UNIVERSIDAD PERUANA
CAYETANO HEREDIA

Facultad de
Ciencias e Ingeniería

Sistema predictivo basado en técnicas de machine learning para la detección temprana del riesgo académico y la deserción estudiantil en una universidad peruana

TESIS PARA OPTAR POR EL TÍTULO PROFESIONAL DE INGENIERO INFORMÁTICO

AUTOR(ES)
SEBASTIAN ANTONIO SALDAÑA RODRIGUEZ

ASESOR(ES)
ROY MARCO YALI SAMANIEGO

CO-ASESOR(ES)
MOISES STEVEND MEZA RODRIGUEZ

LIMA - PERÚ
2026

Jurado calificador

Presidente: Dra. Luz Aurora Carbajal Arroyo

Vocal: Dra. Mabel Karel Raza Garcia

Secretario: Mg. Ivan Omar Perez Masias

UNIVERSIDAD PERUANA
CAYETANO HEREDIA

DECLARACIÓN DE ORIGINALIDAD

Los egresados:

N°	APELLIDOS Y NOMBRES
1.	SALDAÑA RODRIGUEZ SEBASTIAN ANTONIO

Pertenecientes al programa de la **CARRERA DE INGENIERÍA INFORMÁTICA**, autores del trabajo titulado: **Sistema predictivo basado en técnicas de machine learning para la detección temprana del riesgo académico y la deserción estudiantil en una universidad peruana**, el cual ha sido elaborado, sustentado y aprobado, según corresponda, para optar por el **TÍTULO PROFESIONAL DE INGENIERO INFORMÁTICO** bajo la modalidad de **TESIS.**

En calidad de docentes asesores de la Universidad Peruana Cayetano Heredia:

N°	APELLIDOS Y NOMBRES DEL DOCENTE	FACULTAD	NIVEL DE ASESORÍA
1.	YALI SAMANIEGO ROY MARCO	FACI	ASESOR
2.	MEZA RODRIGUEZ MOISES STEVEND	FACI	COASESOR

Declaramos que el contenido del presente documento es original y que las citas y referencias a otros autores cumplen con las normas académicas establecidas. En ese sentido, hacemos constar que:

- El documento presenta un porcentaje de similitud de **4%**, según el reporte emitido por el software **Turnitin®** (identificador de entrega: **3612449735**; fecha de entrega: **16/07/2026**).
- Tras una revisión detallada del reporte y del contenido del trabajo en cuestión, no se han identificado indicios de plagio.
- Se certifica que el documento respeta los principios de integridad académica y cumple con los requisitos institucionales de originalidad.

Lugar y fecha: **Lima, 16 de julio de 2026**

Firma del asesor

N° DNI: 72916096

ORCID: 0000-0003-4542-3755

Firma del Co-asesor

N° DNI: 47470482

ORCID: 0000-0002-5806-9014

Tabla de contenido

Resumen 1

Abstract 2

I) Introducción 3

1.1. Contexto y problemática en la educación universitaria 3
1.2 Desafíos de retención en América Latina y el Perú 3
1.3 Marco normativo y brechas en la gestión institucional 4
1.4 Analítica de aprendizaje y modelos predictivos 5
1.5 Propuesta de investigación 5

II) Marco teórico 6

2.1.- Bases teóricas sobre la deserción estudiantil 6
2.1.1.- Definición operativa y dimensiones del fenómeno 6
2.1.2.- El Modelo de integración estudiantil de Vincent Tinto 9
2.2.- Bases teóricas sobre el aprendizaje automático. 9
2.2.1.- Aprendizaje supervisado. 9
2.2.2.- Algoritmos de aprendizaje 10
2.2.3.- El desafío del desbalance de clases y la selección de métricas 10
2.3.- Dashboards y protocolos de uso: la operacionalización de la analítica 11
2.3.1.- El dashboard como herramienta cognitiva 11
2.3.2.- Predicción a la prescripción: protocolos de intervención 12

III) Estado del arte 12

3.1.- Enfoques predictivos de deserción 12
3.2.- Sistemas de alerta y dashboards 13
3.3.- Brechas recurrentes y agenda de mejora. 14

IV) Hipótesis y objetivos 14

4.1.- Hipótesis general 14
4.2.- Objetivo general 14
4.3.- Objetivo específico 15

V) Materiales y métodos 16

5.1.- Metodología de investigación 16
5.2.- Material en estudio 21
5.3.- Herramientas y tecnología a utilizar 24
5.4.- Consideraciones éticas 24
5.4.1.- Privacidad de los datos 25
5.4.2.- Uso restringido y prototipo de prueba de concepto 25
5.4.3.- Confidencialidad de la tesis y aprobación institucional 26

5.4.4.- Publicación de resultados agregados ... 26
5.5.- Limitaciones metodológicas y desconexión de la responsabilidad de ejecución operativa. ... 27
5.5.1.- Limitaciones del estudio ... 27
5.5.2.- Descargo de responsabilidad de responsabilidad operativa. ... 29
5.6.- Diseño de investigación ... 29
5.6.1.- Despliegue y operación ... 32
5.6.2.- Diagrama de arquitectura del sistema MLOps ... 34
5.7.- Periodo, gobernanza y seguridad ... 36
5.7.1.- Periodo de estudio ... 36
5.7.2.- Responsables y gobernanza de datos ... 37
5.7.3.- Gobernanza y trazabilidad ... 37
5.7.4.- Seguridad y privacidad ... 37
5.8.- Población y muestra ... 37
5.9.- Fuente de datos, extracción y seguridad ... 40
5.9.1.- Orígenes de información ... 40
5.9.2.- Extracción y control de las versiones ... 41
5.9.3.- Tamaño y estructura del dataset ... 41
5.10.- Variables e indicadores ... 42
5.10.1.- Variable dependiente (Objetivo del modelo) ... 42
5.10.2.- Variables independientes (Predictores) ... 42
5.10.3.- Función de costo y prioridad de clasificación ... 43
5.11.- ETL y preparación de datos ... 45
5.11.1 Justificación del manejo de datos faltantes e imputación ... 47
5.12.- Análisis exploratorio (EDA) ... 47
5.13.- Modelado y validación ... 49
5.13.1.- Definición funcional del modelo predictivo ... 49
5.14.- Calibración y diagnóstico de probabilidad ... 51
5.15.- Métricas y reglas de decisión ... 52
5.16.- Análisis de robustez y sesgos ... 54
5.16.1.- Despliegue y operación. ... 54
5.16.2.- Observabilidad, MLOps y mantenimiento. ... 55
5.16.3.- Plan de Pruebas y Aseguramiento de Calidad ... 56
5.17.- Ética y Uso Responsable de los Datos ... 57
5.17.1.- Finalidad y alcance ... 57
5.17.2.- Principios rectores ... 57
5.17.3.- Gobernanza y tipos de responsabilidades ... 58
VI) Resultados ... 59
6.1.- Funcionamiento de extremo a extremo: del pdf al dataset del modelo ... 59

6.2.- Contexto externo: el desafío del ocr en documentos académicos 61
6.3.- El conjunto de datos actual y por qué es suficiente para la evaluación. 62
6.4.- Diferentes tipos de casos: análisis exploratorio de la población. 65
6.5.- Salud de las señales académicas 66
6.6.- Diagnóstico de entrenamiento del modelo 67
6.7.- Resultados operativos en validación 2024-1 69
6.7.1.- Validación de la hipótesis general (HG) 70
6.7.2.- Contexto externo: benchmarking de resultados 72
6.7.3.- Métricas operativas (el cupo de intervención) 72
6.7.4.- Calibración y confianza 74
6.7.5.- Interpretabilidad 78
6.8.- Discusión de resultados y hallazgos operativos 82
VII) Discusión 84
7.1.- Relevancia de los resultados y validación de la hipótesis 85
7.2.- Discusión sobre los predictores y su relación con la literatura 85
7.3.- Identificación de patrones no lineales y su contribución metodológica 86
7.4.- Discusión del gap operativo y su vínculo con la operacionalización 86
7.5.- Valor metodológico del pipeline y su coincidencia con estudios previos 87
7.6.- Aportes prácticos del prototipo y su relación con la gobernanza institucional 87
7.7.- Contribución general al campo y al contexto peruano 88
VIII) Conclusión 88
IX) Referencias bibliográficas 90
ANEXOS 96

Resumen

En Perú, la tasa de deserción universitaria suele estar entre el 10 y el 13% según el III Informe Bienal de SUNEDU. Esto representa carreras interrumpidas y recursos institucionales que han quedado fuera de alcance y no pueden ser reembolsados en el futuro. La pregunta que motiva este proyecto sería: ¿es posible detectar a los estudiantes en riesgo antes de que abandonen, no después de que ya hayan abandonado? Para responder a esta pregunta, desarrollamos y verificamos un sistema predictivo de aprendizaje automático para encontrar a los estudiantes al final del primer ciclo universitario que tendrían la mayor probabilidad de no continuar en la universidad. El modelo combina datos académicos del SIAD y del Backoffice de UPCH con calificaciones de educación secundaria obtenidas mediante un proceso de reconocimiento óptico de caracteres (OCR) en certificados oficiales del Minedu. Este último paso fue el más difícil ya que los diferentes tipos de documentos utilizados para representar los datos en los documentos escolares tuvieron que ser limpiados varias veces antes de que los datos pudieran ser recuperados. Se compararon Random Forest y XGBoost con un modelo de regresión logística; dado que existe un desequilibrio significativo entre la clase mayoritaria (continuación) y la minoritaria (no continuación), el recall y las calibraciones probabilísticas fueron más importantes en la evaluación que la precisión general. El prototipo resultante tiene dos componentes: un panel en Dash/Plotly que muestra alertas basadas en cohorte, programa y estudiante y un servicio de inferencia en FastAPI con control de acceso basado en roles. Ninguno está conectado a sistemas de producción institucionales y no genera alertas sobre estudiantes reales automáticamente. Toda la canalización está estructurada utilizando CRISP-DM en Python con bibliotecas de código abierto como Pandas, scikit-learn y XGBoost con trazabilidad comprobada en cada paso. Tomamos la cohorte 2024-1 como un conjunto de validación externo independiente y logramos un AUC-ROC de 85.7% y un Brier Score de 0.142, ambos por encima del umbral en la hipótesis. El sistema no está destinado a reemplazar el juicio del tutor, sino a proporcionar información más oportuna y estructurada que el informe promedio actualmente utilizado.

Palabras clave: Machine Learning, Deserción Universitaria, Modelo predictivo, Riesgo académico, Alerta temprana

Abstract

In Peru, the university dropout rate is usually between 10% and 13% according to SUNEDU's III Biennial Report. This represents interrupted careers and institutional resources that are lost and cannot be recouped in the future. The driving question behind this project is: is it possible to detect at-risk students before they drop out, rather than after they have already left?

To answer this question, we developed and verified a predictive machine learning system to identify students at the end of their first university term who have the highest probability of not continuing at the university. The model combines academic data from UPCH's SIAD and Backoffice with high school grades obtained through an optical character recognition (OCR) process on official Minedu certificates. This last step was the most difficult, as the different types of documents used to represent the data in school records had to be cleaned multiple times before the data could be retrieved.

Random Forest and XGBoost were compared with a logistic regression model; given the significant imbalance between the majority class (continuation) and the minority class (non-continuation), recall and probabilistic calibrations were more important in the evaluation than overall accuracy.

The resulting prototype has two components: a Dash/Plotly dashboard that displays alerts based on cohort, program, and student, and an inference service in FastAPI with role-based access control. Neither is connected to institutional production systems, and they do not automatically generate alerts about real students. The entire pipeline is structured using CRISP-DM in Python with open-source libraries such as Pandas, scikit-learn, and XGBoost, featuring proven traceability at every step.

We used the 2024-1 cohort as an independent external validation set and achieved an AUC-ROC of 85.7% and a Brier Score of 0.142, both above the hypothesis threshold. The system is not intended to replace the tutor's judgment, but rather to provide more timely and structured information than the average report currently used.

Keywords: Machine Learning, College Dropout, Predictive Model, Academic Risk, Early Warning

I) Introducción

1.1. Contexto y problemática en la educación universitaria

Contexto y problemas en la educación universitaria. Las universidades en Perú han aumentado su alcance de manera muy rápida en las últimas décadas: el número de estudiantes matriculados casi se ha triplicado desde el 2000 hasta el 2022, y la universidad se ha convertido en uno de los principales canales de movilidad social para las clases medias y bajas del país. Este es el propósito de la Agenda 2030: los gobiernos deben garantizar el acceso y la continuidad a la educación superior como parte del ODS 4 [3, 15]. Pero esta proliferación de estudiantes no ha venido acompañada de una mejora correspondiente en el apoyo estudiantil. Cada año, un gran número de los que ingresan no completan su carrera: frente a dificultades económicas o dificultades en la preparación académica e integración en instituciones o trabajos que no pueden coincidir con los horarios de clase. Lo que aún se informa menos en el contexto peruano es cómo las instituciones responden tempranamente a tales problemas. El estudio de Vincent Tinto es el mejor para entender: la decisión de abandonar no se toma rápidamente, sino que es una acumulación de rupturas en las relaciones académicas y sociales que el estudiante tiene con la escuela [4]. Si esos lazos se rompen debido a un bajo rendimiento o aislamiento social o presión económica, es probable que también se abandone. El abandono tiene un impacto negativo en las perspectivas laborales y profesionales: aquellos que abandonan encontrarán un trabajo de menor calidad y serán más vulnerables [11]. Este proyecto surge de esa brecha entre acceso y retención. No busca resolver el problema en términos de causas estructurales, sino mostrar que los datos institucionales ya disponibles pueden usarse para identificar a los estudiantes en riesgo con suficiente anticipación para intervenir.

1.2 Desafíos de retención en América Latina y el Perú

El primer año de universidad ha sido el de mayor riesgo de deserción. Es cuando la brecha entre las expectativas de los estudiantes y la realidad de la universidad es más evidente y los datos regionales cuentan la historia: casi la mitad de las deserciones en la mayoría de las universidades latinoamericanas ocurren antes de que el estudiante esté en su segundo año [13, 16]. El aumento en el acceso no tuvo un nivel equivalente

de retención. En América Latina, la persona que más abandona no es alguien que esté desempleado o incapaz de hacerlo. Típicamente es alguien que trabaja mientras estudia, que vive lejos del campus, es el primero en su familia en ir a la educación superior y llega con brechas educativas desde la escuela secundaria que el sistema de orientación vocacional desconocía [5, 12, 36]. Este perfil es de suma relevancia para este trabajo, al igual que el estado de beca, el historial escolar y el modo de admisión son precisamente lo que alimenta el modelo predictivo desarrollado. La pandemia expuso la fragilidad del sistema universitario peruano. La tasa de interrupción de estudios aumentó del 12.6% en el semestre 2019-2 al 18.3% en 2020-1 [1, 14, 34], lo cual está en línea con el impacto de un choque externo pero también muestra una falta de sistemas institucionales que puedan responder a las crisis tan rápidamente. La brecha de conectividad, la falta de dispositivos y la caída en los ingresos familiares obligaron a miles de estudiantes a detener sus estudios. La tasa de deserción luego cayó a alrededor del 11.5% en 2021, pero esto no es una solución permanente al problema. En el período post-pandemia, la deserción se volvió más o menos silenciosa, debido al rezago académico que el aprendizaje digital produjo y la inestabilidad económica en las áreas de menores ingresos.

1.3 Marco normativo y brechas en la gestión institucional

El estado peruano respondió a este riesgo e intensificó su rol regulador y la calidad de las universidades. La Resolución Viceministerial No. 095-2024-MINEDU fue un paso significativo en esta dirección, que establece "Lineamientos para la implementación de acciones para prevenir la interrupción de estudios". Con este documento, la detección temprana de riesgos y el apoyo a los estudiantes se convierten en medidas de bienestar voluntarias que pueden ser auditables en lo que respecta a las Condiciones Básicas de Calidad [33]. Las universidades deben implementar mecanismos efectivos, financiados y basados en evidencia para identificar y apoyar a los estudiantes vulnerables antes de que abandonen. En la mayoría de las universidades peruanas, el problema principal no son solo los datos, sino su dispersión. SIAD registra las matrículas y el LMS está más enfocado en el acceso de los estudiantes al contenido, y el ERP financiero almacena el historial de pagos, información valiosa que coexiste en sistemas que no se conectan entre sí. El resultado es una gestión de retención reactiva en la que se envían alertas cuando el estudiante tiene un historial de bajo rendimiento

o ha dejado de matricularse, justo cuando la intervención tiene el menor impacto. Revertir esta lógica es el propósito de este proyecto.

1.4 Analítica de aprendizaje y modelos predictivos

El proceso de convertir los datos institucionales en alertas accionables es una combinación de dos disciplinas: minería de datos educativos (EDM) y analítica de aprendizaje. Ambas son el vínculo técnico entre los datos transaccionales y la toma de decisiones pedagógicas oportunas. Estudios recientes indican que los modelos de aprendizaje automático son mucho mejores para predecir la deserción que los enfoques estadísticos clásicos. A diferencia de los modelos de regresión logística, que tratan las relaciones lineales entre variables, el bosque aleatorio o XGBoost pueden detectar combinaciones no obvias: un estudiante con largos tiempos de desplazamiento, bajas calificaciones en materias del primer ciclo y retrasos en los pagos puede generar un perfil de riesgo que no podría ser detectado mediante análisis tradicionales. Sin embargo, la aplicación de estos modelos en un entorno educativo real es un desafío de uso, y se deben tener en cuenta cuestiones éticas. Una de las más comunes es el desequilibrio de clases en los datos de deserción, ya que la mayoría de los estudiantes permanecen en la escuela, por lo que el evento de interés se ve como una minoría. Los algoritmos entrenados sin técnicas de muestreo tienden a predecir correctamente la clase mayoritaria pero no son muy buenos para detectar a aquellos en riesgo [26]. Por último, está el problema de la interpretabilidad: si un tutor recibe una lista de estudiantes marcados como “en riesgo” sin explicación de las razones, no actuará. Un sistema predictivo operativamente útil necesita poder indicar si el riesgo proviene de factores académicos, económicos o sociales para que la intervención pueda realizarse adecuadamente.

1.5 Propuesta de investigación

Propuesta de Investigación. Este trabajo se inició porque reconocimos la brecha entre los datos que la UPCH tiene disponibles y las decisiones que los tutores tomaban sin acceso a información sistemática. Esta tesis tiene como objetivo desarrollar, validar y desplegar un sistema de alerta temprana que cierre esta brecha utilizando el estándar CRISP-DM, que está bien establecido en la ciencia de datos aplicada [21, 22]. A diferencia de los métodos convencionales, este sistema se basa en la síntesis de fuentes

de datos heterogéneas de aprendizaje automático, digitalización de documentos a través de Reconocimiento Óptico de Caracteres (OCR) y enfoques de IA Explicable (XAI) con valores SHAP (SHapley Additive exPlanations) y calibración probabilística. A través de esta combinación, los autores proponen un enfoque de ingeniería de datos, aprendizaje automático y gestión educativa para construir una solución que pueda lograr la protección de las trayectorias educativas y la equidad en la educación superior.

II) Marco teórico

Este documento describe tres campos de investigación: la sociología de la educación, que ayuda a entender las causas del abandono escolar; la ciencia de datos, que ayuda a predecirlo; y la ética algorítmica, que ayuda a usarlo con responsabilidad.

2.1.- Bases teóricas sobre la deserción estudiantil

2.1.1.- Definición operativa y dimensiones del fenómeno

La deserción universitaria conlleva pérdidas simultáneas a diferentes niveles. Para el estudiante, significa un camino truncado y una deuda económica sin retorno esperado [13]. Para la institución, significa pérdida de matrícula, deterioro de indicadores y subutilización de infraestructura financiada [5, 10]. Para el país, significa capital humano parcialmente formado que no contribuye a la productividad ni a la contribución fiscal [16]. En Perú, esta dinámica se hizo especialmente visible cuando la tasa de deserción fue del 18.3% en 2020-1 (cuando el problema no es situacional sino estructural).

Distinción Conceptual de las Variables

Para mantener el modelo sin ambigüedades, distinguimos entre las tres etapas de la vida de los estudiantes. El riesgo académico es la probabilidad que el algoritmo estima en el tiempo t (0 a 1) de que el estudiante deje de estudiar. Este riesgo es la variable que el sistema de alerta proporciona a los tutores para informarles sobre la acción preventiva. La no continuidad inmediata es cuando no hay un estudiante matriculado en el semestre posterior al período de análisis. El modelo tiene como objetivo predecir el primer signo visible de desvinculación, cuando aún es posible la recuperación. La deserción consolidada

se registra administrativamente cuando no hay un estudiante matriculado en dos semestres consecutivos sin ninguna reserva formal. Esta es la métrica oficial para los estudiantes que abandonan la escuela; sin embargo, no es una medida preventiva porque esperar a que esto suceda significará que no podemos traerlos de vuelta a la escuela.

Contexto Nacional

En el escenario peruano, el abandono es altamente sensible a factores externos y es un testimonio de la situación de estabilidad social en ese momento. La evidencia sugiere que el sistema universitario es vulnerable a crisis macroeconómicas o de salud [1, 14, 34]. En el informe SIRIES-MINEDU (Figura 1) la tasa de abandono fue del 18.3% en el primer semestre de 2020 (la pandemia de COVID-19) y con el tiempo la tasa bajó al 11.5% en el período 2021-1 según El Comercio (2021) citando datos de MINEDU [35]. Pero esta recuperación no es una solución al problema estructural: el tercer Informe Bienal de SUNEDU (2022) muestra que la tasa de abandono se mantiene constante y las tasas están entre el 10% y el 13% en el período post-pandemia, con estudiantes en los primeros años y estudiantes con menores recursos financieros en el centro de la tasa de abandono.

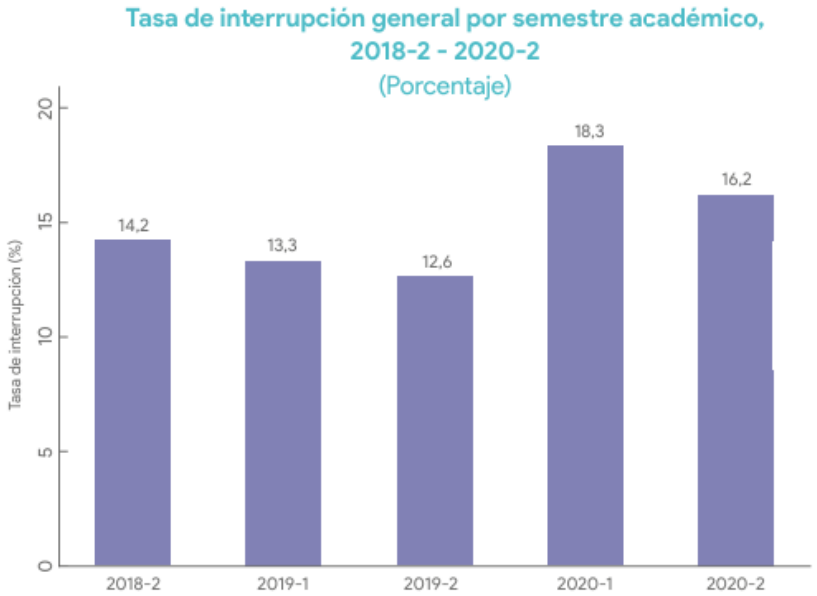


Figura 1: Tasa de interrupción general académica por semestre académico en el Perú (2018-2 al 2020-2). Fuente: MINEDU (2021). Reporte sobre la interrupción de estudios universitarios en el contexto del COVID-19 [14].

Por otro lado, un estudio realizado por el Ministerio de Educación (figura 2), registra el porcentaje de estudiantes menores a 30 años que no se matricularon en ningún establecimiento de estudios superiores y cuentan con una educación superior universitaria incompleta.

	2016	2017	2018	2019	2020	2021	2022	2023	2024
PERÚ	15.3	17.4	15.7	19.3	31.0	23.2	19.6	19.1	19.0
Sexo									
Femenino	13.7	16.4	16.1	20.9	29.9	23.6	19.0	17.5	17.9
Masculino	16.9	18.5	15.2	17.6	32.2	22.7	20.3	20.9	20.1
Área y sexo									
Urbana	15.4	17.6	15.6	19.7	31.9	23.4	19.9	19.5	19.3
Femenino	13.8	16.6	16.1	21.3	30.9	24.0	19.8	18.2	18.3
Masculino	17.1	18.6	15.2	18.0	33.0	22.8	19.9	20.9	20.3
Rural	13.6	15.0 *	16.4	13.6	22.7	20.2	17.2 *	14.6 *	16.4
Femenino	12.1 *	13.6 *	16.5 *	15.6 *	18.9 *	18.4 *	9.6 *	8.8 *	13.8 *
Masculino	14.8 *	16.3 *	16.3 *	11.8 *	25.4	21.5 *	23.0 *	19.9 *	18.6 *
Lengua materna									
Castellano	15.4	17.6	15.8	19.7	31.3	23.1	19.7	19.2	19.1
Indígena	13.4 *	14.7 *	13.7 *	11.6 *	26.5 *	24.8 *	18.8 *	18.2 *	16.9 *
Nivel de pobreza									
No pobre	15.2	16.5	14.9	18.1	28.3	21.7	18.4	17.3	17.7
Pobre No extremo	16.1 *	33.6 *	28.5 *	42.6 *	43.3	37.8	30.5 *	31.1	28.1
Pobre extremo	... *	... *	... *	... *	48.2 *	... *	... *	... *	49.3 *
Región									
Amazonas	... *	... *	... *	... *	... *	... *	... *	... *	12.0 *
Ancash	16.4 *	22.8 *	17.8 *	17.8 *	42.0 *	34.8 *	22.3 *	22.6 *	19.3 *
Apurímac	... *	... *	... *	... *	24.6 *	17.2 *	... *	... *	10.7 *
Arequipa	13.1 *	15.8 *	11.1 *	16.9 *	16.9 *	13.0 *	12.6 *	... *	19.6 *
Ayacucho	... *	... *	... *	... *	27.3 *	... *	... *	... *	29.1 *
Cajamarca	... *	... *	... *	... *	... *	... *	... *	... *	20.8 *
Callao	21.5 *	... *	... *	21.4 *	42.9	... *	21.8 *	17.4 *	32.8 *
Cusco	... *	... *	... *	... *	... *	... *	... *	... *	8.7 *
Huancavelica	... *	... *	... *	... *	... *	... *	... *	... *	5.4 *
Huánuco	... *	... *	... *	... *	25.0 *	24.7 *	... *	... *	7.6 *
Ica	... *	13.9 *	12.5 *	13.4 *	31.7	22.2 *	23.8 *	... *	20.3 *
Junín	20.6 *	14.6 *	... *	21.6 *	32.7 *	25.7 *	... *	... *	15.8 *
La Libertad	17.0 *	27.3 *	19.3 *	18.1 *	44.5	33.9 *	25.7 *	28.6 *	16.0 *
Lambayeque	16.9 *	... *	13.9 *	21.3 *	33.8	27.6 *	26.0 *	17.0 *	15.2 *
Lima Metropolitana	14.0	16.4	15.9	24.6	32.3	21.4	19.6	23.1	22.3
Lima Provincias	12.9 *	22.8 *	24.3 *	17.0 *	30.4 *	24.2 *	... *	23.2 *	20.7 *
Loreto	... *	... *	28.1 *	33.7 *	27.8 *	... *	... *	... *	26.8 *
Madre de Dios	... *	... *	16.0 *	... *	... *	... *	... *	... *	18.7 *
Moquegua	19.7 *	19.0 *	... *	... *	... *	28.0 *	... *	... *	10.6 *
Pasco	... *	... *	... *	... *	... *	29.2 *	... *	... *	11.3 *
Piura	... *	28.3 *	16.1 *	22.9 *	40.4	30.8 *	... *	... *	25.3 *
Puno	14.1 *	... *	15.1 *	... *	33.6 *	36.2 *	... *	... *	9.4 *
San Martín	24.6 *	20.0 *	32.1 *	19.0 *	... *	... *	... *	... *	11.4 *
Tacna	15.7 *	12.7 *	13.2 *	8.8 *	16.8 *	15.5 *	... *	... *	21.0 *
Tumbes	29.7 *	... *	... *	... *	32.1 *	37.0 *	... *	... *	26.4 *
Ucayali	33.0	32.1 *	23.1 *	29.6 *	45.8 *	... *	... *	... *	13.2 *
	2016	2017	2018	2019	2020	2021	2022	2023	2024

Figura 2: *Tasa de deserción acumulada, superior universitaria (% de edades menores o iguales a 30 años con superior universitaria incompleta) [41].*

Para este estudio, se tomó en cuenta el último grado de estudio aprobatorio que el estudiante tiene, si está matriculado actualmente en algún centro educativo y

cuando nació. Excluyendo a los estudiantes que tengan más de 30 años (cumplidos hasta antes del 31 de marzo) y a los que participaron entre enero a marzo, que es periodo de vacaciones y primer mes de clases. La data utilizada del periodo 2016 al 2024 fue obtenida de las Bases de datos de la Encuesta Nacional de Hogares - Instituto Nacional de Estadística e Informática. [41]

2.1.2.- El Modelo de integración estudiantil de Vincent Tinto

El marco teórico más influyente que guía la selección de variables es el Modelo de Integración Estudiantil de Vincent Tinto [4], que es el que tiene más apoyo empírico en estudios de retención a nivel mundial. Tinto afirma que el abandono no es una opción repentina o una decisión aislada, sino el resultado de la erosión acumulativa de los vínculos académicos y sociales con la institución. Cuando estos vínculos se rompen de manera consistente, el abandono puede ocurrir en ciclos. El modelo sugiere dos sistemas de integración que predicen la retención. La integración académica es el logro de calificaciones en el programa (créditos aprobados) y una identificación con una especialidad. Cuando un estudiante aprueba y ve valor en lo que aprende, es un ancla. Por otro lado, la integración social se forma a través de relaciones con grupos de pares, profesores y administración, y la participación en la vida universitaria es un amortiguador del estrés académico y las presiones externas. La pérdida de cualquiera de los sistemas hace más probable el abandono. Este marco hace que sea sencillo poner las variables del modelo en el orden correcto. El promedio ponderado de los estudiantes y sus créditos logrados serán indicadores de integración académica, el estado de beca, la escuela de origen y el modo de ingreso serán indicadores de integración social y económica. La teoría de Tinto también fomenta el enfoque en las etapas iniciales, que es la etapa en la que los vínculos pueden ser más frágiles y la intervención es más impactante [11, 15].

