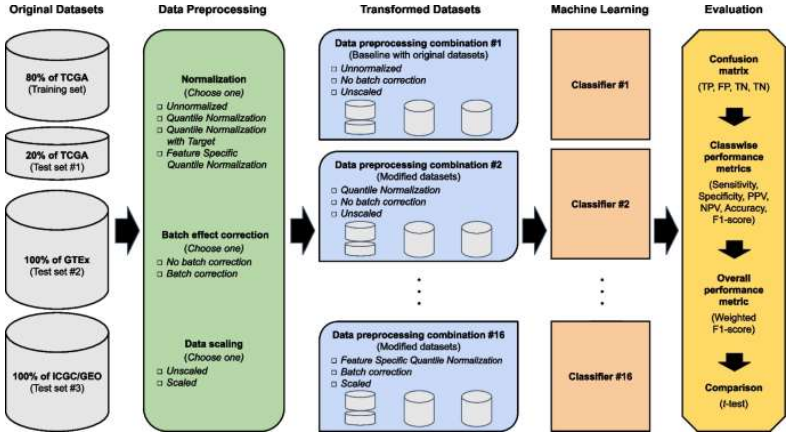
Nota. Fuente: Elaboración propia a partir de datos abiertos del SNIESE (2015–2023).

3.5. Preprocesamiento y preparación de los datos

El conjunto de datos utilizado en la investigación correspondió a registros abiertos de educación superior publicados por el Sistema Nacional de Información de la Educación Superior del Ecuador (SNIESE), administrado por la SENESCYT, y comprendió información del período 2015–2023 en formato agregado, a partir de conteos de matrícula para universidades públicas del Ecuador.

Los datos incluyeron variables de carácter institucional, académicas, demográficas y territoriales, organizadas por combinaciones tipo segmento/cohorte (p. ej., institución–carrera–año–modalidad–sexo–territorio). Dado que la información se presentó como conteos agregados, el riesgo de deserción se aproximó mediante patrones y variaciones temporales de la matrícula a nivel de segmento/cohorte, lo que permitió definir la variable objetivo para el modelado predictivo. La información se analizó en formato agregado y anonimizado, garantizando la protección de datos personales y el cumplimiento de principios éticos en el uso de información pública (Géron, 2022).

Ilustración 5. Pipeline general de preprocesamiento(ML)



Nota. Fuente: Elaboración propia a partir de (Van et al., 2024).

La Ilustración 5 sintetiza el flujo metodológico seguido para preparar los datos y construir el modelo predictivo. El proceso partió del conjunto SNIESE/SENESCYT (2015–2023) en formato agregado, utilizando el campo TOTAL (conteo de matrícula). Luego se realizaron etapas de limpieza, validación y codificación de variables, y se definió la variable objetivo mediante variaciones temporales de matrícula a nivel de segmento/cohorte. Finalmente, se entrenaron distintos modelos de aprendizaje automático y se evaluaron con métricas de clasificación (matriz de confusión, accuracy, precisión, recall, F1-score y AUC) para seleccionar el de mejor desempeño.

3.5.1. Limpieza y depuración de datos

La limpieza y depuración de los datos se llevó a cabo mediante la identificación y tratamiento de registros duplicados, valores faltantes y datos inconsistentes, aplicando criterios acordes con la naturaleza de cada variable. Asimismo, se verificó la coherencia temporal de los registros del período 2015–2023 y la integridad general del conjunto de datos a nivel de segmento/cohorte (combinaciones de variables), con el fin de garantizar su fiabilidad para el análisis posterior.

Estos procedimientos se desarrollaron siguiendo lineamientos ampliamente utilizados en el preprocesamiento de datos para modelos predictivos, los cuales destacan la importancia de contar con información limpia y estructurada para mejorar el desempeño y la estabilidad de los algoritmos de aprendizaje automático (Kuhn & Johnson, 2019).

3.5.2. Transformación y codificación de variables

La transformación y codificación de las variables se realizaron con el objetivo de adecuar la información a un formato compatible con los modelos de aprendizaje automático empleados en la investigación. En esta etapa, las variables categóricas fueron codificadas mediante técnicas apropiadas para permitir su interpretación por parte de los algoritmos predictivos, mientras que la variable TOTAL y las variables numéricas derivadas (cuando correspondió) fueron transformadas para mejorar la estabilidad y el desempeño de los modelos.

Asimismo, cuando correspondió según el algoritmo empleado, se aplicaron procedimientos de escalamiento para evitar sesgos en el entrenamiento y favorecer una representación homogénea de los datos, siguiendo buenas prácticas utilizadas en proyectos de aprendizaje automático (James et al., 2023; Wu et al., 2019).

3.5.3. Selección de características

La selección de características se realizó con el fin de identificar las variables más relevantes para la estimación del riesgo de deserción a nivel agregado (segmento/cohorte) y reducir la complejidad del modelo predictivo. Este proceso se basó en la información institucional, académica, demográfica y territorial disponible en los datos abiertos del SNIESE correspondientes al período 2015–2023, priorizando aquellas variables con mayor relación con los patrones y variaciones temporales de matrícula utilizados para aproximar el riesgo de deserción.

La adecuada selección de variables permitió optimizar el proceso de entrenamiento, disminuir el riesgo de sobreajuste y mejorar la capacidad de generalización de los modelos de aprendizaje automático empleados, siguiendo prácticas ampliamente aceptadas en el desarrollo de modelos predictivos (Brownlee, 2020; Kuhn & Johnson, 2019).

Ilustración 6. Selección de características del modelo predictivo

Nota. Fuente: Elaboración propia a partir de (Brownlee, 2020; Kuhn & Johnson, 2019).

La Ilustración 6 sintetiza la selección de características aplicada a los datos SNIESE/SENESCYT (2015–2023) para identificar variables relevantes en la estimación del riesgo de deserción a nivel agregado (segmento/cohorte). Dado el formato agregado (conteos de matrícula), se priorizaron variables institucionales, académico-programáticas, demográficas y territoriales y su relación con variaciones temporales de matrícula. Este proceso redujo la complejidad del modelo y mejoró el entrenamiento y la generalización de los algoritmos.

3.5.4. Tratamiento del desbalance de clases

El tratamiento del desbalance de clases se justificó porque la variable objetivo RIESGO_ALTO presenta mayor frecuencia en la clase 0 (riesgo bajo) frente a la clase 1 (riesgo alto). En este contexto, un modelo puede alcanzar alta accuracy simplemente prediciendo la clase mayoritaria, pero con baja utilidad preventiva al aumentar los falsos negativos (segmentos en riesgo alto no detectados). Por ello, el modelado se orientó a métricas sensibles a la clase minoritaria (recall y F1-score) y se aplicaron estrategias complementarias: (i) ponderación de clases (p. ej., class_weight='balanced' cuando aplica), (ii) ajuste del umbral de decisión para controlar el trade-off recall–precision según capacidad institucional, y (iii) remuestreo solo en entrenamiento (SMOTE) para mejorar representación de la clase minoritaria. Todas las técnicas

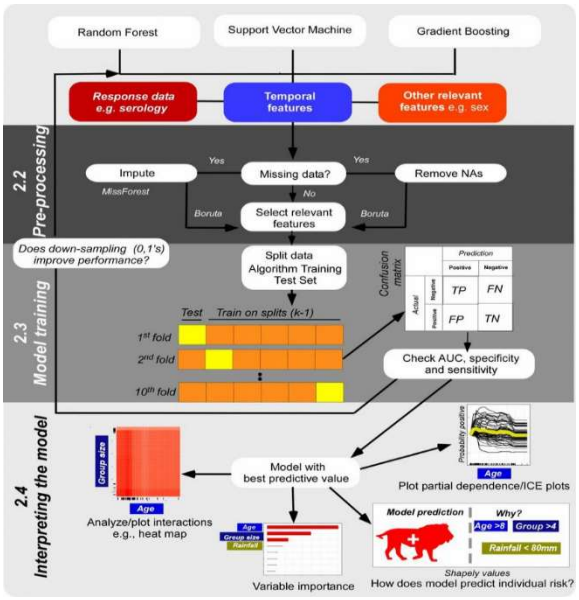
se implementaron sin fuga de información mediante validación temporal, garantizando que las transformaciones y el balanceo se ajusten exclusivamente con datos del periodo de entrenamiento (Fernandez et al., 2018; Krawczyk, 2016).

3.6. Construcción del modelo predictivo

La construcción del modelo predictivo se llevó a cabo utilizando el conjunto de datos previamente preprocesado, transformado y balanceado, con el objetivo de estimar el riesgo de deserción a nivel agregado (segmento/cohorte) en universidades públicas del Ecuador. Para ello, se emplearon técnicas de aprendizaje automático orientadas a problemas de clasificación binaria, considerando como variable objetivo la clase de riesgo (p. ej., alto/bajo) definida a partir de patrones y variaciones temporales de la matrícula en cada segmento/cohorte.

Durante esta etapa, el conjunto de datos fue dividido en subconjuntos de entrenamiento, validación y prueba, respetando el orden temporal de los registros para evitar fuga de información y mantener la coherencia longitudinal del análisis. Adicionalmente, la validación se realizó mediante un esquema de validación cruzada temporal (por bloques de tiempo) aplicado únicamente sobre el conjunto de entrenamiento, con el fin de reducir la variabilidad en los resultados y obtener estimaciones más robustas del desempeño de los modelos (Géron, 2022). Se implementaron algoritmos basados en enfoques de ensemble, tales como Random Forest, Adaptive Boosting y Gradient Boosting, debido a su capacidad para manejar datos heterogéneos y capturar relaciones complejas entre variables. Asimismo, se realizó un proceso de optimización de hiperparámetros, orientado a mejorar el rendimiento predictivo y la capacidad de generalización de los modelos, siguiendo buenas prácticas ampliamente aceptadas en proyectos de aprendizaje automático aplicados al ámbito educativo (Fernandez et al., 2018; James et al., 2023).

Ilustración 7. Proceso predictivo en aprendizaje automático



Nota. Fuente: Elaboración propia a partir de(Fountain-Jones et al., 2019).

La Ilustración 7 resume el flujo general de construcción de modelos predictivos; preprocesamiento (faltantes y selección de características), partición entrenamiento–prueba y validación, entrenamiento de clasificadores (p. ej., Random Forest y Gradient Boosting) y evaluación con matriz de confusión y métricas como AUC, sensibilidad y especificidad. Finalmente, se incluyen técnicas de interpretabilidad (importancia de variables y SHAP) para explicar los resultados. En esta investigación, este flujo se aplicó a datos agregados por segmento/cohorte, usando conteos de matrícula (TOTAL) y sus variaciones temporales para estimar niveles de riesgo.

3.6.1. Definición de la variable objetivo

La variable objetivo se definió con el propósito de representar el riesgo/condición aproximada a nivel agregado (segmento/cohorte) dentro del sistema de educación superior pública del Ecuador. Para ello, se estableció una variable categórica binaria, en la cual se distinguieron segmentos/cohortes con continuidad de matrícula y segmentos/cohortes con comportamiento compatible con abandono durante el período analizado (Tsiakmaki et al., 2020).

La construcción de esta variable se realizó a partir de los conteos agregados de matrícula disponibles en los datos abiertos del SNIESE para 2015–2023. En consecuencia, la deserción

no se midió de forma individual, sino que se aproximó mediante variaciones temporales en la matrícula por segmento/cohorte, lo que permitió formular el problema como una tarea de clasificación y aplicar técnicas de aprendizaje automático para estimar el riesgo (Mduma et al., 2019).

3.6.2. División del conjunto de datos

La división del conjunto de datos se realizó con el objetivo de garantizar un proceso de entrenamiento y evaluación adecuado de los modelos predictivos desarrollados. Para ello, el conjunto de datos fue separado en subconjuntos de entrenamiento, validación y prueba, respetando la secuencia temporal del período 2015–2023, con el fin de evaluar la capacidad de generalización de los modelos de manera controlada.

Adicionalmente, se aplicaron esquemas de validación cruzada temporal sobre el conjunto de entrenamiento, con el propósito de reducir la variabilidad asociada a una única partición y obtener estimaciones más robustas del desempeño. Este procedimiento permitió optimizar el uso de la información disponible y minimizar el riesgo de sobreajuste durante la construcción del modelo predictivo (Géron, 2022; James et al., 2023).

Tabla 11. División del conjunto de datos para el entrenamiento y evaluación del modelo

Subconjunto	Propósito metodológico	Uso dentro del modelo
Entrenamiento	Ajuste de parámetros del modelo y estimación interna del desempeño	Entrenamiento de algoritmos y validación cruzada (si aplica)
Validación	Ajuste y selección de hiperparámetros	Evaluación intermedia para comparar configuraciones
Prueba	Evaluación final e independiente del desempeño	Estimación de la capacidad de generalización (uso único al final)

Nota. Fuente: Elaboración propia Romero Cristhian 2026

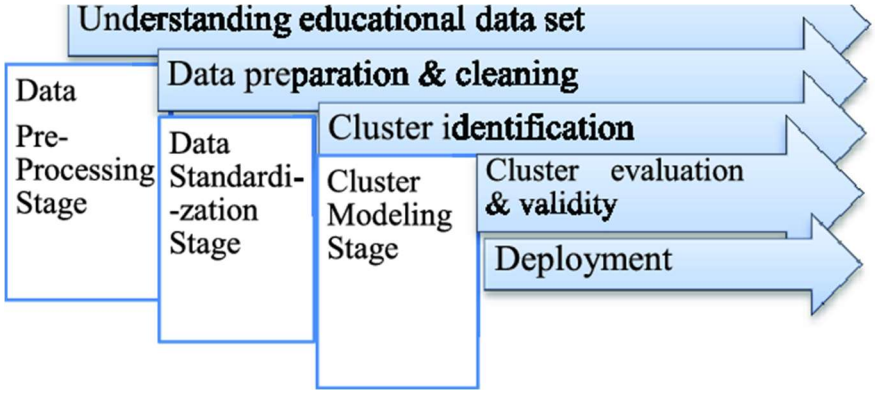
3.6.3. Implementación de los modelos de aprendizaje automático

La implementación de los modelos de aprendizaje automático se realizó a partir del conjunto de datos previamente preparado, con el fin de estimar el riesgo de deserción a nivel agregado (segmento/cohorte) mediante un enfoque de clasificación supervisada. En esta fase se desarrollaron y entrenaron distintos algoritmos predictivos, considerando su desempeño y su capacidad de generalización en un contexto educativo con variables institucionales, académico-programáticas, demográficas y territoriales.

Se implementaron modelos basados en técnicas *ensemble*, específicamente Random Forest, AdaBoost (Adaptive Boosting) y Gradient Boosting, debido a su efectividad para modelar relaciones no lineales y manejar múltiples variables predictoras. Cada modelo fue entrenado utilizando el subconjunto de entrenamiento e incorporó validación cruzada temporal para evaluar su estabilidad y reducir el riesgo de sobreajuste durante el proceso de aprendizaje (Dutt et al., 2017; Jadama & Toray, 2024).

Adicionalmente, se aplicaron procedimientos de optimización de hiperparámetros con el objetivo de ajustar los valores internos de cada algoritmo y mejorar su rendimiento predictivo. Este proceso permitió identificar configuraciones más eficientes para cada modelo, manteniendo un equilibrio entre precisión y capacidad de generalización. La implementación se realizó siguiendo buenas prácticas ampliamente aceptadas en proyectos de aprendizaje automático, priorizando la reproducibilidad y la consistencia metodológica del estudio (Bergstra & Bengio, 2012).

Ilustración 8. Proceso de análisis y modelado de datos educativos



Nota. Fuente: Elaboración propia a partir de(Dutt et al., 2017).

La Ilustración 8 resume un flujo general de análisis de datos educativos aplicado en esta investigación al conjunto SNIESE/SENESCYT (2015–2023) en formato agregado. El proceso incluyó comprensión del conjunto, limpieza y codificación, construcción del modelo predictivo para identificar segmentos/cohortes de mayor riesgo a partir de patrones temporales de matrícula, y evaluación mediante métricas de clasificación. Finalmente, los resultados se plantean como insumo para alertas y decisiones institucionales a nivel agregado.

Tabla 12. Comparación entre técnicas de ensamble: Bagging, Boosting y Stacking

Factor	Bagging	Boosting	Stacking
Enfoque	Considera principalmente aprendices débiles homogéneos.	Considera principalmente aprendices débiles homogéneos.	Considera aprendices débiles heterogéneos (combinación de distintos algoritmos de aprendizaje).
Sesgo y varianza	El objetivo del bagging no es eliminar el sesgo, sino mantenerlo estable y reducir la varianza.	El boosting se utiliza cuando el objetivo es reducir el sesgo del modelo.	El stacking también se emplea cuando el objetivo es reducir el sesgo mediante la combinación de modelos.
Aplicación	La implementación de los siguientes algoritmos permite el uso práctico del bagging: • Random Forest • Bagging canónico	El boosting es una estrategia que puede implementarse mediante los siguientes algoritmos: • AdaBoost • XGBoost	El stacking es un enfoque práctico que puede implementarse mediante los siguientes algoritmos: • Stacking canónico • Blending
Método de selección de subconjuntos	Durante el proceso de bagging, los subconjuntos de entrenamiento se seleccionan aleatoriamente y luego se realizan reemplazos a partir del conjunto de datos original.	En el procedimiento de boosting, se seleccionan subconjuntos adicionales de manera aleatoria a partir del conjunto de datos ponderado, y posteriormente se realizan reemplazos.	En el stacking, cada modelo se entrena utilizando el conjunto de datos completo, aplicando el conocimiento aprendido del conjunto general.
Dependencias	En bagging, cada modelo se desarrolla de manera independiente de los demás.	En boosting, el desempeño de los modelos posteriores depende del rendimiento del modelo anterior.	El stacking ilustra cómo se crea un modelo generalizado integrando múltiples modelos en uno solo.
Conclusión	El bagging consiste en ajustar múltiples árboles de decisión a subconjuntos del mismo conjunto de datos y promediar las predicciones obtenidas.	El boosting consiste en agregar secuencialmente modelos de ensamble que corrigen las predicciones de modelos anteriores, generando un promedio ponderado de las predicciones.	El stacking combina distintos tipos de modelos entrenados sobre los mismos datos y utiliza un modelo secundario para determinar la forma óptima de combinar sus predicciones.

Nota. Fuente: Elaboración propia a partir de(Jadama & Toray, 2024).

3.6.4. Ajuste y optimización de hiperparámetros

El ajuste de hiperparámetros se realizó con el objetivo de mejorar el rendimiento predictivo y la capacidad de generalización de los modelos de aprendizaje automático implementados para la estimación del riesgo de deserción a nivel agregado (segmento/cohorte). Este proceso permitió identificar configuraciones adecuadas de los parámetros externos que

controlan el comportamiento de cada algoritmo, reduciendo el riesgo de sobreajuste o subajuste durante el entrenamiento.

Para ello, se aplicaron estrategias sistemáticas de búsqueda de hiperparámetros, evaluando distintas combinaciones y seleccionando aquellas que ofrecieron un mejor equilibrio entre desempeño y estabilidad. La optimización se desarrolló mediante validación cruzada sobre el conjunto de entrenamiento, lo que permitió obtener estimaciones más robustas y reducir la dependencia de una única partición de los datos. Este procedimiento se llevó a cabo siguiendo buenas prácticas en proyectos de aprendizaje automático, priorizando la reproducibilidad, la consistencia metodológica y la eficiencia computacional (Hutter et al., 2019; Probst et al., 2018).

3.7. Evaluación y validación del modelo predictivo

La evaluación y validación del modelo predictivo se realizó con el propósito de medir su capacidad para estimar de forma confiable el riesgo de deserción a nivel agregado (segmento/cohorte) en universidades públicas del Ecuador. Para ello, se aplicaron métricas de desempeño adecuadas a problemas de clasificación binaria y a contextos caracterizados por el desbalance de clases. El desempeño de los modelos fue evaluado mediante métricas como accuracy, precisión, recall, F1-score y área bajo la curva ROC (AUC), las cuales permitieron analizar de manera integral la calidad de las predicciones generadas. En particular, se priorizó el uso de métricas sensibles a la identificación de segmentos/cohortes con mayor nivel de riesgo, debido a que una clasificación errónea podría afectar la implementación de estrategias institucionales orientadas a la prevención del abandono académico (Kuhn & Johnson, 2019; Powers, 2020).

Adicionalmente, se aplicaron esquemas de validación cruzada, con el fin de reducir la variabilidad de los resultados y garantizar una estimación más robusta del desempeño de los modelos. Este procedimiento permitió evaluar la estabilidad de los algoritmos frente a distintas particiones del conjunto de datos, fortaleciendo la confiabilidad de los resultados obtenidos (Fawcett & Provost, 2013).

Finalmente, los resultados de la evaluación permitieron comparar el desempeño de los distintos algoritmos implementados y seleccionar el modelo con mejor equilibrio entre

precisión, capacidad de generalización e interpretabilidad. Este proceso de validación constituyó un insumo fundamental para respaldar la aplicabilidad del modelo predictivo como herramienta de apoyo a la toma de decisiones institucionales orientadas a mejorar la permanencia estudiantil.

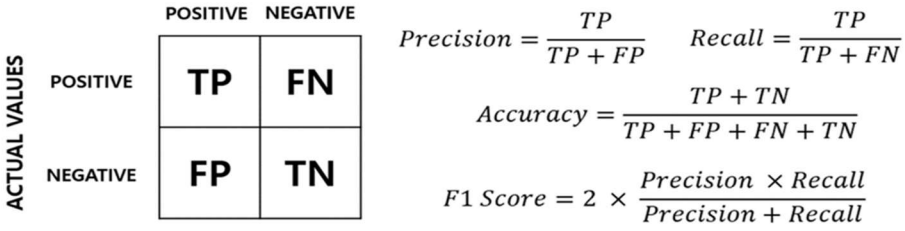
3.7.1. Métricas de evaluación empleadas

Las métricas de evaluación empleadas permitieron analizar de manera integral el desempeño de los modelos predictivos desarrollados para estimar el riesgo de deserción estudiantil a nivel agregado (segmento/cohorte). Dado que el problema abordado correspondió a una tarea de clasificación binaria y presentó un desbalance entre clases, se seleccionaron métricas que ofrecieron una evaluación más robusta que el uso exclusivo de la exactitud global.

En este sentido, se utilizaron las métricas de accuracy, precisión, recall y F1-score, las cuales permitieron evaluar tanto la capacidad general de clasificación como el equilibrio entre la identificación correcta de segmentos/cohortes con mayor nivel de riesgo y la reducción de errores de clasificación. Particularmente, el recall fue considerado una métrica clave, debido a su importancia en la detección de segmentos en riesgo, ya que una baja sensibilidad podría derivar en la omisión de casos relevantes para la toma de decisiones institucionales (Haixiang et al., 2017; Powers, 2020).

Adicionalmente, se empleó el área bajo la curva ROC (AUC) como una métrica global de desempeño, al permitir evaluar la capacidad discriminativa de los modelos independientemente del umbral de clasificación utilizado. El uso conjunto de estas métricas permitió realizar una comparación objetiva entre los algoritmos implementados y seleccionar el modelo con mejor equilibrio entre precisión, sensibilidad y capacidad de generalización (Krawczyk, 2016; Saito & Rehmsmeier, 2015).

Ilustración 9. Matriz de confusión: métricas de precisión, recall, exactitud y F1-score.



Nota. Fuente: Elaboración propia a partir de(Seol et al., 2023).

3.7.2. Estrategia de validación

La estrategia de validación se diseñó con el propósito de garantizar la robustez, estabilidad y capacidad de generalización de los modelos predictivos desarrollados para la estimación del riesgo de deserción a nivel agregado. Dado que el estudio se basó en datos históricos y presentó un desbalance entre clases, fue necesario aplicar procedimientos de validación que permitieran evaluar el desempeño de los algoritmos de manera confiable y evitar conclusiones sesgadas.

En este contexto, se utilizó un esquema de validación cruzada temporal (k particiones en orden cronológico), también conocido como validación tipo forward chaining (TimeSeriesSplit). A diferencia del k-fold aleatorio, este procedimiento respeta la secuencia temporal del período 2015–2023 en cada iteración, el modelo se entrenó con observaciones de años anteriores y se validó con un bloque posterior de tiempo, repitiendo el proceso hasta completar las particiones. Este procedimiento se aplicó sobre el subconjunto de entrenamiento con el fin de estimar el rendimiento del modelo de manera más estable y sin introducir fuga de información.

