REPORTE TÉCNICO 2

HACIA UNA PREDICCIÓN TEMPRANA DEL AUSENTISMO CRÓNICO EN SECUNDARIA

INEEd
Instituto Nacional de Evaluación Educativa

La elaboración de este documento estuvo a cargo de: Gimena Castelao y Vivian Couto

Tutor: Martín Fernández Campos

Corrección de estilo: Mercedes Pérez y Federico Bentancor
Diseño y diagramación: Diego Porcelli

Montevideo, 2025
ISSN: 3046-4390

Edificio Los Naranjos, Planta Alta, Parque de Innovación del LATU
Av. Italia 6201, Montevideo, Uruguay
(+598) 2604 4649 – 2604 8590
ineed@ineed.edu.uy
www.ineed.edu.uy

Cómo citar: Castelao, G. y Couto, V. (2025). *Reporte técnico 2. Hacia una predicción temprana del ausentismo crónico en secundaria.* Recuperado de https://www.ineed.edu.uy/images/publicaciones/reportes/Reporte-tecnico-2-Hacia-una-prediccion-temprana-del-ausentismo-cronico-en-secundaria.pdf

Este reporte trata de adolescentes y adultos mujeres y varones. El uso del masculino genérico obedece a un criterio de economía de lenguaje y procura una lectura más fluida, sin ninguna connotación discriminatoria.

Nota aclaratoria

Este informe técnico presenta el proyecto final de Gimena Castelao y Vivian Couto para el Diploma en Analítica de Datos Aplicada a Proyectos de la Universidad Tecnológica del Uruguay (UTEC). Ambas investigadoras cursaron con becas de la universidad que cubrieron la mitad del costo del programa y una de ellas recibió además un apoyo económico complementario del Instituto Nacional de Evaluación Educativa (INEEd). El estudio ejemplifica la articulación entre academia y política pública, fortaleciendo la capacidad local para diseñar herramientas de análisis de datos adaptadas a la realidad nacional. Su publicación responde al mandato del INEEd de generar y difundir evidencia que fortalezca el sistema educativo uruguayo y cuenta con la autorización expresa de las autoras para su difusión.

Es interés de INEEd publicar este trabajo porque constituye un aporte sustantivo a un problema de política pública cada vez más relevante. Específicamente, el estudio contribuye a predecir, a través de datos administrativos de la Dirección General de Educación Secundaria (DGES), cuáles estudiantes de liceos públicos tienen mayores probabilidades de presentar problemas en su asistencia. Conocer esta información a tiempo permite la intervención temprana. A su vez, el haber identificado factores que contribuyen en mayor medida con la problemática permite también que dichas intervenciones se orienten a las causas.

Los resultados muestran que los patrones de inasistencia se configuran muy temprano en el ciclo lectivo, ya que las ausencias acumuladas en marzo, abril y mayo son el predictor más potente, lo que permitiría realizar intervenciones preventivas y oportunas.

A ello se suman la persistencia individual, evidenciada por el aumento de probabilidad de reincidir en quienes presentaron ausencias crónicas en años anteriores; el efecto de pares, ya que un promedio elevado de faltas a nivel grupal incrementa el riesgo individual, y las condiciones socioeconómicas del estudiante y del centro, que refuerzan el ausentismo y subrayan la dimensión estructural de este problema. Además, variables académicas como la repetición, la extraedad y el bajo desempeño previo completan el perfil de riesgo, mientras que factores organizativos como el turno vespertino y la residencia en zonas rurales agravan aún más la situación.

Por último, cabe destacar que este reporte presenta, a modo de referencia, un prototipo de sistema de alerta temprana construido sobre los datos disponibles, que podría ayudar a los liceos públicos a identificar con antelación a los estudiantes del centro con mayor riesgo de ausentismo antes del cierre del primer trimestre lectivo. Aunque el modelo puede perfeccionarse y adaptar sus umbrales a las prioridades y recursos reales de la DGES, su calibración inicial demuestra cómo umbrales bajos habilitan intervenciones universales de bajo costo (por ejemplo, con mensajes automatizados) y umbrales altos concentran esfuerzos en casos críticos que requieren tutorías o acompañamiento personalizado.

RESUMEN

En Uruguay, el ausentismo crónico en educación media es un problema crítico, asociado a un mayor riesgo de fracaso académico y abandono escolar, que se profundiza en los sectores con mayor vulnerabilidad. Este proyecto busca desarrollar un modelo predictivo que identifique a estudiantes de educación media básica de liceos públicos en riesgo de ausentismo crónico, utilizando técnicas de aprendizaje automático. Para ello, se utilizan datos sociodemográficos, académicos y contextuales de 2016 a 2023, provenientes de registros administrativos de la Administración Nacional de Educación Pública.

Se probaron varios modelos de clasificación, incluyendo la regresión logística y algoritmos XGBoost, LightGBM y CatBoost, siendo Catboost el modelo que obtuvo el mejor desempeño, con una exactitud (*accurracy)* del 84% y un AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*) de 0,918. Este rendimiento demuestra la capacidad del modelo para identificar de manera precisa a los estudiantes en riesgo de ausentismo crónico, facilitando la implementación de intervenciones preventivas oportunas que promuevan la asistencia y mejoren su trayectoria educativa.

INTRODUCCIÓN

En el contexto de la educación secundaria en Uruguay, el ausentismo crónico se ha convertido en una de las principales señales de alerta temprana de problemas académicos y potencial abandono escolar (INEEd, 2020, 2023; London, Sanchez y Castrechini, 2016). La Administración Nacional de Educación Pública (ANEP) utiliza la métrica de "inasistencia ficta" para el seguimiento del ausentismo, la cual considera tanto las inasistencias injustificadas como la mitad de las justificadas[1]. Este indicador subraya la necesidad de intervenciones oportunas que promuevan la asistencia y reduzcan el riesgo de fracaso académico.

El ausentismo crónico, definido como la no asistencia al 10% o más de los días lectivos, es un problema crítico en el ámbito educativo, ya que está asociado con una mayor probabilidad de fracaso académico y abandono escolar (ANEP, 2023b; Kearney y Childs, 2023; UNESCO, 2021). Identificar tempranamente a los estudiantes en riesgo de ausentismo crónico permitiría implementar intervenciones más específicas y oportunas para mejorar las trayectorias educativas.

Este proyecto busca desarrollar un modelo predictivo de ausentismo crónico para estudiantes de educación media básica en liceos públicos de Uruguay, utilizando técnicas avanzadas de aprendizaje automático. Para lograr esto, se integraron datos de variables clave que capturan aspectos sociodemográficos, académicos y contextuales. Entre las variables incluidas en el análisis se encuentran el sexo, la edad, el índice de vulnerabilidad social (IVS) del estudiante, la región y zona geográfica de residencia, así como el desempeño académico en años previos. Además, se consideraron variables de contexto escolar, como el quintil de vulnerabilidad social del centro educativo y el promedio trimestral de inasistencias del grupo, las cuales reflejan las influencias del entorno educativo sobre la asistencia.

Para construir el modelo predictivo, se emplearon técnicas de aprendizaje automático, incluyendo la regresión logística y métodos avanzados de *boosting*, como XGBoost (*Extreme Gradient Boosting*), LightGBM (*Light Gradient Boosting Machine*) y CatBoost. Estos modelos son capaces de manejar la complejidad de múltiples variables interrelacionadas y de identificar patrones de riesgo de ausentismo crónico con alta precisión. La elección de estos modelos se basa en su capacidad para trabajar con grandes volúmenes de datos y su eficacia comprobada en tareas de clasificación en otros contextos educativos.

El objetivo es que el modelo logre predecir con precisión, al finalizar el primer trimestre lectivo, el riesgo de ausentismo crónico al finalizar el año escolar. De esta forma, el modelo

[1] Circular n.° 2704/06 del Consejo de Educación Secundaria.

CatBoost logró predecir el ausentismo crónico con una exactitud (*accuracy)* del 84%, lo que permitirá una intervención oportuna en aquellos estudiantes que se encuentren en riesgo de ausentismo crónico antes de que este se consolide.

Se concluye que el ausentismo crónico está estrechamente relacionado con las variables sociodemográficas, académicas y contextuales analizadas, lo que subraya la importancia de abordar este fenómeno desde un enfoque integral. Este proyecto representa un avance en el desarrollo de herramientas predictivas en la educación uruguaya, brindando la posibilidad de implementar un sistema de alerta temprana que permita reducir las tasas de ausentismo crónico y, en última instancia, mejorar las oportunidades de éxito académico para los estudiantes.

JUSTIFICACIÓN Y OBJETIVOS

El ausentismo crónico es un problema que, si no se aborda de manera temprana, puede convertirse en abandono escolar. Estudios recientes, como el de Kearney y Childs (2023), han demostrado que los estudiantes que son crónicamente ausentes en la educación media tienen mayores probabilidades de abandonar la enseñanza en comparación con aquellos que mantienen una asistencia regular. Además, investigaciones como las de London et al. (2016) sugieren que predecir el ausentismo crónico y actuar preventivamente puede ser crucial para mejorar las tasas de finalización educativa.

Dado que el ausentismo es un indicador temprano del bajo rendimiento académico, la implementación de un sistema de análisis predictivo basado en técnicas de *machine learning* se presenta como una herramienta viable para abordar este problema. Según datos del Monitor Educativo Liceal, la métrica de "inasistencia ficta" ha revelado una tendencia creciente en el ausentismo, con promedios de inasistencias que han subido de 27,1 en 2019 a 32,2 en 2023. Asimismo, el porcentaje de egresados de educación media superior en el tramo de 21 a 23 años ha aumentado de manera lenta en los últimos años y se sitúa en 51,6%, porcentaje muy distante de la meta de 100% planteada por la ANEP (INEEd, 2024). Estos resultados destacan la necesidad de intervenciones efectivas y sostenidas para reducir el ausentismo y apoyar la retención escolar.

Los modelos predictivos han demostrado su eficacia para identificar a los estudiantes en riesgo de deserción en estudios previos (Queiroga, Batista Machado, Paragarino, Primo y Cechinel, 2022), brindando herramientas a los gestores educativos para intervenir de manera oportuna con estrategias focalizadas. Este proyecto se justifica en la necesidad de contar con herramientas de apoyo a la toma de decisiones basadas en datos reales, capaces de predecir con precisión el comportamiento de los estudiantes en términos de asistencia y rendimiento. Al igual que en los estudios sobre predicción de deserción (Macarini et al., 2018; Queiroga et al., 2022), el uso de algoritmos avanzados para la predicción del ausentismo puede generar intervenciones tempranas más eficientes y mejorar las tasas de retención escolar.

El objetivo general del proyecto se enfoca en desarrollar y validar un modelo predictivo capaz de identificar tempranamente a estudiantes de educación media básica de liceos públicos en Uruguay que se encuentren en riesgo de ausentismo crónico. Este modelo permitirá facilitar la implementación de intervenciones preventivas oportunas, contribuyendo a mejorar la retención escolar y a promover trayectorias educativas más estables para los estudiantes en contextos vulnerables.

OBJETIVOS ESPECÍFICOS

1. Analizar la influencia de diversas variables: examinar la relación entre variables sociodemográficas (como nivel socioeconómico, sexo, edad y región geográfica), académicas (rendimiento previo) y de contexto escolar (índice de vulnerabilidad del centro educativo, promedio de inasistencias del grupo) con el ausentismo crónico, con el fin de identificar los factores de riesgo más relevantes asociados a la inasistencia.

2. Desarrollar modelos predictivos: implementar diferentes modelos de aprendizaje automático (regresión logística, XGBoost, LightGBM y CatBoost) para predecir el riesgo de ausentismo crónico en estudiantes de educación media.

3. Comparar el desempeño de los modelos desarrollados: evaluar y comparar los modelos predictivos mediante métricas de desempeño como precisión, sensibilidad y AUC-ROC, para seleccionar el modelo más efectivo en términos de precisión y capacidad de generalización.

4. Validar el modelo predictivo seleccionado: aplicar técnicas de validación cruzada en el modelo seleccionado para garantizar su robustez y confiabilidad en la identificación de estudiantes en riesgo de ausentismo crónico.

REVISIÓN DE LA LITERATURA

El ausentismo crónico ha sido objeto de análisis de varios estudios a nivel internacional debido a sus efectos negativos sobre el rendimiento académico y la continuidad educativa. Diversas investigaciones han abordado factores de riesgo asociados a la inasistencia que proporcionan una línea de base relevante para el desarrollo de modelos predictivos que permitan la identificación temprana de estudiantes en riesgo de ausentismo. Esta revisión bibliográfica examina algunas investigaciones nacionales e internacionales que exploran el ausentismo estudiantil y la aplicación de modelos predictivos en educación, proporcionando un marco comparativo que justifica la pertinencia de este proyecto en el contexto uruguayo.

En el ámbito internacional, el ausentismo crónico ha sido ampliamente estudiado, dado que se trata de un problema que afecta a los sistemas educativos de todo el mundo. Diversas investigaciones han identificado factores de riesgo asociados con la inasistencia, tales como el apoyo familiar, el contexto económico y el entorno escolar. Estudios recientes como el de Kearney y Childs (2023) han destacado cómo el uso de estrategias analíticas avanzadas permite entender mejor las causas subyacentes del ausentismo. Estos autores emplearon una variedad de técnicas de análisis predictivo, como árboles de clasificación y regresión (*classification and regression tree*, CART), análisis de bosque aleatorio y máquinas de vectores de soporte para identificar patrones en grandes conjuntos de datos de asistencia escolar. Entre estos modelos, el análisis de bosque aleatorio resultó ser el más efectivo, alcanzando un AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*) de 0,92 en la predicción de riesgo de ausentismo crónico. Los hallazgos de Kearney y Childs demuestran que, además de factores académicos, variables relacionadas con el entorno familiar y la seguridad escolar son determinantes significativos en la predicción de inasistencia. Este enfoque metodológico facilita el desarrollo de políticas educativas específicas y adaptadas a las condiciones locales, mejorando la implementación de sistemas de alerta temprana y de estrategias de intervención en múltiples niveles de soporte.

En Albania, el estudio de Mukli y Rista (2022) demuestra cómo la aplicación de algoritmos de aprendizaje automático sobre una base de 210.000 registros y 26 variables permitió identificar factores críticos en las tasas de ausencias. Utilizando el modelo Bayes Net, que alcanzó una precisión del 97,5%, los investigadores señalaron los problemas familiares y la insatisfacción con el entorno universitario como determinantes clave en el ausentismo estudiantil. Por otro lado, en Estados Unidos, Lee et al. (2023) examinaron el ausentismo en educación secundaria mediante un enfoque basado en la teoría ecológica, capturando factores de influencia en diferentes niveles: individual, familiar, escolar y comunitario. Emplearon un modelo de bosque aleatorio que alcanzó una precisión del 85%, un AUC-ROC de 0,88 y una sensibilidad (*recall)* del 80%, lo que evidenció cómo las condiciones

socioeconómicas y el apoyo social impactan en la asistencia escolar. Estos estudios internacionales resaltan que el ausentismo crónico responde a una variedad de factores interrelacionados y enfatizan la importancia de adoptar un enfoque integral y personalizado para abordar esta problemática.

En América Latina, algunos estudios también han explorado el uso de modelos predictivos para abordar el ausentismo estudiantil. En Colombia, Arteaga y Tapias (2024) desarrollaron un sistema de recomendación de acciones en la Institución Educativa Cecilia de Lleras que clasificaba a los estudiantes en niveles de riesgo de inasistencia (alto, medio y bajo). Su modelo de bosque aleatorio alcanzó una precisión de 99%, destacando el impacto del estado socioeconómico y las condiciones familiares como factores determinantes en la predicción de ausencias escolares.

En Argentina, Torino (2023) desarrolló un modelo predictivo basado en *machine learning* para abordar la deserción escolar, empleando una muestra representativa de datos socioeconómicos y habitacionales de la Encuesta Permanente de Hogares. Su modelo de *boosting* alcanzó un AUC-ROC del 0,87, identificando factores como el acceso limitado a recursos educativos, condiciones habitacionales deficientes y la necesidad de trabajar desde temprana edad como elementos que contribuyen significativamente al riesgo de abandono.

En Uruguay, el uso de modelos predictivos ha estado mayoritariamente enfocado en la deserción y en el rendimiento académico, dejando un área menos explorada en torno al ausentismo. Aguirre Imbriaco y Veneri (2018), por ejemplo, emplearon datos del programa Compromiso Educativo para modelar factores asociados a la promoción en estudiantes de cuarto año de educación media. Utilizando modelos como regresión logística y bosques aleatorios de árboles condicionales (Conditional random field, CRF), su análisis identificó qué variables (como el historial de exámenes aprobados, la trayectoria sin repetición y las expectativas familiares de educación terciaria) eran predictores de éxito académico. El CRF con muestreo simple mostró un AUC-ROC de 79,5% y una precisión parcial de 66,8%.