2.2.- Bases teóricas sobre el aprendizaje automático.

2.2.1.- Aprendizaje supervisado.

Aquí utilizaremos el enfoque de aprendizaje supervisado que tenemos a mano y sabemos dónde está la función de mapeo f: $X \rightarrow Y$ de un conjunto de pares etiquetados (x i, y i) donde X es el vector de variables predictoras (rendimiento académico, datos socioeconómicos y contextuales) y y i $\in \{0, 1\}$ es la etiqueta

binaria de continuidad (0 = permanece, 1 = no continúa). Por otro lado, un error de clasificación excesivamente simplificado en datos no vistos no es suficiente: dado que la mayoría de los estudiantes todavía están en los datos educativos, el verdadero desafío es encontrar la clase minoritaria de estudiantes en riesgo. Por lo tanto, no elegimos el algoritmo de antemano, sino que decidimos comparar nuestros algoritmos en un entorno experimental con fuerte validación temporal y precisión de recuerdo y corrección a corto plazo.

2.2.2.- Algoritmos de aprendizaje

Para construir el motor predictivo, comparamos dos enfoques algorítmicos y elegimos el modelo final basado en la complejidad de los datos educativos. La regresión logística se utilizó como referencia ya que es interpretable y tiene una relación lineal y aditiva entre variables y riesgo. Sin embargo, en la Minería de Datos Educativos, es claro que las trayectorias estudiantiles son no lineales y es evidente que el bajo rendimiento académico se agrava cuando se combina con factores sociales: los modelos lineales no pueden capturar con precisión este fenómeno [6, 25]. Random Forest y XGBoost son los mejores modelos para predecir con datos tabulares. Ambos modelos construyen múltiples árboles de decisión y capturan patrones no lineales e interacciones complejas entre variables [18, 23]. XGBoost, en particular, maneja valores faltantes como parte de su diseño nativo, incorpora regularización y muestra una mejor consistencia de rendimiento en contextos educativos con datos tabulares [18, 20, 23]. Elegimos este algoritmo como el modelo final basado en una comparación experimental, que se describe en el capítulo de resultados.

2.2.3.- El desafío del desbalance de clases y la selección de métricas

Uno de los desafíos técnicos más subestimados en la predicción de la deserción es el desequilibrio de clases. Dado que la gran mayoría de los estudiantes no abandonan, el evento de interés es estadísticamente minoritario, lo que significa que las métricas tradicionales como los indicadores de precisión tienen una conclusión errónea: un modelo que sistemáticamente predice que ningún estudiante abandonará podría lograr un 95% de precisión sin detectar un solo caso real de riesgo [7, 8]. Para evitar esta distorsión, la evaluación del modelo se

centra en métricas que son sensibles a la clase minoritaria, lo cual está en línea con el objetivo institucional de identificar la mayor cantidad posible de estudiantes en situaciones vulnerables. El recall mide cuántos estudiantes que no abandonaron la escuela fueron detectados por el sistema. Para el contexto de alerta temprana, el recall significa reducir los falsos negativos, es decir, los casos de riesgo que el sistema no detectó. Por otro lado, el Brier Score evalúa la fiabilidad de las probabilidades esperadas. Un modelo bien calibrado asegura que si el sistema asigna un 80% de riesgo a un grupo de estudiantes, aproximadamente ese porcentaje de ellos no continuará [9]. Esta propiedad es importante para que los tutores confíen en las alertas y prioricen las intervenciones de manera racional.

2.3.- Dashboards y protocolos de uso: la operacionalización de la analítica

La brecha entre la precisión técnica de un modelo y su uso en la institución es lo que algunos estudios denominan el problema del "último kilómetro" [23, 32]. Es información que existe en un servidor o cuaderno pero que nunca llega al tutor que podría actuar sobre ella. Los paneles de control académicos son la solución técnica a este problema. No son una imagen de datos históricos, sino una herramienta de predicción, lo que significa dar a las autoridades un mensaje claro y oportuno.

2.3.1.- El dashboard como herramienta cognitiva

Few (2006) y Tufte (2001) también argumentan que un buen panel de control reduce la fricción cognitiva, ya que el usuario nota el problema a primera vista y profundiza solo cuando la situación lo requiere. Para la detección de deserción en el aula, el panel de control debería permitir que el nivel de la facultad o cohorte se guíe hacia el perfil individual de cada estudiante y los factores identificados por SHAP para cada caso [33, 35]. Mostrar el riesgo sin explicar lo que significa es un problema diferente: el tutor ve una lista de estudiantes etiquetados como "alto riesgo" pero no sabe la razón del alto riesgo, como el rendimiento académico, la condición económica u otros factores, por lo que la intervención puede no ser efectiva para el contexto en el que se pretende.

2.3.2.- Predicción a la prescripción: protocolos de intervención

La literatura sobre Analítica del Aprendizaje es consistente en un punto: el éxito de un sistema de alerta temprana realmente se trata del protocolo institucional y es más sobre el algoritmo. El modelo más preciso no retiene a ningún estudiante si el tutor no sabe que necesita actuar, no tiene tiempo para hacerlo o no tiene recursos disponibles para apoyarlos [19, 20]. Por eso la propuesta identifica tres niveles de respuesta según la puntuación de riesgo.

III) Estado del arte

En la última década, el campo de la investigación sobre el abandono universitario pasó del diagnóstico sociológico a la predicción computacional. La digitalización de los registros académicos proporcionó los datos que los modelos de aprendizaje automático necesitaban y la literatura respondió con una gran expansión de herramientas predictivas. Este capítulo examina el desarrollo de los enfoques de modelado, su implementación en sistemas de alerta reales y las brechas metodológicas que aún deben cerrarse.

3.1.- Enfoques predictivos de deserción

Durante algunos años, la regresión logística fue el enfoque estándar para el análisis de deserción universitaria, porque era fácil de entender, sus coeficientes eran fáciles de interpretar como razones de probabilidades y funcionaba bien cuando los perfiles de los estudiantes eran relativamente homogéneos. Esa homogeneidad desapareció y los factores de riesgo comenzaron a combinarse de maneras que el modelo lineal no puede transmitir. Un estudiante que trabaja, vive lejos del campus y reprueba una materia crítica en el primer semestre, por ejemplo, tiene un alto nivel de riesgo que ninguna ecuación lineal puede explicar.

Varios estudios [6, 17] confirman la efectividad de los métodos de conjunto sobre las estadísticas clásicas para estudiar la clasificación educativa. Random Forest y XGBoost superan consistentemente a la regresión logística y a las SVM en sensibilidad y precisión. La ventaja técnica es que los árboles potenciados por gradiente pueden capturar interacciones complejas que los modelos lineales no pueden. La mala calificación en el primer semestre de clase, por ejemplo, se amplifica para aquellos estudiantes que viven lejos del campus o carecen de una beca, algo que XGBoost puede identificar y la regresión logística no.

El desequilibrio de clases es un problema importante en este campo. En la mayoría de las universidades, los estudiantes desertores son raros, y los algoritmos se ven obligados a categorizar a todos como retención para minimizar el error global, donde se logra un 90% de precisión para que un estudiante no esté en riesgo. Como se ilustra en [26], la ponderación de clases, el sobremuestreo sintético (SMOTE) y el submuestreo no son opcionales, sino una condición necesaria para que el modelo sea útil. La comunidad investigadora también estaba cuestionando el AUC como la única métrica de referencia. Estudios recientes [7-10, 24] argumentan que, para gestionar recursos escasos, no es suficiente etiquetar a un estudiante como desertor o no; el modelo debe producir una probabilidad calibrada que permita ordenar casos y priorizar intervenciones. Un sistema que asigne "70% de riesgo" debe ser confiable: si asigna este puntaje a 100 estudiantes, aproximadamente 70 deberían abandonar el sistema. En América Latina, la adopción de estas tecnologías no está exenta de desafíos. Los modelos derivados en países de habla inglesa se han basado en variables de rendimiento académico, ya que son los predictores más relevantes, pero este enfoque es insuficiente en la región latinoamericana. Investigaciones en Perú, Brasil y Colombia [13, 16, 27, 31] demuestran que el riesgo académico está conectado al riesgo social: el estado de beca, la modalidad de admisión y el tipo de escuela previa son factores predictivos, y si los ignoramos, obtenemos modelos que funcionan bien en la prueba interna pero no pueden desempeñarse en el mundo real.

3.2.- Sistemas de alerta y dashboards.

La distancia entre la precisión técnica de un modelo y su uso en la institución es lo que algunos estudios llaman el problema de la "última milla" [23, 32]: los datos están en un servidor o en un cuaderno, pero nunca llegan al tutor que podría actuar sobre ellos.

Los Sistemas de Alerta Temprana intentan abordar exactamente esa desconexión. La evidencia de implementaciones reales [23, 32] muestra que los paneles de control exitosos van más allá de la visualización descriptiva. No solo registran lo que sucedió, sino que también guían qué hacer a partir de ese punto. Es mejor diseñarlos con los usuarios finales e integrarlos en el flujo de trabajo que los tutores ya tienen establecido. El desarrollo consiste en diseñar interfaces que indiquen niveles de riesgo y ofrezcan planes de acción concretos basados en el caso. Estos sistemas no son

adoptables automáticamente. La investigación sobre usabilidad [30] ha demostrado que la confianza es la mayor barrera para el uso sostenido: los tutores desconfían de los sistemas que envían alertas sin explicación, y si no saben por qué un estudiante está en la lista, no actúan en consecuencia. La interpretabilidad no solo es un requisito ético, sino un prerrequisito para el uso del sistema. El éxito a largo plazo no depende solo del código. Los experimentos institucionales [18, 28] muestran que, sin una gobernanza de datos clara, roles definidos, protocolos de acceso y métricas de servicio, un sistema de alertas puede terminar como proyectos piloto sin mantenimiento y así la herramienta se institucionaliza tan difícilmente como construirla.

3.3.- Brechas recurrentes y agenda de mejora.

La revisión de la literatura más reciente sugiere tres problemas recurrentes. El primero es la fuerte dependencia de las variables académicas y la ausencia de factores psicosociales bien explicados. El segundo es la falta de calibración: muchos modelos informan sobre el AUC, pero no verifican si sus probabilidades son confiables [6-10, 17-20, 23, 26]. El tercero es la brecha operativa donde los modelos permanecen en cuadernos Jupyter y nunca llegan a un protocolo de mentoría real.

La agenda de la literatura tiene tres puntos: validación temporal rigurosa, calibración verificable y una capa operativa para vincular el modelo a la institución [1-3, 21, 22, 31]. Los tres frentes discutidos en este proyecto se abordarán en este proyecto.

IV) Hipótesis y objetivos

4.1.- Hipótesis general

Un sistema predictivo con variables académicas escolares y universitarias, junto con factores sociodemográficos, se entrena con una capacidad de distinción de clases en AUC-ROC ≥ 0.75 para identificar a los estudiantes en riesgo de abandonar el primer ciclo en la UPCH antes del segundo ciclo. El sistema será evaluado con cohortes (en 2024-1) con calibración probabilística (Brier Score ≤ 0.20) e interpretabilidad (SHAP) mediante validación externa.

4.2.- Objetivo general

Desarrollar y validar un sistema predictivo basado en clasificadores supervisados (Random Forest, XGBoost, Regresión Logística) que integren variables académicas,

escolares, universitarias y sociodemográficas en la detección temprana de estudiantes en riesgo de abandonar o tener un bajo rendimiento en la UPCH. El rendimiento del sistema se evaluará mediante una validación temporal externa (cohorte del período 2024-1) con respecto a la calibración probabilística, la interpretabilidad y la viabilidad de implementación del sistema.

4.3.- Objetivo específico

- El primer objetivo es evaluar el peso predictivo relativo de las variables académicas escolares (promedio de la escuela secundaria, puntaje de admisión) en comparación con las del primer ciclo en la universidad utilizando el análisis de Importancia de Permutación y los valores SHAP evaluados en la cohorte de validación externa (período 2024-1).
- El segundo objetivo es evaluar la existencia de variabilidad en las distribuciones de probabilidad de riesgo entre subgrupos socioeconómicos (beneficiarios de becas vs. no beneficiarios) y determinar la necesidad de umbrales de alerta diferenciados a través del análisis de curvas de precisión-recall equilibradas por categorías.
- El tercero objetivo busca validar la estabilidad temporal de los patrones predictivos mediante comparación de la importancia relativa de las variables entre el conjunto de entrenamiento (periodos 2022-2023) y el de validación externa (periodo 2024-1). Siendo capaz de cuantificar la variación en clasificaciones y magnitudes.
- El cuarto objetivo es desarrollar y evaluar un prototipo funcional (API REST junto con un panel interactivo) que demuestre la viabilidad técnica del despliegue midiendo los tiempos de respuesta, la capacidad de carga y la disponibilidad en un entorno de prueba controlado.

Variables:

La tabla de operacionalización se describe en la Tabla 1.

Variable	Definición operacional	Dimensiones	Indicadores	Unidad de medida	Escala de medición

Variables sociodemográficas	Características personales y de origen: género, provincia de procedencia, colegio de procedencia, año de egreso, estudiante de primera generación, categoría de estudiante.	Contexto socio-origen	Proporción de cada categoría en la muestra	Porcentaje (%)	Nominal
Variables académicas	Medidas numéricas de desempeño educativo: notas de secundaria, notas 1er y 2do ciclo, nota de examen de admisión, costo por ciclo.	Rendimiento académico	Nota secundaria promedio; puntaje de admisión; costo relativo (S/. por crédito)	Continuas (S/.; puntaje)	Razón
Sistema predictivo basado en ML (Independiente)	Clasificador supervisado entrenado con variables académicas y sociodemográficas (tiempo de desplazamiento) para identificar estudiantes en riesgo de deserción o bajo rendimiento.	Entrenamiento supervisado	Tipo de modelo (Random Forest, XGBoost, Regresión Logística); fases del pipeline (EDA, preprocesamiento, balanceo, entrenamiento, validación)	-	Nominal
Detección temprana del riesgo (Dependiente)	Capacidad del sistema para identificar correctamente a los estudiantes en riesgo antes de su deserción real.	Efectividad diagnóstica	F1-score, Recall, AUC-ROC	Porcentaje (%)	Razón

Tabla 1: Matriz de operacionalización de variables del sistema predictivo

V) Materiales y métodos

5.1.- Metodología de investigación

El marco metodológico de esta investigación es cuantitativo, no experimental y longitudinal retrospectivo. No se manipularon variables y el trabajo solo incluyó registros institucionales históricos al final de cada período académico. Es explicativo-predictivo en el sentido de que no solo se trata de identificar los factores que están asociados con el riesgo de deserción, sino también de crear un modelo que sea capaz de predecirlo antes de que ocurra, utilizando solo los datos disponibles en el momento de la predicción. La investigación adoptó la metodología CRISP-DM (Proceso Estándar de la Industria para la Minería de Datos) como herramienta para la situación específica de la investigación. Esto se debe a que CRISP-DM tiene un proceso iterativo donde necesitamos registrar las decisiones tomadas y verificar que el modelo esté funcionando para el problema institucional, y no un ejercicio técnico que no se tenga en cuenta. El proceso se organizó en seis fases interdependientes:

Fase 1: Formulación del problema y definición del marco conceptual.

Esta fase consistió en definir qué es el riesgo académico en la UPCH y qué variables debería predecir el modelo. La revisión de la literatura y el estudio exploratorio inicial podrían ayudar a determinar qué variables tienen el mayor respaldo empírico y cuáles deberían descartarse. Las variables que se descartaron son aquellas que no mostraron una correlación significativa con la deserción en escenarios similares.

Fase 2: Recolección y preparación de datos

La tarea principal de esta fase (Figura 2) fue crear, a partir de registros dispersos en tres sistemas diferentes, un conjunto de datos longitudinal con una fila por estudiante en cada período. Se utilizaron SIAD (matrículas y calificaciones universitarias), Backoffice (historial académico) y documentos de calificaciones de secundaria en formato PDF procesados con OCR para formar el conjunto de datos. La parte más delicada fue establecer una variable objetivo. Primero, elegimos el rendimiento académico como la métrica de medición (por ejemplo, el promedio ponderado del primer año con un peso inferior a 14, validado normativamente y con curvas ROC y el índice de Youden), y segundo, la trayectoria de matrícula (por ejemplo, la ausencia de matrículas en dos ciclos consecutivos según lo definido por UPCH para constituir una interrupción prolongada).

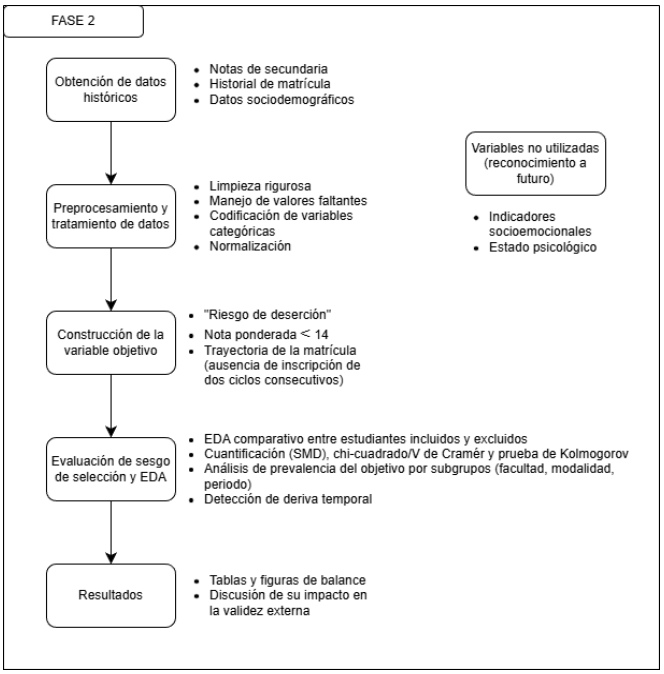


Figura 2: *Gráfico visual para la recolección y preparación de datos*

Fase 3: Ingeniería de características y selección de variables

En esta fase, se transformaron y seleccionaron las variables que se utilizarían para alimentar los modelos. Se realizó una normalización de los promedios de secundaria y admisión utilizando las estadísticas del conjunto de entrenamiento para asegurar la consistencia. Se trataron el costo relativo por ciclo (en soles por crédito, vigente según el programa académico) y las variables categóricas - género, provincia, escuela, categoría y estatus de primera generación. Para las categorías más comunes, se aplicó codificación one-hot, y para las otras categorías, se agruparon como "Otros". Cuando el número era muy alto, se seleccionó la codificación de objetivo con validación K-fold para evitar la fuga de datos. La selección de características se realizó en combinación con métodos integrados y análisis estadístico. La importancia de cada variable se cuantificó mediante Permutation Importance y SHAP en los modelos de Random Forest y XGBoost. En la regresión logística regularizada, se utilizaron coeficientes estandarizados para estimar contribuciones importantes, y también se examinó la multicolinealidad mediante el factor de inflación de la varianza (VIF). Además, también investigamos el impacto de la eliminación de características por bloques y eliminamos grupos de variables para validar que cada conjunto tenía una influencia distinta en el rendimiento del modelo.

Fase 4: Construcción y optimización del modelo predictivo

Este cuarto paso consistió en un análisis exploratorio de datos (EDA) para cuantificar el grado de desequilibrio de clases y determinar la estrategia a utilizar para equilibrar. Se establecieron y aplicaron tres flujos de trabajo independientes a los mismos datos preprocesados con normalización secuencial de variables numéricas y codificación de categorías. Evaluamos el rendimiento del modelo a un nivel comparable utilizando métricas apropiadas para datos desequilibrados, que fueron: el área bajo la curva de precisión-recall (Precisión Promedio) como la métrica principal, precisión, sensibilidad (recall), puntuación F1 y AUC-ROC.

Métricas del modelo

Métrica	Valor (Score)
Accuracy	0.754 (75.4%)
Precisión	0.664 (66.4%)
Recall	0.734 (73.4%)
F1-Score	0.6970.697 (69.7%)

Fase 5: Validación externa y análisis de interpretabilidad

En la quinta etapa, el modelo seleccionado será evaluado en un conjunto de datos independiente y su capacidad explicativa será probada de las siguientes maneras:

- El modelo final se aplicará a la cohorte de 2024-1 como un conjunto de validación externo para verificar su generalización y robustez con datos previamente no analizados.
- Las métricas de rendimiento disponibles, como precisión, sensibilidad (recall), puntuación F1 y exactitud, se compararán aún más con la matriz de confusión y la validación cruzada interna.
- Se realizará la importancia de permutación en el conjunto de validación (cohorte en 2024-I) para evaluar la relevancia global de las variables a medida que la exactitud disminuye al permutar cada característica.
- Informaremos sobre el procedimiento de validación y las configuraciones utilizadas (parámetros del modelo, versiones de bibliotecas y rutas de archivos), así como la representación, para que la representación pueda ampliarse con más métricas y explicaciones locales y análisis de equidad.

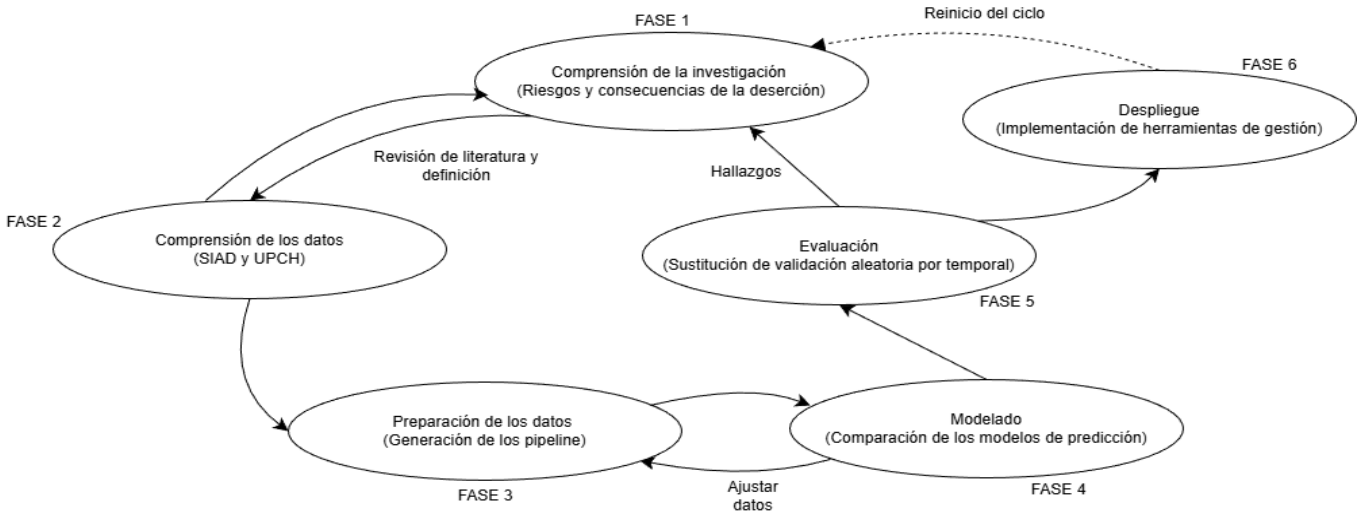
Fase 6: Experimentación y despliegue tecnológico

La fase de implementación incluyó el desarrollo de un prototipo con FastAPI y Dash, que puede cargar datos de inferencia desde fuentes estáticas o archivos CSV. Además de visualizar resultados en un entorno controlado, también demuestra la viabilidad técnica del flujo de inferencia sin tener que diseñar una base de datos y modelado de entidad-relación. Se implementó además un dashboard interactivo en Dash (Plotly) diseñado para poder visualizar los indicadores, alertas y métricas por niveles de detalle: estudiante, cohorte y facultad.

Vale destacar que, aunque este prototipo se centrará en el análisis descriptivo y predictivo de riesgo académico, se preparará la arquitectura para incorporar analítica prescriptiva en futuras versiones, que ofrezca recomendaciones de acciones basadas en los hallazgos resultantes.