La validación cruzada resultó especialmente pertinente en problemas de clasificación con clases desbalanceadas, ya que permitió evaluar la consistencia de los modelos frente a diferentes particiones de los datos y minimizar la variabilidad de los resultados obtenidos. Asimismo, esta estrategia facilitó la comparación objetiva entre los algoritmos implementados, al proporcionar métricas promedio que reflejaron de manera más precisa su comportamiento general (Alpaydin, 2020; Kelleher et al., 2015; Lee, 2020).

En consecuencia, la estrategia de validación adoptada fortaleció la confiabilidad de los resultados del estudio y contribuyó a seleccionar el modelo predictivo con mejor equilibrio entre precisión, sensibilidad y capacidad de generalización, aspectos fundamentales para su aplicación en contextos institucionales orientados a la toma de decisiones en educación superior.

3.8. Herramienta y tecnologías utilizadas

El procesamiento y análisis de la información se realizó utilizando el lenguaje de programación Python, debido a su versatilidad y a la disponibilidad de bibliotecas

especializadas para la manipulación de datos, el modelado predictivo y la evaluación de modelos. La ejecución del código se llevó a cabo mediante Google Colab, un entorno de computación en la nube basado en Jupyter Notebook, que permitió desarrollar y ejecutar notebooks sin instalación local, facilitando la experimentación, la documentación del proceso y la reproducibilidad de los resultados.

Para la preparación y transformación del conjunto de datos se utilizaron Pandas y NumPy, las cuales permitieron realizar tareas de limpieza, estructuración y manejo eficiente de datos tabulares. La construcción y entrenamiento de los modelos predictivos se efectuó con Scikit-learn, que proporcionó implementaciones consolidadas de algoritmos de clasificación, validación cruzada, optimización de hiperparámetros y métricas de evaluación.

Asimismo, para el análisis exploratorio y la representación gráfica de resultados intermedios, se emplearon Matplotlib y Seaborn, lo que facilitó la visualización de distribuciones, tendencias y patrones generales presentes en los datos. En conjunto, estas herramientas permitieron desarrollar un pipeline metodológico coherente, transparente y alineado con buenas prácticas en proyectos de aprendizaje automático aplicados a la educación superior

3.9. Procedimiento metodológico

Asimismo, para el análisis exploratorio de los datos y la representación gráfica de información intermedia, se emplearon las bibliotecas Matplotlib y Seaborn, las cuales facilitaron la visualización de distribuciones, tendencias y patrones generales presentes en los datos. El uso conjunto de estas herramientas permitió desarrollar un pipeline metodológico coherente, transparente y alineado con buenas prácticas en proyectos de aprendizaje automático aplicados a la educación superior.

El procedimiento metodológico de la investigación se desarrolló de manera secuencial y estructurada, siguiendo buenas prácticas de la ciencia de datos y del aprendizaje automático, con el objetivo de construir y evaluar un modelo predictivo de riesgo de deserción estudiantil basado en datos abiertos de educación superior.

En una primera etapa, se realizó la recolección de los datos abiertos correspondientes al período 2015–2023 desde las fuentes oficiales del Sistema Nacional de Información de la

Educación Superior del Ecuador (SNIESE), en formato agregado (conteos de matrícula) por segmento/cohorte. Posteriormente, se efectuó una revisión inicial del conjunto de datos para verificar su estructura, consistencia temporal y completitud, asegurando su idoneidad para el análisis predictivo.

En la siguiente fase, se llevó a cabo el preprocesamiento y preparación de los datos, que incluyó la limpieza y depuración de observaciones agregadas, la transformación y codificación de variables, la selección de características relevantes y el tratamiento del desbalance de clases. Estas actividades permitieron mejorar la calidad de los datos y optimizar su uso en el entrenamiento de los modelos de aprendizaje automático.

Una vez preparados los datos, se procedió a la construcción de los modelos predictivos, formulando el problema como una tarea de clasificación binaria para estimar el riesgo de deserción a nivel agregado (segmento/cohorte), aproximado mediante variaciones temporales de matrícula. Para ello, el conjunto de datos fue dividido en subconjuntos de entrenamiento, validación y prueba, respetando la secuencia temporal del período 2015–2023, y se aplicaron algoritmos basados en enfoques de ensemble, tales como Random Forest, Adaptive Boosting y Gradient Boosting. Adicionalmente, se realizó la optimización de hiperparámetros con el fin de mejorar el desempeño y la capacidad de generalización de los modelos.

Posteriormente, se implementó una estrategia de validación cruzada sobre el subconjunto de entrenamiento, orientada a evaluar la estabilidad de los modelos frente a diferentes particiones del conjunto de datos y a reducir la variabilidad de los resultados.

La evaluación del desempeño se efectuó mediante métricas de clasificación adecuadas a contextos desbalanceados, tales como accuracy, precisión, recall, F1-score y el área bajo la curva ROC (AUC).

Finalmente, los resultados obtenidos permitieron comparar objetivamente los modelos desarrollados y seleccionar aquel con mejor equilibrio entre precisión, sensibilidad y capacidad de generalización. El procedimiento metodológico adoptado garantizó la coherencia interna del estudio y proporcionó una base sólida para el análisis de resultados y la discusión de hallazgos en los capítulos posteriores, orientados a apoyar la toma de decisiones institucionales en educación superior.

CAPÍTULO IV: RESULTADOS Y ANÁLISIS

4.1. Resultados

4.1.1. Caracterización del dataset SNIESE/SENESCYT (2015–2023)

El conjunto de datos utilizado corresponde a la Base estadística del Registro de matrícula de Universidades y Escuelas Politécnicas (UEP), publicada como datos abiertos por la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) mediante el Sistema Nacional de Información de la Educación Superior (SNIESE). La base se encuentra estructurada en formato tabular y contiene registros agregados de matrícula para el período 2015–2023, los cuales se emplean como insumo para el análisis longitudinal del comportamiento de la matrícula en universidades públicas del Ecuador.

En términos estructurales, cada fila representa un segmento/cohorte definido por la combinación de variables institucionales, académicas, demográficas y territoriales; mientras que el campo TOTAL corresponde al conteo agregado de matrículas para dicha combinación en un año determinado. En consecuencia, la unidad de análisis en esta investigación no es el estudiante individual, sino el segmento/cohorte (combinación de atributos) para el cual se observa la evolución temporal del conteo de matrícula (Datos Abiertos, 2019).

Tabla 13. Variables incluidas en la base SNIESE/SENESCYT (2015–2023)

Variable	Descripción	Tipo / escala	Ejemplo
AÑO	Período de referencia del registro agregado	Numérica (año)	2015
NOMBRE_IES	Institución de educación superior (universidad/escuela politécnica)	Categórica nominal	Escuela Politécnica Nacional
TIPO_FINANCIAMIENTO	Tipo de financiamiento de la IES	Categórica nominal	Pública
NOMBRE_CARRERA	Carrera/programa académico	Categórica nominal	Ingeniería Agroindustrial
NIVEL_FORMACION	Nivel de formación	Categórica ordinal/nominal	Tercer nivel de grado
MODALIDAD	Modalidad de estudios	Categórica nominal	Presencial
CAMPO_AMPLIO	Campo amplio de conocimiento	Categórica nominal	Agricultura, silvicultura…
CAMPO_ESPECIFICO	Campo específico de conocimiento	Categórica nominal	Agricultura, silvicultura…

CAMPO_DETALLADO	Campo detallado de conocimiento	Categórica nominal	(según registro)
TIPO_SEDE	Tipo de sede	Categórica nominal	Matriz
PROVINCIA_SEDE	Provincia de la sede	Categórica nominal	Pichincha
CANTON_SEDE	Cantón de la sede	Categórica nominal	Quito
SEXO	Sexo reportado en el registro agregado	Categórica nominal	Hombre / Mujer
ETNIA	Categoría étnica reportada	Categórica nominal	Afroecuatoriano
PUEBLO_NACIONALIDAD	Pueblo/nacionalidad (si aplica)	Categórica nominal	No aplica
DISCAPACIDAD	Condición de discapacidad (si aplica)	Categórica nominal	Ninguna
PROVINCIA_RESIDENCIA	Provincia de residencia reportada	Categórica nominal	Pichincha
CANTON_RESIDENCIA	Cantón de residencia reportado	Categórica nominal	Quito
TOTAL	Conteo agregado de matrículas para la combinación de variables	Numérica (conteo)	2

Nota. Fuente: Elaboración propia a partir de SNIESE/SENESCYT 2015–2023(Datos Abiertos, 2019).

4.1.2. Construcción de segmentos/cohortes (unidad de análisis agregada)

En esta investigación, la unidad de análisis corresponde a un segmento/cohorte agregado, definido como una agrupación de matrícula que comparte características institucionales, académicas y demográficas/territoriales reportadas en los datos abiertos del SNIESE/SENESCYT (p. ej., NOMBRE_IES, NOMBRE_CARRERA, NIVEL_FORMACION, MODALIDAD, SEXO y PROVINCIA_RESIDENCIA). Debido a que el conjunto de datos se encuentra en formato agregado (conteos en el campo TOTAL) y no contiene identificadores individuales ni trayectorias por estudiante, el término “cohorte” se utiliza en sentido operativo como pseudo-cohorte, es decir, un grupo definido por la combinación de variables observables.

Cada segmento se observa longitudinalmente en el período 2015–2023, y se identifica mediante un SEGMENTO_ID único. Adicionalmente, se definió ANIO_COHORTE como el primer año en que el segmento aparece en la serie (min(AÑO)), lo que permite analizar su comportamiento temporal mediante variaciones en TOTAL_SEG y construir la variable objetivo de riesgo de deserción a nivel agregado.

Ilustración 10. Ejemplo de estructura segmento–año y asignación de SEGMENTO_ID y ANIO_COHORTE

```
df_seg shape: (269472, 11)
Años: 2015 - 2023
Segmentos únicos: 76077
```

1 to 5 of 5 entries Filter

index	NOMBRE_IES	NOMBRE_CARRERA	NIVEL_FORMACION	MODALIDAD	SEXO	PROVINCIA_RESIDENCIA	AÑO	TOTAL_SEG	SEGMENTO_STR	SEGMENTO_ID	ANIO_CO HORTE
0	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2016	2	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE EMPRESAS|CUARTO NIVEL O POSGRADO|PRESENCIAL|HOMBRE|NO_REGISTRA	1	2016
1	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2017	5	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE EMPRESAS|CUARTO NIVEL O POSGRADO|PRESENCIAL|HOMBRE|NO_REGISTRA	1	2016
2	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2018	5	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE EMPRESAS|CUARTO NIVEL O POSGRADO|PRESENCIAL|HOMBRE|NO_REGISTRA	1	2016
3	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2019	4	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE EMPRESAS|CUARTO NIVEL O POSGRADO|PRESENCIAL|HOMBRE|NO_REGISTRA	1	2016
4	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2020	4	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE EMPRESAS|CUARTO NIVEL O POSGRADO|PRESENCIAL|HOMBRE|NO_REGISTRA	1	2016

Show 25 per page
Like what you see? Visit the data table notebook to learn more about interactive tables.

Nota. Fuente: Elaboración propia Romero Cristhian 2026

4.1.3. Ingeniería de variables desde TOTAL (variables temporales)

En esta investigación, el campo TOTAL (consolidado como TOTAL_SEG) representa el conteo anual de matrícula por segmento/cohorte. Dado que no se dispone de trayectorias individuales ni de una variable explícita de abandono por estudiante, la ingeniería de variables se orientó a transformar dicho conteo en indicadores temporales que describan el comportamiento longitudinal de cada segmento y permitan aproximar señales asociadas al riesgo de deserción a nivel agregado.
Una vez consolidada la base a nivel segmento–año (una fila por segmento/cohorte y año, 2015–2023), se generaron variables derivadas que capturan memoria histórica, cambio, intensidad, tendencia y estabilidad, respetando la secuencia temporal y utilizando únicamente información de años previos (rezagos y ventanas móviles) para evitar fuga de información en el entrenamiento/validación del modelo(D'Agostino, 2024).

Las variables derivadas del campo TOTAL incluyeron:

- **Rezagos (lags) de matrícula:** valores previos del segmento (TOTAL_L1, TOTAL_L2, TOTAL_L3) para incorporar memoria histórica del comportamiento del segmento.
- **Variaciones absolutas:** diferencias de matrícula entre períodos (DIF_1, DIF_2) para cuantificar incrementos o disminuciones netas.
- **Variaciones relativas (tasa interanual):** proporción de cambio respecto a la matrícula previa (TASA_1) para normalizar el cambio y comparar segmentos de distinto tamaño.
- **Indicadores de continuidad temporal:** Un flag de salto (FLAG_SALTO) y reglas para anular diferencias/tasas cuando no existe continuidad anual, evitando interpretaciones sesgadas por faltantes de años.

- **Tendencia y estabilidad histórica:** Estadísticas móviles calculadas solo con años previos (MEDIA_3, STD_3) y un z-score histórico (Z_3) para representar desviaciones respecto al comportamiento reciente del segmento.
- **Transformación de escala:** LOG_TOTAL = log(1 + TOTAL_SEG) para reducir asimetrías y el efecto de valores extremos.
- **(Opcional) Participación relativa dentro de la institución:** SHARE_IES, que expresa el peso del segmento respecto al total de matrícula de su IES en el mismo año.

En conjunto, estas variables permitieron representar el comportamiento temporal de cada segmento/cohorte como un conjunto de características predictivas (features), constituyendo la base para la etapa posterior de operacionalización del riesgo y definición de la variable objetivo (alto/bajo), así como para el entrenamiento y evaluación de modelos de aprendizaje automático (Random Forest, AdaBoost y Gradient Boosting) bajo un esquema de validación temporal.

Ilustración 11. Variables temporales generadas desde TOTAL_SEG (ejemplo de salida)

	NOMBRE_IES	NOMBRE_CARRERA	NIVEL_FORMACION	MODALIDAD	SEXO	PROVINCIA_RESIDENCIA	AÑO	TOTAL_SEG	SEGMENTO_STR	SEGMENTO_ID	...	TOTAL_L3	DIF_1	DIF_2	TASA_1	FLAG_SALTO	MEDIA_3	STD_3	Z_3	LOG_TOTAL	SHARE_IES
0	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2016	2	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE...	1	...	NaN	NaN	NaN	NaN	0	NaN	NaN	NaN	1.098612	0.000192
1	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2017	5	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE...	1	...	NaN	3.0	NaN	1.5	0	NaN	NaN	NaN	1.791759	0.000569
2	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2018	5	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE...	1	...	NaN	0.0	3.0	0.0	0	3.500000	2.121320	0.707107	1.791759	0.000633
3	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2019	4	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE...	1	...	2.0	-1.0	-1.0	-0.2	0	4.000000	1.732051	0.000000	1.609438	0.000559
4	ESCUELA POLITECNICA NACIONAL	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	HOMBRE	NO_REGISTRA	2020	4	ESCUELA POLITECNICA NACIONAL|ADMINISTRACION DE...	1	...	5.0	0.0	-1.0	0.0	0	4.666667	0.577350	-1.154701	1.609438	0.000567

5 rows × 26 columns

Nota. Fuente: Elaboración propia Romero Cristhian 2026

De acuerdo con la ilustración 11 los valores NaN aparecen porque se generaron variables temporales con rezagos (lags) y ventanas móviles. En el primer año de cada segmento/cohorte no existe un año previo, por lo que no se pueden calcular TOTAL_L1, DIF_1, TASA_1, MEDIA_3, STD_3 (y derivados), y quedan como NaN. En el segundo año ya se completan algunas (lags y diferencias), pero las móviles pueden seguir en NaN hasta contar con suficientes años anteriores. Por tanto, estos NaN son esperables y reflejan la falta de historial, no un error de procesamiento.

Ilustración 12. *Proporción de valores faltantes (NaN) en variables temporales*

Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la ilustración 12 nos muestra la salida resume de la proporción de valores faltantes (NaN) por variable: TOTAL_SEG y FLAG_SALTO no presentan faltantes (0%), mientras que TOTAL_L1, DIF_1 y TASA_1 registran ≈30.67% de NaN debido a que son variables con rezago que no se pueden calcular en el primer año de cada segmento o cuando existen saltos temporales; por su parte, MEDIA_3 y STD_3 presentan ≈48.62% de NaN porque dependen de una ventana móvil histórica (varios años previos), y no todos los segmentos cuentan con suficiente continuidad temporal para estimarlas.

4.1.4. Operacionalización del riesgo y variable objetivo (alto/bajo)

Como no se dispone de una variable directa de "abandono" por estudiante, el riesgo de deserción se operacionalizó a nivel de segmento/cohorte a partir de la variación interanual de la matrícula. En este enfoque, se consideró alto riesgo cuando un segmento presenta una disminución significativa de su matrícula respecto al año anterior, ya que este patrón es compatible con una reducción de la permanencia en ese grupo a nivel temporal.

Para construir la variable objetivo-binaria se empleó el cambio interanual relativo (TASA_1) calculado en la etapa de ingeniería de variables, definiendo un umbral de caída (percentil o valor fijo) para clasificar los segmentos:

- **Riesgo Alto (1):** Segmentos con caída interanual igual o mayor al umbral definido (p. ej., percentil 25 de TASA_1 o una caída ≥ 20%).
- **Riesgo Bajo (0):** Segmentos con variación estable o crecimiento, o con caídas menores al umbral.

Adicionalmente, para evitar sesgos, la etiqueta solo se asignó cuando existe continuidad temporal real (sin salto de años y con TASA_1 disponible). Los registros sin información previa (primer año del segmento) o con discontinuidades se mantuvieron sin etiqueta para el

entrenamiento (o se excluyeron), garantizando coherencia longitudinal y evitando fuga de información.

Ilustración 13. Umbral por percentil

```
...  (np.float64(-0.1111111111111111),
      RIESGO_ALTO
      0.0    139902
      NaN     82659
      1.0     46911
      Name: count, dtype: int64)
```

Nota. Fuente: Elaboración propia Romero Cristhian 2026

La ilustración 13 muestra el conteo de registros por clase de la variable objetivo RIESGO_ALTO, construida a partir de un umbral de caída interanual de la matrícula (TASA_1, umbral ≈ −0.111). Se observan 139,902 segmentos clasificados como riesgo bajo (0) y 46,911 como riesgo alto (1). Además, 82,659 registros aparecen como NaN porque no cuentan con TASA_1 válida (por ejemplo, primer año del segmento o discontinuidades temporales), por lo que no fueron etiquetados.

Ilustración 14. Alternativa umbral fijo

...

RIESGO_ALTO	count
0.0	146892
NaN	82659
1.0	39921

dtype: int64

Nota. Fuente: Elaboración propia Romero Cristhian 2026

En cambio, en la ilustración 14 se presenta la distribución de la variable objetivo RIESGO_ALTO: se registran 146,892 casos clasificados como riesgo bajo (0) y 39,921 como riesgo alto (1). Adicionalmente, 82,659 observaciones aparecen como NaN porque no fue posible calcular la variación interanual (TASA_1) por falta de año previo continuo (primer año del segmento o discontinuidades), por lo que quedaron sin etiqueta.

En este trabajo se utilizó como criterio principal un umbral fijo de caída interanual (definido sobre TASA_1) para asignar la etiqueta RIESGO_ALTO, debido a su interpretación directa y estabilidad operativa para la toma de decisiones. El umbral por percentil se reporta como análisis de sensibilidad, con el fin de contrastar la estabilidad de la etiqueta frente a un criterio relativo. En ambos casos, la etiqueta solo se asignó cuando existe continuidad temporal (sin saltos y con TASA_1 calculable); por ello, el primer año del segmento y las observaciones con discontinuidades se mantienen sin etiqueta (NaN) y no se incluyen en el entrenamiento del modelo.

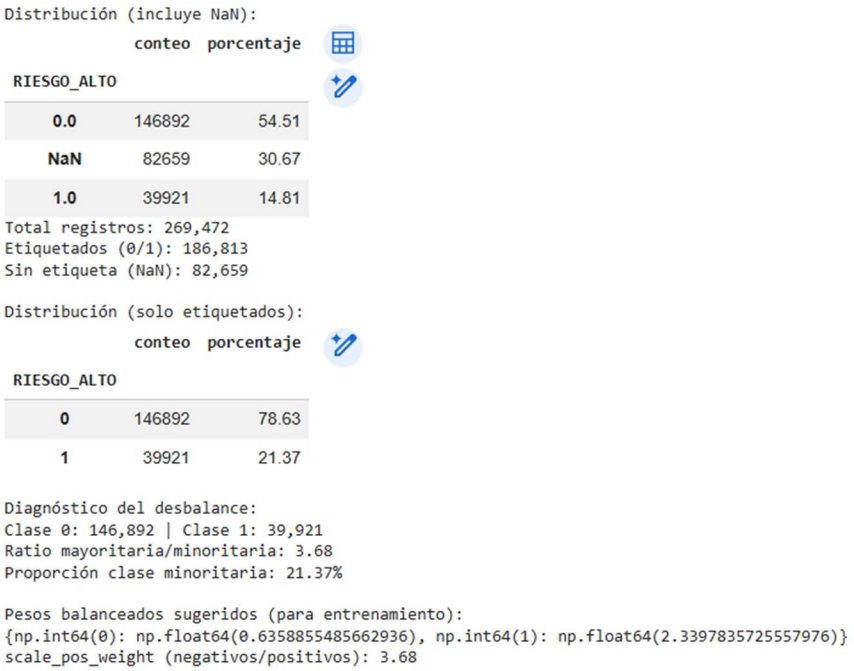
4.1.5. Distribución de clases y diagnóstico del desbalance

Una vez definida la variable objetivo RIESGO_ALTO (alto/bajo) a partir del cambio interanual de matrícula (TASA_1) y aplicado el criterio de continuidad temporal, se analizó la distribución de clases para identificar posibles problemas de desbalance. Este diagnóstico es relevante porque, en tareas de clasificación, una distribución desigual entre clases puede sesgar el aprendizaje del modelo hacia la clase mayoritaria y reducir la capacidad de detección de los segmentos con alto riesgo, que constituyen el principal interés institucional.

En este estudio, la distribución observada evidenció que la clase RIESGO_BAJO (0) concentra la mayor proporción de registros etiquetados, mientras que RIESGO_ALTO (1) representa una fracción menor. Adicionalmente, se identificó un conjunto de registros sin etiqueta (NaN) debido a la ausencia de información previa continua (por ejemplo, primer año del segmento o discontinuidades temporales), por lo que no se consideran para el entrenamiento del modelo (o se excluyen en la fase de modelado).

Este comportamiento confirma la presencia de desbalance de clases, condición que se tomó en cuenta en la etapa posterior de modelado mediante estrategias como: (i) selección de métricas sensibles a la clase minoritaria (precisión, recall y F1-score), (ii) validación temporal, y (iii) técnicas de balanceo en el conjunto de entrenamiento (por ejemplo, submuestreo de la clase mayoritaria o métodos de re-muestreo controlado), con el fin de mejorar la identificación de segmentos/cohortes en alto riesgo y evitar conclusiones sesgadas.

Ilustración 15. Distribución de NaN de riesgo alto



Distribución (incluye NaN):

RIESGO_ALTO	conteo	porcentaje
0.0	146892	54.51
NaN	82659	30.67
1.0	39921	14.81

Total registros: 269,472
Etiquetados (0/1): 186,813
Sin etiqueta (NaN): 82,659

Distribución (solo etiquetados):

RIESGO_ALTO	conteo	porcentaje
0	146892	78.63
1	39921	21.37

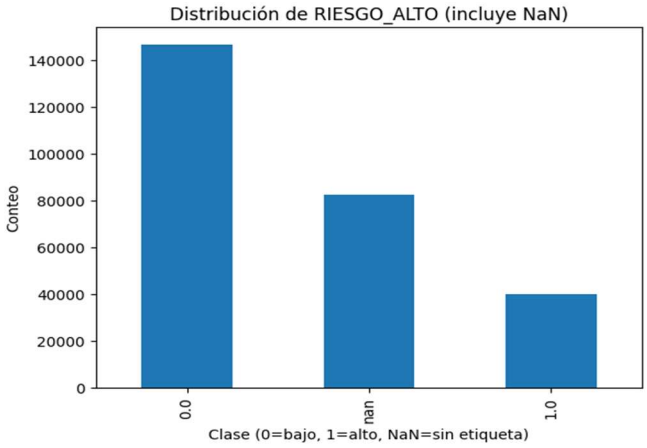
Diagnóstico del desbalance:
Clase 0: 146,892 | Clase 1: 39,921
Ratio mayoritaria/minoritaria: 3.68
Proporción clase minoritaria: 21.37%

Pesos balanceados sugeridos (para entrenamiento):
{np.int64(0): np.float64(0.6358855485662936), np.int64(1): np.float64(2.3397835725557976)}
scale_pos_weight (negativos/positivos): 3.68

Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la ilustración 15, se visualiza la salida que muestra la distribución de la variable objetivo RIESGO_ALTO (umbral TASA_1 ≈ −0.111): del total, 54.51% se clasifica como riesgo bajo (0), 14.81% como riesgo alto (1) y 30.67% queda sin etiqueta (NaN) por falta de año previo/continuidad. Considerando solo los etiquetados, la clase 1 representa 21.37% frente a 78.63% de la clase 0, evidenciando un desbalance moderado (ratio ≈ 3.68:1 a favor de la clase 0). En consecuencia, en la fase de modelado se priorizó el uso de métricas sensibles a la clase minoritaria (recall y F1-score), junto con estrategias como ponderación de clases (p. ej., class_weight='balanced'), ajuste de umbral y remuestreo en entrenamiento (SMOTE), evitando fuga de información mediante validación temporal.