Otro estudio relevante en Uruguay es el de Cardozo et al. (2022), que desarrollaron un sistema predictivo en educación primaria utilizando datos de la Evaluación Infantil Temprana para predecir el riesgo de repetición escolar. Aplicaron modelos de regresión logística, redes bayesianas y CART y lograron una precisión del 80%, con sensibilidades de entre el 62% y el 64% y una especificidad del 83%, destacando cómo las habilidades cognitivas y socioemocionales pueden predecir el rezago académico. Aunque este estudio no se enfocó en la educación media ni en el ausentismo, demuestra el potencial de los modelos predictivos para anticipar trayectorias educativas y orientar intervenciones en los primeros años escolares.

En el nivel de educación media, Macarini et al. (2018) desarrollaron un sistema de alerta temprana en Uruguay para identificar estudiantes en riesgo de abandono escolar, empleando un modelo de bosque aleatorio para analizar datos de rendimiento, asistencia y variables sociodemográficas. Este modelo alcanzó un AUC-ROC superior a 0,90, permitiendo clasificar a los estudiantes en niveles de riesgo bajo, moderado y alto.

Por su parte, Queiroga et al. (2022) desarrollaron un sistema predictivo en colaboración con la ANEP para detectar estudiantes en riesgo de fracaso académico o abandono en educación media básica en Uruguay. Utilizando modelos de *random forest y redes neuronales multicapa* (*multi-layer perceptron*, MLP) con datos de 261.446 estudiantes, lograron un AUC-ROC superior a 0,90 para segundaria y 0,95 para técnico profesional, mostrando un excelente desempeño en la identificación temprana de riesgo. Este sistema de alerta temprana se encuentra en proceso de ajuste para su implementación en el monitoreo de centros educativos, permitiendo intervenciones adaptadas desde el inicio de la trayectoria educativa de los estudiantes.

Finalmente, Alvez Legelén (2024) desarrolló un estudio sobre las trayectorias educativas en educación media en Uruguay, siguiendo a estudiantes de sexto grado de primaria durante ocho años (2013-2021) para identificar aquellos en riesgo de desvinculación. A través de modelos Logit, el estudio mostró que factores como el sexo (masculino), un bajo nivel socioeconómico, el rezago escolar y el bajo rendimiento en matemática y lectura están correlacionados con el riesgo de desvinculación. El modelo final alcanzó una precisión del 61,1%, una sensibilidad del 73,4%, una especificidad del 86,9% y un AUC-ROC de 83,9%.

Aunque en Uruguay se han desarrollado varias investigaciones sobre modelos predictivos aplicados al rendimiento y abandono escolar, no se encontraron estudios que se hayan centrado en predecir específicamente el ausentismo. Los estudios existentes, como los apoyados en datos de Aristas (INEEd, 2020, 2023) y el Programa para la Evaluación Internacional de Alumnos (PISA, por su sigla en inglés) (ANEP, 2023b), señalan que la asistencia escolar es una problemática relevante en la educación media uruguaya, especialmente en los sectores más vulnerables.

La revisión de la literatura deja en evidencia la efectividad de los modelos predictivos para anticipar el ausentismo y otros desafíos académicos en diversos contextos. En América Latina, varios estudios destacan factores socioeconómicos y familiares como determinantes en la asistencia escolar. En Uruguay, aunque se han implementado modelos predictivos en áreas como la deserción, el ausentismo sigue siendo tratado principalmente como un indicador temprano de riesgo de abandono o bajo rendimiento. En lugar de enfocarse en predecir directamente el ausentismo, muchos estudios lo utilizan para señalar otros problemas, limitando la posibilidad de entender y abordar los factores que lo desencadenan.

Este proyecto busca dar un paso atrás al centrar su análisis en el ausentismo en sí mismo, lo que permitirá anticipar el problema y orientar intervenciones hacia los factores específicos que afectan la asistencia. Así, al mejorar la identificación temprana de estudiantes en riesgo de ausentismo crónico, el proyecto contribuirá a implementar estrategias preventivas más precisas, promoviendo una educación más inclusiva y accesible.

MARCO TEÓRICO Y CONCEPTUAL

EL AUSENTISMO CRÓNICO COMO INDICADOR DE RIESGO Y SUS FACTORES ASOCIADOS

El ausentismo es un indicador clave de riesgo en el ámbito educativo, ya que afecta tanto el rendimiento académico como la continuidad escolar de los estudiantes (de Melo, Failache y Machado, 2015; London et al., 2016). Definido generalmente como la ausencia al 10% o más de los días lectivos, este fenómeno no solo es un reflejo de problemas inmediatos de asistencia, sino un síntoma de trayectorias no óptimas y de posibles desvinculaciones del sistema educativo (Kearney y Childs, 2023). En el contexto uruguayo, donde solo el 51,6% de los jóvenes entre 21 y 23 años ha culminado la educación media superior (INEEd, 2024), el ausentismo crónico se convierte en una problemática urgente, especialmente en sectores de mayor vulnerabilidad socioeconómica (ANEP, 2023a). Los adolescentes que no asisten regularmente presentan mayor probabilidad de desvincularse, una situación particularmente acentuada en aquellos provenientes de sectores socioeconómicos bajos, donde la asistencia escolar enfrenta obstáculos significativos relacionados con el trabajo juvenil y la falta de recursos familiares (ANEP, 2022). Desde la perspectiva de la oportunidad de aprendizaje, el ausentismo crónico limita las posibilidades de logro académico, reduciendo el tiempo efectivo de aprendizaje y aumentando el riesgo de fracaso escolar. En términos de integración escolar, este fenómeno incrementa las trayectorias educativas no óptimas, como la repetición de grado y el abandono escolar.

Desde la perspectiva de la teoría ecológica del desarrollo (Bronfenbrenner, 1979), este trabajo considera los niveles más directamente relacionados con el entorno inmediato del individuo: el microsistema, el mesosistema y el exosistema. En primer lugar, el microsistema comprende los contextos inmediatos en los que el individuo interactúa directamente, como la familia, la escuela, los compañeros y el vecindario. Estas interacciones son las más influyentes en el desarrollo diario. Por otro lado, el mesosistema refiere a las conexiones e interacciones entre los distintos microsistemas. Por ejemplo, la relación entre los padres y los profesores de un estudiante o cómo la dinámica familiar afecta su desempeño escolar. Por último, el exosistema incluye entornos más amplios que afectan indirectamente al individuo, como las políticas escolares, las condiciones laborales de los padres o los servicios comunitarios disponibles en la zona.

Esta perspectiva es útil para analizar el ausentismo escolar, dado que muestra cómo factores individuales, familiares, escolares y comunitarios interactúan y afectan las trayectorias educativas. Desde esta óptica, el ausentismo escolar puede entenderse como un fenómeno

influenciado por la interrelación de factores presentes en el nivel familiar (microsistema y, en sus relaciones, mesosistema), escolar (microsistema y mesosistema) y comunitario (exosistema), destacando la importancia de los contextos inmediatos y las estructuras más amplias en las trayectorias educativas (Arteaga Ramos y Tapias López, 2024; Dräger, Klein y Sosu, 2023; Liu y Lee, 2022).

En el ámbito familiar, el contexto socioeconómico es un factor determinante, dado que los estudiantes provenientes de familias de bajos ingresos enfrentan barreras adicionales para asistir regularmente, como la necesidad de trabajar, problemas de transporte o de salud sin la debida atención. En el hogar, la estabilidad emocional y el apoyo percibido también influyen en la asistencia; conflictos familiares o un ambiente sin respaldo pueden incrementar el riesgo de ausentismo (Arteaga Ramos y Tapias López, 2024; Dräger et al., 2023).

Dentro del entorno escolar, un clima inclusivo y el sentido de pertenencia pueden actuar como factores protectores del ausentismo. Las escuelas que fomentan la integración y el apoyo a los estudiantes en riesgo de inasistencia reducen significativamente las tasas de ausentismo. En contraste, un ambiente hostil o percibido como excluyente puede disminuir la motivación para asistir. Además, la percepción de apoyo por parte del personal docente y la disponibilidad de recursos escolares son factores clave en la reducción del ausentismo (Gilmore y Newcomer, 2022). Por su parte, Liu y Lee (2022) señalan que los estudiantes con altas tasas de inasistencia suelen reportar percepciones más negativas sobre el clima escolar, reflejando que la desconexión puede estar asociada con un entorno poco acogedor o carente de apoyo.

Por último, los factores comunitarios, como la violencia en el entorno o la falta de servicios básicos, presentan desafíos adicionales para la asistencia regular. Las comunidades con altos índices de inseguridad o con escasos recursos de apoyo social aumentan las dificultades de los estudiantes para asistir a clases.

En conjunto, estos factores ilustran cómo el ausentismo crónico es un reflejo de múltiples desigualdades estructurales. La identificación temprana de estudiantes en riesgo, basada en el conocimiento de estos factores asociados, puede ser una herramienta poderosa para desarrollar intervenciones específicas que reduzcan las barreras a la asistencia y promuevan trayectorias educativas más estables y exitosas.

MARCO CONCEPTUAL[2]

El aprendizaje automático (*machine learning*) es una disciplina del área de estadísticas e inteligencia artificial cuyo objetivo es otorgar a las computadoras la habilidad de aprender y mejorar de forma autónoma, sin necesidad de una programación explícita, a través del análisis de grandes cantidades de datos (Brown, 2021).

[2] Excepto indicación expresa, las fuentes de esta sección son las diapositivas del curso teórico del Diploma en Analítica de Datos Aplicada a Proyectos de la UTEC, elaboradas por la profesora. Alejandra Tabares Pozos (Universidad de los Andes, 2024) y *notebooks* del curso proyecto del mismo diploma, elaborados por el tutor de la tesis, Martín Fernández Campos.

Existen dos grandes subcategorías de aprendizaje automático: supervisado y no supervisado. Los modelos de aprendizaje supervisado se entrenan con conjuntos de datos etiquetados, que permiten a los modelos aprender y ser más precisos con el tiempo (Brown, 2021). Por ejemplo, un algoritmo se entrenaría con fotos de flores y otras cosas, todas etiquetadas por humanos, y la máquina aprenderá a identificar imágenes de flores por sí sola. En el aprendizaje automático no supervisado, un programa busca patrones en datos no etiquetados, es decir, no hay conocimiento *a priori*. El modelo puede encontrar patrones o tendencias que las personas no están buscando explícitamente (Brown, 2021). Por ejemplo, un programa de aprendizaje automático no supervisado podría examinar datos de desempeño escolar e identificar tipos de estudiantes.

MÉTODOS DE CLASIFICACIÓN

Entre los modelos de aprendizaje supervisado se encuentran los métodos de clasificación. Un clasificador es una función que asigna observaciones no etiquetadas a una clase o etiqueta, utilizando las estructuras de datos internas de la observación. El objetivo es que nuestro clasificador funcione bien no solo en los datos de entrenamiento, sino también en las observaciones de prueba que no se usaron para entrenar el clasificador. En este proyecto, nos interesa predecir si un estudiante será ausente crónico a fin de año, en función de variables socioeconómicas y de desempeño académico. La tarea es, por lo tanto, una clasificación binaria (dos clases posibles: ausente crónico o no ausente crónico). Existen diversos métodos de clasificación, basados en distintos campos de investigación. A continuación, se presentan los métodos empleados en nuestro proyecto: regresión logística, XGBoost, LightBoost y CatBoost.

Regresión logística

La regresión logística modela la probabilidad logarítmica de una variable dependiente binaria Y, utilizando una combinación lineal de covariables independientes X. El modelo completo se define matemáticamente de la siguiente manera:

$$p(\mathrm{x}) = \frac{1}{1 + e^{-(\beta_0+\beta_1+\beta_2 x_2+\ldots+\beta_p x_p)}}$$

donde p(x) es la probabilidad de pertenencia a la clase positiva, $\beta_0, \beta_1, \ldots, \beta_p$ son los coeficientes del modelo, y $x_1, x_2, \ldots, x_p$ son las características de entrada.

Los coeficientes de la regresión logística, β, representan el efecto de la variable independiente en la probabilidad de ocurrencia de las clases de estudio. Generalmente son ajustados utilizando máxima verosimilitud. Este modelo se usa como modelo de clasificación cuando se quiere estudiar la probabilidad de un suceso con solo dos posibles resultados. Entre las principales ventajas de la regresión logística se destaca su interpretabilidad; en este sentido, los coeficientes pueden interpretarse en términos de *odds ratios*, lo que permite comprender el impacto de las variables independientes en la probabilidad del resultado. Esta fortaleza proviene de su origen: las ciencias sociales.

Métodos *boosting*

El *boosting* es una metodología de aprendizaje lento de ensamble que combina las predicciones de múltiples modelos débiles –esto es, con un rendimiento de precisión un poco mejor que el azar– para, impulsados, producir una predicción final robusta. Los modelos débiles son típicamente árboles de decisión[3] y las predicciones se combinan a través de un promedio ponderado, donde se da más peso a los árboles que producen un mejor resultado en los datos de entrenamiento (Kashnitsky, 2020). Los pesos se actualizan en cada iteración del proceso de *boosting*, lo que ajusta el enfoque a las muestras que fueron mal clasificadas en la iteración anterior. La predicción final se obtiene tomando un voto mayoritario ponderado de todos los árboles en el conjunto.

Entre los métodos de *boosting* más populares se destacan XGBoost, LightGBM, CatBoost y AdaBoost. Estos algoritmos han ganado una fuerte relevancia por su eficacia en múltiples aplicaciones, desde tareas de clasificación y regresión hasta sistemas de detección y recomendación de anomalías. Nuestro proyecto implementa los primeros tres.

XGBoost es una implementación optimizada y escalable del algoritmo de *gradient boosting*[4] para modelos basados en árboles. Fue lanzada en 2014. Generalmente alcanza gran eficiencia, rendimiento y capacidad para manejar grandes volúmenes de datos. Entre sus principales ventajas se destacan: la capacidad para manejar datos faltantes de forma nativa, las opciones avanzadas de ajuste vía hiperparámetros y la regularización (incluye métodos de regularización para reducir la posibilidad del sobreajuste y mejorar la capacidad de generalización del modelo). Las desventajas de XGBoost son su enorme consumo de recursos, la complejidad y el tiempo que conlleva su parametrización, su dependencia de datos bien preparados (la preparación y normalización de nuestros datos deberían aliviar esta debilidad) y la propensión al sobreajuste (aunque esto ocurre en mayor medida cuando no se ajustan correctamente los hiperparámetros o cuando el conjunto de datos es pequeño, que no es el caso de nuestro proyecto).

LightGBM, lanzado en 2017 por Microsoft, es un *framework* de *boosting* que utiliza algoritmos de aprendizaje basados en árboles. Este método se destaca por su capacidad para soportar grandes volúmenes de datos, alta velocidad de entrenamiento, bajo uso de memoria y mayor precisión.

Por su parte, CatBoost, también lanzado en 2017, es un *framework* avanzado de *boosting* diseñado específicamente para manejar variables categóricas de forma eficiente y precisa. Los datos categóricos pueden integrarse directamente en el modelo, sin la necesidad de transformarlos en variables numéricas con técnicas como codificación *one-hot* (obligatorio en la regresión logística). Esto se logra mediante el uso de una codificación ordenada de las variables categóricas e incorporando la información de orden en el proceso de aprendizaje.

---

[3] Los árboles de decisión son estructuras de datos en el *machine learning* que dividen el conjunto de datos en subconjuntos cada vez más pequeños en función de sus características. Los árboles de decisión pequeños (construidos con poca profundidad) resultan perfectos para esta tarea, al ser malos predictores (*weak learners*), fáciles de combinar y generarse de forma muy rápida (Fernández-Casal, Costa y Oviedo, 2024).

[4] Método de aprendizaje de ensamble que combina las predicciones de múltiples modelos débiles para producir una predicción final robusta.

Este enfoque produce modelos muy precisos y con menor sobreajuste, en comparación con los métodos tradicionales de codificación *one-hot*. Como se detalla en la sección 5, nuestra base de datos final para ajustar los modelos contiene muchas variables categóricas (10 en un total de 16). Aunque CatBoost fue diseñado para ser escalable en grandes conjuntos de datos, corre con la desventaja de tener una baja velocidad de entrenamiento.