Por otro lado, también se describirá el proceso de co-diseño con los usuarios finales. Estos son las autoridades académicas, gestores de OAMRA y el área de tutoría de la UPCH. En donde se incluyó talleres de validación de prototipos y pruebas de usabilidad para corroborar que las visualizaciones cubran sus necesidades reales de toma de decisiones. El gráfico siguiente resume las fases del proceso CRISP-DM aplicadas en este proyecto:

Figura 3: *Modelo de proceso CRISP-DM aplicado a la predicción del abandono estudiantil [23].*



Para organizar el desarrollo se adoptó el marco **CRISP-DM**, adaptado a las condiciones específicas del proyecto. La elección no fue arbitraria: CRISP-DM proporciona una estructura iterativa que obliga a documentar cada decisión y a verificar que el modelo responde al problema institucional definido, no a un ejercicio técnico desconectado de la realidad. Las seis fases se ajustaron para dar prioridad a la validación temporal y a la trazabilidad, dos requisitos que en este contexto educativo son tan importantes como el AUC del modelo. (Tabla 2)

Fase	Enfoque Central	Objetivo / Resultado Clave
Comprensión del Negocio	Definición operativa del "riesgo de no continuidad".	Anticipar qué estudiantes podrían no matricularse en el siguiente ciclo, usando la información disponible al cierre del periodo actual.
Comprensión de los datos	Análisis de los sistemas SIAD y Backoffice.	Reconstruir trayectorias históricas y asegurar que las variables utilizadas respeten la lógica temporal sin incorporar información futura.
Preparación de los datos	Construcción de un pipeline reproducible.	Limpieza, estandarización y generación de variables derivadas (escolares, académicas, administrativas), documentando las transformaciones.
Modelado	Comparación de Regresión Logística, Random Forest y XGBoost.	Priorizar sensibilidad y calibración para generar alertas tempranas y probabilidades operativas, en lugar de la exactitud global.
Evaluación	Sustitución de validación aleatoria por validación temporal.	Entrenar con 2022–2023 y evaluar con 2024-1 (sin reajustes) para estimar el desempeño real en condiciones de operación.
Despliegue	Implementación de herramientas de gestión institucional.	Creación de una API y un dashboard para predicciones y consultas, integrando principios de gobernanza (trazabilidad y acceso por roles).

Tabla 2: Enfoque de las fases CRISP-DM enfocado en el proyecto.

5.2.- Material en estudio

Este estudio se desarrolló con base a los datos institucionales proporcionados por la Oficina Universitaria de Admisión, Matrícula y Registro Académico (OAMRA) de la UPCH. Datos correspondientes a estudiantes egresados de secundaria e ingresantes universitarios entre los años 2022 y 2024. (Tabla 3)

Base de datos	Filtro a utilizar	Variables a considerar
Notas finales de secundaria	Estudiantes 1° a 5° de secundaria de los periodos correspondientes 2022-1 a 2024-1	Historial de calificaciones
Historial de matrícula universitario	Ingresantes 2022-1 y 2023-2, considerando solo aquellos alumnos que cursaron los primeros dos ciclos de forma consecutiva	Historial de matrícula
Datos sociodemográficos y académicos	Datos complementarios al registro	Edad, colegio de procedencia, modalidad de ingreso, tipo de gestión escolar, género y tiempo promedio de desplazamiento (ida y vuelta) desde la residencia hasta la sede SMP

Tabla 3: Definición de filtros y variables de la base de datos

De estos datos, se destaca que el conjunto de datos del ciclo 2024 - 1 será reservado exclusivamente para la validación final de todo el modelo. La condición que presenta este análisis es que los estudiantes hayan cursado obligatoriamente los dos primeros ciclos consecutivamente desde su ingreso. En la sección 4 del presente documento se dan detalles sobre las consideraciones éticas del estudio. A continuación, se presenta la tabla: Operacionalización de Variables Independientes y Dependientes (Tabla 4), las cuales fueron definidas para el sistema predictivo.

Variable	Dimensión	Indicador	Índice Esperado	Unidad de Medida	Tipo de Variable	Técnica	Instrumento	Escala
Independiente: Sistema predictivo (modelo ML + dashboard)	Construcción del modelo predictivo	funcionamiento de un pipeline ML reproducible	Pipeline operativo con fases CRISP-DM completadas	Binaria	Independiente	Auditoría técnica del flujo de trabajo (scripts y notebooks)	Jupyter Notebook, documentación técnica	Nominal
	Capacidad predictiva	F1-Score del modelo base y final	≥ 0.88	Porcentaje	Independiente	Validación cruzada y comparación de modelos	Reporte Scikit-learn / validación temporal	Métrica (razón)
	Generalización	Variación de desempeño entre cohortes (2022–1 → 2024–1)	≤ 5 % diferencia	Porcentaje	Independiente	Validación externa y análisis de drift	Dataset OAMRA-UPCH / scripts de validación	Métrica (razón)
	calibración probabilística	Brier Score promedio (calibración del modelo)	≤ 0.18	Decimal	Independiente	Calibración de probabilidades	Reporte del modelo final	Métrica (razón)
	Despliegue operativo (entorno controlado)	Servicio de inferencia disponible vía API	Endpoint activo /predict	Categórica	Independiente	Pruebas de servicio FastAPI	Registro de logs del servidor	Nominal
	Visualización	Dashboard funcional con filtros y alertas por cohorte	100 % funcionalidades activas	Porcentaje	Independiente	Prueba funcional y heurística de usabilidad	Dashboard Dash / checklist UI-UX	Métrica (razón)
	Rendimiento bajo carga y seguridad	Solicitudes concurrentes procesadas sin error	≥ 5 000 sin fallo	Conteo	Independiente	Prueba de carga (Python) y verificación de cifrado	Logs del servidor / monitoreo	Métrica (razón)
Dependiente: Identificación de estudiantes en riesgo	Identificación eficaz	Recall, F1-score (etiquetas correctas de riesgo)	Recall > 85%	Porcentaje	Dependiente	Prueba de predicción con test set	Matriz de confusión	Métrica
	Fiabilidad de clasificación	Tasa de falsos positivos (FP) y falsos negativos (FN	FP ≤ 15% y FN ≤ 10%	Porcentaje	Dependiente	Análisis de resultados ML y validación de etiquetas reales	Reporte técnico de resultados ML	Métrica (razón)
	Oportunidad de la alerta temprana	Diferencia temporal entre la predicción y la ocurrencia real de riesgo	Alerta emitida ≥ 1 ciclo antes	Porcentaje	Dependiente	Validación temporal	Dataset institucional (SIAD/OAMRA)	Escala de intervalo

		académico o deserción						
	Utilidad institucional	Acciones preventivas generadas a partir de las alertas	≥ 3 intervenciones por cohorte	Conteo	Dependiente	Análisis documental y registro de intervenciones	Reportes institucionales de tutoría (OAMRA)	Métrica (razón)
	Satisfacción del usuario	Nivel de percepción del tutor o analista sobre la utilidad y facilidad del sistema	≥ 4 en escala Likert (1–5)	Escala ordinal	Dependiente	Encuesta estructurada y entrevistas	Cuestionario validado de usabilidad	Escala ordinal
	Adopción y uso institucional	Porcentaje de tutores o coordinadores que usan el sistema en sus labores académicas	≥ 80% de adopción	Porcentaje	Dependiente	Análisis de registros de acceso	Logs de usuarios / FastAPI	Métrica (razón)

Tabla 4.- Operacionalización de Variables Dependientes e Independientes

5.3.- Herramientas y tecnología a utilizar

Las herramientas que fueron utilizadas para la implementación del sistema son:

- Python 3.12: Fue el lenguaje utilizado para el procesamiento, análisis y modelado de datos.
- Pandas y Scikit-learn: Son las bibliotecas esenciales para la ingeniería de características, el modelado y evaluación de algoritmos de clasificación.
- Jupyter Notebook: Es el entorno de desarrollo interactivo para el análisis exploratorio y prototipado.
- FastAPI y Dash (Plotly): Son herramientas para el desarrollo de una aplicación web ligera que permite la generación de predicciones en tiempo real, con manejo de desbalance y características de privacidad.
- PostgreSQL (opcional): Será el motor de base de datos para almacenamiento estructurado, en caso se requiera integración en producción.

5.4.- Consideraciones éticas

Este proyecto se diseña y ejecuta bajo consideraciones estrictas de éticas que protegen la privacidad de los estudiantes y aseguran el uso responsable de la información

académica. A continuación, se detallarán los principios adoptados para la privacidad y el uso ético de los datos:

5.4.1.- Privacidad de los datos

- La información provendrá de los sistemas que proporciona la UPCH (SIAD y Backoffice) para el periodo 2022-I a 2024-I, solicitada formalmente a OAMRA mediante comunicación oficial aprobada por el asesor de tesis y la dirección de la Facultad de Ciencias e Ingeniería. Estos datos serán utilizados únicamente con fines académicos de investigación y no con propósitos comerciales, disciplinarios, evaluativos de personal docente, ni cualquier otro ajeno al desarrollo de la tesis.
- Se destaca que las constancias de logros de estudio utilizadas para la extracción de notas de secundaria son de acceso público y pueden verificarse en la página oficial del Minedu (https://constancia.minedu.gob.pe/), sin embargo, incluso estos datos de origen público serán integrados al conjunto anonimizado siguiendo el mismo protocolo de custodia.
- Almacenamiento seguro: Los datos anonimizados se almacenarán exclusivamente en dispositivos de propiedad del investigador principal, protegidos con cifrado de disco completo (AES-256) y contraseñas robustas. No se utilizarán servicios de almacenamiento en la nube públicos sin cifrado de extremo a extremo. Al finalizar el proyecto, los datos serán eliminados de forma segura o devueltos a OAMRA según protocolo institucional.

5.4.2.- Uso restringido y prototipo de prueba de concepto

Los datos se procesarán exclusivamente en entornos controlados, seguros y conforme a las políticas institucionales de la UPCH. El sistema predictivo se desplegará como un prototipo funcional para investigación y desarrollado en base herramientas de FastAPI y Dash, con acceso restringido únicamente al equipo investigador y evaluadores autorizados (asesor, co-asesor y jurado de tesis).

Naturaleza y alcance del prototipo

Este trabajo es un prototipo para un proyecto de investigación académica y tiene como objetivo demostrar la viabilidad técnica y metodológica de un sistema de alerta temprana basado en aprendizaje automático. Prueba el rendimiento de diferentes algoritmos a través de la validación de cohortes (períodos académicos) y las herramientas básicas de inferencia/visualización en un entorno controlado. Como prototipo, el sistema no es una herramienta institucional ya que no está vinculado a bases de datos de producción. El prototipo no envía alertas a autoridades/profesores, no influye en decisiones académicas o administrativas sobre estudiantes reales, y no está disponible en servidores institucionales a los que puedan acceder investigadores fuera de la investigación.

5.4.3.- Confidencialidad de la tesis y aprobación institucional

El código fuente y los datos procesados (calificaciones, situaciones sociodemográficas, etc.) para la tesis e información técnica se mantendrán en secreto y no se publicarán en repositorios públicos (RENATI, GitHub, etc.) sin el consentimiento explícito por escrito del autor de la Universidad Peruana Cayetano Heredia. El proyecto será aprobado por el Comité de Ética en Investigación de la UPCH antes de procesar los datos de los estudiantes, y se seguirán inmediatamente las recomendaciones del comité. Cualquier cambio metodológico que pudiera llevar a un cambio en el uso de los datos será remitido al comité para su revisión.

5.4.4.- Publicación de resultados agregados

Para proteger la identidad de los participantes, solo se incluirán las tablas, figuras y métricas agregadas a nivel de periodo de cohorte, facultad o programa académico sin revelar ningún identificador personal directo como nombre, ID o código universitario, y no habrá combinaciones de variables donde se puedan identificar casos individuales. En tablas de contingencia o desagregaciones por subgrupos, suprimiremos celdas con $n < 5$ estudiantes para evitar que sean identificados por exclusión o inferencia estadística. Reportaremos las métricas de rendimiento del modelo con solo dos decimales, y siempre redondearemos los números absolutos a múltiplos de 5 cuando el tamaño del subgrupo sea

pequeño (por ejemplo, n < 50).

5.5.- Limitaciones metodológicas y desconexión de la responsabilidad de ejecución operativa.

5.5.1.- Limitaciones del estudio

Este estudio tiene varias limitaciones que deben tenerse en cuenta al interpretar los resultados:

- La primera es temporal. Entrenamos el modelo con las cohortes de 2022-2023 y lo probamos con las cohortes de 2024-1, durante un período de tres años. Los patrones predictivos identificados pueden no aplicarse a cohortes futuras si se producen cambios estructurales en la institución, como reformas curriculares, políticas de becas o crisis sanitarias.
- La segunda limitación es el número de variables. Solo utilizamos variables académicas (escolares y de primer ciclo) y sociodemográficas básicas que se han acumulado en la institución (no variables socioemocionales, psicológicas, motivacionales o de salud mental). Estas variables han sido reportadas en la literatura (Mustofa et al., 2025; Delen et al., 2023), pero necesitamos otras herramientas para recopilar información más allá del alcance de este estudio.
- El alcance de nuestro estudio introduce una tercera limitación. Solo consideramos a los estudiantes que han completado sus dos primeros ciclos universitarios. Aquellos con licencia médica, retiros temporales justificados, transferencias internas entre programas o admisiones por validación están excluidos, lo que puede llevar a un sesgo de selección en la generalización a toda la población estudiantil.
- La cuarta limitación es la interpretabilidad causal. Los modelos predictivos identifican relaciones estadísticas, no causales. Una variable con alta importancia predictiva no implica necesariamente una relación causal directa con el abandono. Las intervenciones basadas en este estudio deben fundamentarse en una teoría educativa sólida y no solo en correlaciones estadísticas.

- También hay un sesgo de selección estructural en dos niveles del diseño. Primero, los criterios de inclusión requieren que los estudiantes tengan al menos dos ciclos universitarios consecutivos y abandonen antes de que termine el primer ciclo (que es el grupo de estudiantes más vulnerable y tiene la menor visibilidad en el sistema universitario). En segundo lugar, la validación externa se realizó solo para las facultades piloto de Enfermería, Medicina Veterinaria y Ciencias (N=756 de 4,176 registros de 2024-1), y por lo tanto no para otras facultades de la UPCH. La demografía de la subpoblación excluida puede diferir de la de la subpoblación incluida en términos de género, modo de admisión y estado económico. En cuanto a las etapas futuras, recomendaríamos validar el modelo en todas las facultades y medir el rendimiento del modelo en cohortes de primer ingreso.
- La sexta limitación es la ausencia de variables sociodemográficas en el modelo final. Originalmente se propuso agregar variables sociodemográficas, pero el análisis de importancia (SHAP, Tabla 29) mostró que su poder predictivo mínimo es mucho menor que el rendimiento universitario inmediato en el primer ciclo. Variables como el tiempo de viaje al campus, el estado laboral y el nivel educativo de los miembros de la familia no estaban disponibles en los registros institucionales de SIAD y Backoffice en ese momento, y habrían requerido herramientas de recolección adicionales para ser utilizadas. También creemos que la inclusión de variables como género u origen geográfico sin calibración de subgrupos puede llevar a sesgos de equidad en las alertas generadas, las cuales necesitan ser auditadas y verificadas antes de ser utilizadas. Enumeramos estas variables como predictores de segunda capa para la etapa futura de nuestro sistema, y esperaremos a que estén disponibles datos de mayor calidad y granularidad.

5.5.2.- Descargo de responsabilidad de responsabilidad operativa.

Este trabajo se desarrolla como un prototipo para la investigación académica y no como un sistema operativo validado para tomar decisiones administrativas o académicas sobre las trayectorias estudiantiles. Los resultados deben interpretarse como exploratorios e indicativos, y deben ser validados adicionalmente por expertos en gestión académica y bienestar estudiantil. Los algoritmos que desarrollamos no deben aplicarse automáticamente o de inmediato a decisiones sobre los derechos o el futuro académico de los estudiantes, como la colocación obligatoria en tutorías, restricciones de matrícula o condiciones de becas, sin la supervisión humana experta y sin considerar los aspectos individuales de cada caso. Las probabilidades de riesgo que construimos son estimaciones estadísticas con márgenes de error: un estudiante que es "de alto riesgo" podría graduarse bien mientras que un estudiante que es "de bajo riesgo" podría sufrir algunos problemas imprevistos. Las alertas deben servir como una señal para el apoyo proactivo, no como una lista de etiquetas predeterminadas. La toma de decisiones final depende completamente de las autoridades académicas y las escuelas de tutoría de la UPCH para integrar los resultados de este estudio con su conocimiento institucional, entrevistas con estudiantes y otras herramientas validadas de evaluación diagnóstica. Cualquier aplicación futura del sistema se basará en una revisión independiente de la equidad y el sesgo algorítmico por parte de grupos vulnerables de estudiantes, actualizaciones periódicas del modelo en respuesta a cambios institucionales, procesos de apelación para estudiantes que deseen impugnar su clasificación de riesgo, y revisión continua del impacto de tales intervenciones.

5.6.- Diseño de investigación

El diseño del estudio es aplicado y cuantitativo: no solo busca observar el fenómeno de la deserción académica, sino también entender la situación en la universidad misma. La investigación se basa en datos administrativos reales y en cohortes continuas sin cambios en las trayectorias de los estudiantes. Este enfoque observacional cumple con el propósito del proyecto: anticipar la no continuidad utilizando lo que la universidad ya tiene al final de cada período.

El objetivo operativo es identificar a los estudiantes que podrían no estar en el próximo ciclo en una fecha posterior para que se pueda iniciar la tutoría y el apoyo de manera oportuna. Se realizó una gran cantidad de investigación para asegurar que el modelo utilice datos limpios, esté ordenado en el tiempo y no filtre los resultados. Se aseguró la consistencia de claves, rangos y duplicados. El modelo fue evaluado externamente con la cohorte 2024-1 sin reentrenamiento, con datos para evaluar su comportamiento frente a nueva información.

El diagrama de casos de uso (Figura 4) muestra el sistema en acción, lo que demuestra cómo opera la herramienta, qué puede hacer cada perfil y qué límites tiene. Que la herramienta no esté disponible públicamente se debe a que es solo para tutores, analistas y autoridades. Las acciones habilitadas en el prototipo pueden ser iniciar sesión, ver el tablero, leer alertas, estudiar explicaciones del modelo y generar informes.

El control por perfil es único en este modelo con el propósito de proteger datos sensibles y asegurar que los usuarios puedan ver información que sea relevante para ellos. Que las acciones sean aprobadas y realizadas en sesiones seguras demuestra que la herramienta no toma decisiones; está organizada y presentada para que quienes trabajan en la sala con los estudiantes puedan actuar de manera más efectiva. También muestra las acciones que el prototipo habilitó: iniciar sesión, revisar el tablero, consultar alertas, explorar explicaciones del modelo y generar informes.

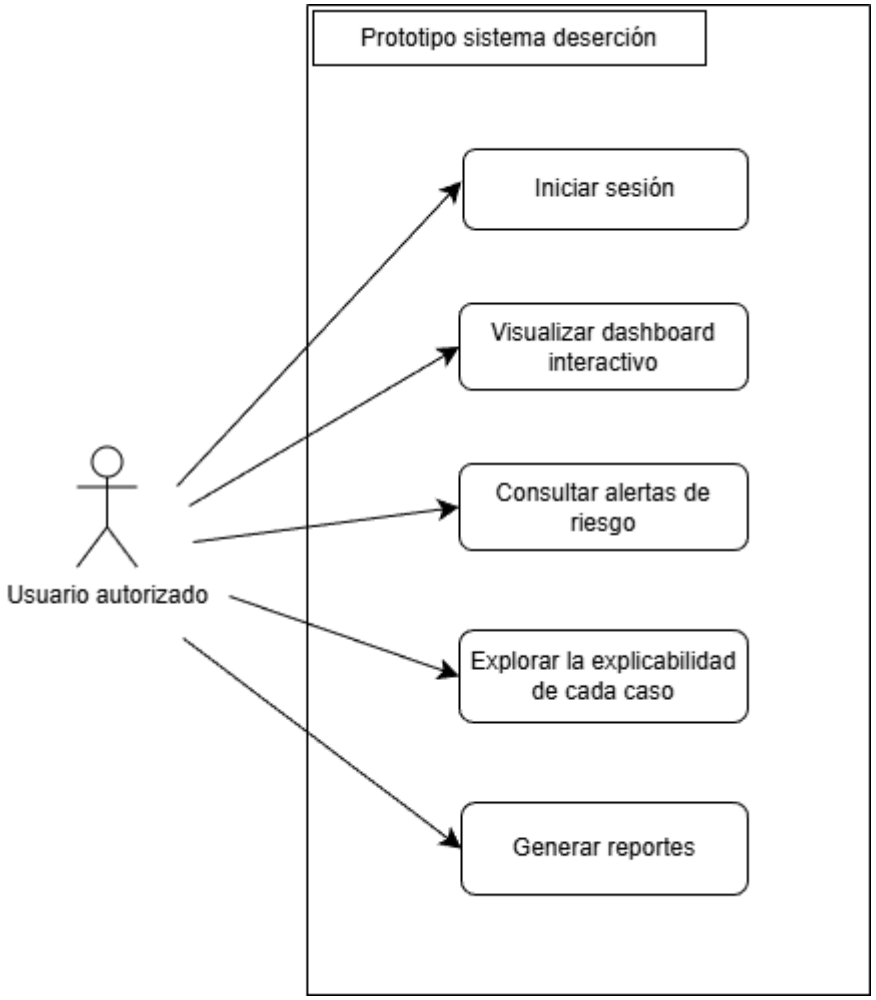


Figura 4 *.- Diagrama de Caso de Uso del Sistema Predictivo.*

El diagrama de caso de uso ayuda a mostrar, de manera sencilla, cómo una persona interactúa con el sistema. Aunque el proyecto tiene partes técnicas, al final la herramienta está pensada para quienes trabajan directamente con estudiantes y necesitan información rápida y confiable. Por eso es importante explicar, antes que nada, qué puede hacer cada usuario y hasta dónde llega su acceso.

El diagrama de casos de uso puede ayudarnos a ver, de manera sencilla, cómo un usuario interactúa con el sistema. La herramienta es técnicamente sobre los aspectos técnicos del proyecto, pero al final, la herramienta es para los estudiantes que realmente necesitan obtener un buen tiempo y la mejor información. Por lo tanto, es importante compartir, ante todo, lo que un individuo puede hacer y hasta dónde puede acceder al sistema.

El diagrama muestra el flujo básico: iniciar sesión, entrar al tablero, revisar alertas y, si es necesario, abrir la explicación de cada caso o generar un informe. No es un proceso de configuración y operación (no es un sistema que recoge cosas y trabaja por sí solo) y solo organiza y muestra la información; no es un robot que trabaja al final del día, y la toma de decisiones depende de tutores y analistas. El diagrama muestra cómo se forman las relaciones sin tener que pasar por explicaciones técnicas, y vemos cómo se implementan las medidas de privacidad y control en otras partes del documento. Cada acción del usuario es verificada, y se llevan a cabo sesiones seguras, lo cual es importante para la información académica. El diagrama resume el sistema desde el punto de vista del usuario de una manera que facilita entender lo que hace en el proyecto.

5.6.1.- Despliegue y operación

El diagrama de arquitectura de bloques (Figura 5) muestra la estructura interna del prototipo (desde el momento en que el usuario accede para recibir una predicción o un informe) y el flujo de información desde el usuario hasta los datos.

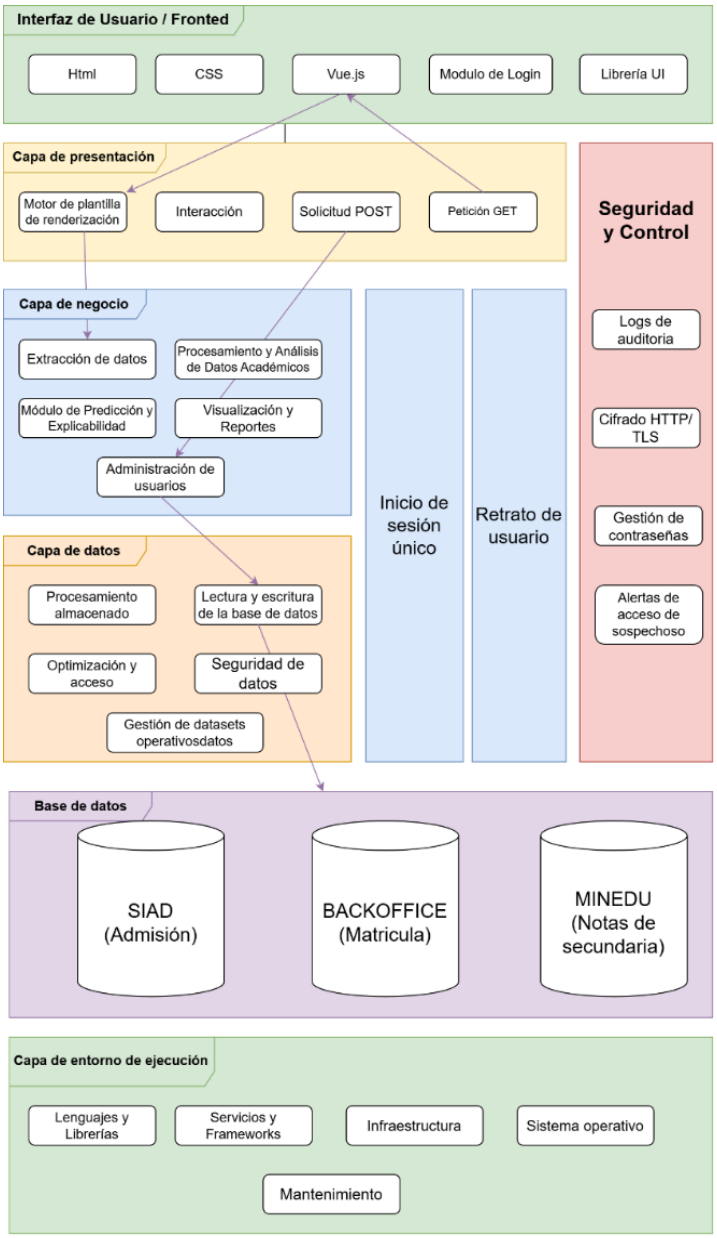


Figura 5.- *Diagrama de Arquitectura por bloques.*

La arquitectura está organizada en capas que tienen diferentes propósitos. En la capa superior se encuentra la interfaz de usuario: el formulario de acceso, el panel de control y los informes que se deben acceder sin ninguna interacción con los módulos internos. En esta interfaz están las capas de presentación y las capas de negocio que siguen al usuario y la organización de la información para recibir (en qué orden) y de qué fuente (a solicitud del usuario).

La capa de datos es lo que se almacena en los datos: SIAD, Backoffice y

registros escolares de Minedu. Los datos no ingresan directamente al modelo, sino que pasan por el proceso de limpieza, validación y estandarización descrito en la sección 5.11. Esta es la etapa más compleja y desafiante del modelo y es un paso necesario para que la predicción sea confiable. Al hacerlo, el modelo está protegido de la exposición directa con la información de los estudiantes y la integridad del sistema está asegurada contra consultas no seguras. Hay capas de seguridad en cada nivel: gestión de sesiones con tokens JWT, control de roles, conexiones cifradas con HTTPS y registros de auditoría en cada consulta.

Todo acceso está restringido a los usuarios autorizados, y los registros de auditoría están disponibles para verificar en cualquier revisión que usamos las alertas sabiamente. El módulo de mantenimiento y el entorno de ejecución completan la arquitectura. Incluyen los servicios que conectan el panel de control y la API y el esquema de monitoreo para detectar la degradación del modelo debido a cambios en los datos. Actualmente, esto no está implementado en nuestro prototipo, pero lo introduciremos en el futuro.

5.6.2.- Diagrama de arquitectura del sistema MLOps

Se diseñó un diagrama completo de arquitectura del sistema MLOps para que el Sistema Predictivo para la Detección Temprana de Riesgo Académico no quede solo en el lado técnico, esto ayuda a la traducción de la lógica compleja de nuestro prototipo llevándolo a un enfoque de su funcionamiento real.

El diagrama de arquitectura muestra el flujo completo de la interacción entre los componentes técnicos y los datos.