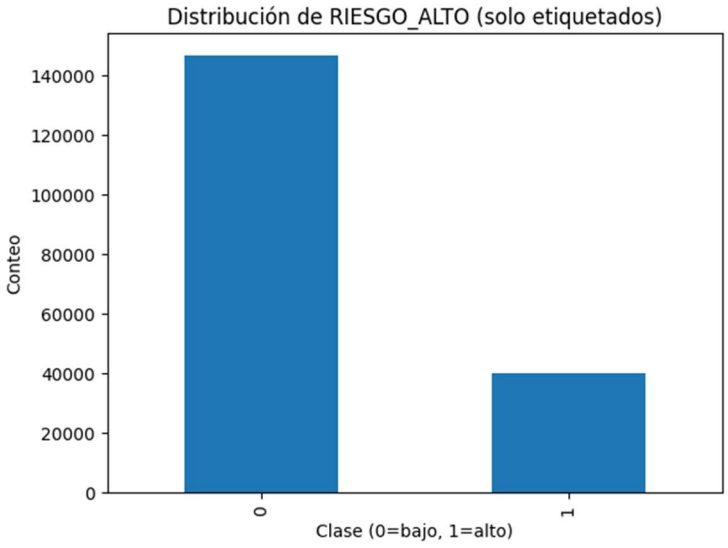
Ilustración 16. Distribución de riesgo alto(NaN)



Nota. Fuente: Elaboración propia Romero Cristhian 2026

La ilustración 16 muestra, mediante un gráfico de barras, la distribución de RIESGO_ALTO incluyendo los registros sin etiqueta (NaN): predomina la clase 0 (riesgo bajo), seguida por los NaN (segmentos sin variación interanual calculable por falta de continuidad temporal) y, en menor proporción, la clase 1 (riesgo alto), evidenciando un desbalance hacia la clase 0.

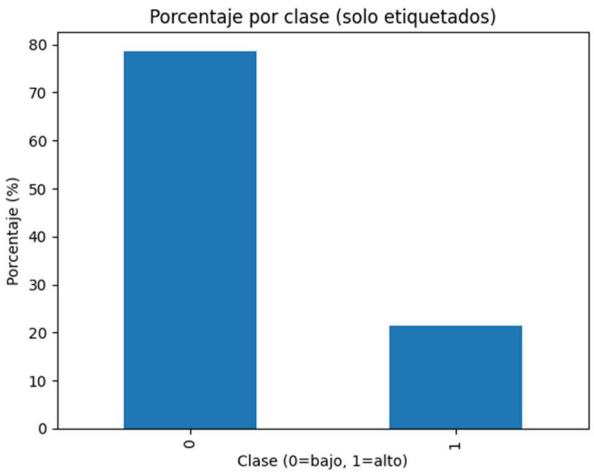
Ilustración 17. Distribución de riesgo alto(etiquetados)



Nota. Fuente: Elaboración propia Romero Cristhian 2026

La ilustración 17 muestra, mediante un gráfico de barras, la distribución de RIESGO_ALTO considerando solo los registros etiquetados: la clase 0 (riesgo bajo) es claramente mayoritaria frente a la clase 1 (riesgo alto), lo que confirma un desbalance de clases (más casos de bajo riesgo que de alto riesgo).

Ilustración 18. Porcentaje por clase (etiquetados)



Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la Ilustración 18 se observa el porcentaje por clase (solo etiquetados): la clase 0 (riesgo bajo) concentra aproximadamente 79% de los registros, mientras que la clase 1 (riesgo alto) representa cerca de 21%, evidenciando un desbalance de clases a favor del riesgo bajo.

Además de destacar análisis claves para la Distribución de clases y diagnóstico del desbalance entre las cuales tenemos:

Desbalance por año (AÑO)

Ilustración 19. Gráfico desbalance por año

Conteos por año:

AÑO	RIESGO_ALTO 0	1
2016	13118	3219
2017	13551	3047
2018	14894	4066
2019	17804	4750
2020	19795	5674
2021	22823	5216
2022	22439	6503
2023	22468	7446

Porcentajes por año:

AÑO	RIESGO_ALTO 0	1
2016	80.30	19.70
2017	81.64	18.36
2018	78.55	21.45
2019	78.94	21.06
2020	77.72	22.28
2021	81.40	18.60
2022	77.53	22.47
2023	75.11	24.89

Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la Ilustración 19 se presenta la distribución anual de la variable RIESGO_ALTO (solo registros etiquetados) para el período 2016–2023, mostrando los conteos y porcentajes de las clases 0 (riesgo bajo) y 1 (riesgo alto). En todos los años predomina la clase 0 (aprox. 75.11%–81.64%), mientras que la clase 1 se mantiene como minoritaria (aprox. 18.36%–24.89%), con una tendencia de incremento del riesgo alto hacia 2023 (24.89%), lo que sugiere mayor frecuencia de caídas interanuales de matrícula en los años más recientes.

Relación del desbalance con discontinuidad (FLAG_SALTO)

Ilustración 20. Proporción por Flag(Salto)

Proporción por FLAG_SALTO (fila=FLAG_SALTO):

FLAG_SALTO	RIESGO_ALTO 0	1
0	0.786	0.214

Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la Ilustración 20 se presenta la proporción de RIESGO_ALTO según FLAG_SALTO (fila = FLAG_SALTO). Para los registros con FLAG_SALTO = 0 (es decir, con continuidad

anual y sin saltos temporales), 78.6% se clasifica como riesgo bajo (0) y 21.4% como riesgo alto (1), confirmando que, aun con continuidad temporal, la clase 0 sigue siendo mayoritaria y el desbalance se mantiene.

Ilustración 21. Porcentaje sin etiqueta por año

```
% sin etiqueta (NaN): 30.67%
Porcentaje sin etiqueta por año:
```

AÑO	NAN_LABEL
2015	100.00
2016	20.17
2017	25.90
2018	27.23
2019	23.94
2020	24.11
2021	23.51
2022	24.85
2023	29.24

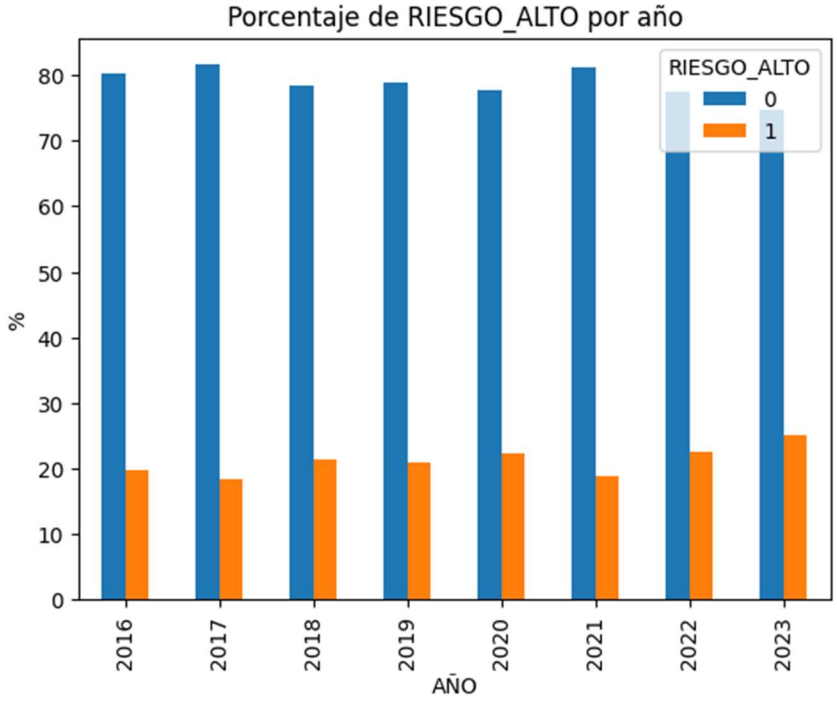
dtype: float64

Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la Ilustración 21 se muestra la proporción de registros sin etiqueta (NaN) en la variable objetivo. En total, 30.67% de las observaciones no pudieron ser etiquetadas; por año, 2015 presenta 100% sin etiqueta (al no existir un año previo para calcular la variación interanual), mientras que entre 2016–2023 el porcentaje de NaN se mantiene aproximadamente entre 20% y 29%, asociado principalmente a segmentos sin historial continuo o con discontinuidades temporales que impiden calcular TASA_1 y asignar la clase (0/1).

Distribución porcentual de la variable objetivo RIESGO_ALTO por año.

Ilustración 22. Desbalance de clases de RIESGO_ALTO por año.



Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la Ilustración 22 se muestra el porcentaje anual de RIESGO_ALTO (solo registros etiquetados) para el período 2016–2023. En todos los años predomina la clase 0 (riesgo bajo) con valores aproximados entre 75% y 82%, mientras que la clase 1 (riesgo alto) se mantiene como minoritaria (≈ 18%–25%), observándose un aumento gradual del riesgo alto hacia 2023.

4.1.6. Resultados del preprocesamiento

19

En esta sección se presentan los resultados del preprocesamiento aplicado a la base de matrícula, con el objetivo de asegurar consistencia estructural, semántica y temporal antes del análisis longitudinal y el modelado. En primer lugar, se depuró la estructura del archivo mediante la eliminación de columnas residuales tipo *"Unnamed"* y filas completamente vacías. Posteriormente, se estandarizaron los campos categóricos, normalizando el texto (mayúsculas y eliminación de espacios) y homologando representaciones de valores faltantes.

A continuación, se validaron los tipos y rangos de las variables clave, convirtiendo AÑO y TOTAL a formato numérico y descartando registros sin información válida o con valores inconsistentes (por ejemplo, totales negativos). Como resultado, se obtuvo un dataset limpio (df) con calidad suficiente para consolidación y análisis.

Posteriormente, la información se consolidó a nivel segmento–año, generando df_seg como unidad de análisis agregada y calculando TOTAL_SEG como el conteo anual consolidado por segmento. Finalmente, se verificó la continuidad temporal por segmento (número de años disponibles) y se aplicó un criterio mínimo de historia para garantizar estabilidad longitudinal en las series utilizadas en etapas posteriores.

4.1.6.1. Tamaño del dataset por etapa (trazabilidad)

El tamaño del dataset por etapa se usa como evidencia de trazabilidad del preprocesamiento: permite verificar cuántas observaciones y variables se conservan tras cada transformación. En este estudio, el cambio principal en el número de filas no responde a pérdida de información, sino a la redefinición de la unidad de análisis al consolidar los datos a nivel segmento–año; posteriormente, el número de registros se mantiene y solo aumenta el número de variables por la incorporación de características derivadas para el análisis y el modelado.

Posteriormente, se reporta el tamaño del dataset resultante, con el fin de documentar el volumen de datos disponible para el análisis.

Ilustración 23. Tamaño del dataset resultante

2.8.1.6.1 Tamaño por etapas

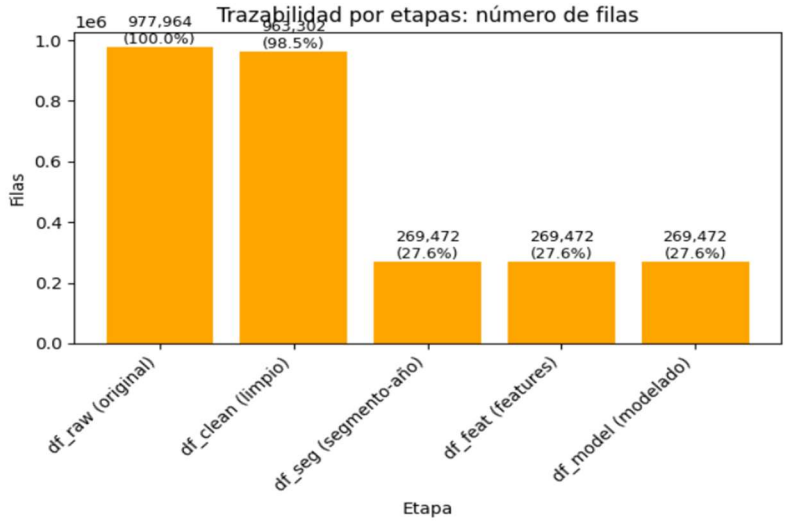
	etapa	filas	columnas	% vs original
0	df_raw (original)	977964	19	100.00
1	df_clean (limpio)	963302	19	98.50
2	df_seg (segmento-año)	269472	11	27.55
3	df_feat (features)	269472	26	27.55
4	df_model (modelado)	269472	27	27.55

Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la ilustración 23 se muestra el tamaño del dataset por etapas. df_raw (original) conserva 977,964 filas y 19 columnas (100%). df_clean (limpio) mantiene 963,302 filas y 19 columnas (98.50%). La reducción principal ocurre al consolidar a segmento–año (df_seg), quedando 269,472 filas (27.55%) y 11 columnas. Luego, df_feat (features) y df_model (modelado) mantienen esas 269,472 filas y aumentan columnas (26–27) por la incorporación de variables derivadas y la variable objetivo.

En las siguientes ilustraciones se podrá a detallar la trazabilidad por etapas tanto para columnas y filas.

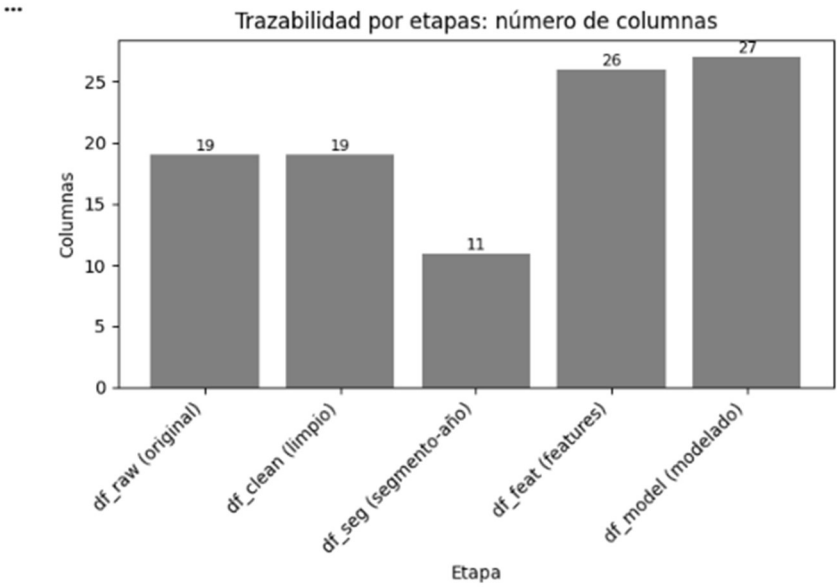
Ilustración 24. Trazabilidad por etapas: número de filas



Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la ilustración 24 se presenta la trazabilidad por etapas del número de filas del dataset: se observa que el conjunto original registra 977,964 filas en df_raw y pasa a 963,302 tras la limpieza en df_clean (98.5%). Posteriormente, se reduce a 269,472 (27.6%) en la etapa de segmentación/agregación por segmento–año (df_seg), manteniéndose constante en la generación de variables (df_feat) y en el dataset final para modelado (df_model).

Ilustración 25. Trazabilidad por etapas: número de columnas



Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la ilustración 25 en cambio podemos visualizar la. Trazabilidad por etapas del número de columnas del dataset: se mantiene en 19 variables en el conjunto original y limpio (df_raw y df_clean), disminuye a 11 columnas en la etapa de segmentación/agregación por segmento-año (df_seg) debido a la consolidación de campos, y posteriormente aumenta a 26 y 27 columnas en las fases de ingeniería de características (df_feat) y dataset final de modelado (df_model), reflejando la incorporación de variables derivadas para el análisis predictivo.

4.1.6.2. Calidad de datos (variables clave)

La calidad del dataset se evaluó a partir de variables clave para el análisis por segmentos/cohortes: AÑO, TOTAL y las variables que conforman la unidad segmento–año (p. ej., institución, modalidad, carrera, territorio). Se verificó la completitud (nulos en AÑO y TOTAL), la validez (AÑO en 2015–2023 y TOTAL numérico con TOTAL ≥ 0), la integridad (duplicados y unicidad de la combinación segmento–año) y la consistencia de las categorías. Estos controles aseguraron un dataset confiable para la ingeniería de características y el modelado longitudinal.

Ilustración 26. Verificación de calidad del dataset: completitud e integridad (df_clean)

```
2.8.1.6.2 Calidad de datos (df_clean)
Rango de AÑO: 2015 - 2023
Duplicados exactos: 0
Faltantes totales: 0

Top 10 columnas con más faltantes:
```

	0
AÑO	0
NOMBRE_IES	0
TIPO_FINANCIAMIENTO	0
NOMBRE_CARRERA	0
NIVEL_FORMACION	0
MODALIDAD	0
CAMPO_AMPLIO	0
CAMPO_ESPECIFICO	0
CAMPO_DETALLADO	0
TIPO_SEDE	0

dtype: int64

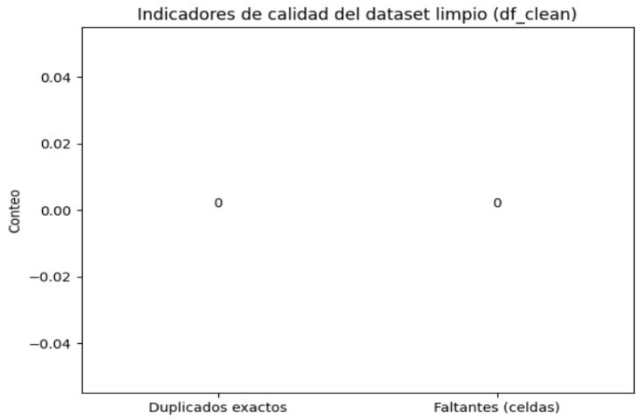
Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la ilustración 26 se observa el resumen de calidad de datos del dataset limpio (df_clean): el período analizado se mantiene entre 2015 y 2023, no se identifican duplicados

exactos (0) ni valores faltantes en las variables evaluadas (0 faltantes totales y 0 por columna), lo que evidencia alta completitud e integridad en el conjunto de datos.

Se puede detallar con exactitud el resumen correspondiente de los resultados obtenidos en cada etapa del preprocesamiento.

Ilustración 27. Indicadores de calidad del dataset limpio (df_clean)



Nota. Fuente: Elaboración propia Romero Cristhian 2026

La ilustración 27 resume dos métricas de calidad posteriores a la limpieza: el número de duplicados exactos y la cantidad de valores faltantes (celdas vacías). En el gráfico se observa que ambos indicadores se mantienen en 0: no hay duplicados exactos y los faltantes son nulos (0 celdas), lo que evidencia alta completitud e integridad del dataset tras el preprocesamiento.

4.1.6.3. Resultados de segmentación

La segmentación generó el dataset df_seg bajo la unidad de análisis segmento–año, asegurando unicidad por combinación (0 duplicados en SEGMENTO_ID–AÑO). Se identificaron 76.077 segmentos únicos y el período se mantiene entre 2015–2023. Además, el número de segmentos activos por año muestra una tendencia creciente, lo que evidencia un aumento en la cobertura y diversidad de combinaciones registradas.

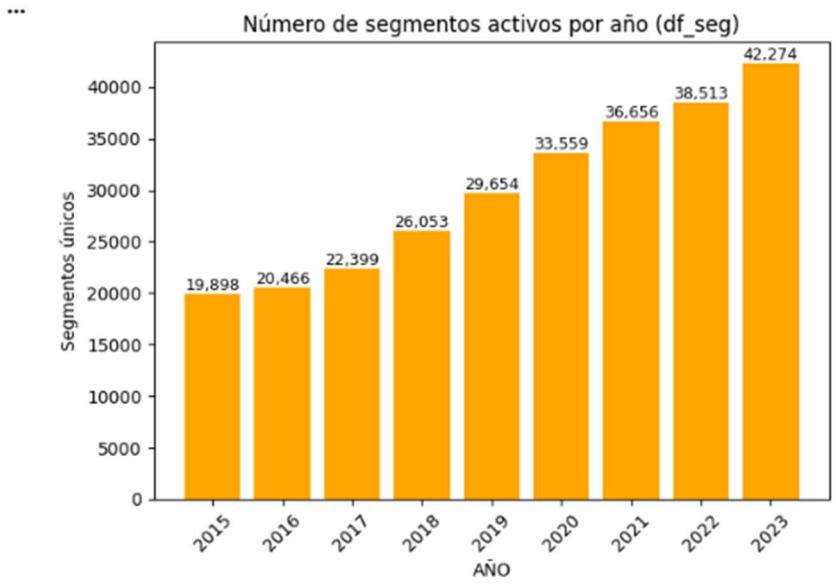
Ilustración 28. Resultados de segmentación

```
...
    2.8.1.6.3 Segmentación (df_seg)
    Segmentos únicos: 76077
    Duplicados (SEGMENTO_ID, AÑO): 0
    Rango de AÑO en df_seg: 2015 - 2023
```

Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la Ilustración 29 se presenta el número de segmentos activos por año en la base consolidada df_seg, entendido como la cantidad de SEGMENTO_ID únicos con registro en cada período. Se observa una tendencia creciente a lo largo del intervalo 2015–2023, pasando de 19.898 segmentos en 2015 a 42.274 en 2023. Este comportamiento evidencia una mayor cobertura y diversificación de los segmentos reportados en los años más recientes, lo que fortalece la disponibilidad de información para el análisis longitudinal.

Ilustración 29. Número de segmentos activos por año



Nota. Fuente: Elaboración propia Romero Cristhian 2026

4.1.6.4. Continuidad temporal

En esta etapa se evaluó la continuidad histórica de los segmentos en el tiempo, midiendo para cada SEGMENTO_ID la cantidad de años con registros disponibles. Este análisis permite verificar la consistencia del panel segmento–año e identificar segmentos con series temporales suficientes para el cálculo de rezagos, tasas y demás variables longitudinales, evitando el uso de trayectorias incompletas en el modelado.

Ilustración 30. Continuidad temporal

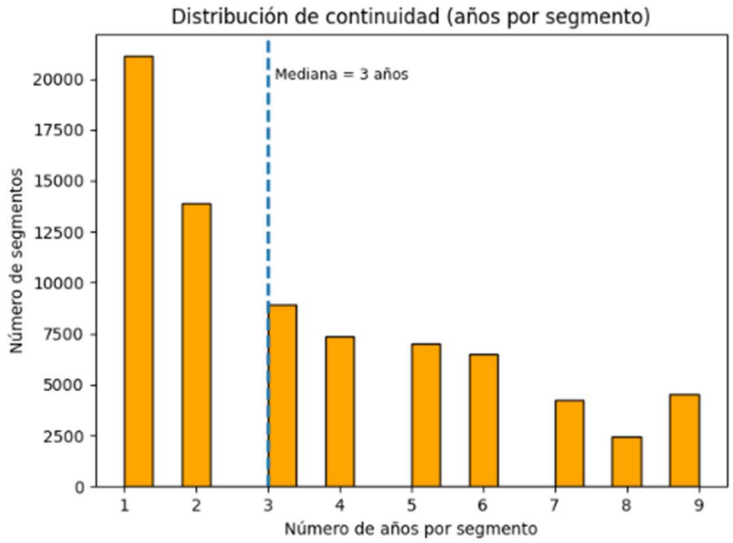
	valor
min_años_por_segmento	1.000000
mediana_años_por_segmento	3.000000
max_años_por_segmento	9.000000
promedio_años_por_segmento	3.542096

Nota. Fuente: Elaboración propia Romero Cristhian 2026

La ilustración 30 se presenta un resumen estadístico de la continuidad temporal por segmento, es decir, cuántos años de información están disponibles para cada SEGMENTO_ID. Se observa que el mínimo de años registrados por segmento es 1, la mediana es 3, el máximo alcanza 9 años y el promedio es 3,54 años por segmento. Estos resultados evidencian que la mayoría de los segmentos cuenta con una trayectoria temporal moderada, aunque existen segmentos con series cortas y otros con historial más extenso.