TUNING DE HIPERPARÁMETROS

Los hiperparámetros son elementos muy relevantes de los modelos de *machine learning.* La correcta configuración de hiperparámetros, como la penalización en regresión logística o el número de estimadores en modelos de *boosting*, es fundamental para optimizar el rendimiento de los modelos.

En el modelo de regresión logística, para mejorar la estimación y prevenir el sobreajuste, la penalización se aplica con técnicas de regularización mediante Ridge (también conocida como regularización L2) y Lasso (regularización L1). La regularización introduce un término de penalización en la función de costo para restringir la magnitud de los coeficientes. El objetivo es equilibrar el ajuste al conjunto de entrenamiento y la complejidad del modelo. La formulación general de regularización es la siguiente:

$$\hat{\beta} = \arg\min_{\beta} \{L(\beta)+\lambda(\beta)\}$$

donde L(β) es la función de pérdida (por ejemplo, error cuadrático), P(β) es el término de penalización y $\lambda \geq 0$ es el parámetro de regularización que controla la fuerza de la penalización.

Ridge penaliza grandes valores de β, reduciendo la varianza del estimador. Lasso, por su parte, induce esparsidad en β, por lo cual algunos coeficientes se vuelven exactamente 0; en este sentido, realiza selección de variables automáticamente. La regresión Lasso es útil cuando se sospecha que solo un subconjunto de variables es relevante; esta sospecha es válida para nuestro proyecto. La regularización afecta el balance entre sesgo y varianza del modelo. Al aumentar λ, se incrementa el sesgo, pero se reduce la varianza. El objetivo es encontrar λ que minimice el error cuadrático medio (ECM).

Por su parte, los hiperparámetros de los modelos de *boosting* incluyen, por ejemplo, el número de estimadores (número de árboles en el modelo), la tasa de aprendizaje (factor por el cual se multiplica cada árbol antes de añadirlo al modelo general), la profundidad máxima de los árboles (limita la profundidad de cada árbol), el mínimo de observaciones por hoja (número mínimo de observaciones que una hoja debe tener), el fraccionamiento de columnas (proporción de características a considerar en cada árbol) y el submuestreo (proporción de datos de entrenamiento utilizada para cada árbol). Estos hiperparámetros ayudan a controlar el sobreajuste y mejorar la generalización.

En este trabajo, la técnica de optimización de hiperparámetros empleada es *random search* con *cross-validation* (validación cruzada).

Random-search con *cross-validation*

El *random search* es un enfoque donde se prepara un set de hiperparámetros candidatos del que aleatoriamente se seleccionan algunos subconjuntos para luego ejecutar *K-Fold cross-validation*. Las ventaja de este método es que permite controlar el tiempo computacional al elegir el número de búsquedas de parámetros. La desventaja es que, si el espacio de hiperparámetros es muy grande, puede que existan ciertos parámetros no explorados suficientemente (Bergstra y Bengio, 2012).

La *cross-validation* se utiliza en la búsqueda de hiperparámetros para asegurar que el modelo se entrene y se evalúe de manera robusta y generalizable, evitando que el rendimiento dependa de una sola partición de los datos (por ejemplo, 80-20). Para evaluar la calidad de cada combinación de hiperparámetros, se aplica *cross-validation*, donde el conjunto de datos se divide en varios subconjuntos (*folds*). En cada iteración, uno de los subconjuntos se reserva para evaluar el modelo, mientras que el resto se usa para entrenarlo. Al final, se obtiene una medida de desempeño promedio para cada combinación de hiperparámetros, garantizando que la selección final no dependa de una única partición del conjunto de datos, sino que generalice bien en distintas particiones. En efecto, la *cross-validation* evita que los hiperparámetros se ajusten demasiado a una única partición de los datos, lo que aumenta la probabilidad de que el modelo generalice bien en datos nuevos.

En resumen, la *cross-validation* en la búsqueda de hiperparámetros ayuda a seleccionar configuraciones que sean efectivas y generalizables, minimizando el riesgo de sobreajuste y logrando un modelo con mejor desempeño en distintos conjuntos de datos.

MÉTRICAS DE EVALUACIÓN DE LOS MODELOS

En clasificación binaria es crucial evaluar cuán bien un modelo clasifica las instancias en dos clases. Las métricas de evaluación comunes incluyen: la matriz de confusión, precisión (*precision*), sensibilidad (*recall*), especificidad (*specifity*), F1-Score, Curva ROC y AUC (AUC-ROC). Estas métricas proporcionan diferentes perspectivas sobre el rendimiento del modelo.

Matriz de confusión

Es un resumen de las predicciones del modelo comparadas con las etiquetas reales. La estructura de la matriz de confusión para clasificación binaria es la siguiente:

Verdaderos positivos (VP)	Falsos positivos (FP)
Falsos negativos (FN)	Verdaderos negativos (VN)

La matriz de confusión contiene la siguiente información:

VP: instancias predichas correctamente como positivas.
FP: instancias predichas incorrectamente como positivas.
FN: instancias predichas incorrectamente como negativas.
VN: instancias predichas correctamente como negativas.

A partir de estas medidas se construyen las métricas que se describen en los puntos posteriores.

Métrica de exactitud (*accuracy*)

Esta métrica indica el número de observaciones clasificadas correctamente (VP y VN), en comparación al número total de observaciones clasificadas en cualquier clase (VP, VN, FP y FN).

$$accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

Esta métrica tiene la desventaja de ser poco representativa cuando las clases están muy desequilibradas, pues algunas tendrán muchos más elementos que otras.

Métrica de precisión (*precision*)

Es la proporción de predicciones positivas correctas sobre todas las predicciones positivas. La medida *precision* responde a la pregunta: ¿cuántos de los elementos clasificados como positivos son realmente positivos?

$$precision = \frac{VP}{VP + FP}$$

Métrica de sensibilidad (*recall*)

El *recall* es la proporción de predicciones positivas correctas sobre todas las instancias positivas reales. Esta métrica responde otra pregunta: ¿cuántos de los elementos realmente positivos fueron correctamente clasificados?

$$recall = \frac{VP}{VP + FN}$$

Métrica de especificidad (*specificity*)

La especificidad es la proporción de predicciones positivas negativas correctas sobre todas las instancias negativas reales. Esta métrica es importante en escenarios donde es crucial evitar falsos positivos.

$$specificity = \frac{VN}{VN + FP}$$

F1-Score

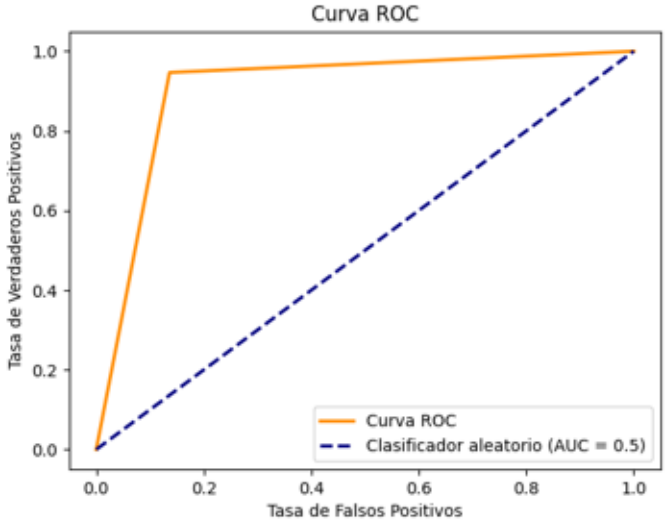
El F1-Score es la media armónica entre *precision* y *recall*, balanceando ambos. Es una medida útil cuando se busca un balance entre *precision* y *recall.*

$$1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Curva ROC y *Area Under the Curve* (AUC)

La Curva ROC (*Receiver Operating Characteristic*) es la gráfica que muestra la tasa de verdaderos positivos (TPR o *recall*) frente a la tasa de falsos positivos (FPR) para diferentes umbrales. Se utiliza para evaluar, comparar y seleccionar clasificadores con base en su desempeño.

GRÁFICO 1
CURVA ROC



Nota: la diagonal de línea punteada de color azul representa un clasificador aleatorio, donde los puntos que están por arriba representan resultados de clasificación mejor que el azar, mientras los puntos que están por debajo simbolizan a los peores clasificadores.

El punto de origen del espacio ROC (0,0) corresponde a un clasificador que nunca predice una observación positiva. Esto significa que no detecta correctamente ninguna observación positiva (tasa de verdaderos positivos, TVP = 0) y tampoco clasifica incorrectamente observaciones negativas como positivas (tasa de falsos positivos, TFP = 0). En el otro extremo, el punto (1,1) es un clasificador que predice todas las observaciones como positivas, lo que maximiza tanto los verdaderos positivos (TVP = 1) como los falsos positivos (TFP = 1). Este comportamiento indica que el clasificador no logra distinguir adecuadamente entre clases. Un clasificador perfecto se situaría en el punto (0,1), donde la TVP (*recall*) es 1 y la TFP es 0. Si bien no es realista esperar obtener un clasificador perfecto, es un punto al cual se quiere acercar para encontrar el mejor clasificador.

El AUC (área bajo la curva) es una medida agregada del rendimiento en todos los posibles umbrales de clasificación.

En el gráfico 1, el AUC mide toda el área bidimensional por debajo de la curva ROC completa de (0,0) a (1,1). Cuando el AUC es igual a 1, el modelo es perfecto. Cuando el AUC es igual a 0,5, el modelo no es mejor que el azar (clasificador aleatorio en el gráfico 1). Cuando el AUC es mayor a 0,5, el modelo tiene un rendimiento mejor que el azar. En breve, cuanto más cerca de 1, mejor es el rendimiento del modelo.

METODOLOGÍA

DISEÑO DEL ESTUDIO

El objetivo del proyecto es desarrollar un modelo de clasificación que permita predecir el ausentismo crónico de los estudiantes de educación media básica en Uruguay en el mes de junio de un determinado año, es decir, luego del primer trimestre del año lectivo. El corte en el primer trimestre se sustenta en que una correcta implementación de un modelo robusto de *machine learning* en el mes de junio permitiría implementar estrategias de intervención oportunas que puedan modificar el estatus de ausente crónico del estudiante. Esta suposición ha sido observada en un plan reciente de la ANEP: el Plan Asiste 2023 (ANEP, 2023a) tiene el propósito de mejorar los niveles de asistencia en educación inicial y primaria mediante estrategias de acción para aquellos estudiantes que, en el primer trimestre del año lectivo, presentan 10% o más de inasistencias y son, por lo tanto, potenciales ausentes crónicos.

Una característica importante de este proyecto es el considerable tiempo invertido en la limpieza y preparación de los datos. Las tres bases con las que iniciamos el proyecto corresponden a registros administrativos que no fueron concebidos para fines estadísticos y, menos aún, para su procesamiento en un modelo de aprendizaje automático. Además, atendiendo a la literatura disponible, muchas de las variables potencialmente predictoras del ausentismo crónico tuvieron que ser construidas para su incorporación en el modelo.

En efecto, nuestra metodología continúa con subsecciones enfocadas en mejorar los datos, con la meta última de obtener modelos robustos. En este capítulo se describen los conjuntos de datos primarios, se detalla el extenso proceso de limpieza y preparación de estos datos, se presenta la base final obtenida y se muestran las transformaciones aplicadas a las variables para su inclusión en los modelos de *machine learning*, como la transformación de potencias y el uso del método *one-hot encoder*, entre otras.

DESCRIPCIÓN DEL CONJUNTO DE DATOS

Se trabajó con datos administrativos del período 2016-2023. Los datos principales provienen de los registros administrativos de la ANEP, brindados por el Instituto Nacional de Evaluación Educativa (INEEd), bajo un acuerdo de confidencialidad. El manejo de los datos se realizó con estrictas medidas de seguridad y anonimización, respetando las normativas de protección de datos personales y garantizando la confidencialidad de la información.

Las bases de datos utilizadas incluyen:

- **Base de inscripciones**. Contiene información de los estudiantes inscriptos en liceos públicos desde 2016 hasta 2023, con un total de 2.159.283 observaciones y 74 columnas. Algunas de las variables más relevantes de esta base son: la fecha de nacimiento del estudiante, el curso y turno en el cual está inscripto, el grado que está cursando y el centro educativo al cual asiste, con datos de su localización geográfica (por ejemplo, departamento, zona).

- **Base de resultados**. Contiene 542.819 observaciones y 8 columnas. Registra información sobre el desempeño de los estudiantes en términos de resultados académicos, por ejemplo, la calificación por reunión y la calificación al final, etc. Contiene múltiples tipos de evaluación en la educación media básica, en función del plan educativo en el que esté matriculado el estudiante.

- **Base de inasistencias**. Contiene 46.361.991 observaciones y 5 columnas. Esta base registra las fechas en las que los estudiantes no asistieron a clase. Los datos no contemplan, por lo tanto, la totalidad de los días hábiles correspondientes al año lectivo y grado. En breve, para la obtención de la tasa de ausentismo crónico fue necesario crear otras bases de datos. Este proceso se detalla más adelante.

PREPARACIÓN DE LOS DATOS

PREPARACIÓN DE DATOS DE LA BASE DE INSCRIPCIONES

En este proyecto se realiza una primera selección entre los estudiantes, eligiendo aquellos matriculados en educación media básica (séptimo a noveno grado) en los planes educativos Reformulación 2006 y Plan de Educación Media Básica 2023. Estos planes, siendo el segundo una actualización del primero tras la reforma educativa aprobada en 2020, abarcan la mayor proporción de estudiantes en la base de datos, con un total de 1.293.430 y 95.999 inscripciones, respectivamente, lo cual representa aproximadamente dos tercios del total de inscripciones en secundaria.

Para el análisis se emplea un enfoque por cohortes, iniciando con la identificación de los estudiantes que cursan séptimo grado de secundaria por primera vez en cada año siguiendo su trayectoria educativa hasta 2023. Se excluyen del análisis aquellos estudiantes que

cada año formaban parte de otras cohortes. La tabla 1 presenta el número de estudiantes inscriptos por año lectivo y cohorte en educación media básica, con un total de 179.480 estudiantes.

TABLA 1
ESTUDIANTES DE SECUNDARIA INSCRIPTOS EN EDUCACIÓN MEDIA BÁSICA POR COHORTE Y AÑO LECTIVO
REFORMULACIÓN 2006 Y PLAN DE EDUCACIÓN MEDIA BÁSICA 2023

Cohorte / año lectivo	2016	2017	2018	2019	2020	2021	2022	2023	Total
2016	30.361	28.377	26.091	5.963	2.189	624	40	8	93.653
2017	0	31.167	29.018	27.100	5.364	1.866	409	22	94.946
2018	0	0	29.096	27.174	25.622	4.481	1.489	285	88.147
2019	0	0	0	29.323	27.728	26.622	4.038	1.260	88.971
2020	0	0	0	0	29.686	28.371	26.760	3.771	88.588
2021	0	0	0	0	0	29.847	28.331	27.016	85.194

Nota: número de estudiantes inscriptos por cohorte y año lectivo para los planes Reformulación 2006 y Plan de Educación Media Básica 2023.

Cada cohorte comprende entre 85.000 y 95.000 casos aproximadamente, para un total de 179.480 estudiantes. Los estudiantes comienzan la educación media por primera vez en cada cohorte. Para identificar a los estudiantes de la cohorte 2016 que cursan séptimo grado de secundaria por primera vez, se verifica (con datos del Sistema de Información Integrada del Área Social [SIIAS] del Ministerio de Desarrollo Social) quiénes cursaron el último año de la educación primaria en 2015. De los 38.514 estudiantes que iniciaron séptimo en 2016, 30.361 cumplen con esta condición. Para las cohortes siguientes, se corrobora que los estudiantes no formen parte de la cohorte anterior.

Atendiendo a la literatura sobre ausentismo crónico, se incorporan a la base datos variables relevantes para los modelos predictivos, incluyendo: edad, extraedad, situación de extraedad al inicio del trayecto educativo y variables de vulnerabilidad social. Dado que estas variables no estaban disponibles en las bases originales, fueron creadas mediante distintas metodologías, las cuales se describen a continuación.

Edad y extraedad

Para el cálculo de la edad de cada estudiante, se parte de la fecha de nacimiento. Los datos en esta columna son de tipo integral y su conversión al tipo *fecha* es inconsistente, sin perjuicio del formato que se ingrese (es decir, brinda fechas erróneas en todas las variantes de formato). Luego de varios análisis se identifica que el integral corresponde al número de días transcurridos desde el 1 de enero de 1970. La fecha de corte utilizada para el cálculo de la edad es el 30 de abril del año lectivo. Por su parte, la columna extraedad contiene los años que exceden la edad teórica correspondiente al grado de cada estudiante en cada año lectivo. Las edades teóricas en séptimo, octavo y noveno son 12, 13 y 14 años, respectivamente. Se agrega también la columna situación de extraedad al inicio de la educación media básica, que toma el valor True cuando el estudiante al ingresar a la educación media (es decir, en séptimo) tiene más de 12 años.