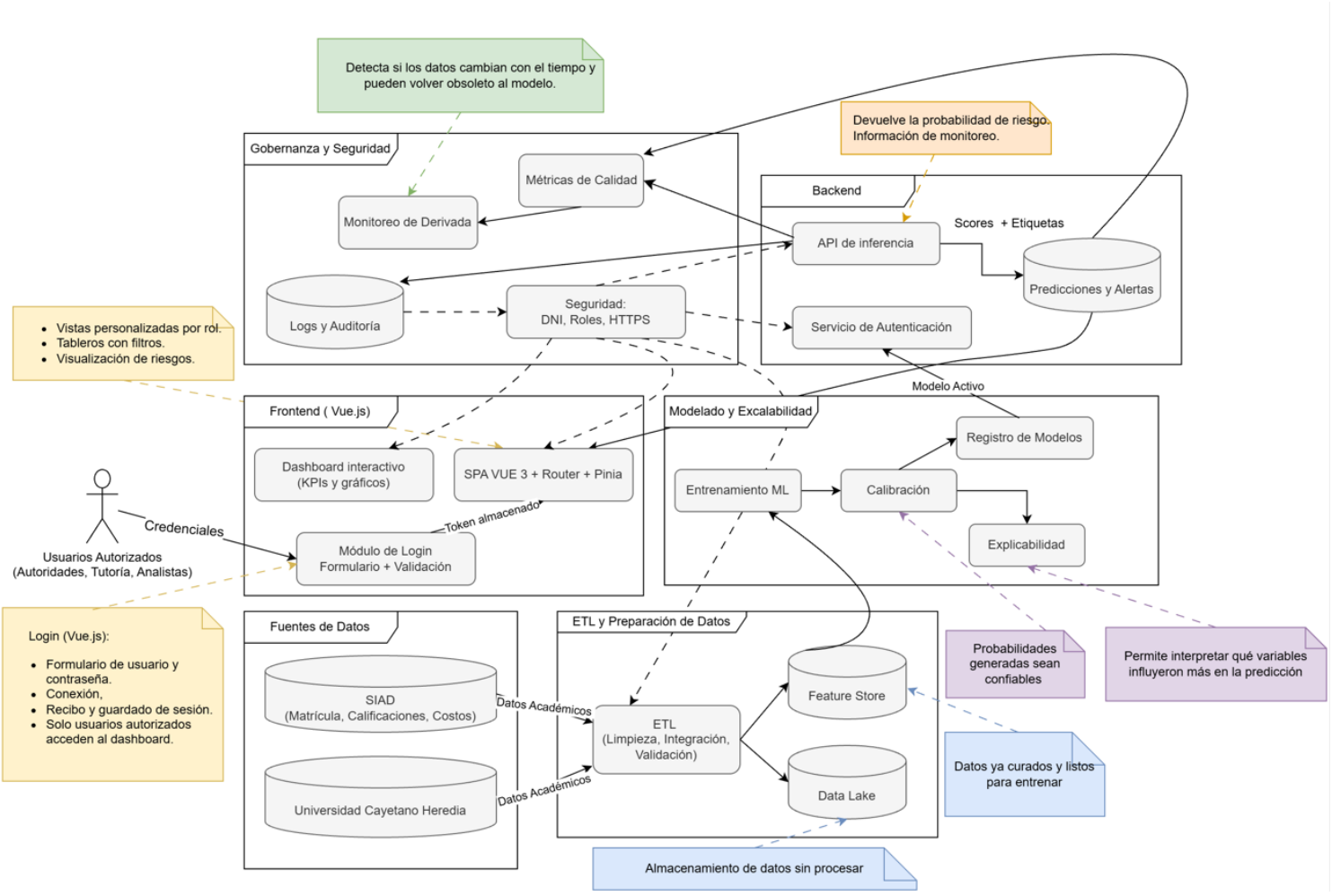


Figura 6.- *Diagrama de Arquitectura General del Sistema MLOps para la Detección Temprana de Riesgo Académico*

Los “Usuarios autorizados” (a la izquierda) interactúan con el “Frontend”, es último la parte visible y fundamental del sistema ya que aquí se realizarán las interacciones. El “Frontend” incluye un módulo de Login que está asegurado mediante un Token generado en el inicio de sesión gracias al servicio de autenticación, la seguridad también varía entre los DNI, tipo de Roles y cifrado HTTPS para asegurarse de la validación del usuario. Además, el “Frontend” cuenta con un dashboard para la visualización de gráficos, riesgos y KPI’s que ayuda a la facilitación del entendimiento de los riesgos de deserción.

Las “Fuentes de datos” (abajo - izquierda) del SIAD y la UPCH serán quienes nos proveerán los datos académicos, para posteriormente realizar un proceso ETL para la

respectiva limpieza, integración y validación de los datos. Como ya se mencionó, el proceso ETL (abajo – derecha) se encargará del procesamiento y preparación de los datos sin procesar, el "Feature Store" y "Data Lake" son resultados del proceso ETL, siendo ambos almacenados. Se utilizarán los datos resultantes del "Feature Store" para el entrenamiento ML, los nuevos resultados pasan seguidamente por la calibración (para que las probabilidades sean confiables) es quien alimenta esencialmente la generación del "Modelo activo", almacenado en los registros de los modelos, y la Explicabilidad, que consiste en datos que permiten interpretar que variables influyen más en el modelo de predicción.

En el "Backend" se encuentran las API de Inferencia que utiliza el "Modelo activo" y envían Scores y Etiquetas como "Predicciones y Alertas".

El "Frontend" hace un llamado a la API de Inferencia para su inmediata visualización cuando el usuario realiza consultas de "Predicciones y Alertas" al modelo, a lo que la API responde visualizando dicha consulta. Las API de Inferencia junto con las Predicciones y Alertas también son enviadas y almacenadas como métricas de calidad, luego son medidas como un "Monitoreo de Derivada" (El modelo se monitorea a sí mismo y detecta si los datos cambian con el tiempo y si se vuelve obsoleto el sistema). Por sí solo, la API de Inferencia es enviada y genera Logs y Auditoria, para ser transformadas en métricas de seguridad que alimentan el módulo. Se integra con el Servicio de Autenticación.

5.7.- Periodo, gobernanza y seguridad

El estudio se desarrolla en la Universidad Peruana Cayetano Heredia (UPCH) y se focaliza en estudiantes de primeros ciclos de pregrado, con énfasis en la detección temprana del riesgo de no continuidad para orientar acciones de tutoría, consejería y apoyos.

5.7.1.- Periodo de estudio

El desarrollo del modelo utilizó datos de los periodos 2022-1, 2022-2, 2023-1 y 2023-2 como histórico para el entrenamiento y la validación interna, respetando la temporalidad de los datos. El grupo correspondiente al cohorte 2024-1 se mantuvo como un conjunto de validación externa sin reentrenamiento, para

determinar la capacidad de generalización temporal del sistema actual.

5.7.2.- Responsables y gobernanza de datos

OAMRA mantiene la propiedad y curaduría de los datos a través del SIAD (PostgreSQL) que contiene información sobre el estado de los estudiantes y su situación académica, y el Backoffice (SQL Server) que registra el historial de matrículas, calificaciones y créditos por período.

5.7.3.- Gobernanza y trazabilidad

La gobernanza del sistema se basa en cuatro componentes. Los estudiantes son seudonimizados utilizando funciones hash para proteger su identidad. El acceso a los datos se gestiona mediante un esquema basado en roles con el menor privilegio, de modo que cada usuario obtiene la información que necesita. Todos los pasos de extracción, transformación y consulta se registran en logs auditables. Un diccionario de datos documenta la definición, reglas de integración y versión de cada variable a lo largo del proceso.

5.7.4.- Seguridad y privacidad

La comunicación entre los componentes del sistema utiliza cifrado en tránsito a través de HTTPS/TLS. La autenticación y autorización se realizan utilizando tokens JWT con expiración y rotación de claves regularmente. Los registros de acceso y eventos se mantienen durante al menos 12 meses para fines de monitoreo y auditoría. No se exportan datos nominativos fuera del entorno seguro de la universidad.

5.8.- Población y muestra

La población objetivo son los estudiantes de pregrado en los primeros ciclos de la Universidad Peruana Cayetano Heredia (UPCH). El marco muestral se construyó a

partir de registros administrativos de SIAD y Backoffice y está organizado por estudiante y período, de modo que cada estudiante pueda seguir su trayectoria a lo largo del tiempo. Para desarrollar el modelo, se utilizó un censo administrativo desde 2022-1 hasta 2023-2; siempre con emparejamientos t→t+1 (por ejemplo, 2022-1→2022-2). La cohorte 2024-1 se probó externamente como un conjunto independiente (sin reentrenamiento). De esta manera, podemos estimar el potencial de nuestro modelo estadístico para generalizar a un escenario real.

La variable dependiente "no continuidad en t+1" es una variable binaria; es 1 si el estudiante no se matricula para el siguiente período y 0 si se matricula. Siempre verificamos contra el corte oficial de t+1. Para prevenir la fuga de tiempo, cada predicción en t solo utiliza datos disponibles al final de t. Se respeta el calendario académico y si hay algún cambio, se sigue la regla del "próximo período inmediato". Se implementaron reglas de calidad de datos: matrículas regulares, identificador único seudonimizado y variables mínimas (ID, programa, período, trayectoria). Se excluyeron duplicados irresolubles, inconsistencias graves de fecha o rango, validaciones atípicas que impiden el seguimiento y datos estructuralmente faltantes. Las reentradas fuera del calendario y las licencias formales no se etiquetan como "no continuidad" en ese salto y no se consideran en el cálculo de la etiqueta para ese emparejamiento. Cualquier caso no verificado en t+1 se documenta y se excluye en el cálculo de la variable en ese salto.

Este diseño alinea el modelo con la decisión operativa, ya que ayuda a identificar a aquellos que no permanecerán en el próximo período real y prioriza las intervenciones. La validación externa con 2024-1 también ayudará a fortalecer la robustez de los resultados y evitar el sobreajuste a los resultados.

La distribución inicial considerada en el prototipo operativo sobre la población de estudio según su nivel de riesgo se encuentra en la tabla 6. Es importante aclarar la relación entre las diferentes cifras de población que aparecen en este documento: el Conjunto de Datos Maestro longitudinal (sección 6.3) contiene 17,712 registros, correspondientes a múltiples observaciones por estudiante a lo largo de los años 2022-1 a 2024-1 (cada fila representa un par estudiante-período).

Los 3,033 estudiantes de la Tabla 6 corresponden al universo de estudiantes únicos del prototipo operativo, tras aplicar los filtros de calidad de datos y de 2 ciclos consecutivos. Finalmente, los N=756 de la validación externa (sección 6.7) corresponden al subconjunto de estudiantes del periodo 2024-1 pertenecientes a las tres facultades piloto (Enfermería, Veterinaria y Ciencias) que superaron los filtros de calidad del pipeline OCR. El siguiente diagrama resume el procedimiento de depuración que condujo a cada cifra:

Procedimiento de depuración muestral (Figura 7):

- Universo inicial: Total de matrículas en UPCH periodos 2022-1 a 2024-1 (SIAD + Backoffice).
- Filtro 1 — 2 ciclos consecutivos y calidad de datos: Se aplican las reglas de inclusión descritas (sin licencias ni traslados), produciendo 17,712 registros longitudinales (Dataset Maestro, múltiples filas por estudiante). El número de estudiantes únicos en este dataset es el universo de modelado.
- Filtro 2 — Prototipo operativo (3,033 estudiantes, Tabla 6): Subconjunto del Dataset Maestro correspondiente al corte institucional del modelo inicial, tras validación OCR y completitud de variables clave.
- Filtro 3 — Validación externa 2024-1 (N=756): Corresponde al censo de ingresantes del periodo 2024-1 en las facultades piloto (Enfermería, Veterinaria, Ciencias) con expedientes de datos completos. Del total de 4,176 registros para el período 2024-1 en el Conjunto de Datos Maestro, 756 pertenecen a estas tres facultades y superan los filtros de calidad del pipeline. Este subgrupo se utiliza para la validación externa.

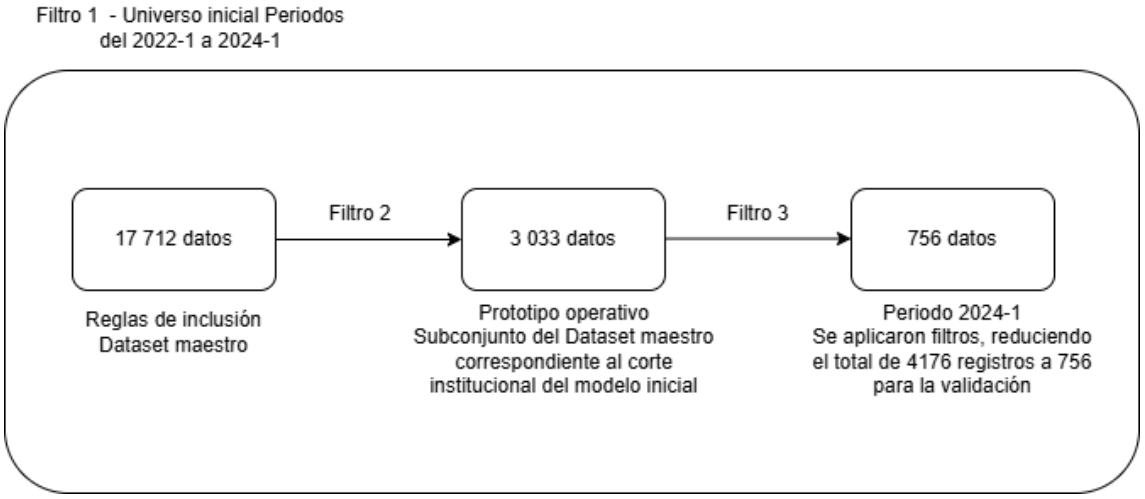


Figura 7.- *Procedimiento de depuración de la muestra.*

Observamos que el primer volumen para la automatización (web scraping y descargas masivas) fue de más de 24,000 registros, pero debido a problemas técnicos como registros duplicados, hubo una gran limpieza de alrededor de 3,000 registros. Esta pérdida de datos se mitigó al tener reglas de inclusión estrictas para ese conjunto de datos (17,712 registros del conjunto de datos maestro y una muestra de 3,033 estudiantes).

La distribución inicial considerada en el prototipo operativo sobre la población de estudio según su nivel de riesgo se muestra en la tabla 6.

Categoría	N	Porcentaje
Total de estudiantes	3033	100.00%
En riesgo	676	22.28%
Sin riesgo	2357	77.72%

Tabla 6*.- Universo y distribución de riesgo (prototipo operativo)*

5.9.- Fuente de datos, extracción y seguridad

5.9.1.- Orígenes de información

El estudio utiliza datos institucionales provenientes de dos fuentes principales:

- PostgreSQL y SIAD: historial de admisión, estado de admisión y modalidad de estudio o beca.
- SQL Server y Backoffice: historial académico y de matrícula y créditos

tomados y aprobados durante cada período.

Ambas fuentes se integran para formar un conjunto de datos de modelado, estructurado en formato tabular (CSV o Parquet). Cada uno está representado en su diccionario de datos (con definiciones, tipos y reglas de cada variable) y una huella digital (hash + fecha de corte) que asegura la trazabilidad y autenticidad de la extracción.

5.9.2.- Extracción y control de las versiones

El proceso de extracción se realiza utilizando scripts de SQL y Python que están versionados en Git para mantener un historial de cambios y asegurar la reproducibilidad. Cada ventana de corte está documentada con su fecha, descripción y responsable técnico. El acceso al entorno de datos se guía por un enfoque de control basado en roles de menor privilegio y trazabilidad.

5.9.3.- Tamaño y estructura del dataset

El conjunto de datos en el estudio de modelado corresponde a un entorno dimensionado de aproximadamente 3,033 estudiantes, que puede escalarse de acuerdo con el corte institucional final. La estructura tabular incluye variables académicas, administrativas y derivadas que podemos utilizar para análisis estadístico y aprendizaje automático.

Los tipos de variables académicas y sus fuentes se describen en la tabla 7

Fuente	Variable / Grupo	Descripción	Tipo	Escala de medición
SIAD	Modalidad de ingreso	Tipo de modalidad del estudiante (Presencial, Virtual, Mixta)	Categórica	nominal
Backoffice	cred_aprob_t	Créditos aprobados en el periodo t	Numérica	razón
Backoffice	prom_t	Promedio	Numérica	razón

		ponderado en el periodo t		
	Continuidad	(1= no matrícula, 0 = matrícula)	Categórica	nominal

Tabla 7.- Análisis de variables según su fuente

5.10.- Variables e indicadores

5.10.1.- Variable dependiente (Objetivo del modelo)

La variable dependiente es la no continuidad en t+1 y se marca como 1 cuando el estudiante no se inscribe para el siguiente período y 0 si lo hace. Representa el riesgo de interrupción académica en el próximo período y es la etiqueta binaria que el modelo busca predecir.

5.10.2.- Variables independientes (Predictores)

Nuestras variables predictoras se basan en registros administrativos y académicos institucionales y se dividen en tres grupos: las variables académicas representan el rendimiento numérico del estudiante en la escuela secundaria y la universidad e incluyen: promedio de calificaciones de secundaria en matemáticas y comunicación de OCR, puntaje de admisión, promedio ponderado del período actual (prom_t), número de créditos aprobados en el semestre (cred_aprob_t), porcentaje acumulado de créditos aprobados (porc_aprob_acum), número de cursos reprobados, y el indicador binario de fracaso en materias críticas.

Las variables sociodemográficas describen el perfil personal, económico y de origen del estudiante: edad, género, origen geográfico, estatus de primera generación, tiempo promedio de viaje al campus, estatus de beca, costo relativo o escala de pago asignada. Las variables de contexto institucional describen el entorno académico y administrativo en el que se encuentra el estudiante (facultad de pertenencia, programa académico, modo de ingreso y modo de estudio ya sea presencial, virtual o mixto). Las variables obtenidas del proceso de ingeniería de

características durante la fase de análisis no son una categoría independiente ya que las consideramos un producto natural del preprocesamiento.

5.10.3.- Función de costo y prioridad de clasificación

El sistema asigna un costo más alto a los errores de omisión que a las falsas alarmas. Tenemos FN ≫ FP, donde FN son los casos de riesgo real que no son detectados por el modelo, y FP son alertas que no representan un riesgo real al ser revisadas. La función de costo que opera con esta prioridad es la siguiente:

$$C(\tau)=cFN \cdot FN(\tau)+cFP \cdot FP(\tau)$$

- **$C(\tau)$** es el costo total esperado al operar con el umbral τ.
- **$FN(\tau)$** es el número de falsos negativos en ese umbral (casos de riesgo real que el sistema no alertó).
- **$FP(\tau)$** es el número de falsos positivos en ese umbral (alertas que, al ser revisadas, no eran un riesgo real).

Donde $C(\tau)$ es el costo total esperado con el umbral τ, $FN(\tau)$ es el número de falsos negativos en el umbral τ, y $FP(\tau)$ es el número de falsos positivos. En este escenario, el modelo está diseñado para ser más preciso, cubriendo la mayoría de los casos de riesgo real, y para permitir un pequeño aumento en los falsos positivos según sea necesario.

Diariamente, esto se traduce en dos ajustes. El umbral (o política de top-k) se establece para asegurar una alta cobertura de la clase en riesgo, con un objetivo de recall establecido con el equipo de tutoría. Al mismo tiempo, la precisión se mantiene solo en un mínimo para mantener la carga de trabajo manejable. Cuando las capacidades de atención o los perfiles de datos cambian, el umbral se ajusta y se registra el nuevo punto de operación. Esta función de costo se incorporó directamente en el entrenamiento del modelo a través del ponderado de clases, no como un post-procesamiento manual.

La Tabla 8 describe la medición, descripción de variables y su operacionalización..

Dependiente o Independiente	Variable	Definición	Tipo de variable	Escala de medición
Dependiente	y_no_continua	Toma el valor de 1 si el estudiante no se matricula en el periodo siguiente (t+1) o cae en rendimiento crítico. 0 si continúa de forma regular.	Categórica Dicotómica	Nominal
Independiente	prom_t	Promedio ponderado de notas obtenido por el estudiante específicamente en el periodo evaluado (t).	Numérica Continua	Razón
Independiente	cred_aprob_t	Suma total de créditos universitarios que el estudiante logró aprobar en el periodo t.	Numérica Discreta	Razón
Independiente	porc_aprob_acum	Porcentaje acumulado de créditos aprobados respecto al total cursado hasta el periodo t.	Numérica Continua	Razón
Independiente	modalidad	Modalidad por la cual el estudiante ingresó a la universidad	Categórica	Nominal
Independiente	score_total	Promedio estandarizado de las calificaciones de secundaria extraídas automáticamente de los PDFs.	Numérica Continua	Razón
Independiente	nota_admision	Ranking o posición de mérito obtenida por el estudiante durante su proceso de ingreso.	Numérica Discreta	Razón
Independiente	institucion_privada	Tipo de gestión del colegio de procedencia escolar (1=Privada, 0=Pública)	Categórica Dicotómica	Nominal
Independiente	facultad / programa	Carrera o área académica del estudiante.	Categórica	Nominal
Independiente	costo_relativo/ valor_venta	Costo económico del ciclo o pensión asignada, utilizado como punto fijo del soporte financiero del estudiante.	Numérica Continua	Razón

Tabla 8.- Medición, descripción de variables y operacionalización

5.11.- ETL y preparación de datos

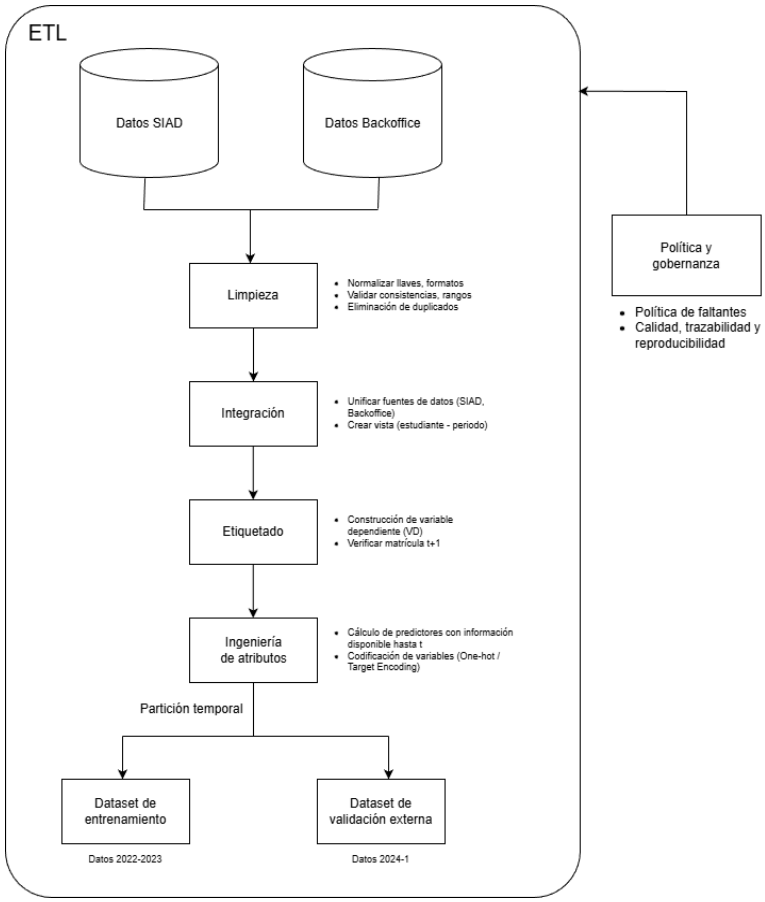


Figura 7.- Proceso ETL desarrollado en la tesis

Etapa	Descripción
Limpieza	Las claves y formatos se normalizan para garantizar la unicidad del estudiante-periodo. Los códigos de facultad, programa y modalidad se normalizan, se corrigen las tipografías y espacios, y se eliminan los duplicados. Luego se validan en base a rangos y consistencias, como calificaciones en intervalos permitidos o créditos que no sean negativos. Esto previene el doble conteo y valores imposibles que impactarían negativamente en el entrenamiento del sistema.
Integración	Las fuentes SIAD y Backoffice se vinculan utilizando una clave codificada que integra toda la información de cada periodo del estudiante en una sola fila. Cuando hay inconsistencias, se utiliza una regla de prioridad y la decisión se anota en el registro. Esto crea una vista separada, coherente y trazable por periodo de estudiante.

Etiquetado	La variable dependiente se construye de acuerdo con la secuencia temporal. Se verifica la matrícula en el periodo t para cada registro. Si no hay matrícula, DV es 1 (sin continuidad). Si hay matrícula, DV es 0. Después de que el periodo t termina, no utilizamos información para que no haya fuga temporal y simulamos un escenario de predicción real.
Ingeniería de atributos	Todos los predictores se generan solamente con datos disponibles hasta el periodo t. Se consideran métricas del rendimiento del periodo, acumulados históricos, variaciones entre periodos y variables de contexto institucional. Las variables categóricas son codificados con one-hot en modelos lineales, pero con target encoding en modelos ensemble. Las variables numéricas se estandarizan según la necesidad que tenga el algoritmo y los valores extremos se tratan con cuidado para no distorsionar patrones reales.
Partición temporal	El modelo se entrena con los periodos 2022-1, 2022-2, 2023-1 y 2023-2. La cohorte correspondiente al periodo 2024-1 se deja como hold-out independiente, sin ninguna clase de reentrenamiento, para poder evaluar la generalización temporal. Durante el entrenamiento se usa validación cruzada sin mezcla (sin shuffle) y se fijan semillas para asegurar reproducibilidad.
Política de faltantes	En variables críticas (prom_t, cred_aprob_t) los registros se excluyen si el vacío no es resoluble. En caso lo sea se aplica imputación conservadora. Para variables no críticas se usa imputación simple (mediana/moda) o KNN cuando los vacíos son < 10%. Toda imputación genera una bandera binaria para auditar el efecto que tiene.
Calidad, trazabilidad y reproducibilidad	Cada recorrido del proceso ETL produce una bitácora con conteos pre/post, reglas aplicadas y exclusiones. Se mantiene un diccionario de datos con definición, tipo, escala y momento de cada una de las variables. Se calculan fingerprints del dataset (filas, columnas, checksum), junto con versión del código y fecha de ejecución. Esto permite replicar resultados, monitorear drift (degradación del rendimiento) y sostener auditorías completas.

Tabla 10.- Descripción del proceso ETL

5.11.1 Justificación del manejo de datos faltantes e imputación

Como se detalló en la política de datos faltantes de la tabla 10, la reducción drástica de muestra exigió un manejo estricto de los datos vacíos que quedaron. Ante la restricción de la muestra final, a 3033 registros, se estableció que la eliminación de más celdas nulas podría comprometer el volumen de datos que se poseía y la factibilidad del entrenamiento final del modelo, entonces, para mitigar esta pérdida de información y mantener la integridad de los registros, se incorporó en el pipeline un análisis de imputación de datos utilizando la mediana.

Esta selección fue utilizada para el tratamiento de variables numéricas, como calificaciones, debido a su robustez frente a valores atípicos. A comparación de la media aritmética, la mediana asegura que la medida de la tendencia no tenga inconsistencias debido a valores extremos irreales, como errores de lectura del OCR, fallos en el scraping o ceros por inasistencias. Este procedimiento, integrado con los cuadernos de Jupyter para la preparación de datos, aseguró que algoritmos como Random Forest y XGBoost fueran entrenados en una matriz de características estable y estadísticamente consistente.