La siguiente ilustración 31, presenta la distribución de continuidad de los segmentos, medida como el número de años en que cada SEGMENTO_ID registra información dentro del período 2015–2023. Se observa una mayor concentración de segmentos con baja continuidad (1 y 2 años), y una mediana de 3 años, lo que indica que al menos la mitad de los segmentos aparece en tres años o menos. En contraste, una proporción menor mantiene series más largas (hasta 9 años), evidenciando que coexisten segmentos intermitentes y segmentos con trayectorias temporales más estables, aspecto relevante para el análisis longitudinal y la construcción de variables temporales para el modelado.

Ilustración 31. Distribución por continuidad



Nota. Fuente: Elaboración propia Romero Cristhian 2026

4.1.6.5. Salidas listas para modelado

En esta fase se consolidó el dataset para modelado, incorporando la variable objetivo RIESGO_ALTO y verificando la distribución de clases. Los resultados muestran que el 54,44% de los registros corresponde a clase 0 (bajo riesgo) (146.713 casos), el 14,88% a clase 1 (alto riesgo) (40.100 casos) y el 30,67% permanece sin etiqueta (NaN) (82.659 casos), principalmente asociado a observaciones sin información suficiente para calcular la tasa o continuidad temporal requerida.

Ilustración 32. Salidas para el modelado

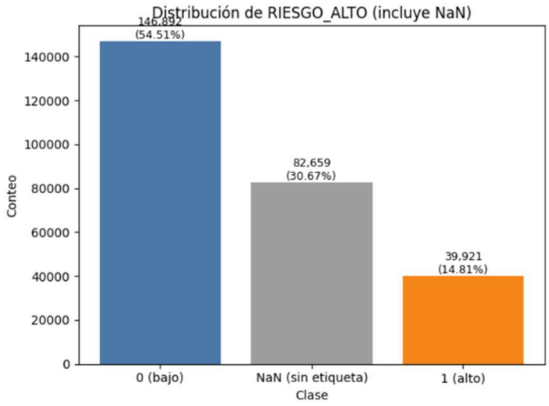
RIESGO_ALTO	conteo	porcentaje
0.0	146892	54.51
NaN	82659	30.67
1.0	39921	14.81

Nota. Fuente: Elaboración propia Romero Cristhian 2026

En la ilustración 32 podemos visualizar la distribución de la variable objetivo RIESGO_ALTO, incluyendo los casos sin etiqueta (NaN). Se observa que la mayoría de los registros corresponde a la clase 0 (bajo riesgo) con 146,892 casos (54.51%). En segundo lugar, aparecen los registros NaN (sin etiqueta) con 82,659 casos (30.67%), y finalmente la clase 1

(alto riesgo) con 39,921 casos (14.81%). En conjunto, la figura evidencia un desbalance de clases a favor del bajo riesgo y una proporción relevante de observaciones sin etiqueta.

Ilustración 33. Distribución de riesgo alto



Nota. Fuente: Elaboración propia Romero Cristhian 2026

La ilustración 33 se presenta un gráfico de barras con la distribución de RIESGO_ALTO incluyendo los registros sin etiqueta (NaN). Predomina la clase 0 (bajo riesgo) con 146,892 casos (54.51%), seguida por los NaN (sin etiqueta) con 82,659 casos (30.67%), y en menor proporción la clase 1 (alto riesgo) con 39,921 casos (14.81%). En conjunto, el gráfico evidencia un desbalance de clases a favor del bajo riesgo y una proporción relevante de observaciones sin etiqueta.

4.1.7. Configuración experimental y estrategia de validación

Para la evaluación de los modelos se definió una configuración experimental con enfoque temporal, evitando fuga de información. Se utilizó únicamente el conjunto etiquetado en la variable objetivo RIESGO_ALTO (0=bajo, 1=alto), obteniéndose 186,813 registros (146,892 en clase 0 y 39,921 en clase 1).

Como variables predictoras se emplearon 14 features: seis categóricas que definen el segmento (NOMBRE_IES, NOMBRE_CARRERA, NIVEL_FORMACION, MODALIDAD, SEXO y PROVINCIA_RESIDENCIA) y ocho numéricas derivadas del historial del segmento (TOTAL_L1, TOTAL_L2, TOTAL_L3, MEDIA_3, STD_3, LOG_TOTAL, SHARE_IES y Z_3), calculadas solo con información previa. El preprocesamiento incluyó imputación (mediana para numéricas y valor más frecuente para categóricas) y codificación ordinal para categóricas.

La validación se realizó mediante split temporal: entrenamiento 2016–2020 (99,918 registros), validación 2021–2022 (56,981) y prueba 2023 (29,914). Adicionalmente, se aplicó validación cruzada temporal tipo expanding window dentro del entrenamiento, con 2 folds. El desempeño se midió con accuracy, precision, recall, F1-score y AUC.

Ilustración 34. Configuración experimental con enfoque temporal

```
RIESGO_ALTO
0    146892
1     39921
Name: count, dtype: int64
Categóricas: ['NOMBRE_IES', 'NOMBRE_CARRERA', 'NIVEL_FORMACION', 'MODALIDAD', 'SEXO', 'PROVINCIA_RESIDENCIA']
Numéricas (solo pasado): ['TOTAL_L1', 'TOTAL_L2', 'TOTAL_L3', 'MEDIA_3', 'STD_3', 'LOG_TOTAL', 'SHARE_IES', 'Z_3']
Total features: 14

Split temporal:
Train: (99918, 14) Valid: (56981, 14) Test: (29914, 14)
Nº folds (CV temporal expanding): 2
```

Nota. Fuente: Elaboración propia Romero Cristhian 2026

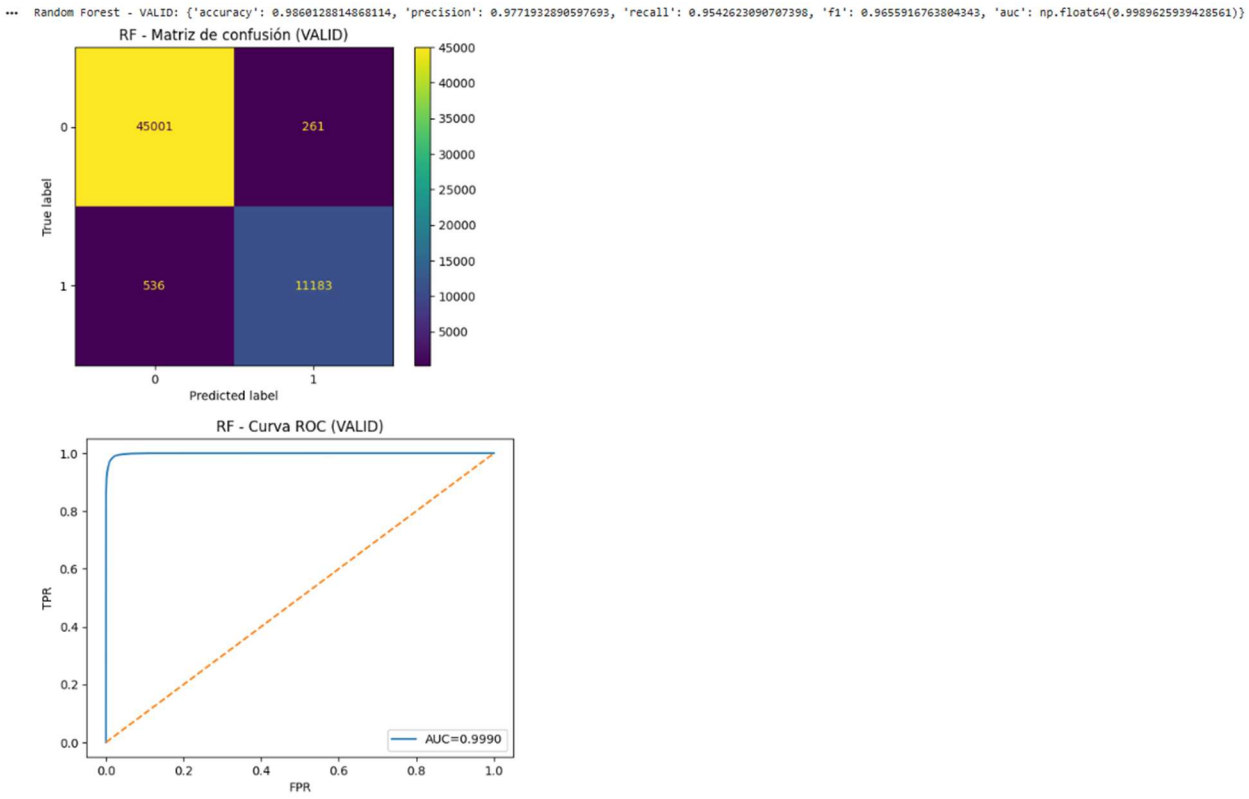
4.1.8. Desempeño de modelos sin balanceo (línea base)

Se evaluaron tres modelos de clasificación sin balanceo (Random Forest, AdaBoost y Gradient Boosting) usando el mismo pipeline de preprocesamiento y el mismo split temporal definido en 2.8.1.7. El desempeño se reportó con accuracy, precision, recall, F1 y AUC, además de matriz de confusión y curva ROC en el conjunto de prueba (2023). Debido al desbalance de clases, la comparación se priorizó principalmente con recall, F1 y AUC, por ser más informativas para la detección de la clase minoritaria (riesgo alto).

4.1.8.1. Random Forest: métricas + matriz de confusión + ROC/AUC

El modelo Random Forest, entrenado sin técnicas de balanceo y bajo el esquema de validación temporal definido, mostró un desempeño alto tanto en validación como en prueba. En validación, alcanzó valores elevados de *accuracy, precision, recall* y *F1-score*, lo que evidencia una adecuada capacidad para discriminar entre segmentos de bajo riesgo (0) y alto riesgo (1). En el conjunto de prueba (2023), el modelo mantuvo resultados consistentes, confirmando estabilidad temporal y capacidad de generalización. La curva ROC muestra una separación clara respecto a la línea base aleatoria, lo que refuerza la robustez del modelo para la identificación temprana del riesgo de deserción a nivel agregado.

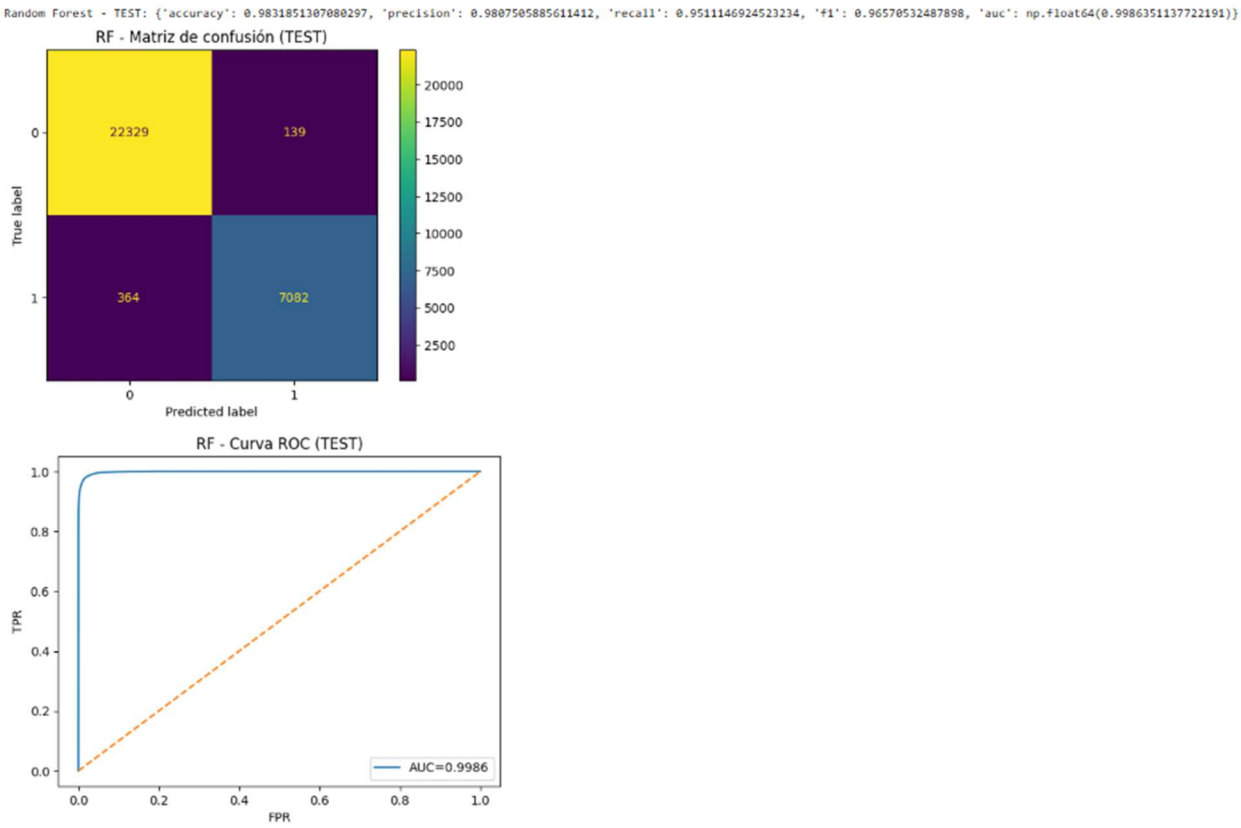
Ilustración 35. Matriz de confusión(VALID)



Nota. Fuente: Elaboración propia Romero Cristhian 2026

Interpretación de la matriz de confusión (Validación – Random Forest). En validación, el modelo clasifica correctamente 45.001 segmentos de bajo riesgo (TN) y detecta 11.183 segmentos de alto riesgo (TP). Los errores se distribuyen en 261 falsos positivos (FP), es decir, segmentos alertados como riesgo alto cuando en realidad son bajo riesgo, y 536 falsos negativos (FN), que corresponden a segmentos con riesgo alto que no fueron detectados. Dado el propósito de alerta temprana, los FN representan el error más crítico porque implican no priorizar segmentos potencialmente vulnerables. El AUC cercano a 0,999 confirma una alta capacidad discriminativa del modelo.

Ilustración 36. Matriz de confusión(TEST)



Nota. Fuente: Elaboración propia Romero Cristhian 2026

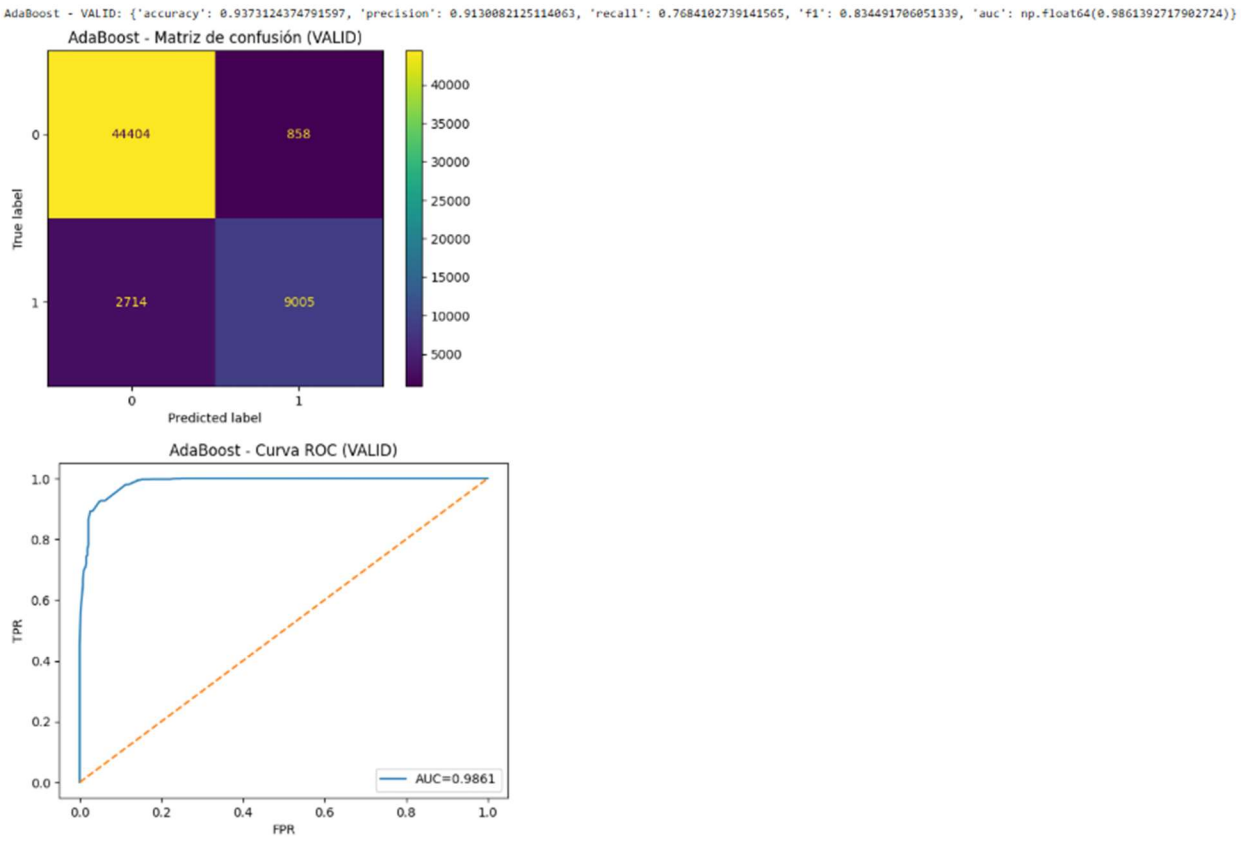
Interpretación de la matriz de confusión (Prueba 2023 – Random Forest). En el conjunto de prueba (2023), el modelo clasifica correctamente 22.329 segmentos de bajo riesgo (TN) y detecta 7.082 segmentos de alto riesgo (TP). Los errores se distribuyen en 139 falsos positivos (FP), es decir, *segmentos sin caída interanual relevante que el modelo alerta como riesgo alto*, y 364 falsos negativos (FN), correspondientes a *segmentos con riesgo alto que no fueron detectados*. En el contexto de alerta temprana, los FN representan el error más crítico, ya que implican omitir segmentos potencialmente prioritarios para intervención preventiva. La curva ROC reporta un AUC ≈ 0,9986, lo que confirma una capacidad discriminativa excelente y una buena generalización temporal del modelo.

4.1.8.2. AdaBoost: métricas + matriz de confusión + ROC/AUC

El modelo AdaBoost se evaluó como línea base sin balanceo usando el mismo preprocesamiento y split temporal. Se reportaron accuracy, precision, recall, F1 y AUC en validación y en prueba (2023). Además, la matriz de confusión permitió identificar errores

(FP/FN) y la curva ROC resumió la capacidad discriminativa, priorizando recall, F1 y AUC por el desbalance de clases.

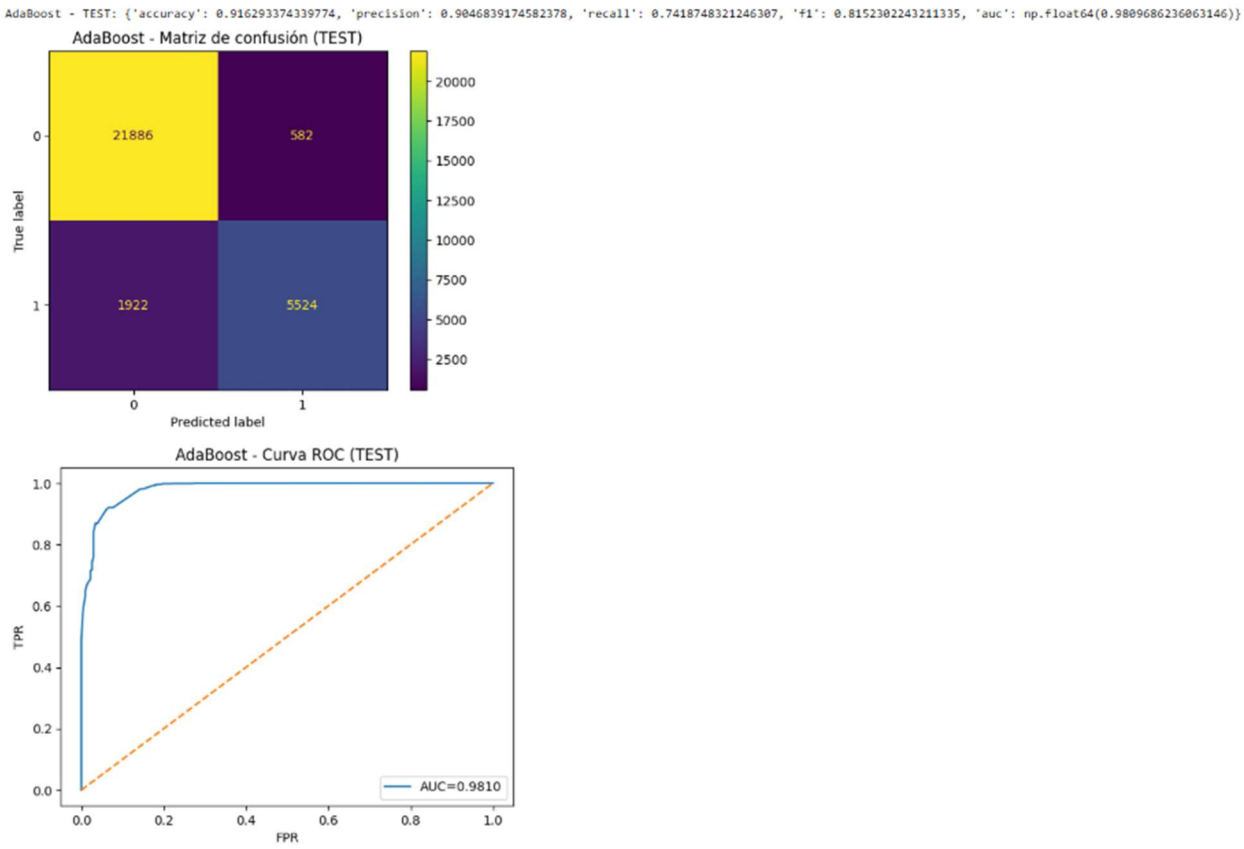
Ilustración 37. Matriz de confusión(VALID)



Nota. Fuente: Elaboración propia Romero Cristhian 2026

Interpretación de la matriz de confusión (Validación – AdaBoost). En validación, el modelo clasifica correctamente 44.404 segmentos de bajo riesgo (TN) y detecta 9.005 segmentos de alto riesgo (TP). Los errores se distribuyen en 858 falsos positivos (FP), es decir, *segmentos sin caída interanual relevante que el modelo alerta como riesgo alto*, y 2.714 falsos negativos (FN), correspondientes a *segmentos con riesgo alto que no fueron detectados*. En un enfoque de alerta temprana, los FN constituyen el error más crítico, ya que implican omitir segmentos potencialmente prioritarios para intervención. La curva ROC reporta un AUC ≈ 0,986, lo que evidencia buena capacidad discriminativa, aunque inferior a los modelos con mejor sensibilidad.