Variables de vulnerabilidad social

Para incorporar información socioeconómica y cultural en el análisis, se genera la variable IVS, siguiendo la metodología empleada por la ANEP (2022), adaptándola a nuestro universo de estudio. Para la construcción del índice se extrajeron datos del SIIAS del Ministerio de Desarrollo Social. El IVS permite medir la vulnerabilidad tanto de los estudiantes como de los centros educativos, proporcionando una herramienta fundamental para analizar las desigualdades educativas. Se calcularon e incluyeron a la base de inscripciones dos índices:

- IVS a nivel de estudiante (IVS_estudiante). Mide la vulnerabilidad de cada estudiante considerando variables como acceso a prestaciones sociales (asignaciones familiares del Plan de Equidad —AFAM—, Tarjeta Uruguay Social —TUS—) y cobertura de salud. El índice oscila entre 0 y 9, donde los valores más altos indican mayor vulnerabilidad. Se excluye la variable residencia en hogares del Instituto del Niño y Adolescente del Uruguay (INAU) para adecuar el índice a nuestro universo específico.

- IVS a nivel de centro educativo (IVS_centro). Se calcula promediando los valores de IVS_estudiante de todos los estudiantes de un centro, permitiendo clasificar los centros en quintiles, donde el quintil 1 agrupa a los centros más vulnerables.

Para el cálculo del IVS a nivel de centro, se incluyeron todos los estudiantes matriculados en educación media básica entre 2016 y 2023, reflejando el contexto socioeconómico completo de cada centro.

PREPARACIÓN DE DATOS DE LA BASE DE RESULTADOS

La base de resultados requirió mínimas transformaciones. A partir de la información del fallo final de cada estudiante, se construyó una variable que agrupa los distintos valores de la columna según categorías amplias y únicas representando el desempeño final del estudiante en términos de aprobación o no del año lectivo. Aunque la nomenclatura de los fallos difiere según el plan, ambos reflejan conceptos equivalentes. Por ejemplo, los fallos *promovido* y *acredita el año* indican que el estudiante promueve el año (el primer fallo era utilizado en el período 2016-2022 y el segundo en 2023); por lo tanto, estos valores se recategorizan en la etiqueta *aprueba*. Las etiquetas *fallo en suspenso*, *acreditación pendiente* y *pase estudios libre* se agrupan en la categoría *aprobación condicional*. Para los fallos *repite por inasistencia* y *recursa por desvinculación* se crea la categoría *no aprueba por inasistencias*, mientras que *repite por rendimiento* se clasifica bajo la nueva categoría *no aprueba por rendimiento*.

PREPARACIÓN DE DATOS DE LA BASE DE INASISTENCIAS

La base de inasistencias presenta una estructura simple aunque con ciertas anomalías típicas de los registros administrativos. En particular, se observan datos inconsistentes cuando la columna booleana "falta" toma el valor "FALSE" (es decir, en teoría el estudiante

asistió a clase). Por ejemplo, en el segundo día de clases de 2016, según los registros, solo tres estudiantes habrían asistido a clase, lo cual resulta poco plausible. Por esta razón, se eliminan las observaciones que contienen "FALSE" en la columna "falta".

Dado que la base de inasistencias únicamente contiene las fechas en las que el estudiante no asistió a clase, y que para calcular la tasa de ausentismo es necesario conocer tanto las fechas de inasistencia como aquellas en la que el estudiante debería haber asistido, fue necesario incorporar a la base de inasistencias todas las fechas en las que se dictaron clases, según año y grado. A estas fechas les denominamos días hábiles. Los días hábiles varían por año y por grado. A este respecto, caben dos aclaraciones: i) las fechas de inicio y fin del año lectivo cambian por año y por grado, y ii) las vacaciones y mayoría de los feriados varían por año. Las fechas de inicio y fin de año lectivo, vacaciones y feriados se obtuvieron de distintas circulares publicadas por la ANEP en el período 2016-2023.

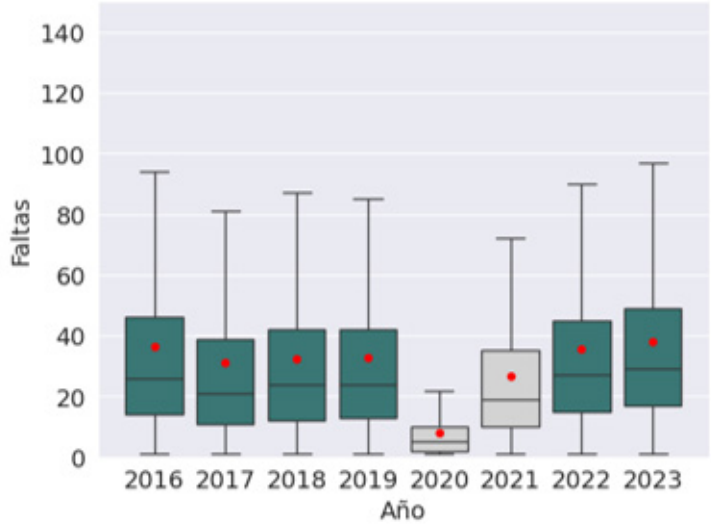
A partir de esta información, se creó una base denominada *sábana* por año lectivo, que contiene las fechas de los días hábiles entre 2016 y 2023. Esta base fue creada empleando la librería *datetime*. La base sábana fue integrada a la base de inasistencias, lo que permitió obtener para cada estudiante las fechas de asistencia e inasistencia y, por lo tanto, su tasa de ausentismo por día. Esta base de inasistencias diarias acumuló cerca de 76 millones de observaciones, por lo que se decidió generar una base más pequeña que agrupara la información por semana del año. Con esta transformación, la base de datos de inasistencias se redujo a 13,5 millones de observaciones.

El registro de la tasa de ausentismo en la última semana de clases, comprendida en esta base, permite crear la variable *target*: si el estudiante se ausenta 10% o más de los días lectivos al momento de la última semana del año, se lo clasificará como ausente crónico. Por lo tanto, la base de datos de inasistencias se redujo nuevamente, para incluir únicamente las observaciones de la última semana del año, y obtener, para cada estudiante, su tasa de ausentismo al finalizar el año lectivo. A esta nueva base de inasistencias (anual) se le agregan columnas con el número de inasistencias de cada estudiante para cada mes del año (esto se realizó *pivoteando* las observaciones de la base), atendiendo a que las faltas en el primer trimestre pueden ser importantes predictoras del ausentismo crónico. Con todas estas transformaciones, la base final de inasistencias culmina con 357.090 observaciones y 20 columnas.

Una decisión de suma relevancia con respecto a la base de datos de inasistencias es la eliminación de los registros correspondientes a los años 2020 y 2021. En este sentido, a mediados de marzo del 2020, Uruguay dispuso la suspensión de clases presenciales en todos sus niveles por 14 días, la que fue experimentando varias prórrogas (ANEP, 2020). En junio de 2020, la ANEP decidió que no se contabilizarían las inasistencias de los estudiantes en dicho año lectivo hasta nueva resolución (2021). En 2021 el dictado de clases continuó sufriendo cambios debido a la emergencia sanitaria. De hecho, el Monitor Educativo Liceal del 20 de agosto de 2024 no incluye datos de inasistencias de 2020 y 2021.

A modo ilustrativo, el gráfico 2 muestra la distribución promedio de las inasistencias por año entre 2016 y 2023: se puede observar que los datos son largamente inconsistentes en 2020 y poco confiables en 2021. Por las razones expuestas, los datos de inasistencias (y solo ellos) de ambos años fueron eliminados. Incluirlos en el modelo generaría un sesgo significativo; su omisión garantiza una mayor precisión y validez del modelo.

GRÁFICO 2
BOXPLOT DE LAS INASISTENCIAS PROMEDIO DE ESTUDIANTES EN EDUCACIÓN MEDIA BÁSICA
AÑOS 2016 A 2023



Nota: la distribución de las inasistencias promedio en 2020 y 2021 muestran un comportamiento atípico, en un contexto de pandemia de COVID-19.

Finalmente, las tres bases de datos (inscripciones, resultados e inasistencias) transformadas se pegan para culminar con una base completa de 357.090 registros y 42 variables.

ANÁLISIS Y RESULTADOS

ANÁLISIS EXPLORATORIO DE DATOS

El objetivo de esta sección es explorar en profundidad los datos utilizados en este proyecto con el fin de identificar y analizar patrones de ausentismo crónico en los estudiantes de liceos públicos en Uruguay. A partir de este análisis, intentaremos tener una primera comprensión de las variables a utilizar para la predicción del ausentismo crónico, no solo identificando la relevancia de cada una, sino detectando posibles problemas de calidad que puedan afectar la precisión del análisis y la modelización posterior.

Como se mencionó anteriormente, la base de datos utilizada proviene de la integración de registros administrativos que capturan información de seis cohortes de estudiantes entre los años 2016 y 2021. Cada registro representa la inscripción de un estudiante en un año específico y contiene variables que describen su trayectoria académica, sus características personales y el contexto de su centro educativo.

El conjunto final de datos cuenta con un total de 357.090 registros y 42 variables. Estas variables incluyen la información que aparece en la tabla 2.

TABLA 2

BASE DE DATOS COMPLETA

Variable	Descripción
estudiante_id	Identificador único de cada estudiante en la base de datos
anio_lectivo	Año académico correspondiente al registro de inscripción
grupo_id	Identificador del grupo en el que está asignado el estudiante
faltas_anual	Cantidad total de faltas acumuladas en el año
habiles_anual	Días hábiles totales en el año académico
tasa_ausentismo_anual	Tasa anual de ausentismo del estudiante
faltas_mes_3	Número de faltas del estudiante en marzo
faltas_mes_4	Número de faltas del estudiante en abril
faltas_mes_5	Número de faltas del estudiante en mayo
faltas_mes_6	Número de faltas del estudiante en junio
faltas_mes_7	Número de faltas del estudiante en julio
faltas_mes_8	Número de faltas del estudiante en agosto
faltas_mes_9	Número de faltas del estudiante en setiembre
faltas_mes_10	Número de faltas del estudiante en octubre
faltas_mes_11	Número de faltas del estudiante en noviembre
faltas_mes_12	Número de faltas del estudiante en diciembre
inas_trimestral_prom	Promedio de inasistencias trimestrales del grupo
fecha_nacimiento	Fecha de nacimiento del estudiante
sexo	Sexo del estudiante (M/F)
edad	Edad del estudiante al inicio del año lectivo
departamento_dsc	Departamento (región) donde reside el estudiante
localidad_dsc	Localidad específica donde reside el estudiante
ivs_estudiante	IVS del estudiante
ivs_estudiante_norm	Índice normalizado de vulnerabilidad social del estudiante
afam_ivs	Apoyo AFAM para el IVS
tus_ivs	Cobertura TUS del IVS
cobertura_ivs	Cobertura del IVS para el estudiante
nivel1a6	Nivel educativo o grado de educación media básica (séptimo/octavo/noveno)
centro_cod	Código del centro educativo en el que está inscripto el estudiante
jurisdiccion_dsc	Jurisdicción administrativa del centro educativo
zona	Zona geográfica del centro educativo (urbana/rural)
turno_dsc	Turno de estudio asignado al estudiante (mañana/tarde/noche)
cohorte	Cohorte a la que pertenece el estudiante (por año de primera inscripción)
ivs_centro	IVS del centro educativo
ivs_centro_quintil	Quintil de vulnerabilidad social del centro educativo (1 a 5)
quintil_centro	Quintil de vulnerabilidad social del centro educativo (1 a 5)
prop_vul_centro	Proporción de estudiantes vulnerables en el centro educativo
anios_base	Cantidad de años que el estudiante permanece en la base de datos
extra_edad	Indica si el estudiante tiene edad superior a la media para el nivel
extra_edad_primaria	Indica si el estudiante tiene extraedad desde primaria
calificacion_general	Calificación general del estudiante en el sistema
fallo_final	Indicador de fallo final del estudiante en el curso académico
ausente_cr	Indica si el estudiante es a fin del año lectivo ausente crónico o no
ratio_ac_emb	Relación entre el número de años con ausentismo crónico y los años que figura en educación media básica (ausente_cr/anios_base)

ANÁLISIS GENERAL DE VARIABLES

Análisis exploratorio de las variables numéricas

Los histogramas generados para las variables numéricas proporcionan una visión clara de la distribución y las características principales de cada una, resaltando patrones importantes que podrían influir en el análisis del ausentismo crónico (gráfico 3).

Las faltas de asistencia, tanto a nivel anual (faltas_anual) como mensual (faltas_mes_3 a faltas_mes_12), exhiben una distribución fuertemente sesgada hacia valores bajos. Esto sugiere que la mayoría de los estudiantes presentan pocas faltas, mientras que un pequeño grupo acumula ausencias muy elevadas. Meses como agosto, octubre y noviembre muestran una mayor dispersión en la cantidad de faltas, destacándose como posibles períodos críticos en el año escolar en los que el ausentismo tiende a incrementarse. Aunque este sesgo hacia los valores bajos podría llevar a subestimar la gravedad del problema, es importante considerar el impacto real de las ausencias. Por ejemplo, acumular cinco faltas en un mes equivale a haberse ausentado al 25% de las clases dictadas en ese período, lo que afecta la continuidad académica del estudiante. De la misma manera, superar las 20 faltas en el año en un contexto de aproximadamente 200 días hábiles significa que el estudiante estuvo ausente más de un 10% de los días dictados.

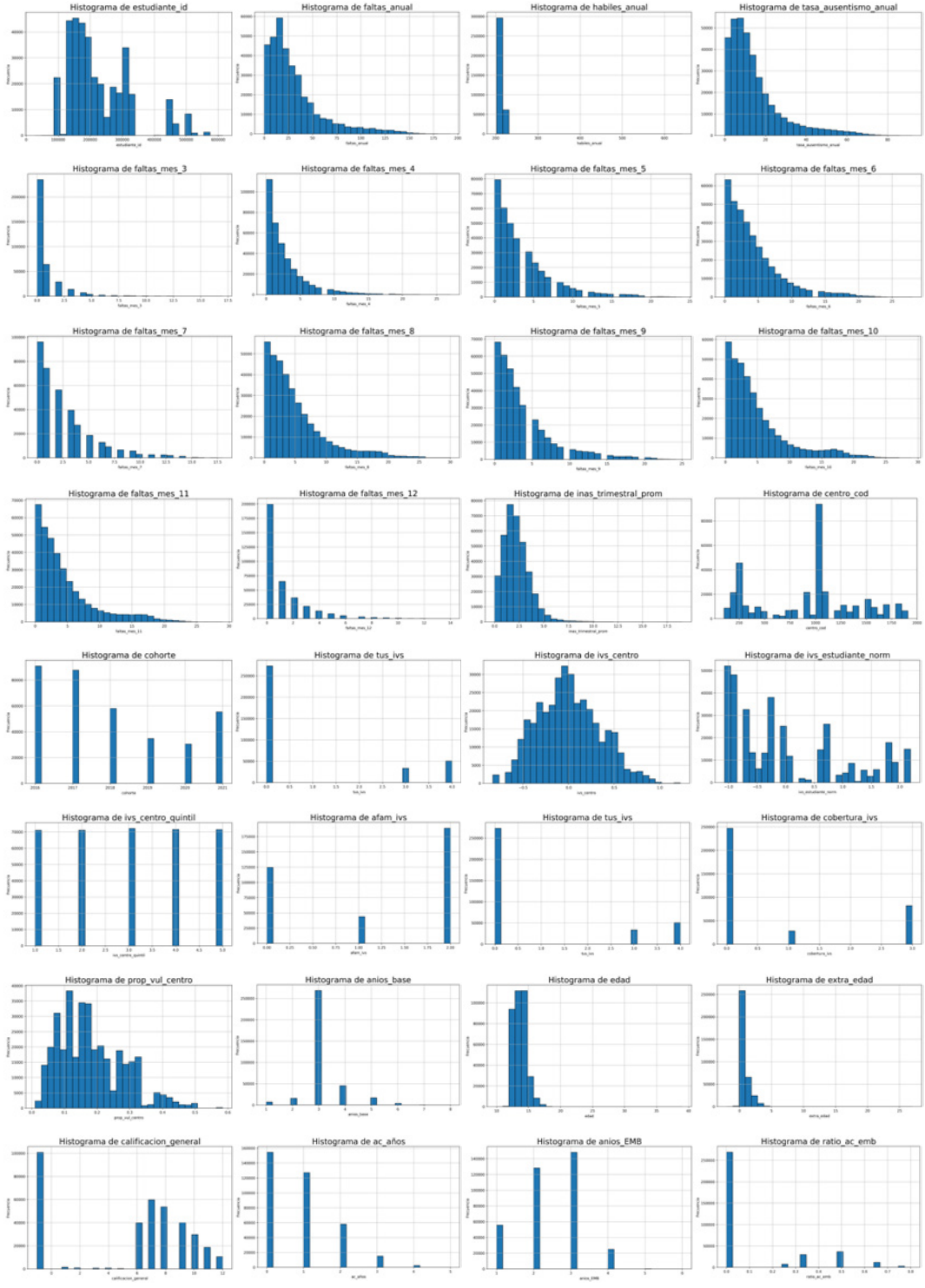
La variable tasa_ausentismo_anual refuerza los patrones observados en las faltas mensuales, mostrando que la mayoría de los estudiantes tienen tasas de ausentismo bajas, mientras que un grupo reducido alcanza tasas extremadamente altas (hasta un 92% de ausentismo). Esta variabilidad sugiere la existencia de subgrupos dentro de la población estudiantil con distintos niveles de riesgo de ausentismo crónico.