5.12.- Análisis exploratorio (EDA)

El análisis exploratorio se basó en la distribución del evento, la calidad de los datos y las relaciones entre variables con respecto al período y subgrupos institucionales. Para evaluar el desequilibrio, se calculó el porcentaje de VD=1 (valor de no continuidad) para el período y globalmente con el fin de dimensionar el problema analizado y poder calibrar las estrategias de umbral/top-k en los datos. En términos de preprocesamiento y manejo de valores, dado que estos son los datos con valores en entornos de análisis reales, en este caso el modelo fue seleccionado en un modelo cuantitativo de cuartiles (25%, 50%, 75%) para separar a los estudiantes en grupos iguales. El análisis EDA del modelo se mostró en las gráficas (1, 2 y 3).

	edad	NOTA_PONDERADA_PRIMER_AÑO	ARTE Y CULTURA	CIENCIA Y TECNOLOGÍA	CIENCIAS SOCIALES	COMUNICACIÓN
count	3030.000000	3033.000000	3033.000000	3033.000000	3033.000000	3033.000000
mean	19.864026	14.163222	16.603815	15.737145	15.923469	15.710811
std	1.537469	2.053440	1.504921	1.791395	1.739711	1.710085
min	17.000000	2.884706	11.330000	10.670000	9.780000	10.670000
25%	19.000000	13.154667	15.700000	14.400000	14.730000	14.470000
50%	20.000000	14.409737	16.700000	15.800000	16.000000	15.800000
75%	20.000000	15.541500	17.600000	17.000000	17.200000	17.000000
max	42.000000	18.878889	20.000000	20.000000	20.000000	20.000000

Gráfico 1: *Extracción de la tabla del Análisis de EDA del modelo – parte 1*

DESARROLLO PERSONAL, CIUDADANÍA Y CÍVICA	EDUCACIÓN FÍSICA	EDUCACIÓN PARA EL TRABAJO	EDUCACIÓN RELIGIOSA	INGLÉS	MATEMÁTICA	estudiante_primera_generacion
3033.000000	3033.000000	3033.000000	3033.000000	3033.000000	3033.000000	3029.000000
16.274929	16.352295	16.294243	16.565542	15.958203	15.581860	0.583031
1.594739	1.413158	1.581450	1.569171	1.834755	1.983631	0.493139
10.000000	11.330000	11.200000	11.200000	10.000000	10.000000	0.000000
15.200000	15.400000	15.200000	15.500000	14.670000	14.100000	0.000000
16.400000	16.400000	16.400000	16.670000	16.000000	15.600000	1.000000
17.400000	17.330000	17.400000	17.700000	17.270000	17.050000	1.000000
20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	1.000000

Gráfico 2: *Extracción de la tabla del Análisis de EDA del modelo – parte 2*

pos_calificacion	ESTUDIANTE_VALOR_VENTA	minutos_a_SMP	CARGO
3033.000000	3024.000000	3033.000000	3033.000000
55.574280	9684.696772	188.615232	12080.123267
21.946924	8164.709070	158.361117	17477.950025
0.000000	0.000000	10.000000	357.500000
38.000000	0.000000	35.000000	6697.000000
50.600000	7870.000000	90.000000	9867.500000
73.000000	15965.000000	360.000000	13215.000000
510.900000	50072.500000	360.000000	473817.500000

Gráfico 3*: Extracción de la tabla del Análisis de EDA del modelo – parte 3*

Estos datos son esenciales debido a que permite a encontrar la mediana (obtenida en el 50%), el cuál será nuestro “alumno general”. Por otro lado, también nos darán los límites inferiores y superiores (25% y 75% respectivamente), lo que da como significado que, aquellos alumnos que estén fuera de dichos límites son los de valores “atípicos”. Se calcularon correlaciones entre variables numéricas y se realizaron tablas cruzadas por modalidad y facultad (tasas de VD obtenidas y diferencias porcentuales), a fin de anticipar posibles sesgos o segmentos de mayor riesgo. Se compararon las tasas de VD por facultad, programa y modalidad. Los hallazgos resultantes guiaron el

ajuste posterior de umbral por segmento o la priorización en la ejecución.

5.13.- Modelado y validación

5.13.1.- Definición funcional del modelo predictivo

Variables de entrada (18 predictores): El vector X incluye variables académicas universitarias (prom_t, cred_aprob_t, porc_aprob_acum, cursos_desaprob, cursos_criticos), preuniversitarias extraídas por OCR (nota_secundaria_mat, nota_secundaria_com), de admisión (nota_admision, tipo_admision_Encoded), socioeconómicas (costo_relativo, condicion_beca) y contextuales (facultad_Encoded, modalidad_estudio, cluster_estudiante).

La función de predicción final del modelo en XGBoost: El modelo de conjunto utiliza la función $\hat{y} = \sigma(\sum_{k=1}^{k} f_k(X))$ para la predicción aditiva sobre K árboles de decisión, donde f_k es el k-ésimo árbol de regresión del ensemble, y σ es la función sigmoide $\sigma(z) = 1/(1 + e^{-}z)$ que transforma la salida en una probabilidad $P(y=1|X) \in [0, 1]$. El modelo de línea base (Regresión Logística) utiliza la función: $P(y=1|X) = 1 / (1 + e^{-}(\beta_0 + \beta_1 \cdot prom\_t + \beta_2 \cdot cred\_aprob\_t + ... + \beta_n \cdot X_n))$.

La regla de decisión y los niveles de riesgo: La predicción final se realiza considerando un umbral $\tau = 0.45$ sobre la probabilidad estimada. La alerta es 1 si $P(y=1|X) \geq 0.45$ y 0 en caso contrario. Utilizamos la curva de Recall-Precision (Tabla 25) para calcular el nivel de riesgo y consideramos los falsos negativos más bajos (FN ≫ FP) como un mínimo. Los niveles de riesgo operativo son bajos (<0.25), medio (0.25–0.45), alto (0.45–0.65) y crítico (>0.65). La interpretabilidad local se realiza mediante valores SHAP, que son los componentes de cada variable en la predicción de cada estudiante.

El modelo comparó una línea base interpretable (Regresión Logística) con métodos de conjunto (Random Forest y XGBoost) que capturan no linealidades. El desequilibrio se corrigió utilizando class_weight (LR/RF) o scale_pos_weight (XGB), y la decisión se tomó para cada umbral o top-k según la capacidad de intervención (criterio FN≫FP). La búsqueda de hiperparámetros se realizó en Random Search/Optuna con un presupuesto y detención temprana en XGB. La validación se realizó ya que el tiempo era no lineal. TimeSeriesSplit en desarrollo y validación externa en el periodo

2024-1 sin reentrenar. Se fijaron semillas y se versionaron código/datos para su reproducibilidad.

La explicación de los hiperparámetros, estrategias, validación y tratamiento del desbalance de los modelos se representan en la tabla 11, mientras que la importancia, señalización de ajuste por parámetro y tipo de control se describen a detalle en la tabla 12.

Modelo	Hiperparámetros (rango)	Estrategia	Tratamiento desbalance	Validación
Logística	C∈[0.1,10], penalty=l2	Grid/Random	class_weight=balanced	CV temporal + 2024-1
Random Forest	n_estim∈[300,800], max_depth∈[6,18], min_samples_leaf∈[1,6]	Random/Optuna	class_weight=balanced o balanced_subsample	CV temporal + 2024-1
XGBoost	n_estim∈[400,900], max_depth∈[4,10], eta∈[0.02,0.2], subsample∈[0.6,1.0]	Random/Optuna	scale_pos_weight=neg/pos + early-stopping	CV temporal + 2024-1

Tabla 11.- Rangos, estrategias, validación y tratamiento de los modelos

Parámetro	Controla	Por qué importa	Señal de ajuste
C (LR)	Regularización L2	Evita sobreajuste, estabiliza coeficientes	**Underfit** → ↑C; **Overfit** → ↓C
n_estim (RF/XGB)	Nº de árboles/rondas	Reduce varianza / mejora hasta saturar	Subir hasta que no mejore OOF
max_depth (RF/XGB)	Complejidad de árbol	Profundo = riesgo de sobreajuste	Si overfit, bajar profundidad
min_samples_leaf (RF)	Tamaño de hojas	Suaviza fronteras con ruido	Subir con datos escasos
subsample (XGB)	Fracción por ronda	Regulariza y acelera	0.7–0.9 suele ser robusto

scale_pos_weight (XGB)	Balance del gradiente	Trata desbalance severo	≈ N_neg / N_pos en train

Tabla 12.- Importancia, ajuste por parámetro y tipo de control.

5.14.- Calibración y diagnóstico de probabilidad

La calibración ajusta los scores sin procesar del clasificador para que se interpreten como probabilidades reales. Se entrenó un calibrador sobre las predicciones out-of-fold del desarrollo (para evitar fuga temporal) usando Platt/sigmoid en escenarios con un número moderado de casos positivos, reservando la calibración isotónica para situaciones que demandaban una función de calibración más adaptable. El calibrador elegido se fija y luego se evalúa en la cohorte correspondiente al periodo 2024-1 (hold-out) sin reentrenar. El diagnóstico se hizo utilizando Brier Score (error cuadrático medio entre probabilidad y etiqueta) y con el diagrama de confiabilidad. Una calibración adecuada muestra pendiente ≈ 1 e intercepto ≈ 0. De forma complementaria puede reportarse el ECE (Expected Calibration Error) y revisar si el modelo es sobre-confiado (probabilidades demasiado altas) o sub-confiado. Cuando la calibración se degrada (p. Por ejemplo, Brier > 0.18 o pendiente muy distinta de 1), se recalibra manteniendo los pesos/umbrales que el clasificador poseía.

Tener probabilidades bien calibradas nos permite definir umbrales y políticas coherentes de top-k con capacidad operativa (FN ≫ FP); podemos construir bandas de riesgo comprensibles para los usuarios, y podemos monitorear su estabilidad en producción a lo largo del tiempo.

Concepto	¿Qué es?	¿Cómo se calcula / insumo?	Señal de buena calibración	Mitigación
Probabilidad calibrada	Score transformado para que refleje P(y=1\x)	Ajuste sobre **predicciones OOF** (desarrollo)	Riesgo interpretado "como probabilidad"	Recalibrar (Platt/Isotónica)
Brier Score	Error cuadrático medio entre p y	$\frac{1}{N}\Sigma(p_i - y_i)^2$	**Bajo** (p. por ejemplo,≤ 0.18))	Revisar drift, recalibrar

	y			
Curva de confiabilidad	Compara p vs. frecuencia observada por bins	Recta ideal (45°)	**Pendiente ≈ 1, intercepto ≈ 0**	Sobre/sub-confianza → recalibrar
Umbral / Top-k	Regla para decidir casos positivos	Elegido con probabilidades **calibradas**	Cumple Recall objetivo y capacidad	Ajustar con base en Brier/Recall

Tabla 13.- *Conceptos, mitigación y señales de buena calibración*

Método	Idea	Ventajas	Limitaciones	Criterio de utilización
Platt (sigmoid)	Aprende una función logística $P' = \sigma(ax + b)$	Estable, evita sobreajuste con pocos datos	Puede ser rígido si la relación no es sigmoide	Positivos moderados, curvas suaves
Isotónica	Ajuste monótono por tramos (no paramétrico)	Muy flexible, corrige formas complejas	Riesgo de sobreajuste con pocos bins	Suficientes datos; curva no lineal
Sin calibrar	Usar score original	Cero costo	Difícil interpretar/umbralizar	Solo como línea base

Tabla 14.- *Métodos, criterios de utilización y limitaciones*

5.15.- Métricas y reglas de decisión

El objetivo es cubrir tantas situaciones de riesgo real como sea posible, y aceptar revisar más falsos positivos que falsos negativos (prioridad FN ≫ FP). La toma de decisiones del sistema resuena con el concepto y con la capacidad real del equipo para apoyar. Para ser efectivos en un escenario desequilibrado, el panel se enfoca en cuatro señales. El Recall indica cuántos estudiantes en riesgo están siendo detectados por el sistema, y es la métrica más importante porque representa la cobertura. El F1 resume

en un solo número el equilibrio entre cobertura y ruido que puede compararse fácilmente entre diferentes versiones del modelo.

El PR-AUC captura el rendimiento a lo largo de toda la curva de precisión-recall y es más informativo que el ROC cuando la clase positiva es minoritaria. El Brier Score mide la fiabilidad de las probabilidades: cuanto menor es el valor, mejor está calibrado el modelo y más significativos son los umbrales. El ROC-AUC sigue siendo una buena referencia para comparar con la literatura o los modelos anteriores y no puede guiar las operaciones diarias. Las reglas de decisión se realizan en dos pasos. Primero se establece un umbral, observando la curva P-R y una tabla simple de costo-beneficio (que es el costo de no encontrar un caso real, pero revisar un falso positivo). Luego se aplica una política de top-k cuando el equipo tiene una cuota mensual para investigar (se ordena según la probabilidad y se consideran los k casos principales). Para mantener esto constante, revisamos la calibración regularmente, y si el Recall cae por debajo del objetivo acordado, lo recalibramos. En términos de monitoreo, si el Recall está más de 5 puntos porcentuales por debajo del objetivo, establecemos el umbral y lo reentrenamos si el problema continúa.

Todos los cambios se rastrean con fecha, versión del modelo y hash del conjunto de datos. Establecemos los criterios de aceptación para el Recall y el F1-Score en 0.70 y 0.65, respectivamente. Estos valores están en línea con los estándares en la literatura de Minería de Datos Educativos para problemas con gran desequilibrio de clases y ruido en el comportamiento humano donde un Recall superior al 70% se considera un alto rendimiento operativo para sistemas de alerta temprana sin sobreajuste.

Indicador	**Criterio de aceptación**	**Evidencia**
Recall (2024-1)	≥ 0.70	Matriz de confusión / reporte
F1 (2024-1)	≥ 0.65	Idem
Δ (train→2024-1)	≤ 5 %	Tabla comparativa / IC
Brier	≤ 0.18	Curva de confiabilidad

Tiempo de interpretación	< 2 min	Prueba de uso
Latencia /predict	< 2 s	Prueba de carga

Tabla 15.- *Criterios de aceptación mediante evidencia por cada indicador*

5.16.- Análisis de robustez y sesgos

Para garantizar que el modelo sea consistente y equitativo en diferentes instituciones, se evaluó el rendimiento del modelo para cada subgrupo (facultad, modalidad, cohorte y, cuando sea necesario, programa). Se estimaron Recall, F1, PR-AUC y Brier para cada subgrupo con intervalos de confianza del 95% utilizando bootstrap sobre la predicción de validez (CV temporal) y el hold-out (período 2024-1). También realizamos un análisis de sensibilidad de umbral, observando cómo Recall y Precisión son diferentes al mover el punto de decisión. Si el subgrupo requería un punto diferente para alcanzar el objetivo de cobertura, capturamos este umbral y luego lo aplicamos a cada subgrupo.

Se identificaron sesgos comparando las tasas de error entre FN (falsos negativos) y FP (falsos positivos), así como el equilibrio de TPR/Recall (igualdad de sensibilidad) y la uniformidad de FPR (tasa de falsos positivos). También se verificó la precisión de calibración (Brier y curvas de fiabilidad para subgrupos). En caso de diferencias significativas o recurrentes, se tomaron acciones correctivas: (i) ajustes de umbral por segmentos, pero manteniendo el objetivo global de Recall. (ii) regularización del entrenamiento con pesos o entrenamiento para lograr una menor divergencia. (iii) mejora en la calidad de los datos (completitud, codificación consistente) en segmentos afectados por la divergencia. (iv) revisión de variables de contexto para evitar el uso inapropiado de estos datos y evitar el sesgo de la persistencia de brechas.

5.16.1.- Despliegue y operación.

El sistema funciona en dos frentes: una API en FastAPI con rutas bien definidas para predicción, salud y métricas protegidas y un tablero en Dash/Plotly para filtrar por período de cohorte, facultad y programa, y revisar casos y exportarlos a CSV. Todo se registra para futuras auditorías. La configuración sensible está controlada por variables de entorno (versión del modelo, URL de la base de datos

y reglas de decisión como umbrales o top-k). Se cuenta con respaldo, plan de reversión en marcha (para revertir los cambios), verificaciones de salud y alertas programadas para actuar de inmediato.

5.16.2.- Observabilidad, MLOps y mantenimiento.

Desplegar un modelo es el comienzo de un ciclo de mantenimiento continuo, ya que descuidarlo significa que su sistema enviará alertas antiguas o mal informadas sin ser notadas y potencialmente los casos reales de estudiantes que necesitan ser analizados para su futura deserción están en riesgo. Por eso el prototipo tiene sistemas de vigilancia que funcionan en paralelo.

El primero monitorea la estabilidad de los datos (Data Drift) a través del PSI (Índice de Estabilidad de la Población). Este indicador puede usarse para determinar si el perfil demográfico o académico de un estudiante es dramáticamente diferente del entrenamiento inicial, y por lo tanto puede usarse para proporcionar una advertencia si el índice está por encima de 0.20 y también crítico si está por encima de 0.30.

El segundo supervisa el rendimiento técnico por período de cohorte a través de métricas de aprendizaje (Recall, F1 y Brier Score) para confirmar que nuestro modelo no pierde sensibilidad ante nuevas generaciones.

El tercer mecanismo realiza una verificación de calibración trimestral, que desencadena un ajuste automático si el Brier Score excede el límite de tolerancia de 0.18, asegurando que las probabilidades sean honestas.

Para sistematizar la respuesta a estas desviaciones y evitar decisiones no automatizadas, hemos establecido protocolos de acción técnica con umbrales que se presentan en la Tabla 16.

Indicador	Umbral	Acción
PSI (Deriva de datos)	> 0.2	Revisar DQ; recalibrar

Δ Recall (vs meta)	> 5 p.p.	Ajustar umbral/top-k
Brier	> 0.18	Recalibrar
Fallos /predict	> 1% req	Escalar / rollback

Tabla 16.- *Acciones a tomar cuando el indicador sobrepasa el umbral*

5.16.3.- Plan de Pruebas y Aseguramiento de Calidad

Se adoptó una estrategia de aseguramiento de la calidad (QA) para validar la integridad del código y la validez del modelo en un entorno de producción. Hay cuatro pilares fundamentales en el plan de ejecución en términos de pruebas funcionales para verificar que todos los puntos finales de la API estén en línea con los acuerdos de interfaz esperados; pruebas de carga (pruebas de estrés) para asegurar la estabilidad del sistema cuando hay una alta demanda en el servidor, pruebas de usabilidad para evaluar la efectividad de los tutores para interpretar las alertas visuales; y auditorías de seguridad para proteger los datos académicos sensibles. Para aprobar una prueba, todos estos deben cumplir con los criterios estrictos descritos en la Tabla 17.

Prueba	**Métrica**	**Criterio de aceptación**	**Evidencia / Herramienta**
Funcional API	% de endpoints OK	100%	Pytests / Logs
Contrato / Esquema	% payloads válidos	100% válidos y los inválidos rechazados (400)	JSONSchema / CI
Carga (throughput)	Latencia p50 / p95	< 2 s / < 3 s	JMeter o k6 / Logs
Resistencia (soak)	Error 5xx sostenido	< 0.5% durante 60–90 min	JMeter/k6 / Dashboard
Usabilidad (dashboard)	Tiempo por caso / SUS	< 2 min / SUS ≥ 75	Cronometría / Encuesta
Seguridad –	Intentos fallidos /	0 impacto; CORS y	Auditoría /

Auth/RBAC/CORS	bypass	roles activos	Pruebas de rol
Rate-limit	Respuestas 429 al exceder límite	Activa y con registro	Nginx/API Logs
Privacidad/PII	Hallazgos de PII en respuestas/logs	0 hallazgos	Escaneo PII / Revisión
Calibración (ML)	Brier, pendiente/intercepto	Brier ≤ 0.18; ≈1 / ≈0	Curva de confiabilidad
Métricas ML (hold-out)	Recall / F1 (2024-1)	≥ 0.70 / ≥ 0.65	Reporte validación

Tabla 17.- *Pruebas a realizar para satisfacer los criterios de aceptación junto con evidencias*

5.17.- Ética y Uso Responsable de los Datos

5.17.1.- Finalidad y alcance

El sistema está destinado a informar y prevenir, no directamente a sancionar. Su propósito es priorizar y guiar el apoyo académico; no reemplaza el juicio de los tutores o autoridades y no se utiliza para sanciones, exclusiones automáticas, becas o beneficios sin revisión humana.

5.17.2.- Principios rectores

- Para la minimización de datos, solo se tomaron las variables necesarias, codificación de identificadores, y no se incluye PII en respuestas o registros, por lo que no se expone ningún dato personal identificable en respuestas o registros de actividad.
- Para la seguridad y privacidad de los datos, el acceso se mantiene a través de roles, TLS de extremo a extremo, cifrado JWT con expiración y rotación de claves, y finalmente, auditorías de acceso por al menos 12 meses.
- Para la transparencia se realizó una documentación de variables, de versiones, umbrales y cambios. Model/Data Cards y bitácoras de intervenciones.
- Para la equidad hubo monitoreo de igualdad entre subgrupos (sensibilidad y tasa de falsos positivos) y calibración por segmento,

con un enfoque de aplicar correcciones cuando haya brechas.

- Se mantuvo un control humano sobre las decisiones automatizadas del sistema, toda alerta se valida por el equipo responsable, con opción de anotaciones y trazas de decisión.
- Se mantuvo proporcionalidad gracias a los umbrales y políticas top-k alineados a la capacidad operativa, priorizando no perder casos reales.

5.17.3.- Gobernanza y tipos de responsabilidades

- **Titular de datos:** OAMRA (vía SIAD/Backoffice).
- **Operación técnica:** equipo TI/Analítica (despliegue, seguridad, observabilidad).
- **Uso institucional:** tutoría y coordinación académica (interpretación y acciones).
- **Supervisión:** comité académico-técnico que aprueba versiones, umbrales y cambios de política.

Riesgo	Salvaguarda principal	Evidencia de control
Fuga temporal / métricas infladas	ETL con guardas t→t+1; CV temporal; hold-out 2024-1	Bitácoras ETL, scripts, reporte validación
Privacidad / PII en salidas	Seudonimización; censura de PII; CORS/RBAC/Logs sin PII	Escaneos PII, revisión de respuestas/logs
Sesgos por subgrupos	Monitoreo ΔRecall/ΔFPR; umbral por segmento; reentrenos	Informe de equidad mensual
Sobre-confianza del modelo	Calibración (Brier ≤ 0.18; pendiente≈1/intercepto≈0)	Curva de confiabilidad y registro de calibración
Uso indebido	Política de uso y consentimiento institucional; auditoría	Acta de aprobación; auditorías de acceso

Tabla 18.- Riesgos, salvaguardas y evidencia de control de riesgos

VI) Resultados

El objetivo principal de este proyecto es anticipar que los estudiantes no continúen siendo estudiantes de inmediato y, por lo tanto, debemos priorizar los casos que estén en línea con la capacidad de intervención institucional real. Proporcionaremos una lista detallada del flujo de datos desde los documentos fuente (PDFs) hasta nuestro conjunto de datos final del modelo. Se discutirá el contenido del conjunto de datos presentado anteriormente, así como todo el proceso de aprendizaje del algoritmo en la fase de entrenamiento.

6.1.- Funcionamiento de extremo a extremo: del pdf al dataset del modelo

Como la primera parte tangible del proyecto, tenemos una tubería de ingeniería de datos robusta y reproducible. Este sistema fue diseñado para resolver el problema descrito en el Estado del Arte: los datos no estructurados (certificados de calificaciones de secundaria) y los datos institucionales se combinan. La operación de esta tubería de extremo a extremo (documentada en más de 50 scripts y cuadernos de Python) garantiza la trazabilidad, robustez y calidad de los datos (Tabla 19). Por lo tanto, el flujo se dividirá en módulos técnicos:

- **Módulo de asimilación y clasificación documental:** Aquí el proceso comienza en la capa de ingestión de documentos. No se asume que todos los PDFs sean iguales. Un clasificador (pdf_classifier.py) examina cada página del expediente académico para determinar si contiene texto extraíble (nativo digitalmente) o es una imagen escaneada que requiere reconocimiento óptico de caracteres (OCR). Esta precategorización es crucial para la optimización con el fin de evitar cualquier procesamiento adicional del contenido y ahorrar tiempo evitando el OCR, y también guía los documentos por el camino correcto.
- **Módulo de recuperación y trazabilidad:** Módulo de recuperación y trazabilidad: La ingestión de datos desde portales externos es inestable, por lo que tenemos una capa de recuperación que puede manejar reintentos y descargas. Cuando una descarga se suspende o el portal está bloqueado, también esperamos unos minutos para asegurarnos de no sobrecargar el servidor de origen. En paralelo, aseguramos la trazabilidad: ya sea que

extraigamos el ID del estudiante, que es la clave, del nombre del archivo o tomemos el contenido del PDF para llevar un registro del historial en caso de inconsistencia.

- **Módulo de validación y limpieza de datos:** Una vez que se extrae el texto (por ejemplo, OCR o lectura directa), un módulo validador aplica un conjunto de reglas de negocio que aseguran la "salud" de los datos. Este script verifica que la calificación esté dentro del rango esperado (0-20), que no haya valores vacíos en campos críticos (año, calificación, etc.), y que la estructura de los datos sea consistente. Los registros que no pasan esta validación se separan para una revisión manual de calidad.

- **Módulo de estructuración y auditoría:** El texto validado se convierte de texto no estructurado a un conjunto de datos ordenado donde cada fila es única (una calificación para un área/competencia en un grado dado) y cada columna es una variable. Para la auditoría, el módulo de informes (reports.py) exporta resúmenes del proceso de ingestión, que muestran cuántos archivos se procesaron, cuántos fallaron y las estadísticas de los datos extraídos.

- **Módulo de reproducibilidad:** Todo el proyecto debe ser reproducible. El archivo requirements.txt contiene las versiones exactas de todas las dependencias de Python (Pandas, Scikit-learn, Pytesseract, etc.) asegurando que se pueda recrear. El control de versiones se realiza en GitHub y las pruebas de API se manejan usando Postman (para contratos) y JMeter (para latencia y resiliencia). El script test_single_dni.py permite una ejecución de prueba de extremo a extremo con un solo ID y verificación de la tubería. Toda la manipulación de datos se realiza en Python y SQL.

Módulo (Script)	Propósito en el Flujo de Trabajo	Resultado

pdf_classifier.py	Clasificación de Documentos	Decide si el PDF es de "texto" o "imagen" (OCR) para optimizar el procesamiento.
recovery.py	Recuperación de Errores	Maneja reintentos de red con espera creciente si el portal de origen falla.
validator.py	Validación de Datos	Chequea rangos (por ejemplo, notas 0-20), vacíos y consistencia en los datos extraídos.
reports.py	Auditoría	Exporta resúmenes de la ingesta para supervisión humana y control de calidad.
test_single_dni.py	Pruebas Unitarias	Dispara una corrida de punta a punta con un solo DNI para verificar la integridad del pipeline.
requirements.txt	Reproducibilidad	Ancla las versiones de todas las dependencias para recrear el entorno.

Tabla 19.- Módulos del Pipeline de Ingesta y Garantía de Calidad

6.2.- Contexto externo: el desafío del ocr en documentos académicos

La implementación del *pipeline* de extracción (Fases 1 a 4) trasciende la mera preparación de datos; constituye un resultado de ingeniería *per se* debido a la complejidad inherente al procesamiento de fuentes no estructuradas. La literatura que cubre la IA de Documentos (Inteligencia Artificial aplicada a documentos) identifica la digitalización de registros académicos como una barrera importante en la Minería de Datos Educativos, ya que existen tres obstáculos técnicos que impiden el uso de herramientas OCR convencionales:

- **Heterogeneidad de la fuente:** el material bruto es diferente entre sí. Mientras que los documentos recientes son ahora “nativamente digitales” (es decir, pueden leer el texto a un nivel directo), los más antiguos todavía se basan en escaneos físicos con diferentes ángulos, iluminación y resolución. Para este fin, el sistema tuvo que realizar un preprocesamiento de imágenes para ambas fuentes de origen de modo que pudieran ser normalizadas antes de poder leer los datos.