Ilustración 38. Matriz de confusión(TEST)



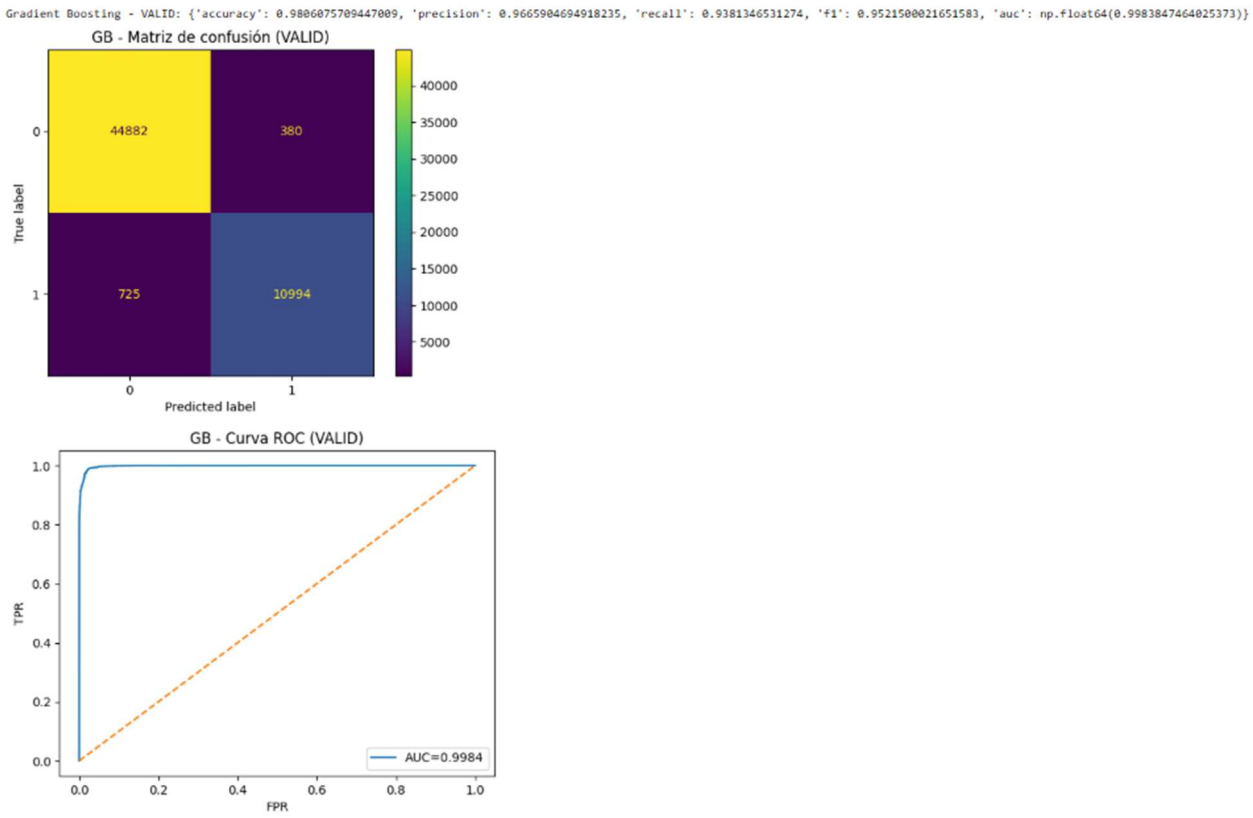
Nota. Fuente: Elaboración propia Romero Cristhian 2026

Interpretación de la matriz de confusión (Prueba 2023 – AdaBoost). En el conjunto de prueba (2023), el modelo clasifica correctamente 21.886 segmentos de bajo riesgo (TN) y detecta 5.524 segmentos de alto riesgo (TP). Los errores se distribuyen en 582 falsos positivos (FP) y 1.922 falsos negativos (FN). Este patrón confirma que la principal limitación del modelo se concentra en los FN, lo que reduce su capacidad para identificar segmentos en riesgo alto cuando el objetivo institucional es preventivo. La curva ROC alcanza un AUC ≈ 0,981, indicando una capacidad discriminativa alta, aunque menor que la obtenida por Random Forest y Gradient Boosting.

4.1.8.3. Gradient Boosting: métricas + matriz de confusión + ROC/AUC

El modelo Gradient Boosting se evaluó en validación y prueba mediante accuracy, precision, recall, F1 y AUC. Adicionalmente, se reportó la matriz de confusión y la curva ROC, evidenciando su capacidad para discriminar entre segmentos de bajo riesgo (0) y alto riesgo (1) sin aplicar técnicas de balanceo.

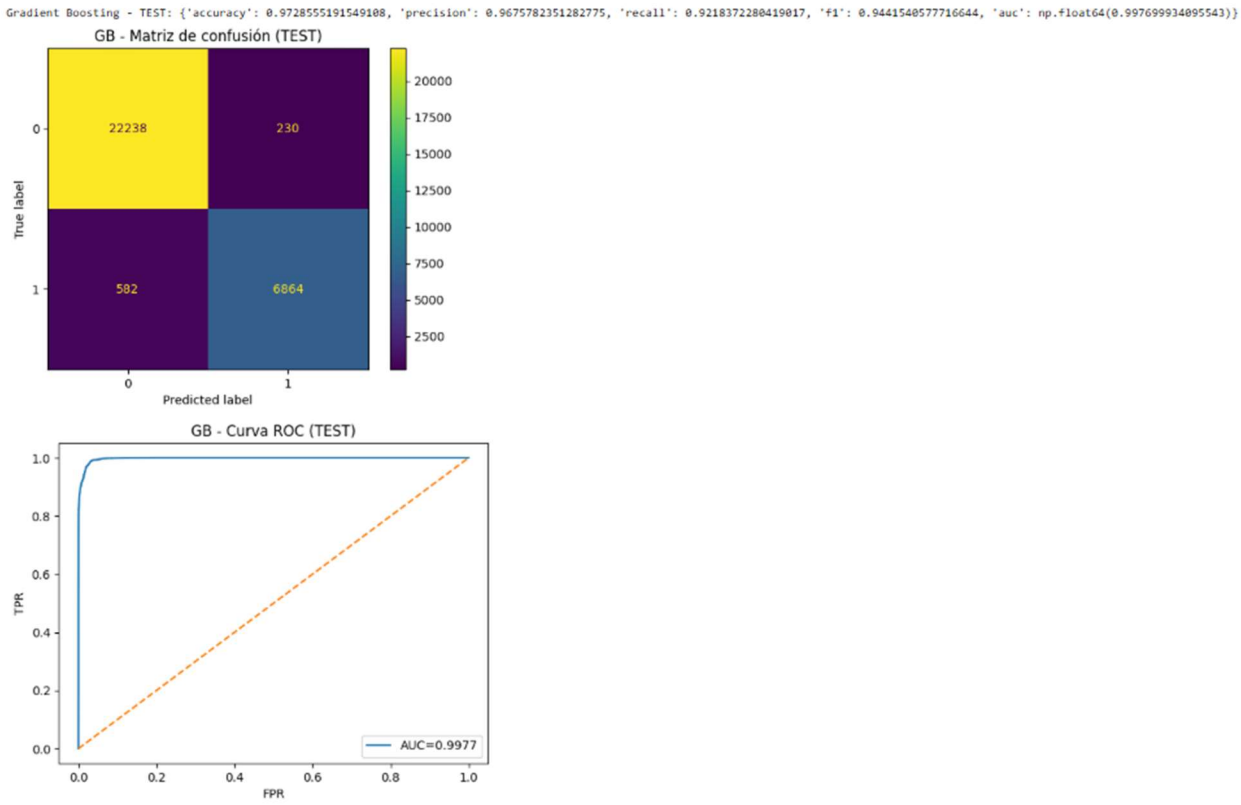
Ilustración 39. Matriz de confusión(VALID)



Nota. Fuente: Elaboración propia Romero Cristhian 2026

Interpretación de la matriz de confusión (Validación – Gradient Boosting). En validación, el modelo clasifica correctamente 44.882 segmentos de bajo riesgo (TN) y detecta 10.994 segmentos de alto riesgo (TP). Los errores se distribuyen en 380 falsos positivos (FP) y 725 falsos negativos (FN). En términos de alerta temprana, el volumen relativamente bajo de FN sugiere una detección consistente de segmentos en riesgo alto, manteniendo a la vez control sobre falsas alertas. La curva ROC reporta un AUC ≈ 0,998, lo que confirma una capacidad discriminativa excelente.

Ilustración 40. Matriz de confusión(TEST)



Nota. Fuente: Elaboración propia Romero Cristhian 2026

Interpretación de la matriz de confusión (Prueba 2023 – Gradient Boosting). En el conjunto de prueba (2023), el modelo clasifica correctamente 22.238 segmentos de bajo riesgo (TN) y detecta 6.864 segmentos de alto riesgo (TP). Los errores se distribuyen en 230 falsos positivos (FP) y 582 falsos negativos (FN). Dado el objetivo preventivo, los FN siguen siendo el error más relevante; sin embargo, su magnitud se mantiene moderada, evidenciando buena generalización temporal. La curva ROC alcanza un AUC ≈ 0,9977, lo que respalda un desempeño discriminativo alto y estable en comparación con AdaBoost.

4.1.8.4. Identificación operativa de segmentos/cohortes en riesgo y no riesgo (desde la matriz de confusión)

Para traducir los resultados del modelo a una lectura institucional, cada predicción del conjunto de validación/prueba se asocia al identificador del segmento (SEGMENTO_ID) y a su probabilidad estimada de riesgo alto. Con ello, la matriz de confusión deja de ser solo un resumen numérico y se convierte en cuatro listados operativos:

- **TP (verdaderos positivos):** Segmentos correctamente identificados como riesgo alto (prioridad principal de intervención).
- **FP (falsos positivos):** Segmentos alertados como riesgo alto que en realidad son riesgo bajo (alertas adicionales; pueden revisarse con criterios operativos o ajustando umbral).
- **FN (falsos negativos):** Segmentos con riesgo alto real que el modelo no detectó (error más crítico, porque implica omitir segmentos potencialmente vulnerables).
- **TN (verdaderos negativos):** Segmentos correctamente clasificados como bajo riesgo (sin priorización inmediata).

En la práctica, para cada periodo (p. ej., Test 2023) se construye una tabla final con; SEGMENTO_ID, etiqueta real (0/1), predicción (0/1), probabilidad estimada y categoría (TP/FP/FN/TN). Esto permite: (i) generar el ranking Top-N con los TP (y eventualmente FP de alta probabilidad), (ii) documentar segmentos no riesgo (TN) como grupo de comparación, y (iii) analizar casos frontera para ajustar el umbral según la capacidad institucional. Se recomienda incluir en anexos un extracto de estas tablas (Top-N y ejemplos de FN/FP) para evidenciar trazabilidad y uso operativo del sistema

4.1.9. Desempeño de modelos con balanceo (comparación antes/después)

Bajo el esquema de validación temporal, se comparó el desempeño de Random Forest, AdaBoost y Gradient Boosting antes y después de aplicar balanceo de clases. Primero se evaluó la línea base sin balanceo con el mismo pipeline y el mismo split (Valid/Test). Luego se incorporó SMOTE dentro del pipeline solo en entrenamiento, evitando fuga de información, y se reentrenaron los tres modelos en las mismas condiciones.

En la línea base, Random Forest y Gradient Boosting muestran métricas globales más altas y mejor detección de la clase minoritaria que AdaBoost. Tras aplicar SMOTE, mejora principalmente la identificación de RIESGO_ALTO (mayor *recall*), especialmente en AdaBoost, mientras que Random Forest y Gradient Boosting mantienen un desempeño alto y estable. En conjunto, la comparación confirma que el balanceo fortalece la detección de segmentos/cohortes en riesgo alto sin alterar el enfoque temporal del experimento.

Ilustración 41. Comparación de sin balanceo(antes)

	modelo	split	accuracy	precision	recall	f1	auc
0	Random Forest	valid	0.986013	0.977193	0.954262	0.965592	0.998963
1	Random Forest	test	0.983185	0.980751	0.951115	0.965705	0.998635
2	AdaBoost	valid	0.937312	0.913008	0.768410	0.834492	0.986139
3	AdaBoost	test	0.916293	0.904684	0.741875	0.815230	0.980969
4	Gradient Boosting	valid	0.980608	0.966590	0.938135	0.952150	0.998385
5	Gradient Boosting	test	0.972856	0.967578	0.921837	0.944154	0.997700

Nota. Fuente: Elaboración propia Romero Cristhian 2026

La Ilustración 41 presenta las métricas comparativas (accuracy, precision, recall, F1 y AUC) de Random Forest, AdaBoost y Gradient Boosting en validación y prueba sin aplicar balanceo. Se observa que Random Forest y Gradient Boosting alcanzan los valores más altos de desempeño global, destacando por su equilibrio entre *precision* y *recall*, y por AUC elevados. En contraste, AdaBoost muestra métricas inferiores, especialmente en la detección de la clase minoritaria (RIESGO_ALTO), reflejado en menores valores de *recall* y *F1* en comparación con los otros dos modelos.

Ilustración 42. Comparación de balanceo(después)

Tabla con balanceo (SMOTE):

	Modelo	Split	accuracy	precision	recall	f1	auc
0	Random Forest	Valid (2021-2022)	0.9852	0.9608	0.9677	0.9642	0.9988
1	Random Forest	Test (2023)	0.9818	0.9595	0.9679	0.9637	0.9984
2	AdaBoost	Valid (2021-2022)	0.9429	0.8166	0.9313	0.8702	0.9855
3	AdaBoost	Test (2023)	0.9279	0.8117	0.9247	0.8645	0.9802
4	Gradient Boosting	Valid (2021-2022)	0.9764	0.9241	0.9646	0.9439	0.9967
5	Gradient Boosting	Test (2023)	0.9678	0.9145	0.9604	0.9369	0.9954

Nota. Fuente: Elaboración propia Romero Cristhian 2026

La Ilustración 42 muestra el desempeño de los mismos modelos tras aplicar balanceo con SMOTE dentro del pipeline, únicamente en el conjunto de entrenamiento. En general, se evidencia un incremento en la sensibilidad hacia la clase minoritaria (mejoras en *recall*), siendo el cambio más notorio en AdaBoost, que mejora su capacidad para identificar casos de riesgo alto, aunque con una posible reducción en *precision* (comportamiento esperado al aumentar la cobertura de la clase minoritaria). Random Forest y Gradient Boosting mantienen métricas globales altas (F1 y AUC elevados) y muestran mejoras puntuales en la detección del riesgo

alto, confirmando que el balanceo contribuye a fortalecer la identificación de segmentos/cohortes priorizados sin afectar la consistencia del esquema temporal de evaluación.

4.1.9.1. Impacto del desbalance de clases y efectividad de las técnicas aplicadas.

En esta investigación, la variable objetivo RIESGO_ALTO presenta una distribución desbalanceada: la clase mayoritaria corresponde a riesgo bajo (0) y la minoritaria a riesgo alto (1). En problemas de clasificación con desbalance, es frecuente que los modelos optimicen el acierto sobre la clase mayoritaria, lo que puede inflar métricas globales como accuracy, pero reducir la capacidad de detectar la clase de interés (riesgo alto). Por esta razón, además de la exactitud, se priorizaron métricas sensibles a la clase minoritaria, particularmente recall (sensibilidad) y F1-score, ya que un sistema de alerta temprana pierde utilidad si omite segmentos/cohortes realmente en riesgo (falsos negativos).
Para mitigar este efecto, se comparó el desempeño antes y después de aplicar técnicas de balanceo. En la línea base (sin balanceo), Random Forest y Gradient Boosting mantienen alta capacidad discriminativa (AUC elevado) y una detección robusta de la clase minoritaria, mientras que AdaBoost muestra mayor dificultad relativa para identificar riesgo alto (mayor proporción de falsos negativos). Posteriormente, al incorporar SMOTE únicamente dentro del conjunto de entrenamiento (evitando fuga de información), se observó una mejora más marcada en la detección de RIESGO_ALTO, reflejada principalmente en incrementos de recall y F1-score, especialmente en AdaBoost, lo cual es consistente con la literatura: el balanceo suele incrementar sensibilidad, aunque puede reducir precisión debido al aumento de alertas (más falsos positivos).
En términos institucionales, este resultado es clave: si el objetivo es no omitir segmentos/cohortes críticos para intervención preventiva, conviene priorizar configuraciones con mayor recall (aceptando algunas alertas adicionales). En cambio, si los recursos de intervención son limitados, puede ajustarse el umbral de decisión para mejorar precisión y complementar la priorización con criterios operativos (capacidad instalada, criticidad del segmento, ventanas de atención). En síntesis, la evaluación confirma que el desbalance afecta la detección del riesgo alto y que el balanceo aplicado fortalece la sensibilidad del sistema sin romper el enfoque temporal del experimento.

Además, el efecto del desbalance se evidencia directamente en la matriz de confusión: cuando el modelo se inclina hacia la clase mayoritaria (0), el error que tiende a crecer es el de falsos negativos (FN), es decir, segmentos con riesgo alto que quedan sin priorización. Esto es crítico en un enfoque preventivo, porque un FN equivale a "no activar" acompañamiento donde sí podría ser necesario. En contraste, el costo de un falso positivo (FP) es generar una alerta adicional que puede gestionarse mediante revisión operativa o ajuste de umbral. Por ello, la comparación antes/después del balanceo debe interpretarse no solo en métricas agregadas, sino en la reducción de FN y el comportamiento del trade-off recall–precision, lo que refuerza la pertinencia de SMOTE aplicado únicamente en entrenamiento y la priorización de recall/F1 para el objetivo institucional del estudio.

4.1.10. Comparación global y selección del modelo final

Con base en los resultados de validación (2021–2022) y, principalmente, de prueba (Test 2023), se compararon Random Forest, AdaBoost y Gradient Boosting mediante accuracy, precision, recall, F1 y AUC (Ilustración 43). Debido al desbalance de clases, se priorizó recall y F1 (clase minoritaria) y AUC.

En Test 2023, Random Forest obtuvo el mejor desempeño (accuracy≈0.983, recall≈0.951, F1≈0.966, AUC≈0.999), seguido por Gradient Boosting (recall≈0.922, F1≈0.944, AUC≈0.998), mientras que AdaBoost fue inferior (recall≈0.742, F1≈0.815). Por ello, se seleccionó Random Forest como modelo final para predecir RIESGO_ALTO.

Ilustración 43. Comparación Global y selección del modelo final

	Modelo	Split	accuracy	precision	recall	f1	auc
0	Random Forest	Valid (2021-2022)	0.9860	0.9772	0.9543	0.9656	0.9990
1	Random Forest	Test (2023)	0.9832	0.9808	0.9511	0.9657	0.9986
2	AdaBoost	Valid (2021-2022)	0.9373	0.9130	0.7684	0.8345	0.9861
3	AdaBoost	Test (2023)	0.9163	0.9047	0.7419	0.8152	0.9810
4	Gradient Boosting	Valid (2021-2022)	0.9806	0.9666	0.9381	0.9522	0.9984
5	Gradient Boosting	Test (2023)	0.9729	0.9676	0.9218	0.9442	0.9977

Nota. Fuente: Elaboración propia Romero Cristhian 2026

Además, en la siguiente ilustración 43 se presenta el ranking final de los modelos evaluados, considerando exclusivamente el desempeño en el conjunto de prueba (Test 2023). La selección del modelo final se realizó priorizando las métricas F1, AUC y recall, evidenciando

a Random Forest como el modelo con mejor rendimiento global para la predicción del riesgo alto, seguido por Gradient Boosting, mientras que AdaBoost mostró resultados inferiores.

Ilustración 44. Elección de la modelo final basada en métricas del conjunto Test

	Modelo	Split	accuracy	precision	recall	f1	auc
1	Random Forest	Test (2023)	0.9832	0.9808	0.9511	0.9657	0.9986
5	Gradient Boosting	Test (2023)	0.9729	0.9676	0.9218	0.9442	0.9977
3	AdaBoost	Test (2023)	0.9163	0.9047	0.7419	0.8152	0.9810

Nota. Fuente: Elaboración propia Romero Cristhian 2026

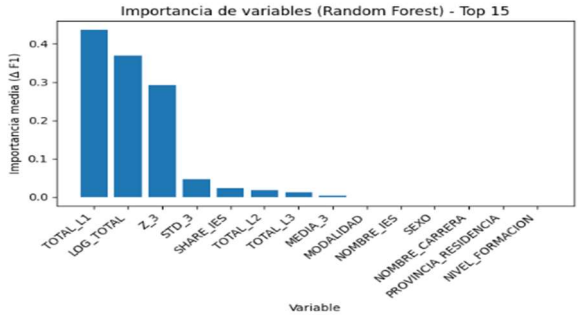
4.1.11. Interpretabilidad del modelo final (importancia de variables / SHAP si aplica)

Para explicar el modelo final (Random Forest), se analizó la importancia de variables, identificando qué predictores influyen más en la clasificación de RIESGO_ALTO. En general, las variables temporales derivadas de la matrícula (rezagos y medidas históricas como *TOTAL_L1, TOTAL_L2, MEDIA_3, STD_3, Z_3 y LOG_TOTAL*) aportan la mayor información, indicando que el modelo se basa principalmente en el comportamiento previo del segmento. De forma complementaria, puede aplicarse SHAP para interpretar las predicciones: a nivel global (variables más influyentes) y local (qué variables explican un caso específico), facilitando una lectura clara y útil de los resultados.

Ilustración 45. Interpretabilidad del modelo final (Random Forest)

Top 15 variables más importantes (Permutation Importance, scoring=F1):

	variable	importancia_media	importancia_std
6	TOTAL_L1	0.437143	0.001208
11	LOG_TOTAL	0.368169	0.002064
13	Z_3	0.292336	0.001904
10	STD_3	0.046912	0.001652
12	SHARE_IES	0.022375	0.000751
7	TOTAL_L2	0.018152	0.000894
8	TOTAL_L3	0.013041	0.000585
9	MEDIA_3	0.003824	0.000521
3	MODALIDAD	0.000024	0.000259
0	NOMBRE_IES	-0.000123	0.000486
4	SEXO	-0.000297	0.000181
1	NOMBRE_CARRERA	-0.000444	0.000208
5	PROVINCIA_RESIDENCIA	-0.000452	0.000181
2	NIVEL_FORMACION	-0.000981	0.000280



Nota. Fuente: Elaboración propia Romero Cristhian 2026

La ilustración 45 muestra el análisis de importancia de variables del modelo Random Forest mediante Permutation Importance (midiendo cuánto disminuye el F1 cuando se altera cada variable). En la tabla y el gráfico de barras se observa que las variables con mayor aporte al desempeño del modelo son TOTAL_L1 (matrícula del año previo), LOG_TOTAL (transformación logarítmica del total) y Z_3 (indicador estandarizado), ya que generan las mayores caídas en F1 cuando se permutan. En un segundo nivel aparecen STD_3 y SHARE_IES, relacionadas con la variabilidad histórica y la participación relativa del segmento dentro de la IES. En contraste, las variables categóricas del segmento (como NOMBRE_IES, PROVINCIA_RESIDENCIA, NOMBRE_CARRERA, MODALIDAD, SEXO y NIVEL_FORMACION) presentan una influencia menor en comparación con las variables temporales, lo que sugiere que el modelo se apoya principalmente en la dinámica histórica de la matrícula para identificar el riesgo alto.

4.1.12. Segmentos/cohortes priorizados: ranking de mayor riesgo (Top N)

A partir del modelo final seleccionado (Random Forest), se estimó la probabilidad de riesgo alto para cada segmento/cohorte utilizando el conjunto de prueba (Test 2023). Con base en estas probabilidades, se construyó un ranking de priorización (Top N), ordenando los segmentos de mayor a menor riesgo estimado. Este ranking permite identificar de manera objetiva aquellos segmentos/cohortes con mayor probabilidad de deserción, constituyendo una herramienta analítica relevante para la focalización de estrategias de intervención institucional, tales como acciones de acompañamiento académico, seguimiento temprano y asignación eficiente de recursos. De esta forma, el modelo no solo demuestra capacidad predictiva, sino también utilidad práctica para la toma de decisiones en el ámbito de la gestión educativa.