El IVS muestra diferencias tanto a nivel de estudiante (ivs_estudiante) como de centro educativo (ivs_centro). La mayoría de los estudiantes tienen un IVS bajo, mientras que el IVS de los centros educativos presenta una distribución más equilibrada. Esta diferencia indica que, aunque muchos estudiantes no se encuentran en condiciones de alta vulnerabilidad, algunos centros concentran poblaciones más vulnerables, lo cual podría estar vinculado a patrones de ausentismo en esos entornos. Las variables vinculadas a la cobertura de programas de apoyo forman parte de la composición del IVS, por lo que no deberían incluirse en la base. Sin embargo, se puede apreciar que la mayoría de los estudiantes no se encuentra en situación de vulnerabilidad extrema, dado que no reciben TUS ni están fuera del sistema de salud. Esto es esperable dado que la población de estudio son aquellos que permanecen matriculados en el sistema educativo.

La variable prop_vul_centro, que representa la proporción de estudiantes vulnerables en cada centro educativo, brinda información sobre la concentración de quienes están en situación de vulnerabilidad dentro de cada centro educativo. La mayoría parece tener una proporción de estudiantes vulnerables alrededor de un rango entre el 10% y el 25%. Esto indica que muchos centros cuentan con una presencia significativa de estudiantes en situación de vulnerabilidad, pero sin que esta proporción sea extremadamente alta en la mayoría de los casos.

GRÁFICO 3

DISTRIBUCIÓN DE LAS VARIABLES NUMÉRICAS



Respecto a la edad, la mayoría de los estudiantes se sitúan en el rango esperado de 12 a 14 años, aunque hay algunos casos de estudiantes mayores, hasta 39 años, lo que podría reflejar repeticiones o trayectorias académicas prolongadas.

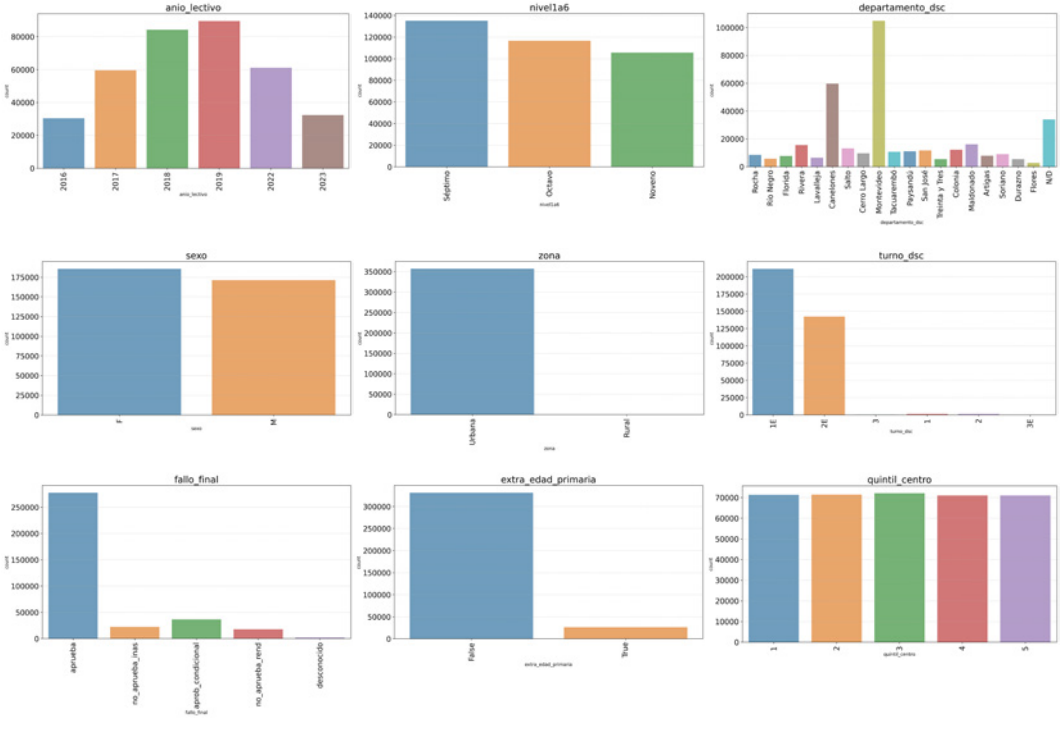
Por otro lado, la distribución de las calificaciones en el año previo se concentra en torno a valores promedio, con algunos casos extremos, con valor en -1 que estaría indicando un faltante de información. Dada la gran cantidad de datos faltantes, no sería adecuado incorporar esta variable en el análisis.

La mayoría de los estudiantes tienen entre uno y tres años en educación media, aunque algunos registran más años cursando educación media básica. Este aspecto es relevante, ya que los que cuentan con trayectorias prolongadas podrían presentar mayores tasas de ausentismo.

Análisis de las variables categóricas

A través de los gráficos generados, se observan patrones en variables como el año lectivo, el grado, la zona de residencia y la situación socioeconómica, que brindan una visión más profunda sobre las condiciones de los estudiantes pertenecientes a las cohortes contempladas en nuestro análisis (gráfico 4).

GRÁFICO 4
DISTRIBUCIÓN DE LAS VARIABLES CATEGÓRICAS



La distribución de registros a lo largo de los años revela un aumento en 2018 y 2019, indicando una mayor concentración de datos en estos períodos. Este incremento se explica por la estructura de la base de datos: al tratarse de cohortes desde 2017 hasta 2021, los años 2018 y 2019 incluyen estudiantes en séptimo, octavo y noveno de secundaria. En contraste, en 2016 solo se registran estudiantes de séptimo y en 2017 se incorporan únicamente los de octavo. De forma similar, como 2021 es la última cohorte considerada, los años 2022 y 2023 incluyen solo a estudiantes de octavo y noveno.

En cuanto a la distribución por niveles educativos, los grados de séptimo, octavo y noveno presentan una proporción relativamente equilibrada, con una leve predominancia en séptimo. Este balance permite realizar un análisis uniforme por grado.

La distribución geográfica muestra una considerable variabilidad entre departamentos: se destacan Montevideo y Canelones debido a su mayor densidad poblacional y, por tanto, a una mayor concentración de estudiantes. En cuanto a la variable zona, predominan los estudiantes de áreas urbanas, con una menor representación de aquellos en áreas rurales. Esto sugiere que la mayoría de los datos provienen de zonas urbanas, probablemente debido a una mayor concentración de centros educativos y a una población estudiantil más numerosa en estas áreas.

La proporción entre estudiantes de sexo femenino y masculino es prácticamente equitativa, lo que permite realizar el análisis sin sesgo de género. Los turnos matutino e intermedio son los más comunes en la muestra, sugiriendo que estos turnos tienen una mayor concentración de estudiantes. Esta distribución sugiere realizar una transformación de la variable en menos categorías.

La gran mayoría de los estudiantes han aprobado el año académico previo, mientras que los casos de no aprobación y aprobación condicional son minoritarios. Esta distribución indica que sería conveniente recategorizar el fallo entre aprobaciones, no aprobaciones y condicionales.

La mayoría de los estudiantes no están clasificados como *extraedad primaria*, lo cual indica que se encuentran en el rango de edad esperada para su nivel cuando ingresan a educación media básica. Sin embargo, existe un pequeño porcentaje que presentan repeticiones o interrupciones en su trayectoria educativa en educación primaria.

Por último, la distribución de los quintiles de centro es uniforme, dado que se construye a partir del IVS_centro, quintilizando a los estudiantes según vulnerabilidad.

Selección de variables e identificación de ajustes a realizar

A partir del análisis exploratorio de las variables numéricas y categóricas, y utilizando el mapa de calor de correlación como una herramienta clave para identificar relaciones entre variables, se han definido las estrategias de selección y los ajustes necesarios para optimizar el modelo de predicción de ausentismo crónico al final del año lectivo. El mapa de calor permite visualizar relaciones fuertes, detectar redundancias y reconocer patrones

que pueden influir en el comportamiento de ausentismo, facilitando así la selección de las variables a incluir en el modelo.

Dado que el objetivo del estudio es predecir en junio el ausentismo crónico a fin del año lectivo (basándonos en los datos del primer trimestre lectivo), solo consideraremos las faltas correspondientes a los meses de marzo, abril y mayo (faltas_mes_3, faltas_mes_4, y faltas_mes_5). Esta selección permitirá detectar patrones tempranos y anticipar intervenciones. Estas variables muestran correlaciones moderadas con la variable objetivo ausente_cr (entre 0,3 y 0,6), lo que sugiere que las ausencias en estos meses pueden ser buenos indicadores tempranos del riesgo de ausentismo crónico.

La variable inas_trimestral_prom, que representa el promedio de inasistencias del grupo en el trimestre, muestra una correlación de 0,4 con ausente_cr. Este valor sugiere que el comportamiento de los pares (reflejado en la media de inasistencias del grupo) tiene una influencia relevante en el ausentismo individual. Incluir esta variable en el modelo puede capturar aspectos de influencia grupal que podrían ser relevantes para comprender patrones de ausentismo en contextos específicos.

El análisis de correlación en el mapa de calor también indica redundancias entre las variables que miden vulnerabilidad, como ser ivs_centro, quintil_centro, prop_vul_centro, tus_ivs, cobertura_ivs y afam_ivs (gráfico 5). Dado que estas variables aportan información similar sobre la vulnerabilidad del entorno educativo, se conservará únicamente quintil_centro, una variable de fácil interpretación que clasifica los centros educativos en función de su nivel de vulnerabilidad. Asimismo, se selecciona ivs_estudiante_norm para capturar la vulnerabilidad a nivel individual, ya que esta versión normalizada proporciona una perspectiva ajustada en función de la población del centro.

Algunas variables presentan limitaciones en su disponibilidad o en su aporte específico para el modelo. Por ejemplo, la variable tasa_ausentismo_anual no estará disponible en junio y podría introducir sesgo en una predicción temprana, por lo que se decide excluirla. De manera similar, hábiles_anual se descarta debido a su falta de variabilidad y su baja relevancia para el modelo predictivo, ya que no aporta información adicional en términos de ausentismo.

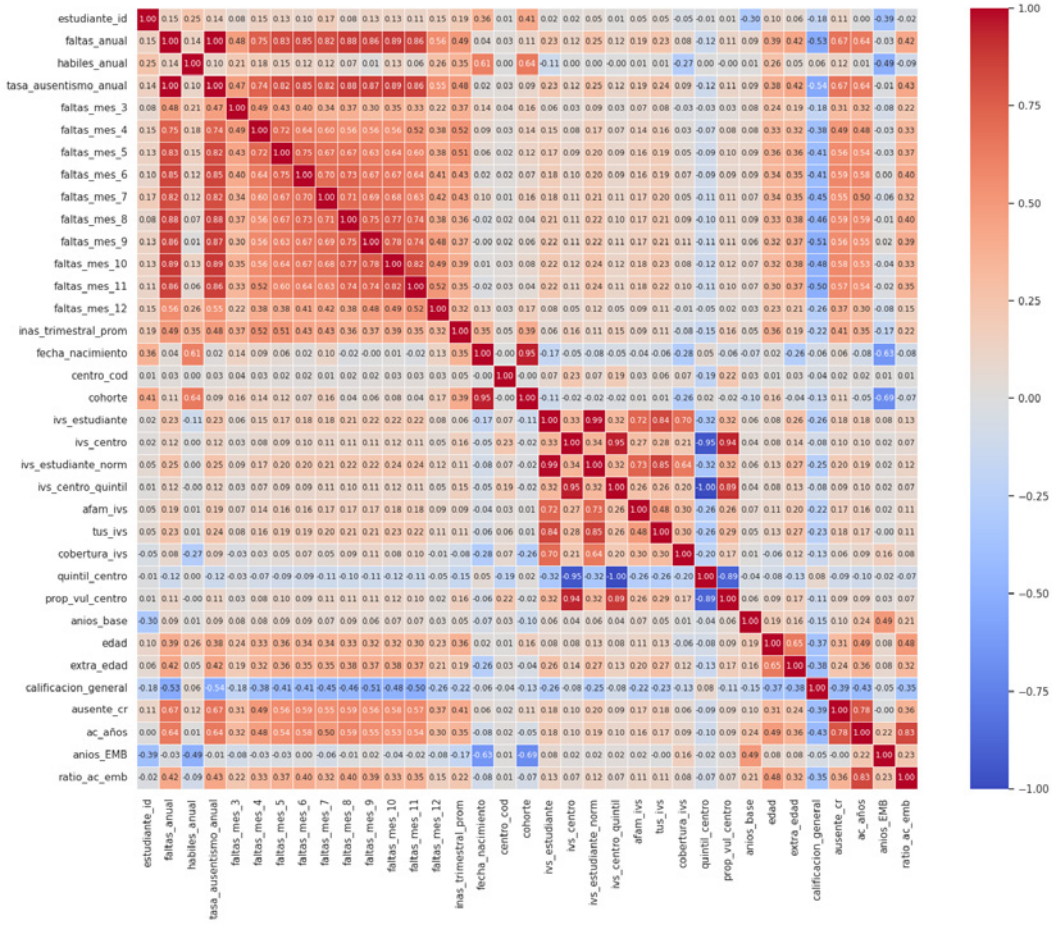
En cuanto al rendimiento académico, la variable calificación_general, que refleja la nota obtenida en el año previo, presenta una alta proporción de datos faltantes, lo que limita su utilidad en el modelo. Sin embargo, la variable fallo_año_anterior se mantendrá como variable categórica, ya que permite captar el desempeño académico de los estudiantes sin las limitaciones de datos faltantes que presenta la calificación numérica.

La trayectoria educativa es otro aspecto importante en el análisis de ausentismo. Las variables ac_años y ratio_ac_emb muestran una alta correlación, dado que ratio_ac_emb se calcula en función de ac_años. Sin embargo, el histograma de ratio_ac_emb indica una proporción significativa de valores en 0, lo que sugiere que la variable puede causar problemas en el análisis. En su lugar, se optará por construir una nueva variable que indique si el estudiante ha sido ausente crónico en años anteriores, proporcionando así un mejor

contexto histórico de la trayectoria de ausentismo y facilitando una visión más clara de quienes están en riesgo.

GRÁFICO 5

MAPA DE CALOR DE CORRELACIÓN DE VARIABLES NUMÉRICAS Y BOOLEANAS



Respecto a la edad, se seleccionan las variables edad y extra_edad_primaria, excluyendo extra_edad por redundante. Estas dos variables permiten capturar la situación de repetición en primaria y con la edad actual del estudiante se capturan atrasos por repetición o abandono posterior.

Para simplificar el análisis y mejorar la representatividad, sería adecuado agrupar las categorías de turno en un número menor de categorías y reorganizar los departamentos en regiones más amplias, de forma que se facilite el análisis sin perder información relevante sobre las variaciones regionales en el ausentismo. La variable de turno, por ejemplo, muestra que los turnos matutino e intermedio concentran la mayoría de los estudiantes, lo que sugiere una mayor uniformidad si se agrupan las categorías menos representadas.