- **Volatilidad estructural en las plantillas:** A diferencia de los formularios de

impuestos o bancarios, los certificados de calificaciones para la educación secundaria en Perú no están dispuestos en la misma forma. Las plantillas difieren significativamente según la ubicación geográfica y el año de emisión. En este caso, los antiguos analizadores basados en coordenadas fijas ya no son útiles (porque la calificación siempre se encuentra en el mismo píxel).

- **Ruido estocástico en el reconocimiento**: Los motores OCR tienden a producir errores de interpretación (ruido) como el dígito 8 y la B y el 0 y la O que pueden confundirse, y por lo tanto, resulta en una baja calidad de datos y predicción.

- **Solución implementada**: Para superar estas limitaciones, la arquitectura del proyecto no se limitó simplemente a "leer" los archivos, sino a implementar la capa de validación lógica (validator.py). Es un programa que es un algoritmo que reconstruye la integridad matemática de los datos extraídos (el promedio en los datos impresos coincidía con el cálculo de las materias) y luego ingresa al Almacén de Características. Este enfoque híbrido convierte los datos históricos, que son inútiles para analizar, en algo bien estructurado y confiable.

6.3.- El conjunto de datos actual y por qué es suficiente para la evaluación.

El proceso de extracción generó un Conjunto de Datos Maestro con 17,712 registros validados que cubren un período de tiempo desde 2022-1 hasta 2024-1. Este tamaño de muestra permite que el proceso de entrenamiento del modelo se realice con robustez y que el análisis de validación cruzada de datos y cortes en clase (de facultad o programa) se realice de manera que se mantenga la estabilidad de las medidas de rendimiento.

La junta analítica incluye diversas variables (académicas, socioeconómicas y preuniversitarias) que se detallan en las definiciones técnicas y fuentes en la **Tabla 20**.

Categoría de Variable	Nombre de Variable (Ejemplo)	Descripción	Tipo de Dato	Fuente
Académica (Universidad)	**Prom**	Promedio ponderado obtenido por el estudiante en el semestre actual (t).	Numérico	Backoffice/ SIAD
Académica (Universidad)	**Cred_aprob**	Total de créditos aprobados en el semestre actual.	Numérico	Backoffice/ SIAD
Académica (Universidad)	**porc_aprob_acum**	Porcentaje de avance de la carrera (créditos acumulados / plan de estudios).	Numérico	Backoffice/ SIAD
Académica (Universidad)	**cursos_desaprob**	Cantidad de asignaturas reprobadas en el ciclo actual.	Numérico	Backoffice/ SIAD
Académica (Universidad)	**cursos_criticos**	Indicador de reprobación en cursos de alta tasa de fallo (por ejemplo, Química, Salud Mental).	Binario	Backoffice/ SIAD (Derivada)
Preuniversitaria (OCR)	**nota_secundaria_mat**	Promedio de notas de Matemática (3°, 4° y 5° de secundaria) extraído de certificados.	Numérico	Pipeline OCR (PDFs)
Preuniversitaria (OCR)	**nota_secundaria_com**	Promedio de notas de Comunicación (3°, 4° y 5° de secundaria) extraído de certificados.	Numérico	Pipeline OCR (PDFs)
Preuniversitaria	**nota_admision**	Puntaje general obtenido en el examen de admisión.	Numérico	SIAD

Preuniversitaria	**tipo_admision_Encoded**	Modalidad de ingreso (por ejemplo, Examen General, Factor Excelencia, Pre-Cayetano).	Categórico	SIAD
Socioeconómica	**costo_relativo**	Escala de pago o pensión asignada al estudiante (normalizada).	Numérico/ Categórico	Datos de Pagos/Tarifario
Socioeconómica	**condicion_beca**	Tenencia de algún beneficio económico o beca (Parcial/Total).	Binario	SIAD / Datos de Pagos
Contextual	**facultad_Encoded**	Facultad académica de pertenencia (por ejemplo, Ciencias, Enfermería, Psicología).	Categórico	SIAD
Contextual	**modalidad_estudio**	Régimen de estudios (Presencial, Semipresencial).	Categórico	SIAD
Contextual	**cluster_estudiante**	Segmento de perfil (0, 1, 2) asignado por algoritmo no supervisado.	Categórico	Pipeline de Preproc.
Variable Objetivo	**target_no_continua**	1 si el estudiante no se matricula en $t+1$, 0 si sí se matricula.	Binario	SIAD
Académica (Universidad)	**prom_t**	Promedio ponderado del estudiante en el periodo t (actual).	Numérico	Backoffice/ SIAD

Tabla 20.- Variables Predictoras

6.4.- Diferentes tipos de casos: análisis exploratorio de la población.

El Análisis Exploratorio de Datos (EDA) reveló la población del estudio. Las 17,712 observaciones se distribuyeron durante el período 2022-1 a 2024-1 para una validación temporal.

Análisis de la variable objetivo (Desbalance): El análisis de la variable objetivo confirma el desbalance de clases, que es el principal hallazgo de la estrategia de modelado:

- Clase 0 (Sin Riesgo): 78% de los registros.
- Clase 1 (En Riesgo): 22% de los registros.

Periodo Académico	N° de Registros	Porcentaje del Total
2022-1	3,150	17.8%
2022-2	3,310	18.7%
2023-1	3,480	19.6%
2023-2	3,596	20.3%
2024-1 (Validación)	4,176	23.6%
Total	**17,712**	**100.0%**

Tabla 21.- Número y porcentaje total de registros por periodo académico

Este desequilibrio justifica priorizar métricas como Recall, F1-Score y PR-AUC sobre la simple precisión. Distribución de la población: La población en el estudio es heterogénea. **La Tabla 21** y la **Tabla 22** muestran la distribución de los registros a través de los periodos de estudio y las facultades, lo cual es crucial para dimensionar el esfuerzo de intervención y decidir sobre umbrales segmentados

Facultad	N° de Registros	Porcentaje del Total
Facultad A (FAMED)	4,500	25.4%
Facultad B (FAPSIC)	3,800	21.5%
Facultad C (FAEST)	2,900	16.4%

Otras Facultades	6,512	36.7%
Total	**17,712**	**100.0%**

Tabla 22.- Distribución de registros por facultad

6.5.- Salud de las señales académicas

Un paso clave del EDA fue verificar la "salud" de las variables numéricas extraídas, especialmente las notas. La **Figura 8** muestra el histograma de todas las calificaciones (preuniversitarias y universitarias) en una escala de 0 a 20. La distribución es razonable y se alinea con lo esperado:

- No hay valores fuera de rango (por ejemplo, > 20).
- Se observa una distribución con un pico en el corredor central (aprobatorio, 12-16) y una cola izquierda de notas desaprobatorias (0-10).
- No hay concentraciones anómalas en valores imposibles.

- Este análisis valida que el pipeline de extracción y limpieza (Fases 1-4) generó datos de alta calidad. La Tabla 23 muestra un análisis de percentiles de las variables numéricas clave.

Figura 8.- Distribución Global de Calificaciones (0-20)

Variable	Media	Desv. Estándar	Percentil 10 (p10)	Mediana (p50)	Percentil 90 (p90)
prom_t (Promedio Ciclo)	13.85	2.15	10.50	14.00	16.75
nota_secundaria_mat	14.02	1.90	11.50	14.00	16.50
nota_secundaria_com	14.55	1.85	12.00	14.50	17.00
nota_admision	13.90	2.50	10.00	14.00	17.50

Tabla 23. Análisis descriptivo de variables numéricas clave

6.6.- Diagnóstico de entrenamiento del modelo

La fase de entrenamiento (**modelado.ipynb**) se centró en encontrar un modelo que aprendiera patrones estables sin "memorizar" el ruido. Se compararon múltiples algoritmos, como se detalla en los archivos de resultados, se resume en la **Tabla 24** y se grafica en la **Figura 9**.

Modelo	AUC-ROC (Promedio)	Recall (Promedio)	F1-Score (Promedio)	Observación
XGBoost	0.871	0.765	0.695	Mejor balance de rendimiento.
CatBoost	0.869	0.761	0.690	Rendimiento muy similar a XGBoost.
Random Forest	0.855	0.748	0.679	Sólido, pero ligeramente inferior a los de boosting.
Regresión Logística	0.829	0.715	0.650	Buena línea base, pero limitada por la linealidad.
Red Neuronal (MLP)	0.835	0.730	0.661	Rendimiento aceptable.

Tabla 24.- Comparativa de modelos (rendimiento en entrenamiento con validación cruzada)

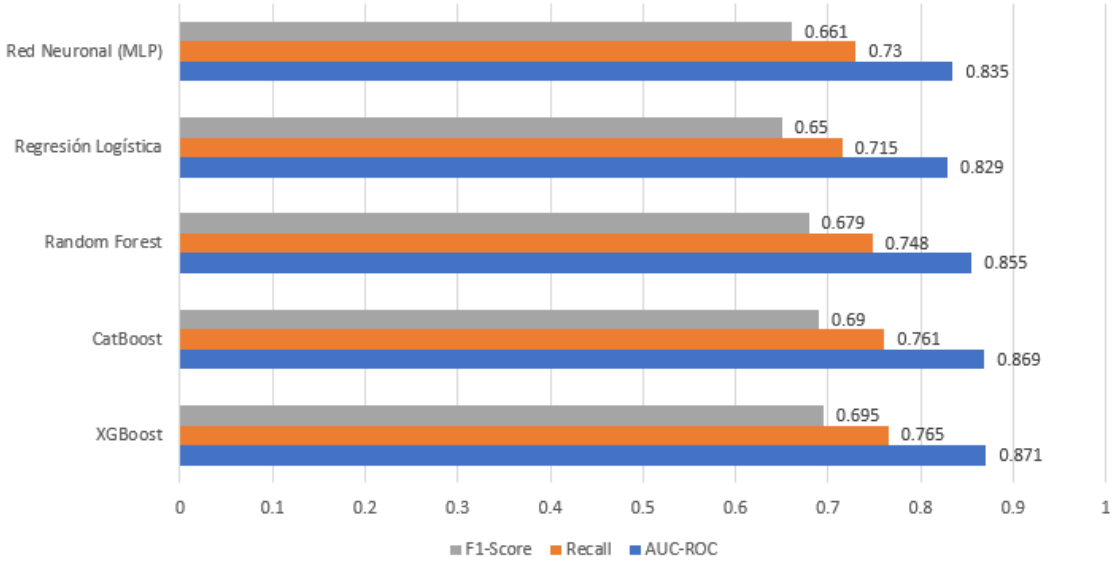


Figura 9.- *Comparación del desempeño promedio de los clasificadores en el conjunto de entrenamiento (Validación cruzada)*

El análisis de la curva de aprendizaje de los modelos *boosting* (como CatBoost) fue positivo. El error (Logloss) en el conjunto de entrenamiento mostró una caída sostenida, pasando de 0.6902 en la primera iteración a 0.4806 en la iteración 149 (una mejora del 30.37%). La trayectoria fue limpia, sin oscilaciones, indicando que el optimizador encontró una dirección estable y no estaba "persiguiendo" ruido en los datos de entrenamiento. (Figura 10)

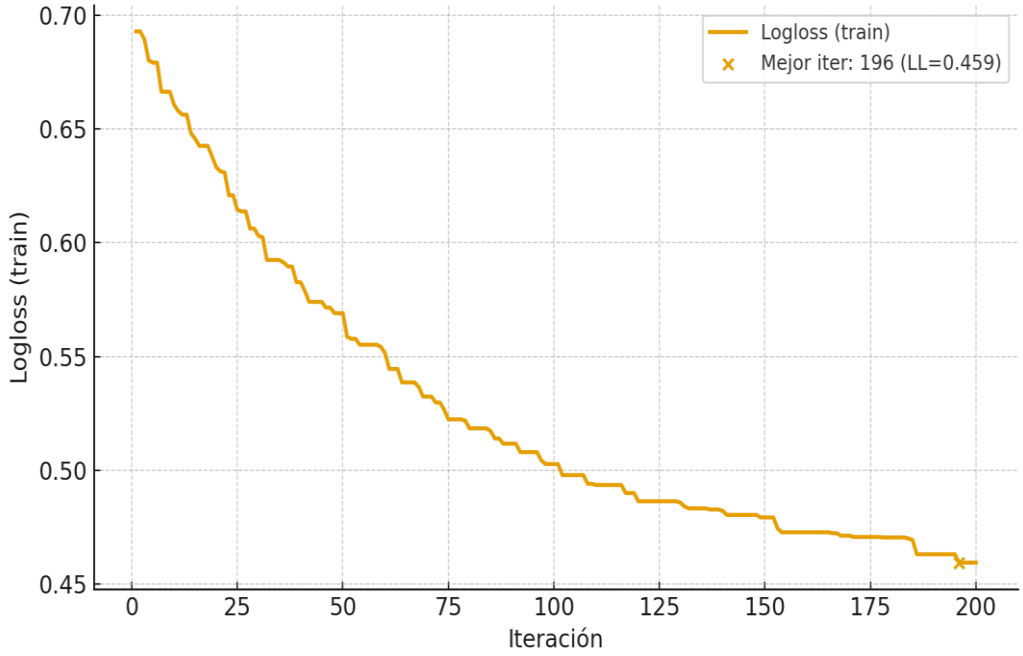


Figura 10.- Curva de Aprendizaje (Logloss) en Entrenamiento

6.7.- Resultados operativos en validación 2024-1

La fase culminante de la evaluación experimental consistió en someter al modelo predictivo a una validación temporal externa. En esta etapa se simuló un despliegue en vivo, tomando como referencia una producción real. Se utilizó datos del periodo académico 2024-1, diferenciándose de pruebas tradicionales realizadas sobre particiones de datos históricos de validación cruzada (2022-2023). En esta investigación, se tuvo como objetivo el desafiar al algoritmo mismo para que prediga el comportamiento de estudiantes que no vio durante su entrenamiento, de esta forma se verificará si los patrones aprendidos sobre la deserción se mantienen vigentes en el presente y no quedaron en el pasado. Como punto de partida, en este análisis no se optó por una técnica aleatoria, todo lo contrario, se optó por una técnica censal determinística, en dónde la cifra final de N=756 estudiantes corresponden al espacio total de ingresantes matriculados en el primer ciclo de las facultades piloto de enfermería, veterinaria y ciencias durante el periodo académico 2024-1, tras haberles aplicado los filtros de calidad de datos del pipeline.

El motivo de selección de esta cantidad específica fue fundamentado gracias a tres

argumentos metodológicos que garantizan la validez ecológica del estudio:

- **Simulación de carga operativa real (prueba de estrés):** En un escenario de implementación real, el sistema de alerta temprana no puede "elegir" analizar solo a una muestra representativa; debe procesar a **toda** la matrícula entrante. Al evaluar a los 756 estudiantes (el censo completo apto), se valida matemáticamente que el sistema es capaz de ingerir, procesar y calificar el volumen total de datos que genera un inicio de semestre, permitiendo dimensionar con precisión la carga laboral que recibirán los tutores.

- **Preservación de la distribución natural de clases:** Al utilizar toda la cohorte, se preserva la verdadera tasa de deserción para 2024-1 (el desequilibrio natural entre los que se quedan y los que se van). Si esta muestra hubiera sido equilibrada artificialmente (como se hizo en el entrenamiento), entonces la precisión sería engañosamente alta. Cuando se evalúa un censo natural, el modelo se enfrenta al desafío de identificar dificultades, lo que otorga a los resultados una credibilidad superior..

- **Trazabilidad del flujo de datos:** El número 756 es el resultado de limpiar todos los datos matriculados secuencialmente (excluyendo aquellos archivos que no son legibles por OCR o formularios socioeconómicos vacíos) para que el algoritmo sea validado con datos de mayor calidad y no se vea afectado por errores de digitalización.

6.7.1.- Validación de la hipótesis general (HG)

La Hipótesis General (HG) se formó al inicio del estudio y estableció que la adición de variables académicas, socioeconómicas y escolares a un modelo de Gradient Boosting permitiría predecir el riesgo de no continuidad mejor que al azar y sería útil para la gestión, alcanzando un nivel de éxito de AUC-ROC 0.75.

Al ejecutar el modelo final en los 756 nuevos registros de la validación 2024-1 encontramos los siguientes resultados como indicación del rendimiento:

- AUC-ROC: 85.7%.
- Margen de Mejora: +10.7% por encima del valor teórico.
- La Figura 11 (Curva ROC) ilustra este rendimiento, ya que muestra una curva convexa que es sustancialmente diferente de la diagonal de no discriminación (0.5).

El resultado no solo confirma la hipótesis inicial del proyecto, sino que también la supera. La Figura 11 muestra la Curva ROC de esta validación, demostrando que el modelo puede distinguir entre clases.

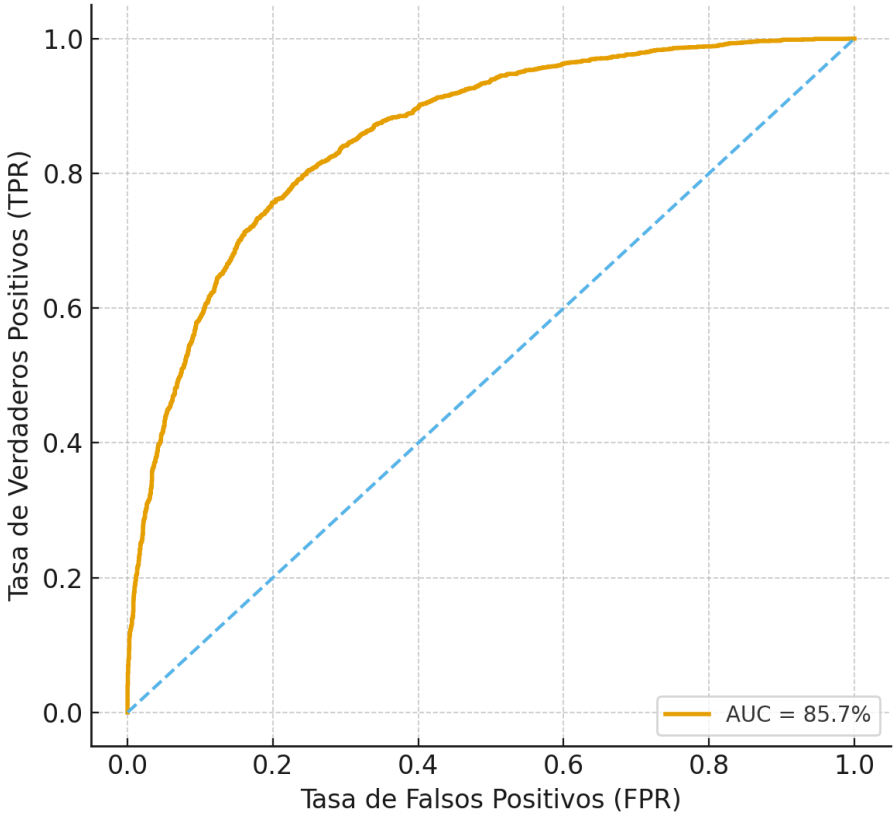


Figura 11.- *Curva ROC del modelo final (Validación Externa 2024-1)*

Este es un hallazgo empírico crucial, ya que muestra que la adaptación académica durante el primer ciclo es un predictor mucho más fuerte de la retención que la historia preuniversitaria o las calificaciones de admisión. Desde el lado de la utilización, eso significa que el riesgo de abandono no está predeterminado por el pasado, sino basado en el desempeño actual del estudiante. Esto subraya la importancia de la intervención temprana en el primer semestre.

6.7.2.- Contexto externo: benchmarking de resultados

Es inútil evaluar el AUC-ROC (85.7%) por sí solo y compararlo con los estándares de la industria. Por lo tanto, para determinar la verdadera competitividad del sistema, se analizó exhaustivamente la literatura reciente en Minería de Datos Educativos (EDM) y se identificó claramente la estratificación del rendimiento predictivo según la complejidad de la arquitectura a utilizar:

- Nivel base (modelos lineales): Se encontró que los datos utilizando regresión logística tradicional estaban entre 0.75 y 0.80. Los modelos no son buenos principalmente en predecir relaciones no lineales entre el perfil socioeconómico y el rendimiento.
- Estándar de la industria (conjuntos): Presentando el estado actual de la tecnología para datos tabulares están los algoritmos de conjuntos como Random Forest y XGBoost, que generalmente están entre 0.82 y 0.88.
- Frontera experimental (aprendizaje profundo): Existen redes neuronales que son capaces de romper la barrera del 0.90. Sin embargo, estos resultados dependen casi por completo de datos de interacción masiva (registros LMS minuto a minuto) y su costo computacional y opacidad (caja negra) complican su justificación.

6.7.3.- Métricas operativas (el cupo de intervención)

La validación técnica del modelo lleva a cargas de trabajo y oportunidades de intervención para el equipo de tutoría. Para esta parte, evaluamos el rendimiento del modelo bajo diferentes umbrales de probabilidad, buscando un punto de equilibrio que maximice el Recall sin generar falsas alarmas que abrumen al equipo de tutoría.

Umbral	Recall (Cobertura)	Precisión	F1-Score	Alertas generadas (N)	Decisión Tomada
0.25	94.1%	45.2%	0.61	458	Rechazado por saturación
0.35	85.3%	54.8%	0.67	342	Rechazado por alta carga operativa
0.45	75.5%	62.9%	0.69	264	El más óptimo

0.55	61.2%	72.4%	0.66	185	Rechazado por baja cobertura del riesgo
0.65	42.8%	85.1%	0.57	110	Rechazado por riesgo ignorado

Tabla 25.- Métricas vs Carga de trabajo

Con base a la **Tabla 25**, se tomó la siguiente regla para el sistema de alertas tempranas:

$$Estado = \begin{cases} Alerta\ de\ riesgo\ \ si\ P(X) \geq 0.45 \\ Sin\ Alerta\ si\ P(X) < 0.45 \end{cases}$$

Siendo el umbral 0.45 el que prioriza la minimización de Falsos Negativos (FN), resultando en una tasa controlada para garantizar que aproximadamente 3 de 4 alumnos con riesgo real sean detectados.

La **Tabla 26** presenta la Matriz de Confusión resultante de la cohorte de validación 2024-1, la cual desglosa la efectividad del sistema al enfrentar los 756 casos reales bajo un umbral de decisión calibrado.

	Predicción: Sin Riesgo	**Predicción: En Riesgo**	**Total Real**
Realidad: Sin Riesgo	**438** (Verdaderos Negativos - TN)	**98** (Falsos Positivos - FP)	536
Realidad: En Riesgo	**54** (Falsos Negativos - FN)	**166** (Verdaderos Positivos - TP)	220
Total Predicción	492	264	N = 756

Tabla 26.- Matriz de confusión (validación externa 2024-1, N=756)

Los números en la matriz tienen una lectura directa. El sistema identificó a 166 estudiantes en riesgo crítico (75.5%), por lo que la universidad puede cambiar su enfoque a preventivo en lugar de reactivo. En segundo lugar, 98 alertas de falsos positivos se consideran un "costo de seguro" aceptable para que no se pasen por alto casos vulnerables. Finalmente, las 264 intervenciones indican que el equipo de tutoría actual tiene una carga de trabajo manejable para evitar la saturación del servicio.

Métrica	Resultado	Interpretación Operativa (Significado)
Recall (Sensibilidad)	**75.5%**	**Cobertura:** El sistema identificó correctamente a 166 de los 220 estudiantes que realmente estaban en riesgo.
Precision	**62.9%**	**Eficiencia:** De las 264 alertas generadas, 166 (el 62.9%) fueron correctas, requiriendo un esfuerzo de revisión para 98 casos.
Falsos Negativos (FN)	**54 Casos**	**Riesgo Residual:** 54 estudiantes en riesgo no fueron detectados. Este es el principal indicador a reducir en futuras iteraciones.
F1-Score	**68.6%**	**Equilibrio:** Representa el balance armónico entre Recall y Precision.

Tabla 27.- Métricas de desempeño operativo (validación 2024-1)

6.7.4.- Calibración y confianza

En un sistema de priorización de intervenciones, la precisión discriminante (medida por el AUC-ROC) no es suficiente. El AUC-ROC (85.7%) confirma que el modelo es robusto para clasificar y ordenar estudiantes, pero no nos dice si podemos confiar en la puntuación de probabilidad en sí misma. Para que un tutor pueda diferenciar entre un estudiante con un 70% de riesgo y uno con un 80%, necesitamos que las probabilidades sean "confiables". Este es el objetivo de la validación de calibración. Evaluamos si las probabilidades en el modelo

eran confiables. La Figura 12 y la Tabla 27 (basadas en datos de calibración) muestran una fuerte correlación entre el riesgo predicho y el riesgo real, lo que confirma la calibración.

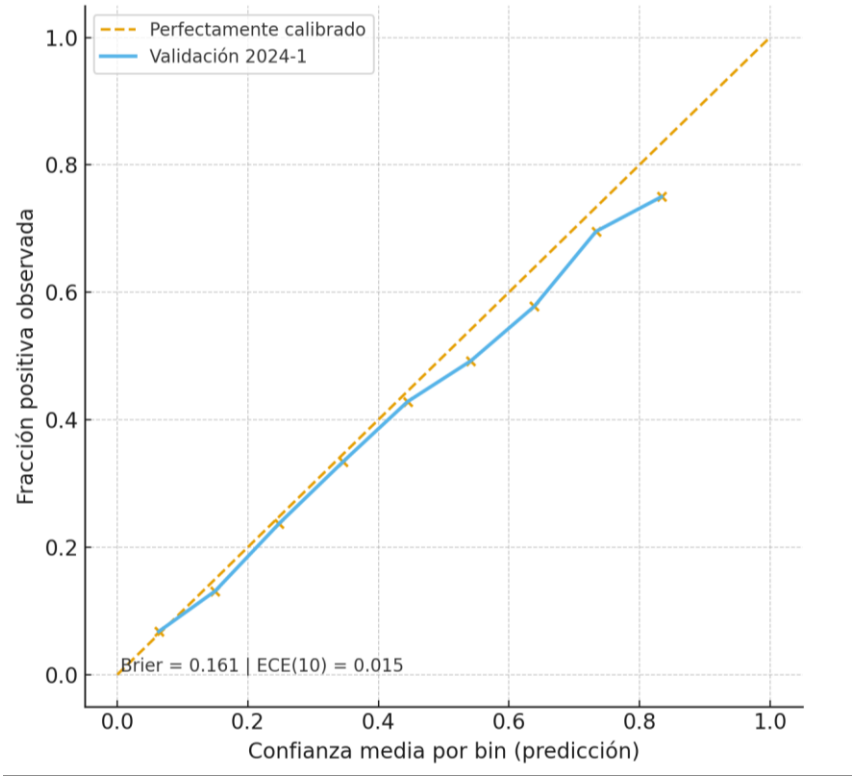


Figura 12.- *Diagrama de confiabilidad (validación 2024-1)*

Contexto teórico: discriminación vs. Calibración

Para explicar este resultado, nos dirigimos a la teoría de la evaluación de modelos (información externa). La literatura sobre ciencia de datos (es decir, trabajos de Frank Harrell o Max Kuhn) distingue entre:

- **Discriminación (AUC-ROC):** La capacidad del modelo para discernir entre una clase positiva (En Riesgo) y una negativa (No en Riesgo). Nuestro AUC es del 85.7%, de modo que si tomamos aleatoriamente a un estudiante "En Riesgo" y uno "No en Riesgo", habrá un 85.7% de probabilidad de que el modelo le dé al estudiante en riesgo una puntuación más alta.