Ilustración 46. Ranking de los segmentos/cohortes con mayor probabilidad

	SEGMENTO_ID	NOMBRE_IES	NOMBRE_CARRERA	NIVEL_FORMACION	MODALIDAD	SEXO	PROVINCIA_RESIDENCIA	PROBA_MEDIA	PROBA_MAX	N_OBS
3591	10636	PONTIFICIA UNIVERSIDAD CATOLICA DEL ECUADOR	TERAPIA FISICA	TERCER NIVEL DE GRADO	PRESENCIAL	MUJER	IMBABURA	1.0	1.0	1
3609	10674	PONTIFICIA UNIVERSIDAD CATOLICA DEL ECUADOR	URBANISMO	CUARTO NIVEL O POSGRADO	EN LINEA	HOMBRE	CHIMBORAZO	1.0	1.0	1
14455	36489	UNIVERSIDAD ESTATAL DE MILAGRO	EDUCACION BASICA	CUARTO NIVEL O POSGRADO	EN LINEA	MUJER	GALAPAGOS	1.0	1.0	1
26566	66629	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	CONTABILIDAD Y AUDITORIA	TERCER NIVEL DE GRADO	A DISTANCIA	HOMBRE	ORELLANA	1.0	1.0	1
14214	36021	UNIVERSIDAD ESTATAL DE MILAGRO	COMUNICACION	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	LOJA	1.0	1.0	1
14212	36019	UNIVERSIDAD ESTATAL DE MILAGRO	COMUNICACION	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	EL ORO	1.0	1.0	1
14741	37136	UNIVERSIDAD ESTATAL DE MILAGRO	INGENIERIA INDUSTRIAL	TERCER NIVEL DE GRADO	PRESENCIAL	MUJER	EL ORO	1.0	1.0	1
14738	37132	UNIVERSIDAD ESTATAL DE MILAGRO	INGENIERIA INDUSTRIAL	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	ZAMORA CHINCHIPE	1.0	1.0	1
14215	36022	UNIVERSIDAD ESTATAL DE MILAGRO	COMUNICACION	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	LOS RIOS	1.0	1.0	1
14696	37026	UNIVERSIDAD ESTATAL DE MILAGRO	GESTION EDUCATIVA	CUARTO NIVEL O POSGRADO	EN LINEA	MUJER	ORELLANA	1.0	1.0	1
26542	66586	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	COMUNICACION SOCIAL	TERCER NIVEL DE GRADO	A DISTANCIA	MUJER	LOJA	1.0	1.0	1
26539	66581	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	COMUNICACION SOCIAL	TERCER NIVEL DE GRADO	A DISTANCIA	MUJER	EL ORO	1.0	1.0	1
26533	66562	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	COMUNICACION SOCIAL	TERCER NIVEL DE GRADO	A DISTANCIA	HOMBRE	MANABI	1.0	1.0	1
26532	66560	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	COMUNICACION SOCIAL	TERCER NIVEL DE GRADO	A DISTANCIA	HOMBRE	LOJA	1.0	1.0	1
26530	66558	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	COMUNICACION SOCIAL	TERCER NIVEL DE GRADO	A DISTANCIA	HOMBRE	GUAYAS	1.0	1.0	1
26527	66555	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	COMUNICACION SOCIAL	TERCER NIVEL DE GRADO	A DISTANCIA	HOMBRE	EL ORO	1.0	1.0	1
26526	66553	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	COMUNICACION SOCIAL	TERCER NIVEL DE GRADO	A DISTANCIA	HOMBRE	CHIMBORAZO	1.0	1.0	1
26524	66551	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	COMUNICACION SOCIAL	TERCER NIVEL DE GRADO	A DISTANCIA	HOMBRE	CARCHI	1.0	1.0	1
3512	10454	PONTIFICIA UNIVERSIDAD CATOLICA DEL ECUADOR	SOCIOLOGIA CON MENCION EN RELACIONES INTERNACI...	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	PICHINCHA	1.0	1.0	1
3740	10921	UNIVERSIDAD AGRARIA DEL ECUADOR	INGENIERIA AGRICOLA MENCION AGROINDUSTRIAL	TERCER NIVEL DE GRADO	PRESENCIAL	MUJER	CHIMBORAZO	1.0	1.0	1
14829	37506	UNIVERSIDAD ESTATAL DE MILAGRO	PEDAGOGIA DE LA LENGUA Y LA LITERATURA	TERCER NIVEL DE GRADO	PRESENCIAL	MUJER	LOJA	1.0	1.0	1
14147	35923	UNIVERSIDAD ESTATAL DE MILAGRO	BIOTECNOLOGIA	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	CHIMBORAZO	1.0	1.0	1
14751	37251	UNIVERSIDAD ESTATAL DE MILAGRO	LICENCIATURA EN ENFERMERIA	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	GUAYAS	1.0	1.0	1
14750	37238	UNIVERSIDAD ESTATAL DE MILAGRO	LICENCIATURA EN DISEÑO GRAFICO Y PUBLICIDAD	TERCER NIVEL DE GRADO	PRESENCIAL	MUJER	GUAYAS	1.0	1.0	1
14749	37233	UNIVERSIDAD ESTATAL DE MILAGRO	LICENCIATURA EN DISEÑO GRAFICO Y PUBLICIDAD	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	GUAYAS	1.0	1.0	1
3758	10989	UNIVERSIDAD AGRARIA DEL ECUADOR	INGENIERIA AMBIENTAL	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	SANTO DOMINGO DE LOS TSACHILAS	1.0	1.0	1
3824	11201	UNIVERSIDAD ANDINA SIMON BOLIVAR	ADMINISTRACION DE EMPRESAS	CUARTO NIVEL O POSGRADO	PRESENCIAL	MUJER	CHIMBORAZO	1.0	1.0	1
3494	10398	PONTIFICIA UNIVERSIDAD CATOLICA DEL ECUADOR	SISTEMAS DE INFORMACION	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	MANABI	1.0	1.0	1
26432	66382	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	CIENCIAS DE LA EDUCACION MENCION QUIMICA Y BIO...	TERCER NIVEL DE GRADO	A DISTANCIA	MUJER	ZAMORA CHINCHIPE	1.0	1.0	1
26427	66368	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	CIENCIAS DE LA EDUCACION MENCION QUIMICA Y BIO...	TERCER NIVEL DE GRADO	A DISTANCIA	MUJER	GUAYAS	1.0	1.0	1

Nota. Fuente: Elaboración propia Romero Cristhian 2026

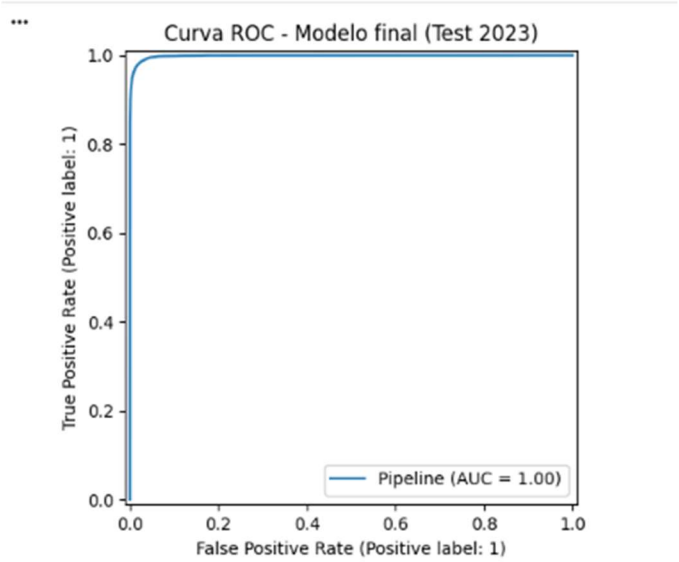
4.2. Análisis

4.2.1. Interpretación general según el objetivo general

En relación con el objetivo general, los resultados confirman que es posible construir un modelo predictivo que estime el riesgo de deserción a nivel agregado (segmento/cohorte) utilizando datos abiertos de matrícula (2015–2023). La estrategia metodológica basada en la consolidación segmento–año, la generación de variables temporales desde el conteo de matrícula (TOTAL_SEG) y la validación con enfoque temporal permitió obtener un desempeño elevado y estable en el conjunto de prueba (2023).

El alto desempeño observado en modelos tipo ensemble (Random Forest y Gradient Boosting) sugiere que la dinámica histórica de matrícula contiene señales consistentes para discriminar segmentos con mayor probabilidad de presentar caídas interanuales relevantes (operacionalizadas como riesgo alto). En términos institucionales, esto respalda la utilidad del enfoque como herramienta de alerta temprana agregada, orientada a priorizar segmentos/cohortes y focalizar recursos de acompañamiento, monitoreo y análisis académico-programático.

Ilustración 47. Curva ROC-Modelo Final



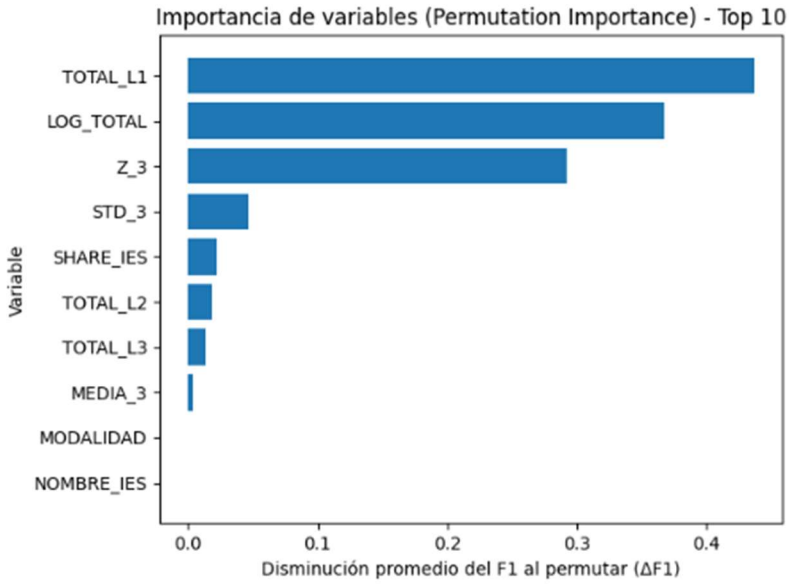
Nota. Fuente: Elaboración propia Romero Cristhian 2026

4.2.2. Análisis por objetivos específicos

4.2.2.1. Variables relevantes y patrones por segmento/cohorte

En cumplimiento del OE1, el análisis evidencia que las variables con mayor aporte para estimar el riesgo de deserción a nivel agregado se relacionan principalmente con la dinámica temporal de la matrícula por segmento/cohorte. La interpretabilidad del modelo final muestra que indicadores derivados del historial (por ejemplo, TOTAL_L1, LOG_TOTAL y Z_3, seguidos por STD_3 y SHARE_IES) concentran la mayor capacidad discriminativa, lo que sugiere que el riesgo alto se asocia a caídas interanuales, inestabilidad y desviaciones respecto al comportamiento reciente del segmento. En contraste, las variables categóricas del segmento (institución, carrera, modalidad, nivel, sexo y territorio) aportan menos y se interpretan como contexto. En síntesis, el riesgo agregado se explica sobre todo por patrones históricos de matrícula, lo que respalda la selección de variables y la operacionalización del riesgo usada en el modelado.

Ilustración 48. Importancia de variables

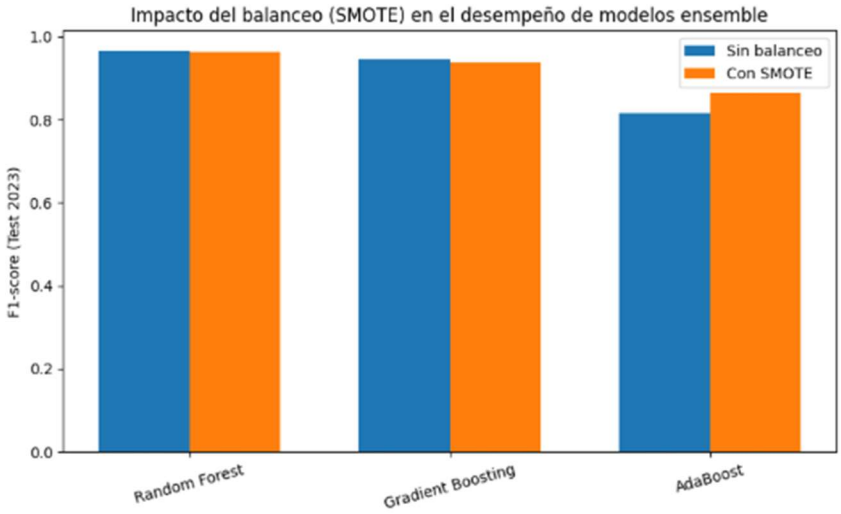


Nota. Fuente: Elaboración propia Romero Cristhian 2026

4.2.2.2. Aporte del pipeline y del balanceo

En cumplimiento del OE2, los resultados muestran que el pipeline de preprocesamiento y el balanceo con SMOTE fortalecen la identificación de segmentos/cohortes con riesgo alto. El pipeline integró imputación, codificación de variables categóricas y estandarización, manteniendo consistencia metodológica y evitando fuga de información mediante validación temporal. La aplicación de SMOTE solo en entrenamiento redujo el desbalance y mejoró principalmente recall y F1-score, sin cambios relevantes en accuracy ni AUC. En particular, Random Forest y Gradient Boosting mantuvieron un desempeño alto y estable en validación y prueba (2023). Aunque AdaBoost mejora con SMOTE, continúa por debajo de los otros modelos, especialmente en precisión y F1. En conjunto, estos hallazgos confirman que el pipeline y el balanceo son claves para robustecer el modelo y apoyar una alerta temprana agregada.

Ilustración 49. Impacto de balanceo

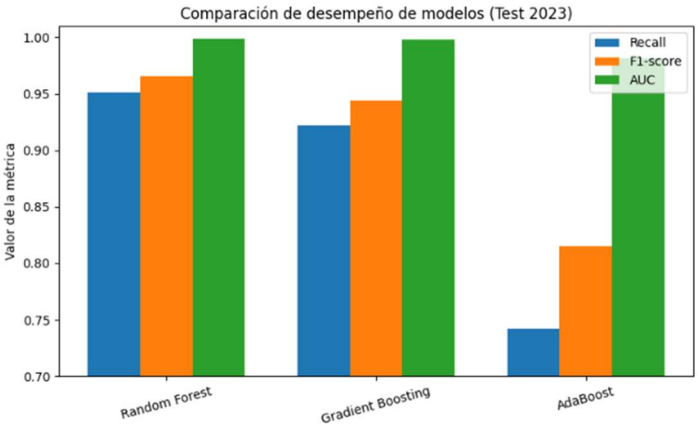


Nota. Fuente: Elaboración propia Romero Cristhian 2026

4.2.2.3. Comparación técnica de los tres modelos (ensemble)

Se compararon Random Forest (bagging), Gradient Boosting (boosting) y AdaBoost (boosting) con el mismo pipeline y validación temporal. En Test 2023, Random Forest obtuvo el mejor desempeño global (F1 = 0.9657; recall = 0.9511; AUC = 0.9986), seguido por Gradient Boosting (F1 = 0.9442; recall = 0.9218). AdaBoost fue inferior, especialmente en recall (0.7419), mostrando menor capacidad para detectar la clase minoritaria (riesgo alto).

Ilustración 50. Comparación de desempeño de modelos

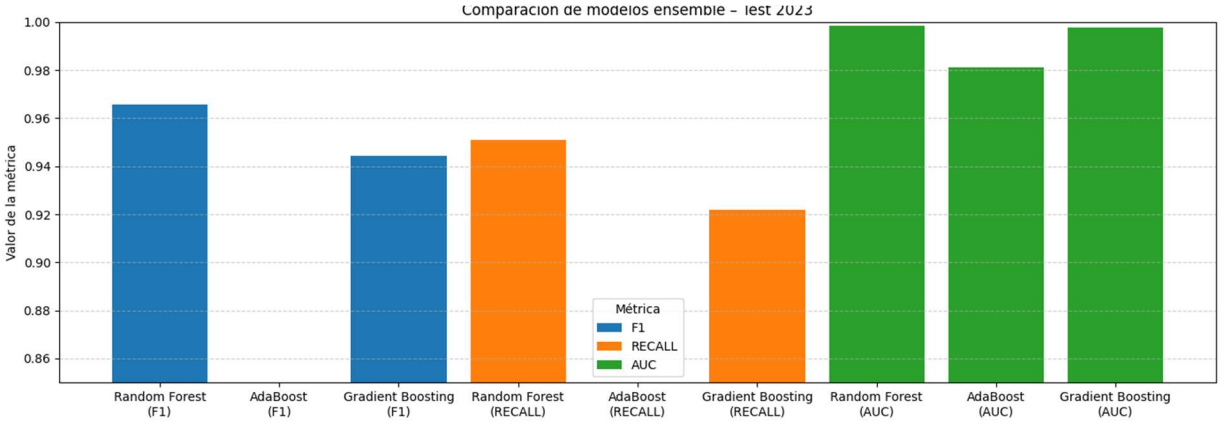


Nota. Fuente: Elaboración propia Romero Cristhian 2026

4.2.2.4. Justificación del mejor algoritmo con base en métricas

Con base en la evaluación comparativa en el conjunto de prueba (Test 2023), el modelo Random Forest se selecciona como el algoritmo más eficiente para estimar el riesgo de deserción a nivel agregado. Esta decisión se fundamenta en que obtuvo el mejor equilibrio global entre métricas, destacando especialmente en indicadores críticos para un problema desbalanceado: recall, F1-score y AUC.En particular, Random Forest alcanzó valores superiores en F1 y recall, lo que evidencia una alta capacidad para identificar correctamente la clase minoritaria (RIESGO_ALTO) y reducir falsos negativos, aspecto esencial para un enfoque de alerta temprana. Adicionalmente, su AUC cercano a 1 confirma una capacidad discriminativa excelente y estable, reflejada también en las curvas ROC observadas. Aunque Gradient Boosting mostró resultados competitivos, se ubicó ligeramente por debajo en F1 y recall, mientras que AdaBoost presentó el menor desempeño relativo, especialmente en sensibilidad al riesgo alto. En síntesis, Random Forest ofrece el mejor desempeño para la finalidad del estudio: priorizar segmentos/cohortes con alta probabilidad de caída interanual de matrícula, con un balance adecuado entre precisión y detección, lo que respalda su selección como modelo final para apoyar la toma de decisiones institucionales.

Ilustración 51. Comparación de mejor algoritmo



Nota. Fuente: Elaboración propia Romero Cristhian 2026

4.2.3. Respuesta a preguntas de investigación

4.2.3.1. Variables con mayor importancia predictiva

Los resultados de interpretabilidad del modelo final evidencian que las variables con mayor importancia predictiva se asocian principalmente a la dinámica histórica de la matrícula

del segmento/cohorte. Destacan variables temporales derivadas del TOTAL como TOTAL_L1, LOG_TOTAL y Z_3, seguidas por medidas de estabilidad/variación como STD_3 y SHARE_IES. En contraste, las variables categóricas del segmento (institución, carrera, modalidad, nivel, sexo y territorio) aportan menor peso relativo y se interpretan como información contextual para caracterizar los segmentos.

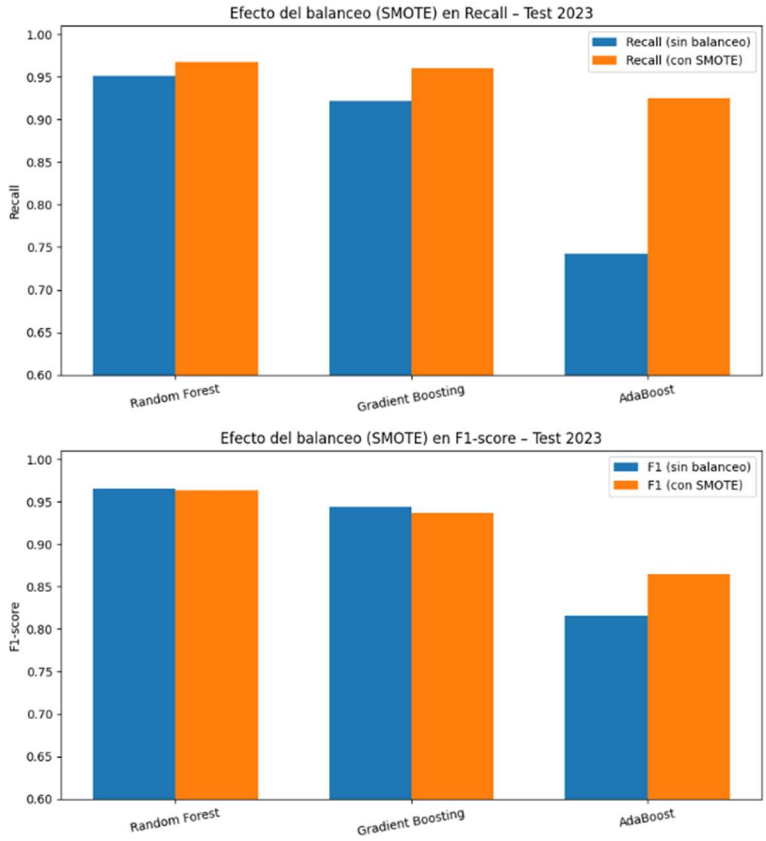
4.2.3.2. Algoritmo con mejor desempeño

De acuerdo con la evaluación en el conjunto de prueba (2023), el algoritmo con mejor desempeño fue Random Forest, al obtener el mejor equilibrio entre métricas relevantes para clasificación con desbalance, especialmente F1-score, recall y AUC. Gradient Boosting presentó resultados competitivos, pero ligeramente inferiores, mientras que AdaBoost mostró el rendimiento más bajo, particularmente en la detección de la clase minoritaria (riesgo alto).

4.2.3.3. Efecto del desbalance y balanceo

El desbalance observado (mayoría de casos en riesgo bajo) puede sesgar el aprendizaje hacia la clase mayoritaria y reducir la detección del riesgo alto. Al aplicar SMOTE únicamente en entrenamiento, se mejora la representación de la clase minoritaria, incrementando principalmente recall y F1-score sin afectar de forma importante accuracy ni AUC. En general, el balanceo fortaleció la capacidad de los modelos para identificar segmentos/cohortes con riesgo alto, siendo más notorio en modelos con menor sensibilidad inicial (como AdaBoost), y manteniendo un desempeño alto y estable en Random Forest y Gradient Boosting.

Ilustración 52. Efecto del desbalance y balanceo



Nota. Fuente: Elaboración propia Romero Cristhian 2026

4.2.4. Discusión con el marco teórico y estudios previos

Los resultados del modelo se interpretan de forma consistente con el marco teórico conceptualiza la deserción como un fenómeno multifactorial asociado a dimensiones académicas, institucionales, demográficas y contextuales (Granda et al., 2024 ; Moreira & Caicedo, 2024). En coherencia con ello, este estudio propone una aproximación desde analítica educativa y aprendizaje automático para estimar riesgo de deserción a nivel agregado (segmento/cohorte) utilizando datos abiertos de matrícula. No obstante, es importante precisar que la evidencia empírica proviene de conteos agregados (TOTAL) y no de trayectorias individuales; por tanto, el modelo estima riesgo respecto de un proxy operativo (caídas interanuales relevantes de matrícula por segmento), y su interpretación debe centrarse en patrones estructurales del sistema más que en causas individuales del abandono. Los resultados muestran que los predictores más influyentes corresponden a variables temporales derivadas del historial de matrícula (rezagos, medidas de dispersión, estandarización

y transformaciones), lo que sugiere que el riesgo agregado se asocia principalmente con volatilidad, inestabilidad y desviaciones respecto del comportamiento reciente del segmento. Este hallazgo es consistente con enfoques de analítica predictiva que sostienen que los patrones longitudinales contienen señales útiles para anticipar escenarios de riesgo, particularmente cuando se emplean modelos de aprendizaje automático con capacidad de capturar interacciones y no linealidades (Bouihi et al., 2024).

13

Metodológicamente, el mejor desempeño de modelos ensemble (Random Forest y Gradient Boosting) refuerza la idea de que la relación entre los predictores y el proxy de riesgo no responde únicamente a patrones lineales; más bien, se explica por combinaciones de señales temporales que, en conjunto, incrementan la probabilidad de clasificar un segmento como riesgo alto. Además, la comparación con y sin balanceo es coherente con recomendaciones habituales en clasificación desbalanceada: no basta con accuracy, sino que deben priorizarse métricas como recall y F1 cuando el objetivo es preventivo y se busca minimizar omisiones (falsos negativos).

En síntesis, los hallazgos respaldan el valor de emplear datos abiertos y analítica educativa como soporte para una alerta temprana agregada, útil para monitoreo institucional, priorización de segmentos/cohortes y focalización de estrategias (tutorías, refuerzos, acompañamiento y revisión académico-programática), sin sustituir el análisis cualitativo o institucional que explique las causas subyacentes.

4.2.5. Implicaciones para toma de decisiones y alerta temprana agregada

Las implicaciones directas para la gestión institucional, especialmente en el diseño de mecanismos de alerta temprana y priorización de acciones. Dado que el estudio se apoya en datos agregados, el aporte principal se orienta a la identificación y monitoreo de riesgo por unidades de análisis (por ejemplo, cohorte, carrera, periodo académico, asignatura, jornada o modalidad), lo que permite decisiones a nivel de planificación y focalización de recursos.Una implementación práctica derivada de este trabajo consiste en transformar la salida del modelo (probabilidad/score) en indicadores de monitoreo, tales como: (i) riesgo promedio por cohorte o programa, (ii) proporción de unidades clasificadas en riesgo alto/medio/bajo, y (iii) tendencias temporales del riesgo entre periodos. Con ello, la institución puede detectar oportunamente "puntos críticos" (por ejemplo, cohortes o periodos donde el riesgo aumenta) y activar estrategias preventivas: tutorías académicas, acompañamiento psicoeducativo, refuerzo en

asignaturas críticas, campañas de retención o ajustes en procesos administrativos que inciden en la continuidad. Adicionalmente, la definición de umbrales de alerta debe responder a la lógica de decisión. Si el costo de no intervenir es alto, el sistema debería priorizar sensibilidad/recall (captar la mayor cantidad de casos de riesgo). En cambio, si la intervención requiere recursos significativos, conviene ajustar el umbral para mejorar precisión y complementar la priorización con criterios operativos (capacidad instalada, severidad del riesgo, ventanas de tiempo disponibles). En ambos casos, la recomendación es que el modelo funcione como apoyo analítico y que su uso esté integrado a rutas de acción claramente definidas, con responsables y seguimiento. Finalmente, para consolidar la utilidad institucional del sistema, se requiere medir impacto: cobertura de intervención, respuesta de los beneficiarios, evolución de indicadores académicos y comparación de resultados entre periodos. Esto evita que la predicción se convierta únicamente en un ejercicio descriptivo y permite retroalimentar el sistema de alerta con evidencia.