Algunas variables adicionales no aportan información relevante para el análisis predictivo. Las variables estudiante_id y centro_cod, al ser simplemente identificadores, no tienen influencia en el ausentismo y, por tanto, serán excluidas. La variable cohorte, aunque

identifica la cohorte a la que pertenece cada estudiante, tampoco muestra una correlación directa con ausente_cr, por lo que también se omitirá del modelo.

Por último, dado que la variable anio_lectivo únicamente indica el año en que se registraron los datos, su valor no aporta información directa para la predicción del ausentismo crónico. Sin embargo, se ha identificado un cambio significativo en los patrones de ausentismo entre los años previos y posteriores a la pandemia de COVID-19, con un incremento notable en los registros pospandemia. Por ello, resulta conveniente crear una variable que distinga entre períodos pre y pospandemia. Esto permitirá al modelo reconocer las diferencias en los patrones de ausentismo antes y después de la pandemia, capturando los efectos contextuales que han impactado en la asistencia escolar en ambos períodos.

Atendiendo a los resultados del análisis exploratorio de datos, la base de datos final contiene 357.090 filas y 17 columnas, que se describen en la tabla 3.

TABLA 3
DESCRIPCIÓN DE VARIABLES DE LA BASE DE DATOS FINAL

Variable	Descripción	Tipo
ausente_cr	Indica si el estudiante es ausente crónico o no (variable objetivo)	Categórica
faltas_mes_3	Número de faltas del estudiante en marzo	Numérica
faltas_mes_4	Número de faltas del estudiante en abril	Numérica
faltas_mes_5	Número de faltas del estudiante en mayo	Numérica
inas_trimestral_prom	Promedio de inasistencias del grupo en el primer trimestre	Numérica
sexo	Sexo del estudiante (M/F)	Categórica
edad	Edad del estudiante al inicio del año lectivo	Numérica
región	Región donde reside el estudiante (Este/Litoral Sur/Centro/ Noreste/ Metropolitana/Litoral Norte)	Categórica
ivs_estudiante_norm	Índice normalizado de vulnerabilidad social del estudiante	Numérica
nivel1a6	Nivel educativo o grado de educación media básica (séptimo/octavo/ noveno)	Categórica
zona	Zona geográfica del centro educativo (urbana/rural)	Categórica
turno_cat	Turno de estudio asignado al estudiante (matutino/vespertino/nocturno)	Categórica
quintil_centro	Quintil de vulnerabilidad social del centro educativo (1 a 5)	Numérica
extra_edad_primaria	Indica si el estudiante tiene extraedad desde primaria	Booleana
fallo_año_anterior	Indicador de fallo final del estudiante en el curso académico del año lectivo anterior (aprueba/no aprueba/aprobación condicional)	Categórica
post_pandemia	Indica si la observación corresponde a los años posteriores a 2020	Booleana
ausente_cr_emb	Indica si el estudiante fue ausente crónico durante la educación media básica en los años lectivos anteriores	Booleana

Nota: la región Centro abarca los departamentos de Durazno, Flores y Florida; la región Este se compone por Lavalleja, Maldonado, Rocha y Treinta y Tres; la región Litoral Norte incluye los departamentos de Artigas, Paysandú y Salto; la región Litoral Sur comprende a Colonia, Río Negro y Soriano; la región Metropolitana abarca Canelones, Montevideo y San José, y la región Noreste incluye los departamentos de Cerro Largo, Rivera y Tacuarembó.

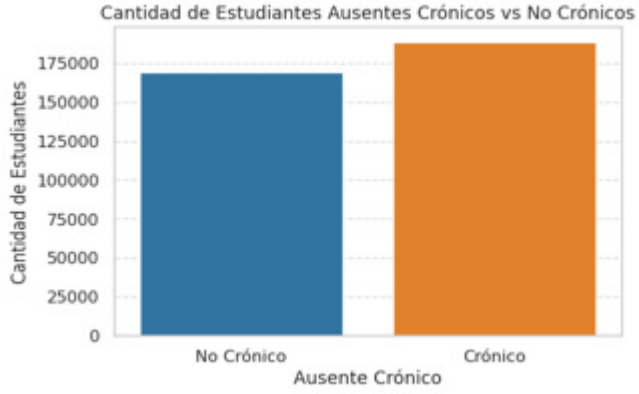
ANÁLISIS DE LA VARIABLE OBJETIVO: AUSENTISMO CRÓNICO

Antes de comenzar el análisis predictivo, es fundamental analizar el comportamiento de la variable que queremos predecir: el ausentismo crónico al final del año lectivo, así como su relación con las variables seleccionadas en la etapa anterior. Esta variable indica si el estudiante ha sido clasificado como crónicamente ausente al final del año lectivo, lo que la convierte en un indicador de riesgo académico o abandono. A continuación, se busca identificar patrones o tendencias para comprender el alcance del problema y orientar el análisis de las variables predictoras.

Distribución general del ausentismo crónico

El gráfico 6 muestra la distribución de estudiantes en las categorías de ausentismo crónico. Se observa una proporción ligeramente mayor de estudiantes crónicamente ausentes (57,3%), lo que resalta la importancia de abordar este problema, que alcanza a 188.325 alumnos en las cohortes analizadas. Asimismo, dado que la diferencia entre ambas categorías no es importante, no será necesario realizar ajustes en términos del balance de clases para el modelado, ya que la distribución de la variable objetivo se considera adecuada para el análisis predictivo.

GRÁFICO 6
DISTRIBUCIÓN DE LA VARIABLE OBJETIVO: AUSENTISMO CRÓNICO



Nota: no hay desbalance en la variable objetivo.

Evolución temporal del ausentismo crónico

El análisis temporal del ausentismo crónico permite observar cómo ha evolucionado esta problemática a lo largo de los años, lo cual es clave para identificar posibles factores externos que afecten la asistencia escolar. El gráfico 7 muestra la tendencia en el ausentismo crónico para cada año lectivo, desde 2016 hasta 2023, diferenciando entre estudiantes crónicamente ausentes y no crónicos.

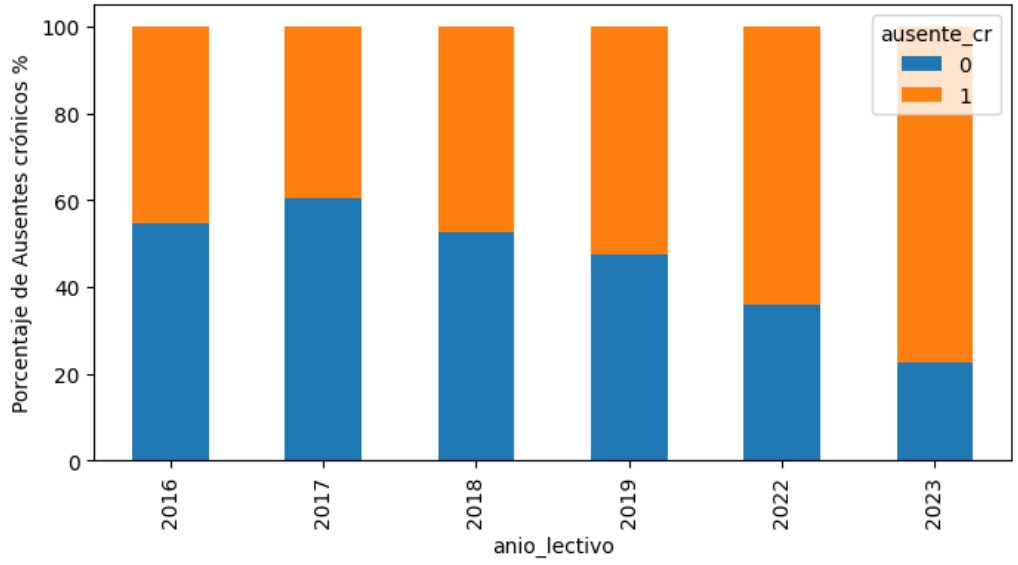
Entre 2016 y 2019, la proporción de estudiantes crónicamente ausentes presenta ciertas fluctuaciones. Si bien disminuye del 45,3% en 2016 al 39,7% en 2017, luego se incrementa

a 47,3% en 2018 y alcanza un 52,4% en 2019. Aunque estos cambios no son abruptos, la tendencia general en el período es al alza, acercándose e incluso superando en 2019 la proporción de estudiantes que no presentan ausentismo crónico.

A partir de 2022, sin embargo, se observa un notable incremento en la proporción de estudiantes crónicamente ausentes: alcanza el 64% ese año y aumenta aún más en 2023, donde el 77% de los estudiantes se clasifican como crónicamente ausentes. Este cambio marca un incremento preocupante en el problema del ausentismo crónico. La tendencia al alza, especialmente pronunciada en 2023, subraya la necesidad de establecer estrategias de intervención específicas para mitigar este fenómeno.

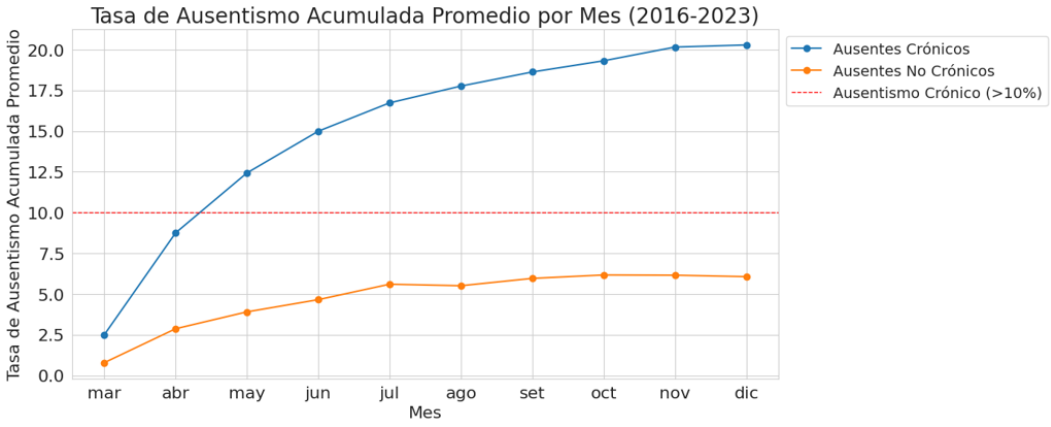
Es importante señalar que el análisis incluye únicamente estudiantes de las cohortes 2016 a 2021, por lo que la base de datos no contiene información de estudiantes ingresantes a la educación media básica en 2022 y 2023. Esta limitación podría sobrerrepresentar el ausentismo crónico en estos últimos años, ya que la base no capta nuevos ingresos, cuya tasa de ausentismo suele ser menor.

GRÁFICO 7
EVOLUCIÓN TEMPORAL DEL AUSENTISMO CRÓNICO POR AÑO LECTIVO
AÑOS 2016-2023



En el gráfico 8 se puede observar que, en promedio, los estudiantes que finalizan el año con ausentismo crónico alcanzan dicha condición cerca de un mes después de haber comenzado las clases y que a partir de entonces exhiben una tendencia al alza muy significativa, superando el 20% de ausentismo hacia fin de año. El comportamiento de los estudiantes sin ausentismo crónico es sustancialmente diferente: en primer lugar, su ausentismo acumulado promedio siempre permanece por debajo del 10% y se estabiliza alrededor del 5% durante la segunda mitad del año; además, la tendencia al alza existe, pero tiene un ritmo sustancialmente más lento. La brecha entre ausentes crónicos y ausentes no crónicos aumenta muy rápidamente y se acentúa.

GRÁFICO 8
TASA DE AUSENTISMO ACUMULADA PROMEDIO POR MES
AÑOS 2016-2023



Relación del ausentismo crónico con las variables seleccionadas

Para analizar el fenómeno del ausentismo crónico, hemos agrupado las variables, siguiendo una versión simplificada de la teoría ecológica del desarrollo, en tres categorías principales: individuales o familiares, del entorno escolar y comunitarias.

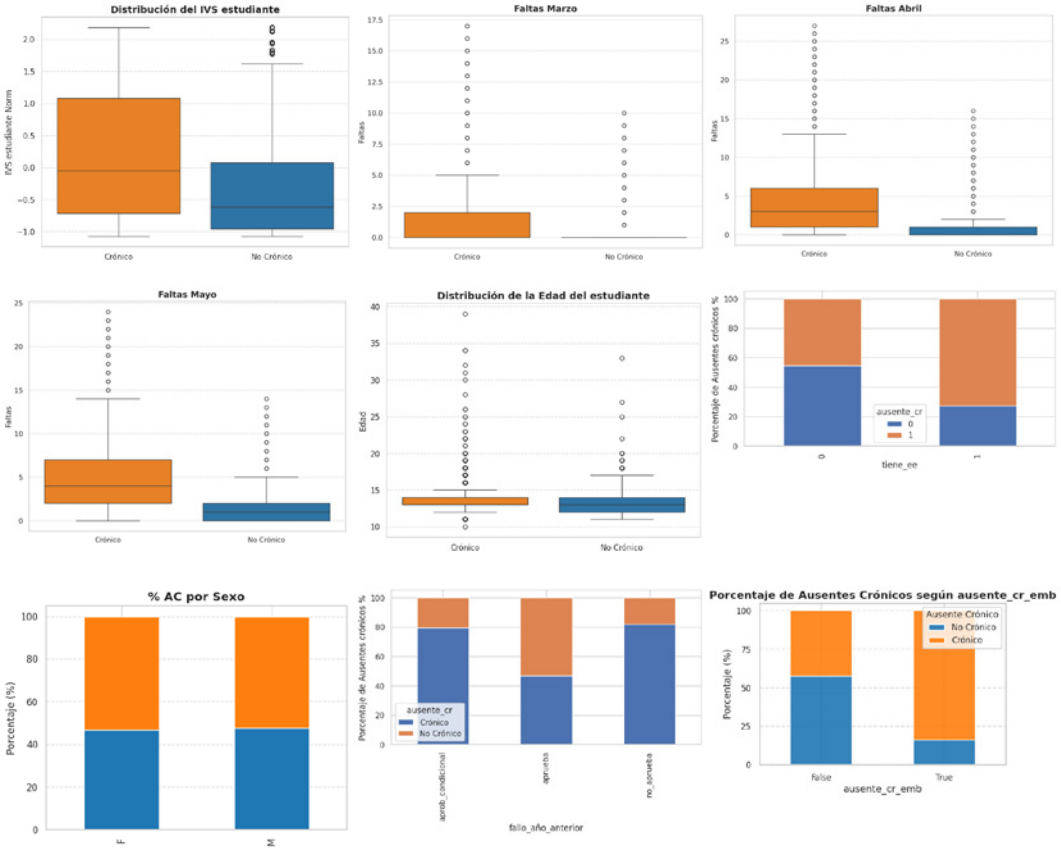
Relación con variables individuales y familiares

Las variables individuales y familiares incluyen aquellas características propias del estudiante y su entorno inmediato. Entre estas, el IVS destaca como un factor relevante. Los estudiantes con ausentismo crónico presentan, en promedio, un IVS más alto, lo que sugiere que aquellos en condiciones socioeconómicas más desfavorables enfrentan mayores dificultades para mantener una asistencia regular (gráfico 9).

Otro aspecto importante es la condición de extraedad en primaria, que se relaciona con una mayor incidencia de ausentismo crónico. Esto podría reflejar desmotivación o dificultades de adaptación que persisten a lo largo de la trayectoria educativa. Además, el historial académico del estudiante tiene incidencia en el ausentismo: aquellos que ya habían presentado ausentismo crónico en años anteriores tienden a repetir este patrón, indicando una persistencia del problema en ciertos casos. De manera similar, los estudiantes con bajo rendimiento en el año anterior, ya sea por fallo o aprobación condicional, muestran una mayor tendencia al ausentismo crónico. Las faltas mensuales acumuladas a lo largo de marzo, abril y mayo también reflejan un patrón sostenido de inasistencia, lo que refuerza la idea de que las dificultades para asistir regularmente están profundamente arraigadas en algunos estudiantes.