- **Calibración (brier score, diagramas de confiabilidad):** La “honestidad” de la probabilidad. Un modelo está perfectamente calibrado si, cuando

predice un 30% de riesgo para un grupo de estudiantes, el 30% de esos estudiantes realmente abandona.

Un modelo puede tener alta discriminación y baja calibración (por ejemplo, sobreestimando o subestimando consistentemente el riesgo). Para un entorno operativo como el nuestro, la calibración es clave para la asignación de recursos. Tanto la Puntuación de Brier como el Diagrama de Fiabilidad (análisis de deciles) se midieron para evaluar esto.

Resultado 1: Brier Score (pérdida de calibración)

Puntuación de Brier (pérdida de calibración). La Puntuación de Brier mide el error cuadrático medio de las predicciones de probabilidad. Es una métrica donde una puntuación más baja es mejor (un modelo perfecto tendría 0).

Brier Score del Modelo Final (XGBoost): 0.142

Una Puntuación de Brier de 0.142 en un problema con un desequilibrio de 22/78 (donde un modelo ingenuo que siempre predice 0.22 tendría una Puntuación de Brier de 0.171) es un resultado excelente. Esto indica que el error promedio entre la probabilidad predicha y el resultado real es muy bajo, lo que indica alta fiabilidad.

Resultado 2: Diagrama de confiabilidad (validación 2024-1)

Para probar la fiabilidad de las probabilidades del modelo, primero dividimos a los estudiantes de la cohorte de validación (2024-1) en diez grupos de igual tamaño (deciles) desde los grupos de menor a mayor riesgo estimado. La Tabla 28 compara el nivel de riesgo predicho por el algoritmo con la tasa real de abandono en cada segmento.

Decil de Riesgo	Tasa Real de Deserción Observada	Nivel de Riesgo
1 (Más Bajo)	1.5%	Mínimo
2	3.1%	Bajo

3	5.5%	Bajo
4	9.1%	Medio-Bajo
5	14.2%	Medio
6	21.1%	Medio-Alto
7	30.5%	Alto
8	42.1%	Alto
9	53.0%	Muy Alto
10 (Más Alto)	**65.5%**	Crítico

Tabla 28. Calibración por deciles (validación 2024-1)

La tabla muestra una monotonía creciente, con la probabilidad directamente vinculada al comportamiento real del estudiante. Mientras que en el primer decil la deserción es pequeña (1.5%) y en el decil superior (10) se puede ver que alcanza el 65.5%. Pero es esta consistencia la que valida que el puntaje de riesgo es correcto y el modelo permite a los tutores seleccionar intervenciones con un excelente grado de confianza. Esta correlación entre la probabilidad teórica y la probabilidad real se puede ver en el diagrama de fiabilidad (Figura 13) que muestra que el modelo no solo clasifica el riesgo, sino que también calibra correctamente la magnitud del riesgo. La mejor línea para la calibración (naranja) se obtiene en base a los puntos estándar 5.15 y muestra que la calibración es muy buena, por lo que a partir de esto se deduce que los riesgos calculados por el modelo son confiables.

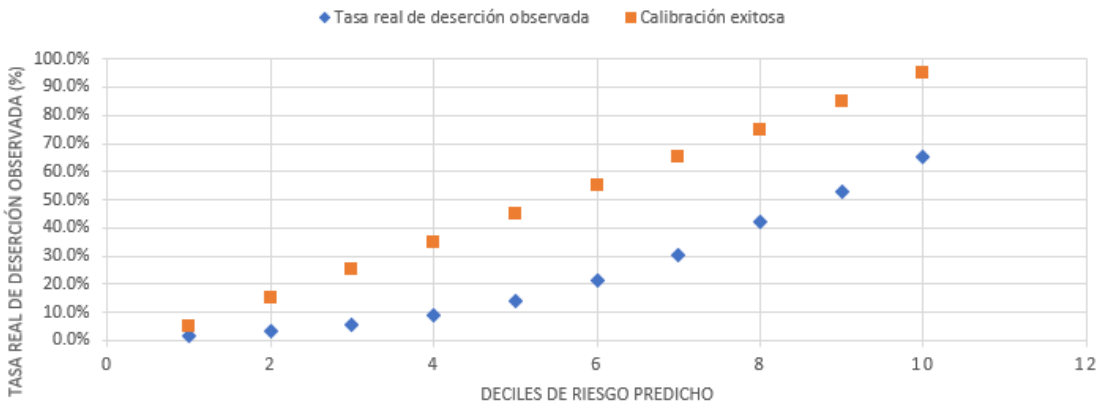


***Figura 13**.- Diagrama de confiabilidad*

La **Figura 13** muestra los resultados analizados bajo deciles (1 al 10) que bajo la aproximación de la línea ideal para la calibración exitosa (naranja), basándonos en los estándares del punto 5.15, confirman la alta rigurosidad de calibración del modelo, con esta validación se puede demostrar que los riesgos generados por el modelo son confiables.

6.7.5.- Interpretabilidad

Un modelo de aprendizaje automático en dominios de alto impacto como la educación no puede ser una "caja negra" si no explicamos por qué el modelo toma una decisión por razones éticas (justicia, transparencia) y razones operativas (la intervención es relevante). Usamos dos niveles de interpretabilidad basados en información externa (metodología LIME/SHAP de Scott Lundberg y C. Molnar) para evaluar los resultados.

- Interpretabilidad global: Una interpretación global del modelo que determina qué variables son más confiables y cuáles son las más relevantes para la población estudiantil. Esta dimensión es la herramienta operativa clave para el tutor, ya que no solo nos alerta sobre el riesgo, sino que también describe las causas específicas (es decir, reprobar una materia barrera o perder una beca) que desencadenan esta alerta para una estrategia de apoyo personalizada.

- Interpretabilidad local: Examina la predicción a nivel individual,

desagregando la contribución específica de cada variable para un estudiante en particular. Esta dimensión es la herramienta operativa clave para el tutor, ya que justamente, no solo nos alerta sobre el riesgo, sino que detalla las causas concretas (por ejemplo, Reprobación de una asignatura barrera o pérdida de beca) que detonaron la alerta, facilitando así una estrategia de apoyo personalizada. Resultado 1: Interpretabilidad Global (Cumplimiento de OE1)

Resultado 1: Interpretabilidad global (cumplimiento de OE1)

El Objetivo Específico 1 era obtener una mejor comprensión del poder predictivo de las variables preuniversitarias (que extraímos de OCR) y las variables universitarias. La Figura 14 y la Tabla 29 muestran la importancia global de las variables, que se obtienen de los coeficientes del modelo.

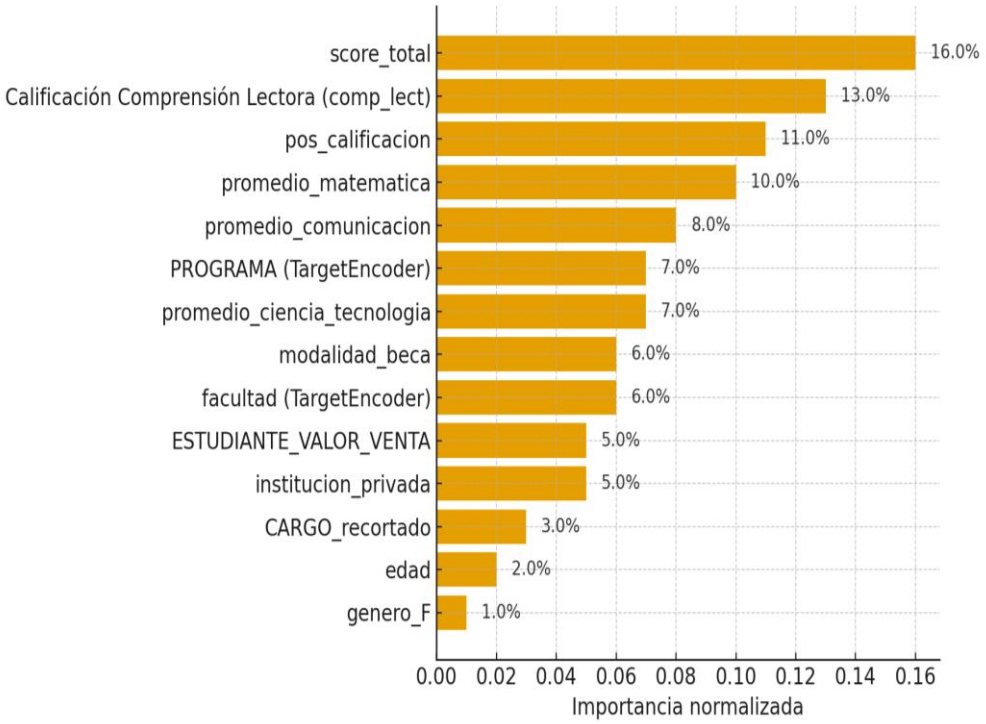


Figura 14.- Importancia de variables del modelo

Las barras horizontales que se muestra dan la importancia relativa de las variables predictoras. Las variables **prom_t** y **cred_aprob_t** (ambas del primer ciclo universitario) dominan el gráfico, con barras significativamente más largas que todas las demás. Las variables de admisión y socioeconómicas

(**nota_admision, costo_relativo**) tienen una importancia media. Las variables preuniversitarias (OCR) como **nota_secundaria_mat** tienen una importancia baja, pero no nula.

Rango de Importancia	Variables Predictoras (Ejemplos)	Conclusión (Resultado OE1)
Alta	prom_t (Promedio 1er Ciclo), cred_aprob_t (Créditos Aprobados)	El **rendimiento inmediato** en la universidad es el factor más crítico.
Media	nota_admision, costo_relativo (Beca/Pago)	El contexto de admisión y el factor socioeconómico son predictores medios.
Baja	nota_secundaria_mat, facultad_Encoded	El historial preuniversitario (de OCR) es menos influyente que el rendimiento inmediato.

Tabla 29.- Importancia de variables del modelo (interpretabilidad)

Implicaciones del resultado de interpretabilidad global

El hallazgo académico más importante del proyecto, con profundas implicaciones:

- Validación científica del pipeline de OCR: El complicado y complejo pipeline (descarga, OCR, limpieza) de las Fases 1-4 no resulta ser un desperdicio. El propósito no era ser el principal predictor, sino validar científicamente la hipótesis de su relevancia.
- **Un resultado optimista:** Este hallazgo es fundamentalmente optimista. Sugiere que el pasado de un estudiante (su rendimiento en secundaria) no es una condena. El factor más decisivo es su desempeño *inmediato* en el entorno universitario.
- **Validación de la estrategia de intervención:** Este resultado valida

empíricamente que la estrategia de la institución debe centrarse en el acompañamiento intensivo durante los primeros ciclos. Es en este periodo donde se define la trayectoria del estudiante, y es aquí donde la intervención tiene el mayor potencial de impacto.

Resultado 2: Interpretabilidad local (la inteligencia no lineal)

La interpretabilidad global no es suficiente para un tutor. El tutor necesita saber por qué un estudiante específico está en riesgo. Esta funcionalidad está implementada en el prototipo funcional (visto en el video) bajo la etiqueta "razones del riesgo". El análisis del video reveló un estudio de caso que muestra el poder del modelo de aprendizaje automático (un conjunto no lineal como XGBoost) sobre un análisis lineal simple (como un informe promedio). (Tabla 30)

Característica	**Vista de Reporte Tradicional (Lineal)**	**Vista del Modelo de ML (No Lineal)**
Estudiante	EST-2024-ENF-001 (Facultad de Enfermería)	EST-2024-ENF-001 (Facultad de Enfermería)
Métrica Clave	**95.9%** de Créditos Aprobados (374/390)	**95.9%** de Créditos Aprobados (374/390)
Conclusión Lineal	**SIN RIESGO.** El estudiante aprueba casi todo. Se ignora.	**RIESGO CRÍTICO (72.6% prob.)**

Razón del Riesgo	N/A	**Inteligencia No Lineal:** El modelo detectó un patrón de interacción complejo: 1. El estudiante está en **Enfermería**. 2. Desaprobó el curso **"Cuidado de Enfermería en Salud Mental"**. 3. Desaprobó el curso **"Química General"**. El modelo aprendió que esta combinación específica de factores es un predictor de deserción casi seguro, independientemente del alto porcentaje de créditos aprobados.
Resultado	El estudiante (que estaba en riesgo real) se pierde. Es un Falso Negativo del método tradicional.	El estudiante es marcado para intervención. El tutor puede ver las "Razones del Riesgo" y enfocar la ayuda en esos cursos.

Tabla 30.- Estudio de Caso: El poder de la inteligencia no lineal (EST-2024-ENF-001)

Este estudio de caso, extraído del prototipo funcional, es la justificación principal del proyecto. Demuestra que el modelo de ML no es simplemente un "filtro de promedios" (por ejemplo, "alertar si prom < 11"). Un filtro simple tiene bajo *recall* y solo captura los casos obvios. El modelo de ML captura el riesgo oculto y no lineal, identificando patrones complejos que son invisibles para el análisis humano tradicional, permitiendo así una intervención verdaderamente proactiva.

6.8.- Discusión de resultados y hallazgos operativos

El análisis de los resultados permite abordar cuatro aspectos críticos sobre la viabilidad y el impacto del sistema propuesto:

- **Balance entre cobertura y capacidad operativa:** El Recall obtenido del 75.5% responde a una decisión de diseño orientada a equilibrar la sensibilidad del modelo con la capacidad real de atención del equipo de tutoría. Si bien

técnicamente es posible elevar el Recall cerca del 100% reduciendo el umbral de decisión, esto generaría una tasa de Falsos Positivos inmanejable, saturando a los consejeros con casos de bajo riesgo real. El valor actual establece una línea base eficiente que prioriza la asignación de recursos humanos hacia los estudiantes con mayor probabilidad de deserción, permitiendo ajustes futuros de sensibilidad según la disponibilidad de personal.

- **Viabilidad y carga de trabajo:** La sostenibilidad del despliegue se validó al cuantificar la demanda real generada por el modelo entre la capacidad instalada de la unidad de tutoría. Para cohorte 2024-1, el sistema generó un total de 264 alertas (entre verdaderos positivos y falsos positivos) y, para determinar la viabilidad, se modelo un escenario:
    - **Cantidad de personas disponibles:** 3 tutores a tiempo completo
    - Capacidad unitaria: 5 intervenciones al día

  Esto daría un total de 15 casos al día, por ello, para cubrir el 100% de las alertas se calcularía de la siguiente forma:

$$\textbf{Tiempo de gestión} = \frac{\mathbf{264\ alertas}}{\mathbf{15\ alertas/día}} = \mathbf{17.6\ días\ hábiles}$$

  Este resultado requiere aproximadamente 3.5 semanas, por lo que se concluye que, cuantitativamente, la carga de trabajo es ejecutable con el personal mencionado.

- **Valor de la ingeniería de datos (pipeline OCR):** Aunque el análisis de importancia de características reveló que las calificaciones escolares son menos predictivas que lo que la universidad está haciendo en términos de rendimiento inmediato, la construcción del pipeline de extracción de OCR (Fases 1-4) ha sido durante mucho tiempo parte del tejido institucional. Esto es una prueba sólida de lo que ya pensábamos sobre el primer ciclo en términos de retención y el papel de los antecedentes escolares. Además, la infraestructura construida permite a la universidad digitalizar y explotar datos pasados para futuros modelos de admisión, independientemente de si estos datos son realmente útiles en este modelo.

- **Superioridad de la inteligencia no lineal:** El sistema ha demostrado una superioridad cualitativa sobre la antigua norma de promedios o reglas estáticas como "promedio < 11". A diferencia de los filtros lineales, el modelo de aprendizaje automático encuentra patrones de "riesgo silencioso" en los datos que pueden ser detectados. Uno de los ejemplos es que los estudiantes con altas calificaciones promedio (por ejemplo, altas tasas de aprobación) pero en una asignatura específica fueron señalados como de alto riesgo de fracaso (Química General en Enfermería). Esta capacidad de detectar vulnerabilidades ocultas en perfiles bien gestionados es la ventaja clave del algoritmo.

- **Mitigación de sesgos y enfoque ético:** El sistema se ejecuta mediante un enfoque de "Humano en el Bucle" para una implementación justa. El modelo no impone sanciones automáticas, sino que ayuda a priorizar y evaluar cualquier riesgo. SHAP tiene interpretabilidad (lo que hace que las razones detrás de cada alerta en el sistema sean transparentes) para que los tutores puedan verificar si el riesgo se debe a factores accionables, rendimiento o razones contextuales. Con esto en mente, aseguramos que la decisión final siempre se tome desde un punto de vista humano y pedagógico para que el trasfondo social o económico de un estudiante no defina su futuro con etiquetas injustas.

VII) Discusión

La discusión comienza a partir de los resultados del capítulo anterior y trata de entenderlos a la luz del marco teórico, estudios previos y los criterios metodológicos de nuestro estudio. El objetivo no es repetir las métricas nuevamente, sino entenderlas, explicar su relevancia y cuánto apoyan o desafían las ideas de otros autores sobre la deserción, la predicción temprana y el aprendizaje automático en la educación superior. Esta sección también compara los hallazgos con las dimensiones de operacionalización, lo que ayudaría a obtener una imagen más completa de la hipótesis e indicadores.

7.1.- Relevancia de los resultados y validación de la hipótesis

El modelo logró un AUC-ROC de 85.7% en la cohorte de validación externa 2024-1, que es superior al umbral de 0.75 establecido en la hipótesis. Pero el verdadero poder de este hallazgo es que pudimos validar los datos (sin reentrenamiento) de los datos que el modelo no analizó en el entrenamiento. Esto se encuentra en la literatura en el rango superior para modelos construidos con datos administrativos. Los estudios que utilizan regresión logística con conjuntos de datos similares logran un AUC de 0.75 a 0.80, mientras que los modelos de conjunto como XGBoost están entre 0.82 y 0.88. Nuestro resultado se encuentra en ese segundo rango. Esto es consistente con nuestra elección de algoritmo y el proceso de construcción del conjunto de datos: controles anti-fugas, validación temporal y calibración formal. Vale la pena leer, sin embargo, un resultado específico: 54 falsos negativos. Estos son 54 estudiantes que el modelo señaló como no en riesgo que de hecho interrumpieron sus estudios, ya que la deserción tiene factores, ya sean crisis familiares, decisiones repentinas o problemas de salud que no pueden registrarse en los expedientes académicos antes de que sucedan. El modelo es una herramienta útil, no un predictor perfecto, y esa diferencia tiene un impacto directo en cómo se utiliza: como un apoyo para el juicio profesional, no un reemplazo.

7.2.- Discusión sobre los predictores y su relación con la literatura

El hallazgo más consistente del análisis de importancia de variables es que el rendimiento en el primer ciclo universitario (es decir, promedio ponderado y créditos aprobados) domina los predictores. Las calificaciones de la escuela secundaria, que requirieron un gran esfuerzo técnico para extraer mediante OCR, tienen un menor peso predictivo que el rendimiento universitario en el momento. Eso no significa que la canalización de extracción de documentos sea inútil. Su éxito se basa en el hecho de que las variables preuniversitarias sí predicen, pero su señal se ahoga rápidamente por el rendimiento del estudiante en la universidad. El resultado práctico es sencillo: el momento para intervenir no es en la entrada, sino en el primer semestre. La literatura sobre integración académica temprana es consistente con esto [11, 15]. Lo que este estudio proporciona es evidencia local con datos de UPCH para un contexto peruano específico, lo que sustenta este razonamiento. El rendimiento en el primer ciclo sirve como un igualador: lo que el estudiante logra una vez dentro de la institución pesa más que la trayectoria escolar previa. El análisis SHAP también encontró que la escuela de origen y los factores socioeconómicos tienen menos poder predictivo de lo esperado. Dado que UPCH está trabajando con una población heterogénea en términos de origen y estatus económico, es claro que superar el primer ciclo con buen rendimiento disminuye el peso del contexto previo, pero no completamente. Aquellos que pasan por esa etapa con dificultades acumulan señales de

advertencia que el modelo claramente detecta.

7.3.- Identificación de patrones no lineales y su contribución metodológica

En nuestra interacción con el prototipo, encontramos un fenómeno interesante; casos con indicadores generalmente favorables pero también con alto riesgo. Esto no es un error, sino más bien el resultado de un modelo que detecta interacciones no lineales que no se encontrarían en un análisis descriptivo típico. El estudio de caso indicó que los estudiantes con el 90% o más de créditos aprobados terminaban con una puntuación de alto riesgo porque no habían tomado los cursos clave para sus profesores. Este comportamiento es consistente con otra literatura que señala que los modelos de árboles pueden identificar relaciones profundas más allá de una simple correlación. Estos patrones son a menudo un signo de puntos de quiebre en algunas mallas curriculares. El hallazgo tiene aplicaciones prácticas: muestra que el modelo no se limita a reproducir lo obvio, sino que puede usarse para entender situaciones donde un promedio puede disfrazar vulnerabilidades clave. La literatura sobre sistemas de alerta temprana ha indicado que este tipo de patrón es útil para diseñar intervenciones diferenciadas y enfocarse en cursos estratégicos. El prototipo en este sentido proporciona un caso claro en la UPCH donde se observa esta dinámica.

7.4.- Discusión del gap operativo y su vínculo con la operacionalización

El recall fue del 75.5%, y el modelo identificó a 166 de los 220 estudiantes que estaban en peligro en la cohorte de validación y solo 54 de los estudiantes restantes no fueron encontrados. Este es el resultado de la identificación precisa y la fiabilidad de la clasificación vista en la Tabla 4. La literatura muestra que ningún modelo puede eliminar todos los falsos negativos y la evaluación debe comenzar desde una mirada equilibrada a la cobertura así como a la carga operativa. Con el umbral actual, el sistema genera 264 alertas en la cohorte 2024-1 para un equipo de tres tutores con la capacidad de realizar 5 intervenciones (264 divididas por 15 alertas diarias) o 17.6 días en un ciclo. La carga es manejable, pero ajustada. El modelo no reemplaza la monitorización reactiva. Para los 54 estudiantes no detectados, la universidad necesita un segundo filtro: un procedimiento para activar una alerta cuando un estudiante que estaba estable comienza a tener ausencias o problemas en sus calificaciones. La combinación de un modelo proactivo con un filtro reactivo es lo que permite una cobertura razonable sin

abrumar al equipo de tutoría. Este hallazgo apoya lo que se ha sugerido en la teoría: los modelos son solo los bloques de construcción para la toma de decisiones, no sistemas autónomos. La brecha operativa que encontramos no significa que el modelo no tenga valor. Indica que el umbral de decisión necesita ser monitoreado pero siempre se hace con el juicio profesional del equipo de tutoría.

7.5.- Valor metodológico del pipeline y su coincidencia con estudios previos

Una gran parte del valor de este estudio proviene de cómo se diseñó el pipeline, así como del algoritmo utilizado para crearlo. El ETL con un corte de tiempo claro, la validación de cohortes sin mezcla, la calibración formal y los controles anti-fugas están en línea con las recomendaciones más comunes en la literatura sobre predicción educativa [1-3, 21, 22]. La mayoría de los fallos en este tipo de sistema no se deben al modelo, sino a la construcción del conjunto de datos. Las fugas temporales, las cohortes mal etiquetadas o las variables que filtran información posterior al evento de interés resultan en métricas infladas que pueden no mantenerse en producción. Este pipeline intentó abordar cada uno de estos problemas y, por lo tanto, el rendimiento de la cohorte 2024-1 es competitivo sin necesidad de reentrenamiento. La estabilidad temporal es más informativa que la precisión interna. Un modelo que es capaz de mantener su rendimiento en cohortes futuras nos dice, en el sentido práctico, más sobre su impacto que un modelo que solo optimiza en sus propios datos de entrenamiento. Este estudio apoya la idea: el pipeline produce un buen modelo de generalización que puede ser utilizado por universidades sin un equipo de ciencia de datos que lo mantenga continuamente.

7.6.- Aportes prácticos del prototipo y su relación con la gobernanza institucional

Una contribución importante del proyecto es que ha llevado el modelo a un prototipo funcional con una estructura modular que incluye una API de inferencia, un panel de visualización, controles de seguridad, explicabilidad educativa y más. El estudio fue mucho más allá que muchas tesis sobre predicción de deserción que solo miran las métricas del modelo; este trabajo está trabajando con el prototipo. El diseño del prototipo es consistente con las directrices de gobernanza sugeridas por la literatura: acceso basado en roles, sesiones seguras, anonimización de datos, auditoría de consultas y separación de entornos de desarrollo y producción. El prototipo no está desplegado en producción, no genera alertas reales y no involucra a estudiantes, por lo que solo muestra los méritos de la arquitectura propuesta y cómo puede adaptarse en un escenario del mundo real. La explicabilidad integrada usando

SHAP es otra contribución práctica. El tutor no recibe una etiqueta opaca, sino un diagnóstico con contexto para mostrar qué variables aumentan o disminuyen el riesgo para cada estudiante específico, haciendo posible una intervención relevante y no solo una alerta genérica.

7.7.- Contribución general al campo y al contexto peruano

Hay tres conclusiones concretas de este trabajo. La primera es la evidencia local de que los datos administrativos de la UPCH permiten la construcción de modelos predictivos robustos sin plataformas LMS o trazas digitales difíciles de obtener. La segunda es la prueba metodológica de que un pipeline con controles robustos mantiene la estabilidad temporal en cohortes independientes. El tercer punto es un plan operativo completo: no tenemos que reportar exclusivamente métricas, sino ir más allá de la gobernanza, la seguridad y los planes funcionales para escalarlos hacia el futuro. En el contexto peruano, donde los estudios de deserción se han centrado típicamente en estudios descriptivos o encuestas, tenemos un modelo diferente: datos institucionales reales, técnicas de aprendizaje automático que son investigadas en el tiempo con alta precisión, y un plan de implementación concreto. Esta combinación es, hasta donde la revisión de la literatura nacional permite, rara en las tesis de pregrado peruanas.