4.2.6. Limitaciones y amenazas a la validez (datos agregados y proxy de deserción)

Este estudio presenta limitaciones que deben considerarse para interpretar adecuadamente los resultados y el alcance de aplicación.

En primer lugar, el uso de datos agregados introduce el riesgo de falacia ecológica, es decir, inferir comportamientos individuales a partir de patrones grupales. Esto limita el uso del modelo para decisiones individualizadas, pero no invalida su utilidad para monitoreo y planificación institucional, que es el propósito principal del enfoque agregado.

En segundo lugar, la variable objetivo se operacionaliza mediante un proxy de deserción (por ejemplo, no matrícula, inactividad u otra definición indirecta). Este enfoque puede incorporar ruido en la etiqueta, ya que el proxy puede capturar situaciones distintas al abandono definitivo (pausas temporales, movilidad académica, cambios administrativos o condiciones externas). Como consecuencia, el modelo estima riesgo respecto del proxy definido y no necesariamente respecto de la deserción confirmada en sentido estricto.

En tercer lugar, existe la amenaza de fuga de información (data leakage) si alguno de los predictores refleja información posterior o demasiado cercana al evento que se pretende

anticipar. Por ello, es clave asegurar consistencia temporal (que las variables utilizadas estén disponibles antes del momento de predicción) y validar con esquemas adecuados.

En cuarto lugar, el desbalance de clases puede afectar el entrenamiento y la interpretación de métricas. Si se aplicaron técnicas de remuestreo (como SMOTE), si bien pueden mejorar la detección de la clase minoritaria, también pueden aumentar el riesgo de sobreajuste o alterar la calibración. Por ello, la interpretación debe apoyarse en desempeño en datos de prueba y, de ser posible, validación por periodos/cohortes no vistas.

Finalmente, como en la mayoría de los estudios observacionales, es probable que existan variables no medidas y posibles sesgos estructurales (por ejemplo, diferencias socioeconómicas capturadas indirectamente por ciertos predictores). En consecuencia, el uso institucional del modelo debe orientarse a apoyo y acompañamiento, evitando enfoques punitivos y promoviendo monitoreo ético del desempeño del sistema.

4.2.6.1. Amenazas a la validez externa

Una primera amenaza a la validez externa corresponde a la generalización temporal; cambios en políticas educativas, reestructuración de oferta académica, condiciones económicas o eventos externos pueden modificar los patrones de matrícula entre periodos (concept drift), afectando la estabilidad del modelo si se aplica en años futuros sin recalibración.
En segundo lugar, existe heterogeneidad institucional y territorial: el comportamiento de la matrícula puede variar entre universidades, carreras, modalidades y provincias; por ello, el desempeño global puede ocultar diferencias por subgrupos y conviene complementar con validaciones estratificadas (por IES, provincia o campo).
En tercer lugar, el uso de datos agregados limita la generalización a decisiones individualizadas; el modelo es válido como herramienta de monitoreo y planificación agregada, no como predictor de deserción por estudiante.
Finalmente, la variable objetivo es un proxy basado en caída interanual de matrícula; por tanto, puede capturar fenómenos distintos al abandono definitivo (movilidad académica, pausas, cambios administrativos, apertura/cierre de cohortes). En consecuencia, el modelo predice riesgo respecto del proxy definido y debe interpretarse dentro de ese alcance.

4.2.7. Síntesis final del apartado

En este apartado se discutieron los resultados del modelo predictivo del proxy de deserción, articulándolos con el marco teórico y hallazgos reportados en estudios previos. Los patrones identificados muestran coherencia con enfoques que explican la deserción como un proceso asociado al desempeño/avance académico y a señales de continuidad o vinculación institucional. El modelo seleccionado demostró el mejor desempeño entre las alternativas evaluadas, respaldando su elección como herramienta analítica para fines de monitoreo y alerta.

A nivel aplicado, los resultados permiten proponer un esquema de alerta temprana agregada que facilite identificar unidades académicas con mayor concentración de riesgo, apoyar la asignación de recursos y orientar intervenciones preventivas. No obstante, la interpretación debe considerar las principales limitaciones del estudio: la naturaleza agregada de los datos, el uso de una etiqueta proxy, y las amenazas asociadas al desbalance, la temporalidad y posibles sesgos. En consecuencia, el modelo se plantea como un insumo para la toma de decisiones basada en evidencia, complementario a la evaluación académica y a la gestión institucional, con necesidad de validación y monitoreo continuo.

26

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

Se reporta el ranking Top N de segmentos/cohortes con mayor probabilidad estimada de riesgo alto (Test 2023). El listado se incluye como insumo para priorización institucional y focalización de acciones de monitoreo y acompañamiento a nivel agregado.

El presente trabajo construyó y evaluó un modelo predictivo para estimar el riesgo de deserción a nivel agregado (segmento/cohorte) utilizando datos abiertos de matrícula del SNIESE/SENESCYT correspondientes al período 2015–2023. A partir de la consolidación segmento–año y la ingeniería de variables temporales basadas en el conteo de matrícula (TOTAL_SEG), fue posible capturar patrones longitudinales y desarrollar un enfoque cuantitativo replicable que permite identificar segmentos con probabilidad elevada de presentar caídas interanuales relevantes, lo cual funciona como una señal operativa de riesgo para fines de monitoreo institucional.

La operacionalización del riesgo mediante un proxy (variación interanual de matrícula TASA_1 y un umbral de caída) permitió formular el problema como una clasificación binaria (RIESGO_ALTO: 1 / RIESGO_BAJO: 0) en un contexto donde no existen trayectorias individuales. Se concluye que este enfoque es útil para una alerta temprana agregada, siempre que su interpretación se centre en tendencias estructurales por segmento/cohorte y no en causalidad individual, debido al carácter agregado de la fuente y a la presencia de observaciones no etiquetadas (NaN) asociadas a falta de continuidad temporal.

Las variables con mayor capacidad predictiva se relacionan principalmente con la dinámica histórica de la matrícula del segmento/cohorte. En particular, indicadores como TOTAL_L1, LOG_TOTAL y Z_3, seguidos por medidas de estabilidad como STD_3 y participación relativa SHARE_IES, concentraron el mayor aporte para discriminar el riesgo alto. En cambio, variables categóricas del segmento (institución, carrera, modalidad, nivel, sexo y territorio) tuvieron un peso menor y se interpretan más como contexto. En síntesis, los resultados muestran que el riesgo agregado está explicado, sobre todo, por señales de volatilidad, caídas e inestabilidad temporal de la matrícula.

Se aplicaron y compararon modelos tipo ensemble, específicamente Random Forest, Gradient Boosting y AdaBoost, manteniendo el mismo pipeline y un esquema de evaluación con enfoque temporal (entrenamiento 2016–2020, validación 2021–2022 y prueba 2023). Los resultados evidencian que Random Forest fue el modelo con mejor desempeño global en Test 2023, destacando en métricas críticas para clasificación desbalanceada (F1-score, recall y AUC), seguido por Gradient Boosting, mientras que AdaBoost mostró menor sensibilidad para detectar la clase minoritaria.

El desbalance de clases tuvo un impacto directo en la detección del riesgo alto. Aunque métricas globales como accuracy pueden verse elevadas, el criterio institucional de utilidad exige priorizar recall y F1-score, ya que el error más crítico es el falso negativo (FN): segmentos realmente en riesgo que no son priorizados. La comparación antes/después confirma que el uso de técnicas como SMOTE aplicado solo en entrenamiento, junto con validación temporal para evitar fuga de información, mejora la sensibilidad hacia la clase minoritaria (principalmente recall y F1), siendo el efecto más evidente en modelos con menor desempeño inicial (AdaBoost).

El análisis con matrices de confusión permitió identificar el tipo de errores y su implicación práctica: los falsos positivos (FP) representan alertas adicionales que pueden demandar más revisión operativa, mientras que los falsos negativos (FN) implican omitir segmentos potencialmente críticos para intervención preventiva. En consecuencia, el modelo seleccionado (Random Forest) no solo muestra alto desempeño estadístico, sino también coherencia con la lógica de una alerta temprana, al mantener bajo nivel de FN y alta capacidad discriminativa (AUC cercano a 1) en un escenario de validación temporal.

Finalmente, se concluye que el enfoque propuesto es aplicable como herramienta de apoyo a decisiones institucionales, permitiendo generar un ranking de segmentos/cohortes priorizados (Top N) según probabilidad estimada de riesgo alto en 2023. Este resultado facilita la focalización de recursos de monitoreo, análisis académico-programático y acciones preventivas a nivel agregado, con la recomendación de que su uso sea complementario a análisis institucionales y no sustituya la interpretación contextual de cada segmento.

5.2. Limitaciones

El presente estudio se desarrolló a partir de datos abiertos agregados de matrícula del SNIESE/SENESCYT (2015–2023), por lo que su principal limitación radica en que la unidad de análisis no corresponde al estudiante individual, sino al segmento/cohorte definido operativamente por combinaciones de variables institucionales, académicas, demográficas y territoriales. En este sentido, los resultados no deben interpretarse como predicciones individualizadas de abandono, sino como señales agregadas para monitoreo y planificación. Esta condición introduce el riesgo de falacia ecológica, es decir, inferir comportamientos individuales desde patrones observados a nivel grupal, lo cual puede llevar a conclusiones inapropiadas si se extrapola el uso del modelo más allá del alcance propuesto.

Otra limitación importante se relaciona con la definición de la variable objetivo. Debido a la ausencia de una etiqueta directa de deserción confirmada, el riesgo se operacionalizó mediante un proxy basado en caídas interanuales relevantes de matrícula (TASA_1 y umbral fijo). Aunque esta aproximación es útil para fines predictivos agregados, puede incorporar ruido en la etiqueta, ya que una disminución de matrícula no necesariamente representa abandono definitivo: también puede reflejar movilidad académica, pausas temporales, cambios administrativos, reestructuración de oferta, apertura/cierre de cohortes o variaciones contextuales externas. Por esta razón, el modelo predice riesgo respecto al proxy definido y no la deserción en sentido estricto.

Adicionalmente, el conjunto presenta discontinuidades temporales a nivel de segmentos, lo cual se refleja en una proporción relevante de observaciones sin etiqueta (NaN) al no existir continuidad anual suficiente para calcular la variación interanual. Si bien se aplicaron reglas de continuidad y criterios para evitar interpretaciones sesgadas, esta situación reduce la cobertura efectiva del modelado y puede limitar la representación de ciertos segmentos con trayectoria incompleta o intermitente, afectando la generalización dentro de subgrupos específicos.

Desde el punto de vista metodológico, existe siempre el riesgo de fuga de información (data leakage) cuando se trabaja con variables temporales. En este estudio se controló mediante el uso de rezagos, ventanas móviles calculadas solo con años previos y un esquema de validación temporal; sin embargo, la amenaza no desaparece por completo si alguna variable

derivada llegara a capturar información demasiado cercana al evento que se pretende anticipar. Por ello, la consistencia temporal debe mantenerse como un criterio obligatorio en futuras actualizaciones del modelo.

Finalmente, el problema de clasificación presenta desbalance de clases (mayoría de riesgo bajo), lo que puede sesgar el aprendizaje hacia la clase mayoritaria e inflar métricas globales como accuracy. Aunque se priorizó el uso de métricas sensibles (recall y F1) y se aplicó SMOTE únicamente en entrenamiento, este tipo de técnicas puede introducir riesgos como sobreajuste o alterar la calibración de probabilidades, especialmente si se pretende interpretar el score como probabilidad "real". En consecuencia, el uso institucional debe apoyarse siempre en validación sobre datos no vistos y en seguimiento de desempeño por períodos.

5.3. Recomendaciones

A partir de los resultados obtenidos, se recomienda que la institución utilice el modelo como una herramienta de alerta temprana agregada, enfocada en la priorización de segmentos/cohortes con mayor probabilidad de caída interanual de matrícula, y no como un mecanismo de clasificación individual. Para su adopción práctica, es conveniente transformar la salida del modelo (score/probabilidad) en indicadores operativos de monitoreo, tales como el porcentaje de segmentos en riesgo alto por programa o cohorte, tendencias temporales del riesgo por año y rankings Top N que permitan identificar puntos críticos que requieren seguimiento académico-programático.

Se recomienda también definir umbrales de alerta alineados con la lógica de decisión institucional. Si el objetivo estratégico es preventivo y el costo de omitir casos es alto, conviene priorizar configuraciones con mayor recall (aceptando algunas alertas adicionales). En cambio, si los recursos para intervención son limitados, se sugiere ajustar el umbral para mejorar precisión y complementar la priorización con criterios operativos, tales como capacidad instalada, criticidad académica del programa, ventanas de atención y disponibilidad de equipos de acompañamiento. En este marco, la matriz de confusión debe seguir siendo un insumo clave para discutir el equilibrio entre falsos positivos y falsos negativos, en función del impacto real de cada tipo de error.

En relación con el desbalance de clases, se recomienda mantener el tratamiento mediante técnicas aplicadas exclusivamente sobre el conjunto de entrenamiento, evitando fuga de información. Además de SMOTE, pueden evaluarse alternativas complementarias como ponderación de clases (por ejemplo, class_weight='balanced') y ajuste de umbral posterior al entrenamiento, comparando siempre los cambios con métricas sensibles a la clase minoritaria (recall y F1), ya que son las más coherentes con un enfoque de alerta temprana. Del mismo modo, se recomienda reportar el desempeño por año y por subgrupos relevantes (IES, provincia, modalidad o campo), ya que el desempeño global puede ocultar diferencias importantes en segmentos específicos.

Para fortalecer la validez externa, se recomienda implementar un esquema de monitoreo de drift o cambios estructurales, considerando que políticas educativas, condiciones económicas o reestructuración de oferta académica pueden modificar los patrones de matrícula en años futuros. Por ello, es recomendable recalibrar o reentrenar el modelo de forma periódica cuando se incorpore nueva información, y mantener pruebas temporales con periodos no vistos. Esta práctica permitiría preservar la estabilidad del sistema y garantizar que las predicciones se mantengan útiles en escenarios cambiantes.

Finalmente, se recomienda ampliar progresivamente el alcance del modelo incorporando, cuando sea posible, variables adicionales que representen mejor el fenómeno de deserción, idealmente integrando datos institucionales internos (por ejemplo, rendimiento académico agregado, reprobación por asignaturas críticas, tasas de retención por periodo, o indicadores administrativos) siempre bajo criterios éticos y de anonimización. Esto permitiría reducir la dependencia de un proxy basado únicamente en matrícula y mejorar la interpretación del riesgo. En paralelo, se recomienda que el uso del modelo esté integrado a rutas de acción claras (responsables, protocolos y seguimiento) y que se mida su impacto en términos de cobertura de intervención, evolución de indicadores de permanencia y mejora de decisiones basadas en evidencia.

.

GLOSARIO

Deserción universitaria: Abandono o interrupción de los estudios por parte del estudiante dentro del sistema de educación superior, asociado a factores académicos, personales, demográficos, institucionales o contextuales.

Deserción inicial o precoz: Tipo de deserción que ocurre cuando el estudiante ha sido admitido en una institución de educación superior, pero no llega a concretar su matrícula ni a iniciar formalmente el primer período académico.

Deserción temprana: Abandono que se produce durante los primeros semestres de la carrera, cuando el estudiante aún se encuentra en etapa de adaptación al entorno universitario.

Deserción tardía: Interrupción de los estudios universitarios que ocurre luego de haber superado una parte significativa del plan curricular, generalmente en niveles intermedios o finales.

Permanencia estudiantil: Continuidad del estudiante dentro del sistema de educación superior hasta la finalización del programa académico, influida por condiciones académicas, institucionales y contextuales.

Factores académicos: Elementos vinculados al proceso formativo (exigencia curricular, evaluación, desempeño, avance académico) que pueden incidir en la continuidad o el abandono.

Factores demográficos: Características del perfil estudiantil y territorial (p. ej., sexo, etnia, discapacidad, residencia) que pueden asociarse con diferencias en permanencia o abandono.

Factores institucionales: Condiciones propias de la institución que influyen en la permanencia (políticas académicas, calidad docente, infraestructura, servicios de apoyo, acompañamiento y gestión).

Datos abiertos: Conjuntos de datos publicados para acceso libre, reutilización y verificación, en formatos que facilitan su uso analítico y su integración en estudios o sistemas de gestión.

Datos abiertos educativos: Información pública del sistema educativo (p. ej., matrícula, oferta académica, características institucionales y territoriales) que permite análisis descriptivo y predictivo sin requerir identificadores personales.

Calidad de los datos: Grado en que los datos son completos, consistentes, válidos y confiables para análisis, asegurando resultados reproducibles y comparables.

Estructura de los datos: Forma de organización y presentación de los datos (campos, formatos, codificación y organización tabular) que facilita su almacenamiento, procesamiento e interpretación.

Ciclo de vida de los datos: Etapas por las que atraviesan los datos desde su generación y publicación hasta su uso, actualización, resguardo y archivado.

Anonimización de datos: Proceso que asegura la protección de identidades eliminando o evitando el uso de identificadores personales; en este estudio se trabaja con información agregada y no individual.

Analítica educativa (Learning Analytics): Uso de métodos analíticos y computacionales para recopilar, analizar e interpretar datos educativos con el fin de apoyar decisiones y mejorar procesos institucionales.

Ciencia de datos: Disciplina que integra estadística, programación y aprendizaje automático para extraer conocimiento a partir de datos y generar modelos útiles para análisis y predicción.

Analítica predictiva: Enfoque analítico que utiliza datos históricos para estimar eventos futuros mediante modelos estadísticos o de aprendizaje automático.

Sistema de alerta temprana (agregada): Herramienta analítica orientada a identificar unidades agregadas (segmentos/cohortes) con mayor probabilidad de riesgo, con base en patrones históricos, para apoyar priorización de acciones preventivas.

Aprendizaje automático (Machine Learning): Rama de la inteligencia artificial que permite a los sistemas aprender patrones a partir de datos y realizar clasificaciones o predicciones.

Modelo predictivo: Representación computacional que estima la probabilidad de ocurrencia de un evento (p. ej., riesgo alto) a partir de variables observadas.

Modelo ensemble: Enfoque que combina varios modelos (generalmente árboles o aprendices débiles) para mejorar precisión, estabilidad y capacidad de generalización.

4

Random Forest: Modelo ensemble tipo bagging que combina múltiples árboles de decisión entrenados con muestras y variables aleatorias, generando predicciones robustas.

Gradient Boosting: Modelo ensemble tipo boosting que construye árboles de forma secuencial, corrigiendo errores del modelo anterior y optimizando una función de pérdida.

AdaBoost: Algoritmo de boosting que ajusta iterativamente los pesos de las observaciones, priorizando los casos mal clasificados para mejorar el desempeño.

Desbalance de clases: Situación en la que una clase aparece con mucha mayor frecuencia que otra (p. ej., riesgo bajo vs. riesgo alto), lo que puede sesgar el aprendizaje del modelo y afectar la detección de la clase minoritaria.

REFERENCIAS

Alpaydin, E. (2020). *Introduction to Machine Learning, fourth edition*. MIT Press.

Alvesson, M., & Sandberg, J. (2011). *GENERATING RESEARCH QUESTIONS THROUGH PROBLEMATIZATION*.

Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: A literature review. *Review of Education*, *8*(2), 614-652. https://doi.org/10.1002/rev3.3202

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, *13*(null), 281-305.

Bertsimas, D., & Kallus, N. (2018). *From Predictive to Prescriptive Analytics* (arXiv:1402.5481). arXiv. https://doi.org/10.48550/arXiv.1402.5481

Bouihi, B., Bousselham, A., Aoula, E., Ennibras, F., & Deraoui, A. (2024). Prediction of Higher Education Student Dropout based on Regularized Regression Models. *Engineering, Technology & Applied Science Research*, *14*(6), 17811-17815. https://doi.org/10.48084/etasr.8644

Brownlee, J. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery.

Castrillón-Gómez, O. D., Sarache, W., Ruiz-Herrera, S., Castrillón-Gómez, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Prediction of main variables that lead to student dropout by using data mining techniques. *Formación universitaria*, *13*(6), 217-228. https://doi.org/10.4067/S0718-50062020000600217

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and Conducting Mixed Methods Research*. SAGE Publications.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. En C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning: Methods and Applications* (pp. 157-175). Springer. https://doi.org/10.1007/978-1-4419-9326-7_5

D'Agostino, A. (2024, septiembre 24). Feature Engineering Techniques for Numerical Variables in Python. *Towards Data Science*. https://towardsdatascience.com/feature-engineering-techniques-for-numerical-variables-in-python-4bd42e8bded7/

Data.gov. (2021). *Data.gov Home*. Data.Gov. https://data.gov/

Datos Abiertos. (2019). *Bienvenida—Datos Abiertos Ecuador*. https://www.datosabiertos.gob.ec/

Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics: The New Science of Winning*. Harvard Business Press.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, *5*, 15991-16005. https://doi.org/10.1109/ACCESS.2017.2654247

Fawcett, T., & Provost, F. (2013). *Data Science for Business*.

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863-905. https://doi.org/10.1613/jair.1.11192

Fischer, C., Hirsbrunner, S. D., & Teckentrup, V. (2022). Producing Open Data. *Research Ideas and Outcomes*, *8*, e86384. https://doi.org/10.3897/rio.8.e86384

Fountain-Jones, N. M., Machado, G., Carver, S., Packer, C., Recamonde-Mendoza, M., & Craft, M. E. (2019). *How to make more from exposure data? An integrated machine learning*

pipeline to predict pathogen exposure (p. 569012). bioRxiv. https://doi.org/10.1101/569012

Fu, Y., & Weng, Z. (2024). Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices. *Computers and Education: Artificial Intelligence*, *7*, 100306. https://doi.org/10.1016/j.caeai.2024.100306

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Gonzalez, N. A. del P., & Chiappe, A. (2024). Learning analytics and personalization of learning: A review. *Ensaio: Avaliação e Políticas Públicas Em Educação*, *32*(122). https://www.redalyc.org/journal/3995/399577691013/html/

Granda, L. E. Á., Yautibug, F. C., & Copa, R. A. (2024). Deserción en la Educación Superior en Ecuador, Causas y Consecuencias. *Ciencia Latina Revista Científica Multidisciplinar*, *8*(3), 11475-11490. https://doi.org/10.37811/cl_rcm.v8i3.12472

Guan, X., Feng, X., & Islam, A. Y. M. A. (2023). The dilemma and countermeasures of educational data ethics in the age of intelligence. *Humanities and Social Sciences Communications*, *10*(1), 138. https://doi.org/10.1057/s41599-023-01633-x

Gutiérrez, R. C., Rivas, H. P. N., & López, E. L. (2024). Reflexiones teóricas del fenómeno de la deserción académica en la educación superior. *Revie - Revista de Investigación y Evaluación Educativa*, *11*(2), 88-109. https://doi.org/10.47554/revie.vol11.num2.2024.pp88-109

Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G., & Gomez-Nieto, E. (2023). Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and

Survival Analysis Methods in the Latin American Context. *Education Sciences*, *13*(2), 154. https://doi.org/10.3390/educsci13020154

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

Hernández Sampieri, R., & Fernandez-Collado, C. F. (2014). *Metodología de la investigación* (P. Baptista Lucio, Ed.; Sexta edición). McGraw-Hill Education.

Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). *Automated Machine Learning: Methods, Systems, Challenges*. Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5

Jadama, A., & Toray, M. (2024). *Ensemble Learning: Methods, Techniques, Application*. https://doi.org/10.13140/RG.2.2.28017.08802

Jagačić, T., Kadoić, N., & Gusić Munđar, J. (2024). Students' Perceptions of the Ethical Aspects of Learning Analytics. *TEM Journal*, *13*, 3526-3535. https://doi.org/10.18421/TEM134-85

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer Nature.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258-268. https://doi.org/10.1080/10580530.2012.716740

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232. https://doi.org/10.1007/s13748-016-0094-0

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

Kunigami, A., & Palomino, N. (2019, marzo 4). Datos abiertos: Conceptos básicos y temas claves. *Abierto al público*. https://blogs.iadb.org/conocimiento-abierto/es/datos-abiertos/

Lee, L. E., Martínez, S. I., Rocha, J. A. C., Villanueva, J. D. T., Menchaca, J. L., Berrones, M. G. T., & Rocha, E. C. (2020). Evaluation of Prediction Algorithms in the Student Dropout Problem. *Journal of Computer and Communications*, *8*(3), 20-27. https://doi.org/10.4236/jcc.2020.83002

Lee, R. S. T. (2020). Data Mining. En R. S. T. Lee (Ed.), *Artificial Intelligence in Daily Life* (pp. 71-118). Springer. https://doi.org/10.1007/978-981-15-7695-9_4

López-Pernas, S., Saqr, M., Conde, J., & Del-Río-Carazo, L. (2024). A Broad Collection of Datasets for Educational Research Training and Application. En M. Saqr & S. López-Pernas (Eds.), *Learning Analytics Methods and Tutorials: A Practical Guide Using R* (pp. 17-66). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54464-4_2

Matute, J. M. C., Franco, A. V., & Segarra, J. I. T. (2023). Factores que inciden en la deserción estudiantil en la unidad académica de Ciencias Sociales de la Universidad Católica de Cuenca. *ConcienciaDigital*, *6*(3), 30-48. https://doi.org/10.33262/concienciadigital.v6i3.2621

Mduma, N., Kalegele, K., & Machuve, D. (2019). A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. *Data Science Journal*, *18*(1). https://doi.org/10.5334/dsj-2019-014

Misiejuk, K., Samuelsen, J., Kaliisa, R., & Prinsloo, P. (2025). Idiographic learning analytics: Mapping of the ethical issues. *Learning and Individual Differences*, *117*, 102599. https://doi.org/10.1016/j.lindif.2024.102599

Moreira, C. R. F., & Caicedo, R. A. A. (2024). Factores que inciden en la deserción estudiantil: Caso Instituto Superior Tecnológico Luis Tello. *Ciencia Latina Revista Científica Multidisciplinar*, *8*(1), 10518-10533. https://doi.org/10.37811/cl_rcm.v8i1.10357

Nabil, A., Seyam, M., & Elfetouh, A. (2022). Predicting students' academic performance using machine learning techniques: A literature review. *International Journal of Business Intelligence and Data Mining*, *20*, 456. https://doi.org/10.1504/IJBIDM.2022.123214

Olive, U., Bosco, M. J., & Enan, N. M. (2025). Predicting Student Dropout in Higher Education: An Ensemble Learning Approach with Feature Importance Analysis. *Journal of Information and Technology*, *5*(4), 31-40. https://doi.org/10.70619/vol5iss4pp31-40

Open Data Charter. (2015). *Open Data Charter*. https://opendatacharter.org/?utm_source=chatgpt.com

Powers, D. M. W. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation* (arXiv:2010.16061). arXiv. https://doi.org/10.48550/arXiv.2010.16061

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. *ACM Comput. Surv.*, *53*(3), 57:1-57:34. https://doi.org/10.1145/3388792

Probst, P., Bischl, B., & Boulesteix, A.-L. (2018). *Tunability: Importance of Hyperparameters of Machine Learning Algorithms*. https://doi.org/10.48550/arXiv.1802.09596

Quiñónez, P. L. A., Quiñónez, E. D. R., Jama, L. E. N., Quiñónez, D. D. P., & Cortez, P. J. M. (2025). Factores Multidimensionales que Influyen en la Deserción Estudiantil en Universidades Públicas del Ecuador. *Ciencia Latina Revista Científica Multidisciplinar*, *9*(2), 7292-7306. https://doi.org/10.37811/cl_rcm.v9i2.17447

Rabelo, A. M., & Zárate, L. E. (2025). A model for predicting dropout of higher education students. *Data Science and Management*, *8*(1), 72-85. https://doi.org/10.1016/j.dsm.2024.07.001

Rebelo Marcolino, M., Reis Porto, T., Thompsen Primo, T., Targino, R., Ramos, V., Marques Queiroga, E., Munoz, R., & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: Insights from moodle log data. *Scientific Reports*, *15*(1), 9840. https://doi.org/10.1038/s41598-025-93918-1

Saeteros, Z. D. (2024). Análisis de la deserción estudiantil y estrategias para incrementar la retención en instituciones de educación superiorAnalysis of student dropout and strategies to increase retention in higher education institutions. *Revista de investigación, formación y desarrollo: Generando productividad institucional*, *12*(1), 88-95. https://doi.org/10.34070/rif.v12.i1.WVVF7996

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Segura, M., Mello, J., & Hernández, A. (2022). Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role? *Mathematics*, *10*(18), 3359. https://doi.org/10.3390/math10183359

SENESCYT. (2010). *SENESCYT - Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación*. https://www.senescyt.gob.ec/consulta-titulos-web/faces/vista/consulta/inicio.xhtml

Seol, D., Choi, J., Kim, C., & Hong, S. (2023). Alleviating Class-Imbalance Data of Semiconductor Equipment Anomaly Detection Study. *Electronics*, *12*, 585. https://doi.org/10.3390/electronics12030585

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Jr, K. C. L. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons.

Siemens, G. (2013). Learning Analytics: The Emergence of a Discipline. *American Behavioral Scientist*, *57*(10), 1380-1400. https://doi.org/10.1177/0002764213498851

Sifuentes, M. C. S. G. C., Pérez, L. G. V., Cantabrana, M. G. N., Acosta, I. I. F. O., Santana, F. A. Á., & Fierro, M. de los Á. S. (2023). Modelo Predictivo de la Deserción Escolar en Educación Superior: Una Aproximación desde la Minería de Datos Utilizando la Metodología CRISP-DM. *Ciencia Latina Revista Científica Multidisciplinar*, *7*(5), 7797-7812. https://doi.org/10.37811/cl_rcm.v7i5.8363

SNIESE. (2010). *Portal Ciudadano SNIESE*. https://infoeducacionsuperior.gob.ec/#/

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2020). Implementing AutoML in Educational Data Mining for Prediction Tasks. *Applied Sciences*, *10*(1), 90. https://doi.org/10.3390/app10010090

Van, R., Alvarez, D., Mize, T., Gannavarapu, S., Chintham Reddy, L., Nasoz, F., & Han, M. V. (2024). A comparison of RNA-Seq data preprocessing pipelines for transcriptomic predictions across independent studies. *BMC Bioinformatics*, *25*(1), 181. https://doi.org/10.1186/s12859-024-05801-x

W3C. (2017, enero 31). *Data on the Web Best Practices*. https://www.w3.org/TR/dwbp/

World Bank. (2020). *World Bank Open Data*. World Bank Open Data. https://data.worldbank.org

Wu, Y., Liu, X., Feng, Y., Wang, Z., Yan, R., & Zhao, D. (2019). *Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs*. 5278-5284.

Zerpa, C., & Rodríguez-Montoya, C. (2024). Disminución de la matrícula universitaria por deserción: Una revisión estructurada. *Ciencia y Educación*, *8*(1), 59-78. https://doi.org/10.22206/cined.2024.v8i1.pp59-78

ANEXOS

Anexo A. Documentación del conjunto de datos (SNIESE/SENESCYT)

Se presenta la documentación técnica del conjunto de datos utilizado en el estudio. Incluye el diccionario de variables originales del SNIESE/SENESCYT (2015–2023), el diccionario de variables derivadas (features) generadas a nivel segmento–año y los metadatos generales del dataset, con el fin de asegurar trazabilidad y reproducibilidad del análisis.

Variables derivadas (features) a nivel segmento–año

Tabla 1. Diccionario de variables derivadas (features)

Feature	Definición	Fórmula/Regla	Tipo
SEGMENTO_ID	Identificador único del segmento (combinación de variables del segmento)	Construido a partir de variables del segmento (concatenación/codificación)	Categórica
ANIO_COHORTE	Primer año en que aparece el segmento	min(AÑO) por SEGMENTO_ID	Numérica
TOTAL_SEG	Matrícula anual consolidada por segmento	Total por SEGMENTO_ID y AÑO	Numérica
TOTAL_L1	Rezago 1 de matrícula	TOTAL_SEG(t−1)	Numérica
TOTAL_L2	Rezago 2 de matrícula	TOTAL_SEG(t−2)	Numérica
TOTAL_L3	Rezago 3 de matrícula	TOTAL_SEG(t−3)	Numérica
DIF_1	Variación absoluta interanual	TOTAL_SEG(t) − TOTAL_SEG(t−1)	Numérica
TASA_1	Variación relativa interanual	(TOTAL_SEG(t) − TOTAL_SEG(t−1)) / TOTAL_SEG(t−1)	Numérica
MEDIA_3	Media móvil histórica (3 años previos)	mean(TOTAL_SEG(t−1), t−2, t−3)	Numérica
STD_3	Desviación estándar móvil (3 años previos)	std(TOTAL_SEG(t−1), t−2, t−3)	Numérica
Z_3	Z-score histórico	(TOTAL_L1 − MEDIA_3) / STD_3 (si STD_3>0)	Numérica
LOG_TOTAL	Transformación logarítmica de escala	log(1 + TOTAL_SEG)	Numérica
FLAG_SALTO	Indicador de discontinuidad anual	1 si hay salto de años; 0 si continuidad	Binaria
SHARE_IES	Participación del segmento dentro de su IES (opcional)	TOTAL_SEG / total_matrícula_IES_en_año	Numérica

Nota. Fuente: Las variables temporales se calculan solo con información previa para evitar fuga de información.
Elaboración propia Romero Cristhian 2026

Metadatos del conjunto de datos

Tabla 2. *Metadatos de datos*

Elemento	Descripción
Fuente institucional	SNIESE – SENESCYT
Cobertura	Universidades públicas del Ecuador
Período	2015–2023
Unidad de análisis	Segmento/cohorte (pseudo-cohorte)
Nivel de agregación	Registros agregados (conteos)
Variable principal	TOTAL (conteo de matrícula)

Nota. Fuente: Elaboración propia con base en SNIESE/SENESCYT.
Elaboración propia Romero Cristhian 2026

Anexo B. Segmentación y unidad de análisis

Se describe la construcción de la unidad de análisis agregada (segmento/cohorte) utilizada en la investigación. Se detallan las variables que conforman el segmento, la regla operativa para la generación del SEGMENTO_ID y los controles aplicados para verificar la unicidad del panel segmento–año.

Variables utilizadas para definir el segmento (pseudo-cohorte)

Tabla 3. Variables utilizadas/ segmento (pseudo-cohorte)

Dimensión	Variables consideradas
Institucional	NOMBRE_IES
Académico-programática	NOMBRE_CARRERA, NIVEL_FORMACION, MODALIDAD
Demográfica	SEXO
Territorial	PROVINCIA_RESIDENCIA

Nota. Fuente: El segmento se define como combinación de variables observables en SNIESE.
Elaboración propia Romero Cristhian 2026

Regla operativa para construcción de SEGMENTO_ID (ejemplo)

Tabla 4. *Regla operativa para construcción de SEGMENTO_ID*

Elemento	Ejemplo
Variables del segmento	NOMBRE_IES + NOMBRE_CARRERA + NIVEL_FORMACION + MODALIDAD + SEXO + PROVINCIA_RESIDENCIA

Ejemplo de concatenación	"EPN|ING_AGROIND|TERCER_NIVEL|PRESENCIAL|MUJER|PICHINCHA"
SEGMENTO_ID	Hash/ID numérico o código resultante de la concatenación

Nota. Fuente: Usar una regla fija y reproducible (concatenación + hash o codificación).
Elaboración propia Romero Cristhian 2026

Control de unicidad del panel segmento–año

Tabla 5. Control del panel segmento–año

Control	Resultado esperado
Duplicados SEGMENTO_ID–AÑO	0
Segmentos únicos	76.077
Período en df_seg	2015–2023

Nota. Fuente: Elaboración propia Romero Cristhian 2026

Anexo C. Variable objetivo (RIESGO_ALTO) y reglas de etiquetado

Se documenta la definición de la variable objetivo RIESGO_ALTO, construida como proxy de riesgo a partir de variaciones interanuales de matrícula. Se incluyen las fórmulas utilizadas, la regla de continuidad temporal (FLAG_SALTO), el criterio principal basado en umbral percentil y una alternativa por umbral fijo como análisis de sensibilidad.

Fórmulas y condición de continuidad

Tabla 6. Fórmulas y condición de continuidad

Elemento	Definición
TOTAL_SEG	Conteo anual de matrícula del segmento
TOTAL_L1	TOTAL_SEG(t−1)
DIF_1	TOTAL_SEG(t) − TOTAL_SEG(t−1)
TASA_1	(TOTAL_SEG(t) − TOTAL_SEG(t−1)) / TOTAL_SEG(t−1)
FLAG_SALTO	1 si el segmento no tiene continuidad anual; 0 si es continuo
Regla de etiquetado	Etiquetar solo cuando FLAG_SALTO=0 y TASA_1 es calculable

Nota. Fuente: Elaboración propia Romero Cristhian 2026

Umbral principal utilizado para modelado (basado en percentil)

Tabla 7. *Umbral principal(basado en percentil)*

Criterio	Valor

Umbral (TASA_1)	≈ −0,111
RIESGO_ALTO = 1	Si TASA_1 ≤ umbral
RIESGO_ALTO = 0	Si TASA_1 > umbral
Sin etiqueta (NaN)	Si no hay continuidad / no existe TASA_1
Distribución (conjunto completo df_model):	
Clase	Conteo
---	---:
0 (bajo riesgo)	146.892
1 (alto riesgo)	39.921
NaN (sin etiqueta)	82.659

Nota. Fuente: a proporción NaN responde a primer año del segmento y discontinuidades (sin año previo).
Elaboración propia Romero Cristhian 2026

Análisis de sensibilidad (umbral fijo)

Tabla 8. Análisis (umbral fijo)

Criterio alternativo	Regla
Umbral fijo	TASA_1 ≤ −0,20 (caída ≥ 20%)
Uso	Validación de estabilidad de la etiqueta (sensibilidad)
Reporte mínimo	Conteos por clase y comparación con el umbral principal

Nota. Fuente: La reducción principal ocurre al consolidar la unidad de análisis a segmento–año.
Elaboración propia Romero Cristhian 2026.

Anexo D. Preprocesamiento, trazabilidad y calidad

Se reporta los resultados técnicos del preprocesamiento aplicado al conjunto de datos. Se presenta la trazabilidad por etapas (df_raw, df_clean, df_seg, df_feat y df_model), los principales controles de calidad (completitud, integridad y validez) y la proporción de valores faltantes generados por la ingeniería temporal (lags y ventanas móviles).

Trazabilidad por etapas (filas/columnas)

Tabla 9. Trazabilidad (filas/columnas)

Etapa	Dataset	Filas	Columnas	% filas vs. df_raw
Original	df_raw	977.964	19	100%
Limpieza	df_clean	963.302	19	98,50%
Segmentación	df_seg	269.472	11	27,55%
Features	df_feat	269.472	26	27,55%
Modelado	df_model	269.472	27	27,55%

Nota. Fuente: La reducción principal ocurre al consolidar la unidad de análisis a segmento–año.
Elaboración propia Romero Cristhian 2026.

Controles de calidad del dataset limpio (df_clean)

Tabla 10. Controles dataset limpio (df_clean)

Control	Resultado
Rango de años	2015–2023
Duplicados exactos	0
Faltantes totales (variables evaluadas)	0
TOTAL inválido (negativo)	0 (tras depuración)

Nota. Fuente: Elaboración propia Romero Cristhian 2026.

Proporción de NaN en variables temporales

Tabla 11. *Proporción de NaN*

Variable	% NaN	Explicación
TOTAL_SEG	0,00%	Conteo consolidado
FLAG_SALTO	0,00%	Indicador de continuidad
TOTAL_L1	≈ 30,67%	Primer año del segmento o discontinuidad
DIF_1	≈ 30,67%	Requiere año previo
TASA_1	≈ 30,67%	Requiere año previo
MEDIA_3	≈ 48,62%	Requiere 3 años previos continuos
STD_3	≈ 48,62%	Requiere 3 años previos continuos

Nota. Fuente: Elaboración propia Romero Cristhian 2026.

Anexo E. Configuración experimental y validación temporal

Se describe la configuración experimental utilizada para entrenar y evaluar los modelos predictivos. Se reporta la partición temporal (train/valid/test), el pipeline de preprocesamiento y modelado, y los criterios de evaluación y selección del modelo final, priorizando métricas adecuadas para contextos con desbalance de clases.

Partición temporal (split)

Tabla 12. Partición temporal (split)

Conjunto	Años	Tamaño (registros)
Entrenamiento	2016–2020	99.918
Validación	2021–2022	56.981
Prueba	2023	29.914

Nota. Fuente: Se usa solo el subconjunto etiquetado (RIESGO_ALTO ≠ NaN). Elaboración propia Romero Cristhian 2026.

Pipeline de modelado (resumen operativo)

Tabla 13. *Pipeline de modelado*

Etapa	Técnica
Imputación numéricas	Mediana
Imputación categóricas	Moda (valor más frecuente)
Codificación categóricas	Ordinal (según pipeline reportado)
Balanceo (cuando aplica)	SMOTE solo en entrenamiento
Modelos evaluados	Random Forest, AdaBoost, Gradient Boosting
Métricas	Accuracy, Precision, Recall, F1, AUC

Fuente: Elaboración propia Romero Cristhian 2026.

Criterios de selección del modelo final

Tabla 14. Criterios de selección del modelo final

Criterio	Justificación
Recall (clase 1)	Prioriza detección de riesgo alto (evitar FN)
F1-score	Balance entre precision y recall en desbalance
AUC	Capacidad discriminativa global independiente del umbral
Desempeño en Test 2023	Evidencia de generalización temporal

Fuente: Elaboración propia Romero Cristhian 2026.

Anexo F. Hiperparámetros finales

Se presenta los hiperparámetros finales seleccionados para el modelo ganador (Random Forest) y, opcionalmente, para los modelos comparados. Su inclusión permite documentar la replicabilidad del experimento y la consistencia de la optimización realizada.

Hiperparámetros finales del modelo ganador (Random Forest)

Tabla 15. Hiperparámetros (Random Forest)

Parámetro	Valor final	Descripción
n_estimators	300	Número de árboles
criterion	gini	Criterio de división
max_depth	None	Profundidad máxima (sin límite)
max_features	sqrt	Nº de variables por división
min_samples_split	2	Mínimo para dividir un nodo
min_samples_leaf	1	Mínimo de muestras por hoja
bootstrap	True	Muestreo con reemplazo
class_weight	None	Pesos por clase
random_state	42	Semilla / reproducibilidad

n_jobs	-1	Uso de núcleos (todos)
max_leaf_nodes	None	Máx. nodos hoja (sin límite)
min_impurity_decrease	0.0	Umbral de impureza
min_weight_fraction_leaf	0.0	Fracción mínima ponderada por hoja
ccp_alpha	0.0	Poda por complejidad
oob_score	False	Validación out-of-bag
verbose	0	Nivel de salida
warm_start	False	Reutilizar árboles previos

Fuente: Elaboración propia Romero Cristhian 2026.

Anexo G. Resultados completos (matrices y métricas)

Se consolida la evidencia cuantitativa del desempeño de los modelos evaluados. Se incluyen matrices de confusión en validación y prueba, así como la tabla comparativa de métricas (accuracy, precision, recall, F1 y AUC), con énfasis en la detección de la clase minoritaria (riesgo alto).

Matrices de confusión (VALID y TEST 2023)

Tabla 16. Matrices de confusión

Modelo	Conjunto	TN	FP	FN	TP
Random Forest	Validación	45.001	261	536	11.183
Random Forest	Prueba 2023	22.329	139	364	7.082
AdaBoost	Validación	44.404	858	2.714	9.005
AdaBoost	Prueba 2023	21.886	582	1.922	5.524
Gradient Boosting	Validación	44.882	380	725	10.994
Gradient Boosting	Prueba 2023	22.238	230	582	6.864

Fuente: Elaboración propia Romero Cristhian 2026.

Métricas por modelo (VALID y TEST 2023)

Tabla 17. Métricas por modelo

Modelo	Conjunto	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	Validación	0,9860	0,9772	0,9543	0,9656	≈0,9990
Random Forest	Prueba 2023	0,9832	0,9808	0,9511	0,9657	≈0,9986
AdaBoost	Validación	0,9373	0,9130	0,7684	0,8345	≈0,9860
AdaBoost	Prueba 2023	0,9163	0,9047	0,7419	0,8152	≈0,9810
Gradient Boosting	Validación	0,9806	0,9666	0,9381	0,9522	≈0,9980
Gradient Boosting	Prueba 2023	0,9729	0,9676	0,9218	0,9442	≈0,9977

Nota Fuente: Accuracy/Precision/Recall/F1 se derivan de las matrices de confusión; AUC se toma de la curva ROC reportada. Elaboración propia Romero Cristhian 2026.

Anexo H. Interpretabilidad (importancia de variables)

Se presenta el análisis de interpretabilidad del modelo final, mediante ranking de importancia de variables. La evidencia permite identificar qué predictores explican en mayor medida la clasificación del riesgo alto a nivel agregado y aporta transparencia al uso del enfoque predictivo.

Ranking de variables más influyentes (Permutation Importance)

Tabla 18. *Ranking de variables más influyentes*

Rank	Variable	Interpretación (qué captura)
1	TOTAL_L1	Matrícula del año previo (memoria inmediata)
2	LOG_TOTAL	Escala del tamaño del segmento (suaviza extremos)
3	Z_3	Desviación respecto al comportamiento histórico reciente
4	STD_3	Variabilidad histórica (inestabilidad del segmento)
5	SHARE_IES	Peso del segmento dentro de la IES
6	MEDIA_3	Nivel promedio histórico reciente
7	TOTAL_L2	Memoria de 2 años previos
8	TOTAL_L3	Memoria de 3 años previos
9	MODALIDAD	Contexto académico-programático
10	NIVEL_FORMACION	Contexto del nivel de estudios

Nota Fuente: Si tienes los valores numéricos de importancia, puedes agregar una columna "Importancia media". Elaboración propia Romero Cristhian 2026.

Anexo I. Segmentos/cohortes priorizados (Ranking – Test 2023)

Se reporta el ranking Top 30 de segmentos/cohortes con mayor probabilidad estimada de riesgo alto en el conjunto de prueba (Test 2023). Este listado se interpreta como priorización agregada para fines de monitoreo y planificación institucional; no constituye inferencia individual.

Ranking Top 30 de segmentos con mayor probabilidad (Test 2023)

Tabla 19. *Ranking Top 30 de segmentos con mayor probabilidad*

SEGMENTO_ID	ID	NOMBRE_IES	NOMBRE_CARRERA	NIVEL_FORMACION	MODALIDAD	SEXO	PROVINCIA_RESIDENCIA	PROBA_MEDIA	PROBA_MAX	N_OBS
3591	10636	PONTIFICIA UNIVERSIDAD CATOLICA DEL ECUADOR	TERAPIA FISICA	TERCER NIVEL DE GRADO	PRESENCIAL	MUJER	IMBABURA	1.0	1.0	1

3609	10674	PONTIFICIA UNIVERSIDAD CATOLICA DEL ECUADOR	URBANISMO	CUARTO NIVEL O POSGRADO	EN LINEA	HOMBRE	CHIMBORAZO	1.0	1.0	1
14455	36489	UNIVERSIDAD ESTATAL DE MILAGRO	EDUCACION BASICA	CUARTO NIVEL O POSGRADO	EN LINEA	MUJER	GALAPAGOS	1.0	1.0	1
26566	66629	UNIVERSIDAD TECNICA PARTICULAR DE LOJA	CONTABILIDAD Y AUDITORIA	TERCER NIVEL DE GRADO	A DISTANCIA	HOMBRE	ORELLANA	1.0	1.0	1
14214	36021	UNIVERSIDAD ESTATAL DE MILAGRO	COMUNICACION	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	LOJA	1.0	1.0	1
14124	36019	UNIVERSIDAD ESTATAL DE MILAGRO	COMUNICACION	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	EL ORO	1.0	1.0	1
14741	37136	UNIVERSIDAD ESTATAL DE MILAGRO	INGENIERIA INDUSTRIAL	TERCER NIVEL DE GRADO	PRESENCIAL	MUJER	EL ORO	1.0	1.0	1
14738	37132	UNIVERSIDAD ESTATAL DE MILAGRO	INGENIERIA INDUSTRIAL	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	ZAMORA CHINCHIPE	1.0	1.0	1
14215	36022	UNIVERSIDAD ESTATAL DE MILAGRO	COMUNICACION	TERCER NIVEL DE GRADO	PRESENCIAL	HOMBRE	LOS RIOS	1.0	1.0	1
14906	37026	UNIVERSIDAD ESTATAL DE MILAGRO	GESTION EDUCATIVA	CUARTO NIVEL O POSGRADO	EN LINEA	MUJER	ORELLANA	1.0	1.0	1

Nota Fuente El ranking se interpreta como priorización agregada para planificación/monitoreo; no es inferencia individual.
Elaboración propia Romero Cristhian 2026.

Anexo J. Notebook de Jupyter (Visual Studio Code / Google Colab)

Notebook de Jupyter