GRÁFICO 9

DISTRIBUCIÓN DE VARIABLES INDIVIDUALES O FAMILIARES ASOCIADAS AL AUSENTISMO CRÓNICO



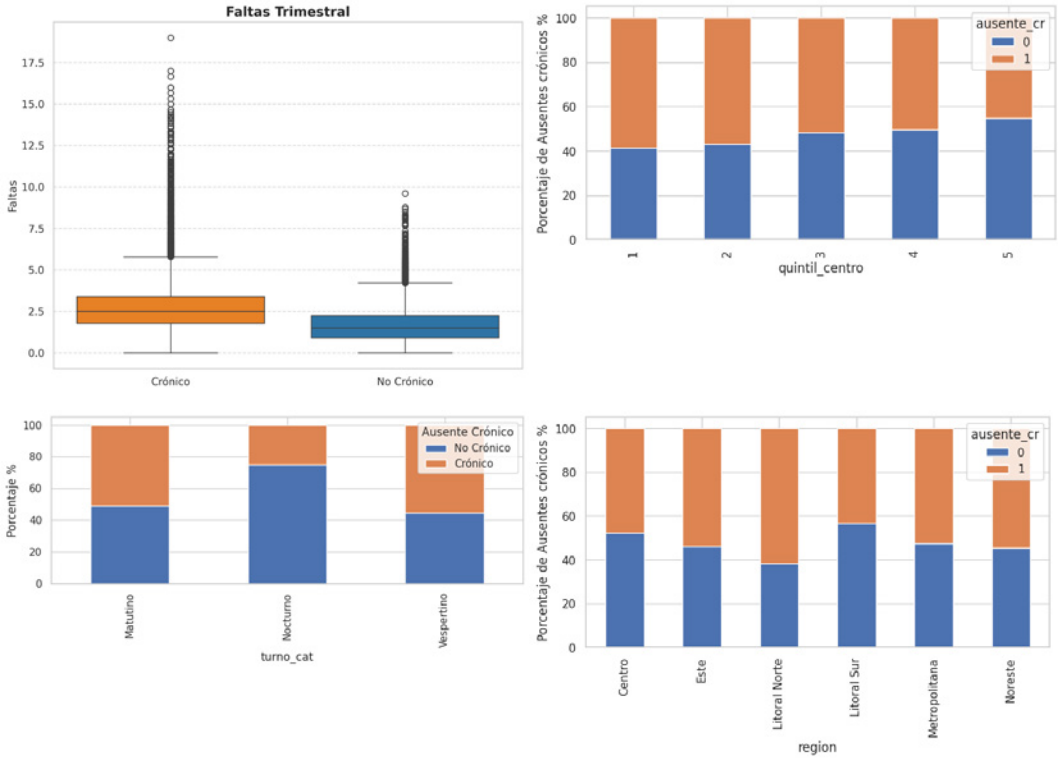
Relación con variables del entorno escolar

Las características del entorno escolar también tienen un impacto significativo en el ausentismo crónico. El quintil socioeconómico del centro educativo es un factor clave, dado que los centros ubicados en los quintiles más bajos registran una mayor proporción de estudiantes con ausentismo crónico, lo que refuerza la influencia del contexto socioeconómico del entorno escolar en la asistencia. El turno también afecta significativamente: los estudiantes del turno vespertino presentan mayores tasas de ausentismo crónico en comparación con los de los turnos matutino y nocturno. Esto podría explicarse por responsabilidades familiares adicionales o diferencias en la motivación y el perfil de los estudiantes en estos turnos.

Además, las dinámicas del grupo juegan un papel importante. El promedio de inasistencias del grupo es mayor en los grupos que incluyen estudiantes ausentes crónicos, lo que sugiere un posible efecto de influencia de pares. Por otro lado, las disparidades regionales también afectan el ausentismo, donde el Litoral Norte muestra mayores tasas de ausentismo crónico, lo que puede estar relacionado con desigualdades en el acceso a recursos educativos y apoyo escolar, así como con las condiciones económicas y sociales de la región.

GRÁFICO 10

DISTRIBUCIÓN DE VARIABLES DEL ENTORNO ESCOLAR ASOCIADAS AL AUSENTISMO CRÓNICO

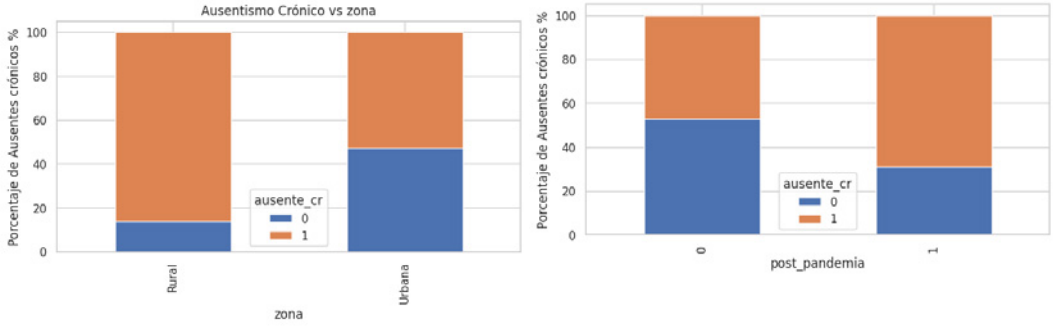


Relación con variables del centro educativo

Las variables comunitarias capturan factores más amplios relacionados con el entorno social y cultural del estudiante. Una de las principales influencias es la zona geográfica, ya sea rural o urbana. Los estudiantes de zonas rurales enfrentan mayores barreras estructurales, como transporte limitado y acceso restringido a servicios educativos, lo que incrementa la incidencia del ausentismo crónico. Por otra parte, el contexto pospandemia ha exacerbado significativamente este fenómeno. Los desafíos para reintegrarse al sistema educativo después de un prolongado periodo de distanciamiento social puede haber incrementado las tasas de ausentismo crónico, destacando el impacto de eventos a gran escala en la educación de los estudiantes.

En conclusión, el análisis de la relación entre el ausentismo crónico y las variables seleccionadas destaca la importancia de incorporarlas en modelos de predicción orientados a estimar la probabilidad de ausentismo crónico. La identificación de patrones consistentes en variables demográficas, socioculturales, del centro educativo y académicas sugiere que estas características pueden aportar información clave en la construcción de modelos predictivos robustos. Variables como la edad, el IVS, la zona de residencia y el historial académico pueden mejorar la precisión de las predicciones al capturar tanto factores individuales como contextuales que inciden en el ausentismo crónico.

GRÁFICO 11
DISTRIBUCIÓN DE VARIABLES COMUNITARIAS ASOCIADAS AL AUSENTISMO CRÓNICO



MODELOS DE *MACHINE LEARNING*

PREPROCESAMIENTO

El preprocesamiento de datos es una fase crucial en el aprendizaje automático. Típicamente, implica limpiar los datos (eliminar duplicados, corregir errores); tratar los datos faltantes (eliminándolos o rellenándolos) y atípicos, y normalizar y transformar los datos, entre otros procesos. Buena parte del procesamiento proviene de los resultados del análisis exploratorio de datos, pues en esa fase se analiza el comportamiento de las variables.

El preprocesamiento mejora la calidad de los datos y garantiza que el modelo de *machine learning* pueda interpretarlos correctamente. El tratamiento que se realiza a los datos depende del modelo. Por ejemplo, para la regresión logística, las variables categóricas deben ser transformadas en variables numéricas o *dummy* a través de la codificación *one-hot*; mientras, en los métodos de *boosting* este paso no es necesario.

Para nuestros modelos de regresión logística (regresión logística y regularización Ridge y Lasso) el preprocesamiento implicó:

- **Variables numéricas**. En primer lugar, calculamos la asimetría de las variables numéricas, empleando la función skew de scipy.stats, que calcula el grado de asimetría de una distribución de datos; esta medida indica si los valores de una variable están distribuidos de manera simétrica alrededor de la media o si tienden a concentrarse en uno de los extremos de la distribución. Las variables numéricas con una asimetría menor a o igual 1 (específicamente: ivs_estudiante_norm, quintil_centro, edad) fueron escaladas (empleando StandardScaler de scikit-learn). Por su parte, las variables numéricas con una asimetría mayor a 1 (específicamente: faltas_mes_3, faltas_mes_4, faltas_mes_5, inas_trimestral_prom) fueron escaladas (empleando StandardScaler) y transformadas (empleando PowerTransformer: método Yeo-Johnson).

El StandardScaler estandariza las variables numéricas para que tengan una media de 0 y una desviación estándar de 1. El puntaje estándar (o z- score) de una muestra χ se calcula como:

$$z = \frac{\chi - \mu}{\sigma}$$

donde μ es la media de las muestras de entrenamiento, y σ es la desviación estándar de las muestras de entrenamiento. El centrado y escalado se realizan de forma independiente en cada característica calculando los estadísticos relevantes en las muestras del conjunto de entrenamiento. La media y la desviación estándar se almacenan para utilizarlas en datos posteriores mediante la transformación. La estandarización de los datos es un requisito común para muchos estimadores de *machine learning;* en particular, los modelos lineales con regularización Ridge o Lasso asumen que todas las variables están centradas en 0 y tienen una varianza comparable.

Adicionalmente, la estandarización permite que datos medidos en escalas o unidades diferentes sean más comparables entre sí. Al estandarizar las variables, se eliminan las diferencias en las escalas y se asegura que todas las variables tengan la misma escala relativa (Tezen, 2024).

El método Yeo-Johnson (aplicado adicionalmente a la estandarización para las variables numéricas con alto sesgo) es uno de los métodos de transformación que acercan los datos numéricos a una distribución normal.

- **Variables categóricas**. Las variables categóricas **no ordinales** de la base son sexo, turno_cat, región y zona. A estas variables se les aplica un codificador *one-hot de sklearn*. Esta técnica convierte datos categóricos en un formato binario en el que cada categoría está representada por una columna independiente con un 1 que indica su presencia y un 0 para todas las demás categorías. Esta transformación permite tratar cada categoría de forma independiente sin implicar falsas relaciones entre ellas (Rojo-Echeburúa, 2024). A las variables categóricas **ordinales** se les aplica OrdinalEncoder de sklearn. Incluyen quintil_centro (cuyo orden de menor a mayor es de 1 a 5, siendo 1 el quintil más vulnerable y 5 el quintil menos vulnerable); nivel_1a6 (cuyo orden de menor a mayor es séptimo, octavo y noveno), y fallo_año_anterior (cuyo orden de menor a mayor es no aprueba, aprobación condicional y aprueba).

Para nuestros modelos de *boosting*, se transforman las variables numéricas únicamente con StandardScaler. En el caso de las variables categóricas, los modelos XGBoost y LightGBM requieren codificador *one-hot* para las no ordinales, y *ordinal* para las ordinales. La transformación de variables categóricas a numéricas en CatBoost **no es necesaria**, ya que este modelo maneja las variables categóricas de forma nativa.

Por último, en la regresión logística, además del preprocesamiento, se emplea el método SelectKBest de sklearn. SelectKBest realiza una evaluación independiente de cada característica en relación con la variable objetivo y elige las que obtengan la puntuación más alta de acuerdo a una función de evaluación determinada (en problemas de clasificación:

f_classif para análisis de varianza [ANOVA]). En nuestro proyecto, se elegirán las cinco características con la puntuación más alta. El empleo de SelectKBest tiene tres ventajas: reducción del sobreajuste (menos datos redundantes significa menos probabilidad de tomar decisiones basadas en datos redundantes o ruido), mejora de la precisión (menos datos engañosos significa que mejora la precisión del modelo) y reducción del tiempo de entrenamiento (menos datos significa que los algoritmos se entrenan más rápido) (Brownlee, 2020).

DEFINICIÓN DE HIPERPARÁMETROS PARA RANDOMIZEDSEARCH

La siguiente tabla define los hiperparámetros de la búsqueda con RandomizedSearch. Para cada combinación, se ejecuta una validación cruzada de 3 pliegues, y la métrica de evaluación que se utiliza es el F1-Score. Si bien es importante predecir con alta precisión los estudiantes ausentes crónicos, lo cual favorecería la selección de la medida *recall*, el F1-Score contribuye a generar un modelo que garantice que la mayoría de las intervenciones se dirijan a estudiantes que realmente necesitan apoyo sin sobrecargar al sistema educativo con casos incorrectamente etiquetados como en riesgo.

TABLA 4
DESCRIPCIÓN DE HIPERPARÁMETROS Y CONFIGURACIÓN PARA LOS MODELOS *BOOSTING*

Hiperparámetro	Definición	Método y configuración
C	Controla la fortaleza de la regularización L2 (Ridge) o L1 (Lasso) cuando se especifica la penalización. Valores más bajos (altos) aplican una regularización más fuerte (débil), ayudando a reducir el sobreajuste (lo que aumenta el riesgo de sobreajuste).	Regresión logística: [0,01; 0,1; 1; 10; 100]
alpha	Alpha es el parámetro de regularización en RidgeClassifier. Controla la fuerza de la regularización, ayudando a reducir la complejidad del modelo. Valores más altos de alpha implican una regularización más fuerte, penalizando grandes coeficientes y reduciendo la varianza del modelo.	Ridge: [0,1; 1; 10; 100; 1000]
C con L1	Con penalty='l1', C regula la fortaleza de la regularización L1 (similar a Lasso). L1 fuerza algunos coeficientes a ser exactamente 0, realizando selección de características. Valores más bajos de C aumentan la regularización y promueven una mayor cantidad de coeficientes 0 (eliminando variables irrelevantes).	Lasso: [0,1; 1; 10; 100; 1000]
n_estimators	Número de árboles potenciados por gradiente. Aumentarlo puede mejorar el modelo, pero puede llevar al sobreajuste.	XGBoost: [100; 150; 200] LightGBM: [100; 150; 200]
max_depth	Profundidad máxima de cada árbol. Controla la complejidad del modelo.	XGBoost: [3; 5; 7] LightGBM: [-1; 5; 10; 15] CatBoost: [4; 6; 7]
learning_rate	Tamaño del paso utilizado para evitar el sobreajuste.	XGBoost: [0,01; 0,05; 0,1] LightGBM: [0,005; 0,01; 0,05; 0,1] CatBoost: [0,01; 0,03; 0,05; 0,1]
subsample	Fracción de observaciones seleccionadas aleatoriamente para cada árbol.	XGBoost: [0,6; 0,8; 1,0] LightGBM: [0,6; 08, 1.0]
colsample_bytree	Fracción de columnas seleccionadas aleatoriamente para cada árbol.	XGBoost: [0.6, 0.8, 1.0] LightGBM: [0.6, 0,8; 1,0]
gamma	Reducción mínima de pérdida requerida para realizar una división.	XGBoost: [0; 0,1; 0,2; 0,3]

min_child_weight	Suma mínima de peso de instancia (Hessiana) necesaria en un nodo hijo.	XGBoost: [1; 3; 5] LightGBM: [0,001; 0,01; 0,1; 1]
reg_alpha	Término de regularización L1 para reducir el sobreajuste.	XGBoost: [0; 0,01; 0,05; 0,1] LightGBM: [0; 0,01; 0,05; 0,1]
reg_lambda	Término de regularización L2 para reducir el sobreajuste.	XGBoost: [0; 0,5; 1,0] LightGBM: [0; 0,5;1,0]
num_leaves	Número de hojas en el árbol completo; afecta a la complejidad del modelo.	LightGBM: [20; 35; 50]
iterations	Número de iteraciones de *boosting*.	CatBoost: [100; 150; 200]
l2_leaf_reg	Término de regularización L2.	CatBoost: [1; 3; 5; 7; 9]
bagging_temperature	Controla la variabilidad de las muestras de los subconjuntos de datos utilizados para el entrenamiento.	CatBoost: [0; 0,5; 1; 1,5; 2]
random_strength	Coeficiente para controlar el impacto de las permutaciones aleatorias de características en el cálculo de la puntuación.	CatBoost: [0,5; 1; 1,5; 2]

EVALUACIÓN DE LOS MODELOS

Para el entrenamiento de los modelos de predicción temprana del ausentismo crónico, se seleccionó el 80% de las observaciones y un 20% para testear.

Para la evaluación de los modelos se obtienen las principales métricas de evaluación (tabla 5) y se grafican las curvas ROC (gráfico 12). Por último, se grafica la distribución de los *scores* obtenidos en la validación cruzada (gráfico 13).