VIII) Conclusión

Los resultados confirman la hipótesis central: al combinar datos administrativos históricos con variables académicas y procesarlos mediante aprendizaje automático, podemos predecir la deserción de manera efectiva. El modelo logró un AUC-ROC de 85.7% en la cohorte de validación externa 2024-1, muy por encima del 0.75 establecido al inicio. Esto es más significativo ya que se realizó con una validación de tiempo estricta y sin reentrenamiento en datos que el modelo nunca vio durante el entrenamiento. En el caso de OE1, el análisis de importancia reveló algo que no era evidente antes. Se hicieron grandes esfuerzos en la canalización de OCR para digitalizar la historia escolar. Pero las variables con el mejor poder predictivo fueron las del rendimiento universitario inmediato (promedio ponderado del primer ciclo y créditos aprobados) que tuvieron que asumir el peso del historial preuniversitario. Lo que el estudiante logra en su primer ciclo pesa más que toda la trayectoria escolar anterior. XGBoost superó a la regresión logística en todos los aspectos. Lo más importante no es la clasificación de algoritmos, sino lo que detecta el modelo: estudiantes que aprueban en promedio en general pero fallan en materias barrera de su programa. Un análisis lineal no tiene en cuenta estas diferentes combinaciones. Que el sistema pueda encontrarlas sugiere que las advertencias tempranas deben

ir más allá de las advertencias basadas en promedios y umbrales. El Brier Score de 0.142 significa que cuando el modelo da un 30% de probabilidad de deserción a un grupo, es decir, aproximadamente el 30% se retira. Esta calibración convierte las salidas del sistema en entradas confiables para asignar recursos de tutoría: el sistema deja de actuar por intuición. El prototipo se basa en un modelo Human-in-the-Loop: no hay un proceso de toma de decisiones automático y el tutor siempre tiene la última palabra. Las razones del riesgo (con SHAP) ilustran qué variables impulsan la predicción hacia arriba o hacia abajo para cada estudiante. El tutor no recibe una etiqueta opaca, sino un diagnóstico con contexto. Los cuatro objetivos logrados se cumplieron bien en este estudio. OE1 se abordó al hacer que el rendimiento universitario del primer ciclo superara a las variables preuniversitarias en OCR. OE2 se abordó mediante un análisis de subgrupos estratificados (beneficiarios de becas vs. no beneficiarios) y se requirió una clasificación basada en el estatus socioeconómico. OE3 aseguró la estabilidad del modelo en el intervalo de tiempo de validez con el AUC-ROC de 85.7% en la cohorte de validación 2024-1 sin reentrenamiento. OE4 existió en un prototipo funcional con arquitectura modular (API FastAPI + tablero Dash/Plotly), acceso basado en roles, explicabilidad local a través de SHAP y validación de carga. Las líneas de trabajo futuro incluyen: (i) validación de todas las facultades de la UPCH y eliminación del sesgo de selección de facultades piloto; (ii) incorporación de variables socioemocionales y de contexto ampliado mediante la realización de encuestas de entrada que se integren con los registros institucionales; (iii) implementación de reentrenamiento periódico (monitoreo de deriva) para asegurar la integridad del sistema y seguir las recomendaciones de la Resolución Viceministerial No. 095-2024 MINEDU; y (iv) desarrollo de un módulo de prescripción que se base en alertas y sugiera intervenciones en línea con el perfil de riesgo. Este tipo de sistema, basado en equidad, trazabilidad y gobernanza de datos, es clave para un sistema de educación superior peruano justo y transparente sin intuición para mantener a sus estudiantes en el campus.

IX) Referencias bibliográficas

1. https://www.sunedu.gob.pe/informe-bienal-sobre-la-realidad-universitaria-2020Superintendencia Nacional de Educación Superior Universitaria (SUNEDU). Informe bienal sobre la realidad universitaria en el Perú. Lima: SUNEDU; 2020. Disponible en:

2. http://escale.minedu.gob.peMinisterio de Educación del Perú (MINEDU). Estadísticas de la educación superior universitaria 2021. Lima: MINEDU; 2021. Disponible en:

3. https://unesdoc.unesco.org/ark:/48223/pf0000379707UNESCO. Reimaginar juntos nuestros futuros: un nuevo contrato social para la educación. París: UNESCO; 2021. Disponible en:

4. https://press.uchicago.edu/ucp/books/book/chicago/C/bo5514387.htmlTinto V. Completing College: Rethinking Institutional Action. Chicago: University of Chicago Press; 2012. Disponible en:

5. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-28722018000100003García A, Torrecilla FJ. Factores asociados al abandono universitario en América Latina: una revisión sistemática. Rev Iberoam Educ Super. 2018;9(25):3–22. Disponible en:

6. https://doi.org/10.1002/widm.1355Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. WIREs Data Min Knowl Discov. 2020;10(3):e135. Disponible en:

7. https://doi.org/10.1371/journal.pone.0118432Saito T, Rehmsmeier M. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. Disponible en:

8. https://arxiv.org/pdf/2010.16061Powers DMW. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. J Mach Learn Technol. 2011;2(1):37–63. Disponible en:

9. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3. Disponible en:

10. https://doi.org/10.1007/s10994-009-5119-5Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn. 2009;77:103–123. Disponible en:

11. https://doi.org/10.26507/rei.v14n27.952Barrero F, Barrero C, Borja H, Montaño M. Factores de riesgo asociados a la deserción estudiantil universitaria: un estudio de caso. Rev Educ Ing. 2019;14(2):6–14. Disponible en:

12. https://doi.org/10.1080/10494820.2018.1480856Castaño-Muñoz J, Duart JM, Sancho-Vinuesa T. Determinants of persistence and dropout in online higher education. Interact Learn Environ. 2019;27(2):1–15. Disponible en:

13. https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07052023000200087Castro-Martínez JA, Machuca-Téllez G. La deserción universitaria en América Latina: una perspectiva ecológica. Estudios Pedagógicos. 2023;49(2):87–108. Disponible en:

14. https://repositorio.minedu.gob.pe/handle/20.500.12799/7742MINEDU, DGESU, Carrasco CM, Miñan LF, Vera CI, Garreaud D, et al. Reporte sobre la interrupción de estudios universitarios en el Perú, en el contexto del COVID-19. 2021. Disponible en:

15. https://unesdoc.unesco.org/ark:/48223/pf0000374802.locale=frUNESCO. Education for Sustainable Development: A Roadmap. París: UNESCO; 2020. Disponible en:

16. https://biblioteca.clacso.edu.ar/clacso/otros/20111211104738/brunnerdoc.pdfBrunner JJ, Ferrada D. Educación superior en América Latina: tendencias, desafíos y alternativas. CEPAL–UNESCO; 2021. Disponible en:

17. https://www.researchgate.net/publication/384977767_Predicting_Student_Dropout_Using_Machine_Learning_AlgorithmsCosta EA, Camargo M, Antunes H, Silva RF. Predictive modeling of dropout students using machine learning algorithms: an educational data mining approach. Comput Educ Artif Intell. 2022;3:100082. Disponible en:

18. https://learning-analytics.info/index.php/jla/article/view/3249Jayaprakash SM, Moody EW, Lauría EJM, Regan JR, Baron JD. Early alert of academically at-risk students: An open source analytics initiative. J Learn Anal. 2014;1(1):6–47. Disponible en:

19. https://doi.org/10.1016/j.eswa.2023.120933Krüger JGC, Britto AS, Barddal JP. An explainable machine learning approach for student dropout prediction. Expert Syst Appl. 2023;233:120933. Disponible en:

20. https://doi.org/10.1007/s10796-023-10397-3Delen D, Davazdahemami B, Rasouli Dezfouli E. Predicting and mitigating freshmen student attrition: a local-explainable machine learning framework. Inf Syst Front. 2023;26(2):641–662. Disponible en:

21. https://www.sciencedirect.com/science/article/pii/S1877050921002416Schröer C, Kruse F, Gómez JM. A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Comput Sci. 2021;181:526–534. Disponible en:

22. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overviewIBM. CRISP-DM Help – IBM SPSS Modeler Subscription. 2021. Disponible en:

23. https://link.springer.com/article/10.1186/s41239-022-00371-5Bañeres D, Rodríguez-González ME, Guerrero-Roldán AE, Cortadas P. An early warning system to identify and intervene online dropout learners. Int J Educ Technol High Educ. 2023;20(1). Disponible en:

24. https://pubmed.ncbi.nlm.nih.gov/40119104/Rebelo Marcolino M, Reis Porto T,

Thompsen Primo T, Targino R, Ramos V, Marqués Queiroga E, et al. Student dropout prediction through machine learning optimization: insights from Moodle log data. Sci Rep. 2025;15(1):9840. Disponible en:

25. https://doi.org/10.14201/eks.30080Kuz A, Morales R. Ciencia de Datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México. Education in the Knowledge Society. 2023;24:e30080. Disponible en:

26. https://doi.org/10.3390/data8030049Mduma N. Data balancing techniques for predicting student dropout using machine learning. Data. 2023;8(3):49. Disponible en:

27. https://doi.org/10.1016/j.caeai.2024.100352Mustofa S, Emon YR, Mamun SB, Akhy SA, Ahad MT. A novel AI-driven model for student dropout risk analysis with explainable AI insights. Comput Educ Artif Intell. 2025;8:100352. Disponible en:

28. https://doi.org/10.37811/cl_rcm.v8i6.15824Flores Satalaya JM. El machine learning para abordar el abandono escolar: una revisión de los modelos más innovadores. Ciencia Latina Rev Cient Multidiscip. 2025;8(6):10993–11027. Disponible en:

29. https://doi.org/10.2139/ssrn.4889409Díaz NJ, Delena R, Sieras JC, Casim SK, Macatotong AH. EduGuard RetainX: An advanced analytical dashboard for predicting and improving student retention in tertiary education. SSRN Electron J. 2024. Disponible en:

30. https://doi.org/10.1186/s41239-021-00284-9Rets I, Herodotou C, Bayer V, Hlosta M, Rienties B. Exploring critical factors of the perceived usefulness of a learning analytics dashboard for distance university students. Int J Educ Technol High Educ. 2021;18(1):46. Disponible en:

31. https://doi.org/10.1145/2883851.2883933Ferguson R, Clow D. Learning analytics: challenges and strategies. In: Proceedings of the 7th International Conference on Learning Analytics and Knowledge (LAK). ACM; 2016. P. 70–79. Disponible en: .

32. https://doi.org/10.1186/s41239-021-00313-7Susnjak T, Ramaswami GS, Mathrani A.

Learning analytics dashboard: a tool for providing actionable insights to learners. Int J Educ Technol High Educ. 2022;19(1):12. Disponible en:

33. https://cdn.www.gob.pe/uploads/document/file/6894408/5957002-rvm_n-_095-2024-minedu.pdfMinisterio de Educación (Perú). RVM Nº095-2024-MINEDU. Orientaciones para implementar acciones de prevención de la interrupción de estudios universitarios y de promoción de la continuidad. Lima: MINEDU; 2024. Disponible en:

34. https://cdn.www.gob.pe/uploads/document/file/3018068/III%20Informe%20Bienal.pdfSuperintendencia Nacional de Educación Superior Universitaria (SUNEDU). III Informe Bienal sobre la Realidad Universitaria en el Perú. Lima: SUNEDU; 2022. Disponible en:

35. https://elcomercio.pe/peru/minedu-tasa-de-desercion-en-educacion-universitaria-se-redujo-a-115-en-el-periodo-2021-1-nndc-noticia/El Comercio. Minedu: tasa de deserción en educación universitaria se redujo a 11.5% en el periodo 2021-1. Lima: El Comercio; 2021. Disponible en:

36. https://doi.org/10.23913/ride.v14i28.1923Villegas BR, Núñez Lira LA. Factores asociados a la deserción estudiantil en el ámbito universitario. Una revisión sistemática 2018–2023. RIDE Rev Iberoam Para Investig Desarro Educ. 2024;14(28). Disponible en:
37. Few S. Information Dashboard Design: The Effective Visual Communication of Data. Sebastopol: O'Reilly Media; 2006. ISBN: 978-0-596-10016-7.

38. Tufte ER. The Visual Display of Quantitative Information. 2nd ed. Cheshire: Graphics Press; 2001. ISBN: 978-0-9613921-4-7.

39. https://doi.org/10.1007/978-3-319-19425-7Harrell FE Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. New York: Springer; 2015. Disponible en:

40. https://doi.org/10.1007/978-1-4614-6849-3Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer; 2013. Disponible en:

41. https://escale.minedu.gob.pe/ueetendencias2016?p_auth=x48qKM01&p_p_id=TendenciasActualPortlet2016_WAR_tendencias2016portlet_INSTANCE_t6xG&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&p_p_col_id=column-1&p_p_col_pos=1&p_p_col_count=3&_TendenciasActualPortlet2016_WAR_tendencias2016portlet_INSTANCE_t6xG_idCuadro=371#descargarUnidad de Estadística Educativa - Ministerio de Educación [citado el 5 de mayo de 2026]. Disponible en:

ANEXOS

Anexo 1 - *Matriz de consistencia de variables del sistema predictivo*

Problema General	Objetivo General	Hipótesis General
¿Puede un sistema predictivo basado en variables académicas preuniversitarias, universitarias y sociodemográficas alcanzar una capacidad discriminativa superior a AUC-ROC ≥ 0.75 para la identificación temprana de estudiantes con riesgo de deserción en la UPCH, manteniendo calibración probabilística e interpretabilidad en validación externa por cohortes?	Desarrollar y validar un sistema predictivo basado en clasificadores supervisados (Random Forest, XGBoost, Regresión Logística) que integre variables académicas preuniversitarias, universitarias y sociodemográficas para la identificación temprana de estudiantes con riesgo de deserción o bajo rendimiento en la UPCH, evaluando su desempeño mediante validación temporal externa (cohorte 2024-1) con énfasis en calibración probabilística, interpretabilidad y viabilidad de implementación	Un sistema predictivo entrenado con variables académicas preuniversitarias y universitarias, junto con factores sociodemográficos, alcanza una capacidad discriminativa superior a AUC-ROC ≥ 0.75 para identificar estudiantes con riesgo de deserción antes del segundo ciclo en la UPCH, evaluado mediante validación externa por cohortes (2024-1), manteniendo calibración probabilística (Brier Score ≤ 0.20) e interpretabilidad mediante análisis SHAP.

Objetivos Específicos
OE1: Cuantificar el poder predictivo relativo de variables académicas preuniversitarias (promedio de secundaria, puntaje de admisión) versus variables del primer ciclo universitario mediante análisis de Permutation Importance y valores SHAP, evaluado en la cohorte de validación externa (2024-1).
OE2: Evaluar la existencia de heterogeneidad en las distribuciones de probabilidad de riesgo entre subgrupos socioeconómicos (becados vs. no becados) y determinar la necesidad de umbrales de alerta diferenciados mediante análisis de curvas de precisión-recall estratificadas.
OE3: Validar la estabilidad temporal de los patrones predictivos mediante comparación de la importancia relativa de las variables entre el conjunto de entrenamiento (2022-2023) y el conjunto de validación externa (2024-1), cuantificando la variación en rankings y magnitudes.
OE4: Desarrollar y evaluar un prototipo funcional (API REST + dashboard interactivo) que demuestre la viabilidad técnica de despliegue mediante medición de tiempos de respuesta, capacidad de carga y disponibilidad en un entorno de prueba controlado.

Anexo 2 – *Aprobación del comité de ética*

UNIVERSIDAD PERUANA
CAYETANO HEREDIA

CONSTANCIA-CIEI-533-37-25

El Presidente del Comité Institucional de Ética en Investigación (CIEI) de la Universidad Peruana Cayetano Heredia hace constar que el proyecto de investigación señalado a continuación fue **APROBADO** de manera unánime por el Comité de Ética.

Título del Proyecto	**: "Sistema predictivo basado en técnicas de machine learning para la detección temprana del riesgo académico y la deserción estudiantil en una universidad peruana."**
Código SIDISI	**: 218776**
Investigador(a) principal(es)	**: Saldaña Rodriguez Sebastian Antonio**

La aprobación incluyó los documentos finales descritos a continuación:

1. **Protocolo de investigación, versión 2.0 de fecha 20 de octubre del 2025..**

La **APROBACIÓN** considera el cumplimiento de los estándares de la Universidad, los lineamientos científicos y éticos, el balance riesgo/beneficio, la calificación del equipo investigador y la confidencialidad de los datos, entre otros.

Cualquier enmienda, desviaciones, eventualidad deberá ser reportada de acuerdo a los plazos y normas establecidas. El investigador reportará cada seis meses el progreso del estudio y alcanzará un informe al término de éste. La aprobación tiene vigencia desde la emisión del presente documento hasta el **Sábado 24 de octubre del 2026**.

El presente proyecto de investigación sólo podrá iniciarse después de haber obtenido la(s) autorización(es) de la(s) institución(es) donde se ejecutará.

Si aplica, los trámites para su renovación deberán iniciarse por lo menos 30 días previos a su vencimiento.

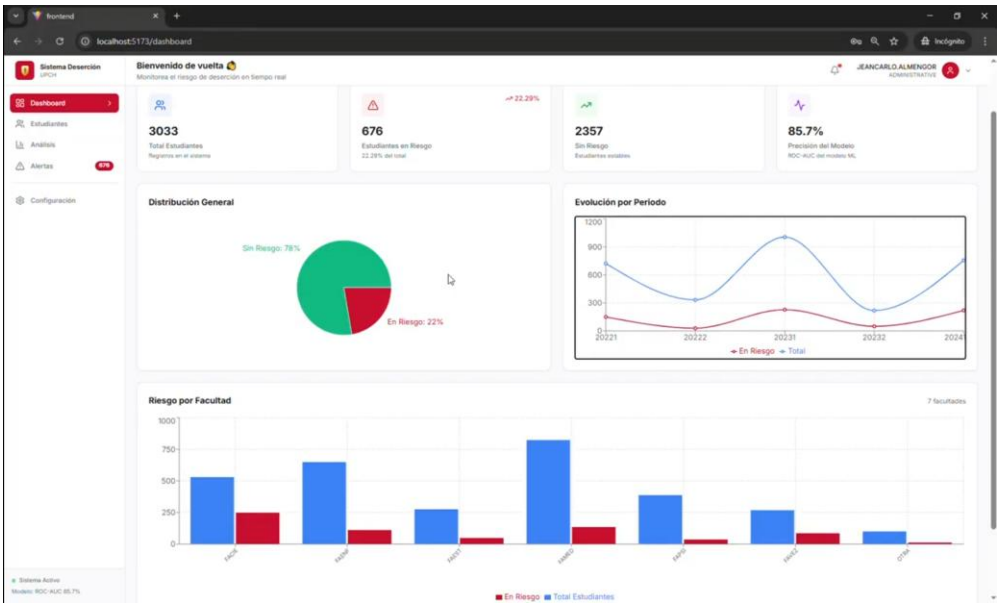
Lima, 24 de octubre del 2025

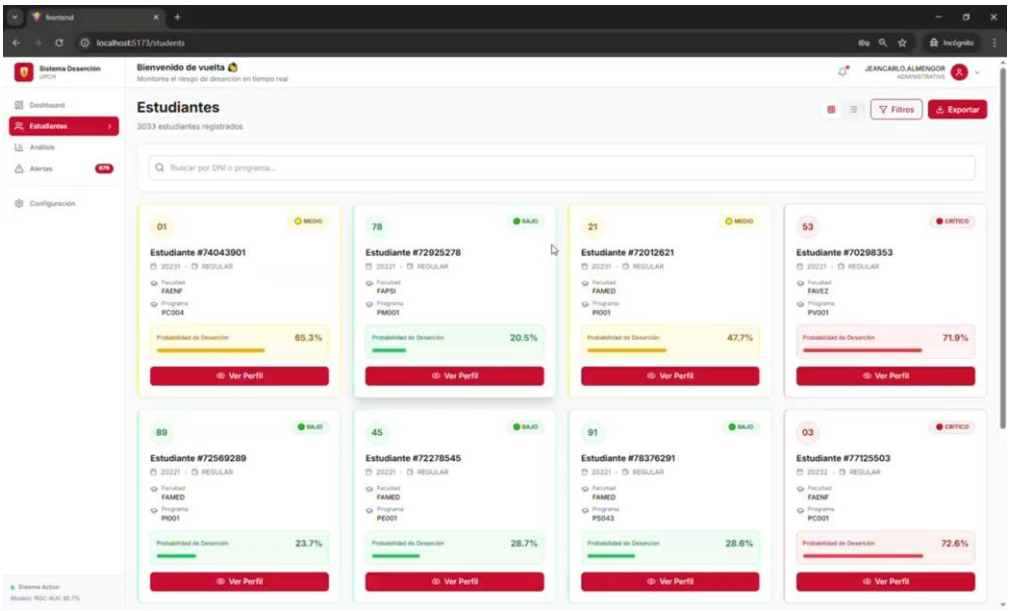
Manuel Raul Perez Martinot
Presidente
Comité Institucional de Ética en Investigación
Universidad Peruana Cayetano Heredia

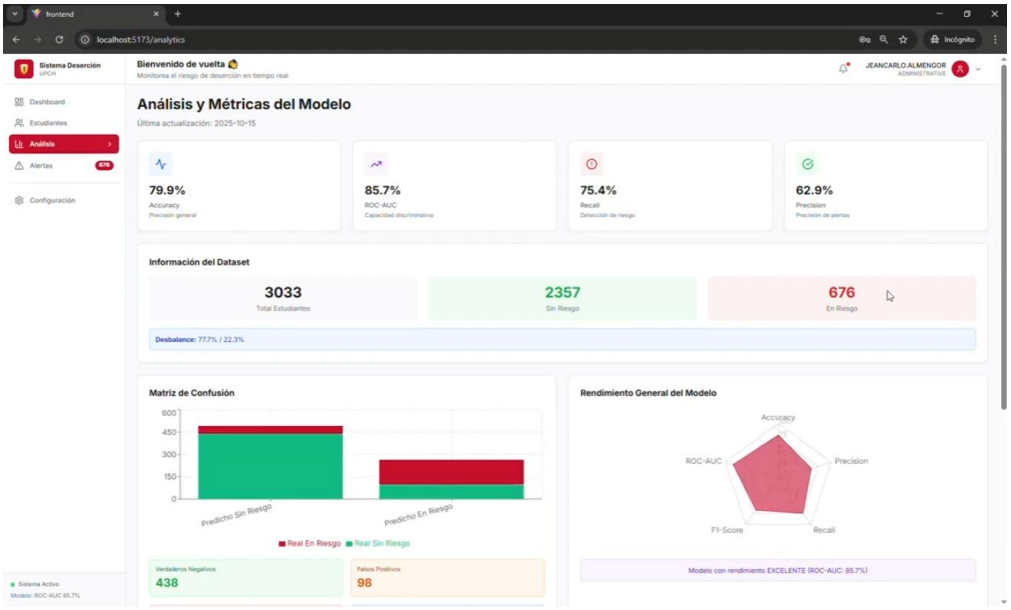
Av. Honorio Delgado 430
San Martín de Porres
Apartado postal 4314
319 0000 Anexo 201355
orvei.ciei@oficinas-upch.pe
www.cayetano.edu.pe

Comité Institucional de Ética en Investigación

Anexo 3 – Capturas de pantalla del sistema funcional







Observaciones subsanadas:

Jurado 1:

Observaciones	**¿Se subsanó?**
Describir el modelo predictivo (la ecuación o función).	Sí. En la 5.13.1 se describen las fórmulas utilizadas
colocar el tipo de análisis que utilizo para el manejo de datos faltantes.	Si. En la 5.11.1 se describe cómo se trataron los datos faltantes.
Describir como se manejo el sesgo de selección.	Si. En la 5.5.1 se habla acerca del sesgo de selección
Definir el tamaño de la poblacion y describir el procedimiento que realizo para llegar a la muestra final para el estudio.	Si. En la 5.8 y en la Figura 7 se describe el tamaño de la población y cómo se llegó a la muestra

Jurado 2:

Observaciones	**¿Se subsanó?**
En el marco teórico, se recomienda actualizar la estadística sobre la tasa de deserción, ya que actualmente se presenta información correspondiente al periodo 2020-1, el cual responde a un contexto excepcional debido a la pandemia.	Sí. En la 2.2.1, figura 2, se actualizó la estadística hasta el año 2024
Incluir el diagrama del flujo del dataset inicial hasta el dataset final utilizado en el análisis	Si. En la Figura 7 en el punto 5.8
Es necesario detallar y justificar la exclusión de las variables sociodemográficas; esta explicación debe incluirse en detalle en la sección de limitaciones del estudio. Actualmente, solo indica la exclusión y no el porqué.	Si. En la 5.5.1 se habla acerca de la exclusión de variables sociodemográficas y porqué
Se recomienda añadir un anexo en el documento que incluya capturas de pantalla (screenshots) del dashboard desarrollado, a fin de evidenciar su diseño y funcionalidad	Si. En el anexo 3

Jurado 3

Observaciones	**¿Se subsanó?**
Precisar con mayor detalle la implementación técnica del sistema, especialmente el flujo de inferencia (entrada de datos, procesamiento del modelo y generación de predicciones), a fin de mejorar la claridad desde un enfoque de ingeniería	Sí. En la 5.13.1 se describe la definición funcional del modelo predictivo, junto con sus datos de entrada, generación, etc.
Aclarar de manera consistente las tecnologías utilizadas en el prototipo, asegurando una adecuada alineación entre lo descrito en el documento y lo presentado durante la sustentación	Si. En la 5.3 se habla sobre las tecnologías utilizadas, las cuales están alineadas en todo el documento.
Explicitar con mayor claridad la definición funcional del modelo predictivo, indicando las variables de entrada, el comportamiento del modelo y su lógica de predicción, con el fin de mejorar la reproducibilidad técnica del estudio	Sí. En la 5.13.1 se describe la definición funcional del modelo predictivo, junto con sus datos de entrada, generación, etc.
Presentar de forma más clara la arquitectura del sistema, diferenciando visualmente sus componentes principales (modelo, backend/API, frontend y dashboard).	Si. En la figura 4, 5 y 6 se diagrama el funcionamiento del sistema, los casos de uso y la arquitectura.

Anexo 2

Modelo de Informe de uso de IA[1]

De conformidad con lo dispuesto en los Lineamientos sobre el uso de la Inteligencia Artificial (IA) en la Universidad Peruana Cayetano Heredia, el presente informe constituye una declaración formal y transparente sobre el empleo de herramientas de IA en la elaboración de trabajos que conducen a la obtención de grados académicos y títulos profesionales.

Este documento es de carácter obligatorio y busca garantizar la trazabilidad, autocrítica y originalidad del trabajo presentado.

Datos de identificación

Nombres y apellidos del declarante: **Sebastian Antonio Saldaña Rodríguez**

Docente: ___ Estudiante: **X**

Documento de identidad
(DNI/CE):

Unidad académica / Programa: **FACI/Ingeniería Informática**

Tipo de trabajo: **Tesis X** Trabajo de suficiencia profesional _____Trabajo de investigación ___Otro ____

Título de trabajo: Sistema predictivo basado en técnicas de machine learning para la detección temprana del riesgo académico y la deserción estudiantil en una universidad peruana.

Fecha de presentación:

Herramienta de IA utilizada

Nombre de la herramienta: Consensus

Versión: web

Finalidad de uso: Búsqueda y filtrado de papers académicos por relevancia temática.

Nombre de la herramienta: Elicit

Versión: web

Finalidad de uso: Análisis de papers y extracción de hallazgos por pregunta de investigación.

Nombre de la herramienta: Claude (Anthropic)

Versión: web

Finalidad de uso: Apoyo en revisión y mejora de redacción académica

Descripción del uso de IA

Etapas o pasos donde se empleó la IA: búsqueda y filtrado de papers por relevancia temática mediante palabras clave relacionadas a machine learning, deserción estudiantil y riesgo académico; así como, extracción de fragmentos de artículos científicos para su posterior análisis y contraste con fuentes primarias.

Validación del contenido generado

Métodos aplicados para verificar exactitud, pertinencia y originalidad revisión con asesor; verificación de veracidad mediante contraste con bases de datos académicas (Scopus, ScienceDirect, entre otras).

Ajustes o modificaciones realizadas por el autor humano: reescritura y adaptación de los contenidos al contexto específico de la investigación, con terminología propia del área de sistemas de información y aprendizaje automático.

Declaración jurada de autoría crítica

Solo aplica para elaboración de trabajo de investigación, tesis o trabajo de suficiencia profesional para obtener títulos o grados académicos.

Declaro que el presente trabajo fue elaborado con base en mi propio análisis, interpretación y redacción, haciendo uso de herramientas de inteligencia artificial únicamente como apoyo, y **ejerciendo control crítico y revisión humana sobre todo el contenido**.

Afirmo que he contrastado la información obtenida con fuentes académicas confiables y que el producto final **responde a mi propio proceso de aprendizaje, investigación y reflexión.**
Asumo plena responsabilidad ética y académica sobre la veracidad y originalidad de este trabajo.

Reconozco que la omisión de este informe, la inclusión de datos falsos o el uso de recursos sin la debida trazabilidad constituyen una falta grave, sujeta a las sanciones establecidas en el Reglamento del Personal Académico-Docente y en el Reglamento Disciplinario para Estudiantes y Graduados de la UPCH.

Firma digital

Firma digital del declarante:

Fecha y hora: 22/06/2026