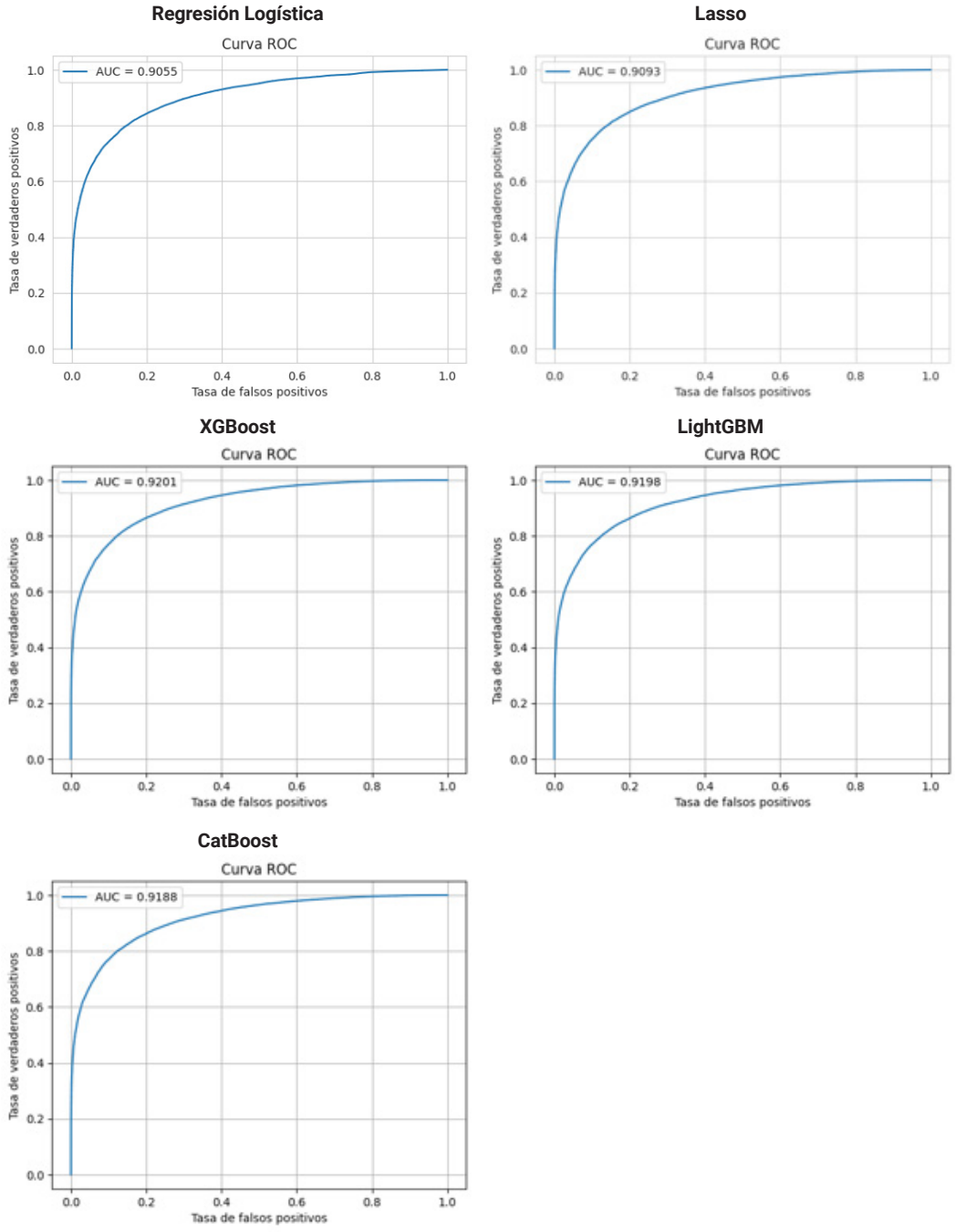
Como se observa en la tabla 5, las métricas de desempeño de los modelos son muy similares. Todos los modelos superan el 82% en su medida de *accuracy*; los modelos de *boosting* mejoran ligeramente esta métrica, superando el 83% en todos los casos. Si bien, en general, los modelos de *boosting* reportan mejores desempeños que los modelos de regresión logística, en el caso de nuestro proyecto, la consistencia en la limpieza, preparación y preprocesamiento de la base de datos permite que la regresión logística sea, a pesar de sus limitaciones, un muy buen modelo.

TABLA 5
MEDIDAS DE RENDIMIENTO DE LOS MODELOS

Modelo	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Regresión Logística	0,82455	0,83537	0,83112	0,83324	0,90551
Ridge	0,82821	0,84960	0,81930	0,83417	N/A
Lasso	0,82751	0,83907	0,83263	0,83584	0,90934
XGBoost	0,83811	0,83911	0,83811	0,83822	0,92006
LightGBM	0,83707	0,83816	0,83707	0,83719	0,91980
CatBoost	0,83667	0,83766	0,83667	0,83678	0,91881

GRÁFICO 12

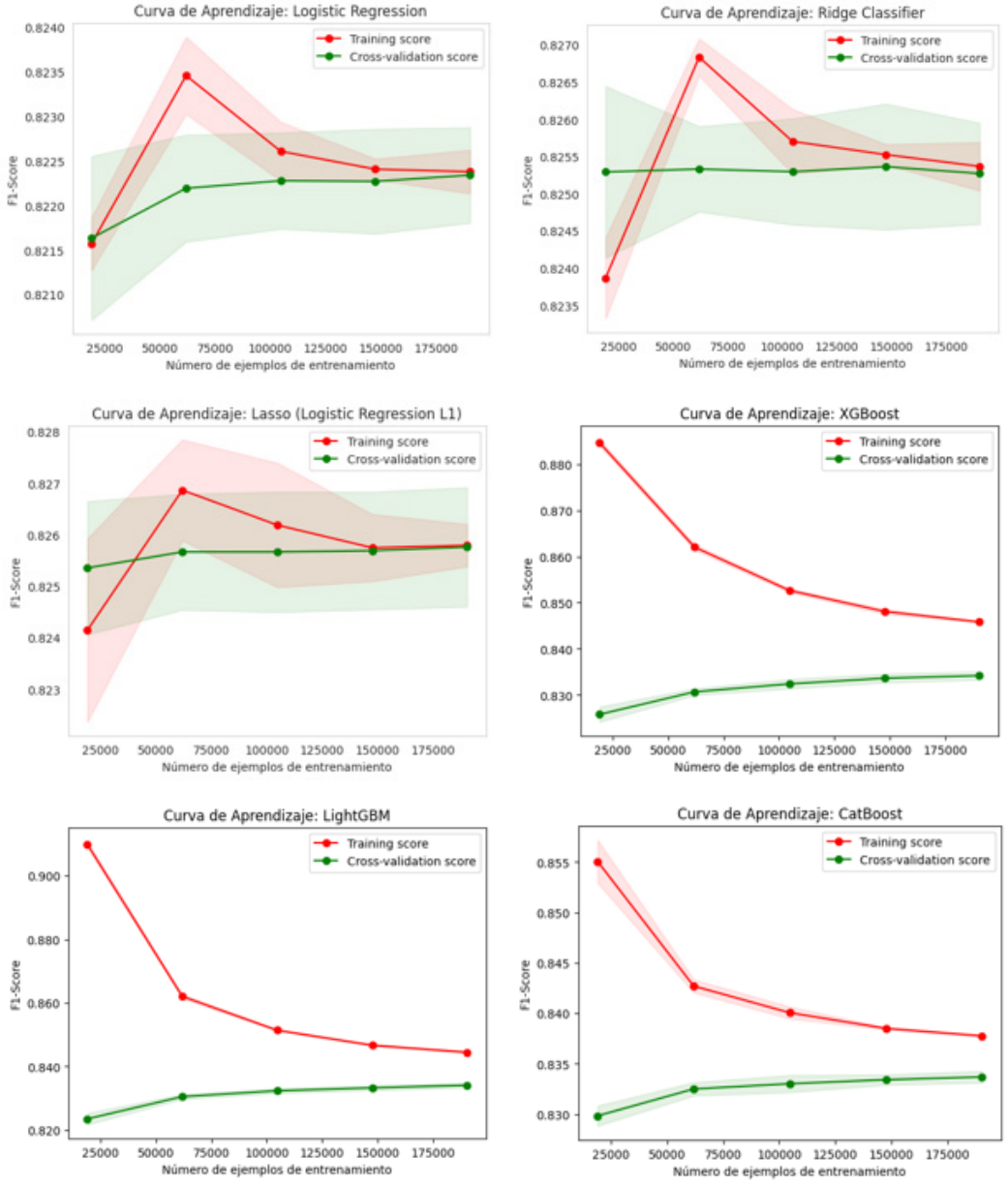
CURVAS ROC DE LOS MODELOS REGRESIÓN LOGÍSTICA (IZQUIERDA) Y LASSO (DERECHA)



Nota: la diferencia entre el valor AUC de los modelos es muy pequeña.

GRÁFICO 13

CURVAS DE APRENDIZAJE DE LOS MODELOS

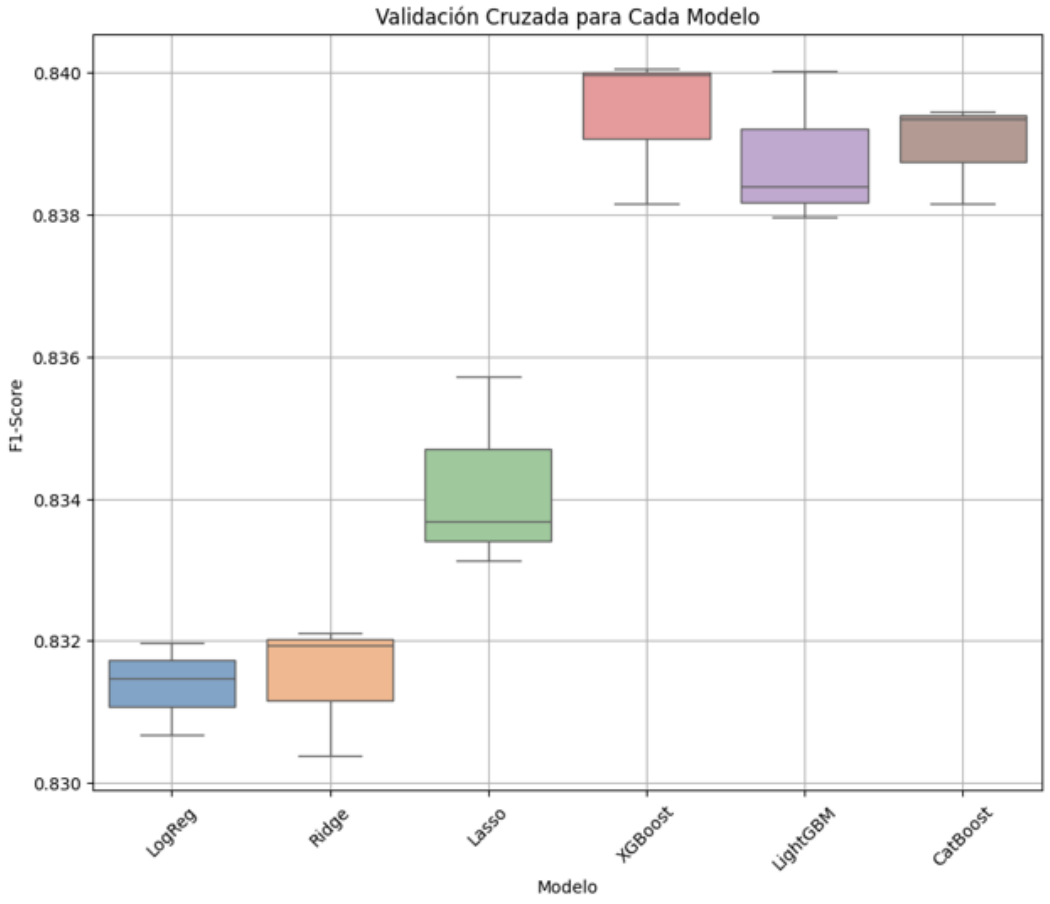


Las curvas de aprendizaje de los modelos de regresión logística sin y con penalización muestran un comportamiento similar (gráfico 13). El F1-Score en validación cruzada aumenta gradualmente y tiende a estabilizarse conforme se agregan más datos. Las líneas de validación cruzada (*cross-validation score*) y entrenamiento (*training score*) tienden a converger conforme aumenta el número de ejemplos en entrenamiento. Esto es señal de que los modelos están generalizando bien. Asimismo, la diferencia entre ambas curvas disminuye con más datos, lo cual sugiere que los modelos no tienen problemas de sobreajuste ni de subajuste. Por otra parte, en los modelos de regresión logística, con y sin penalización, las sombras son comparativamente amplias, lo que indica que hay

una variabilidad relativamente alta en las puntuaciones de entrenamiento y validación cruzada al usar diferentes particiones del conjunto de datos. Las sombras en las curvas de los modelos de *boosting* son muy estrechas, casi inexistentes, lo cual sugiere una mayor consistencia en el rendimiento.

El gráfico 14 muestra los *boxplots* de los *scores* de validación cruzada para cada modelo. Los modelos de *boosting* superan a los modelos de regresión en términos de F1-Score, mostrando tanto un mayor valor promedio como una menor variabilidad en sus resultados. Sin embargo, los modelos de regresión también ofrecen un desempeño razonable (observar la poca variabilidad de los valores en el eje de las Y) y pueden ser opciones válidas cuando la interpretabilidad es prioritaria o cuando los recursos computacionales son limitados.

GRÁFICO 14
PUNTAJES DE VALIDACIÓN CRUZADA PARA LOS DIFERENTES MODELOS

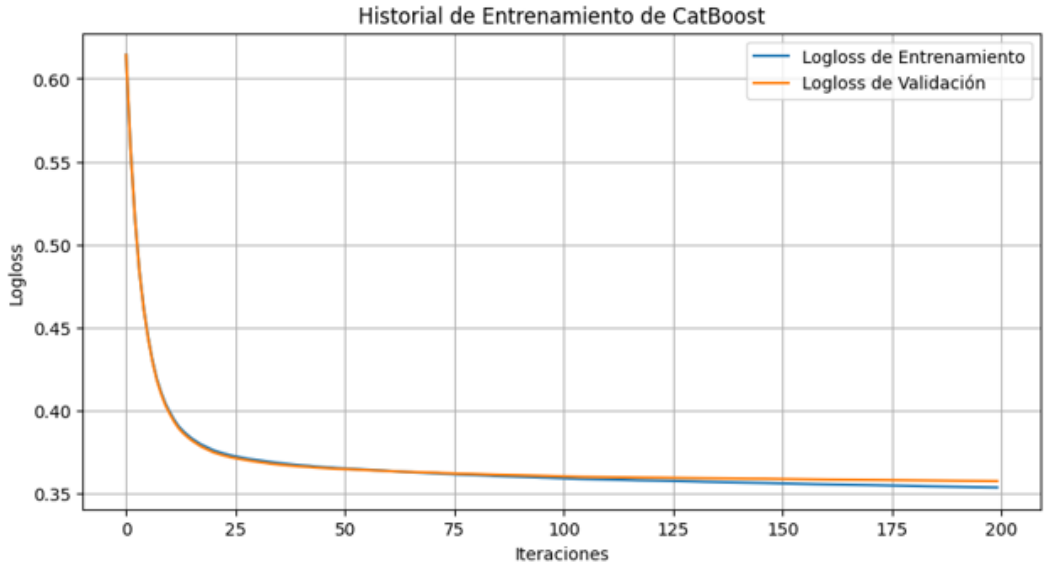


Nota: los modelos de *boosting* tienen un mejor desempeño en el F1-Score (en valor y variabilidad).

Los modelos *boosting* muestran un desempeño más consistente y puntuaciones ligeramente más altas en la validación cruzada. En particular, XGBoost obtiene la mediana más alta en su F1-Score. Sin embargo, las diferencias en las métricas de rendimiento de los tres modelos *boosting* son marginales. De hecho, la prueba de Friedman concluye que los rendimientos de los modelos no son estadísticamente diferentes. En este contexto, entra en la ecuación de la

selección del modelo el conocimiento técnico sobre la materia. En particular, se analiza la importancia de las variables. En los modelos *boosting* las variables que explican la predicción y su intensidad tienen algunas diferencias (ver Anexo). En función de la literatura, los resultados del análisis exploratorio de datos y las virtudes del modelo **CatBoost** para el tratamiento de variables categóricas, este es el modelo seleccionado para continuar el análisis y proponerlo como el mejor modelo de predicción temprana de la ausencia crónica en la educación secundaria básica.

GRÁFICO 15
HISTORIAL DE ENTRENAMIENTO DEL MODELO CATBOOST



Nota: las curvas de pérdida de entrenamiento y validación son muy similares y permanecen casi paralelas a lo largo de las iteraciones (modelo no sobreajusta a datos de entrenamiento).

REPORTE DE CLASIFICACIÓN DEL MODELO CATBOOST

A continuación, se examina el reporte de clasificación del modelo CatBoost. El valor de *accuracy* señala que el modelo clasificó correctamente al 84% de las instancias en el conjunto de prueba, lo que señala un buen rendimiento del modelo. Además, en virtud de las métricas *recall* y *precision*, sabemos que el modelo identificó correctamente al 82% de los ausentes crónicos verdaderos, y que el 86% de las predicciones que el modelo hizo como ausentes crónicos fueron correctas. El F1-Score está equilibrado entre ambas clases (83% en la clase 0 y 84% en la clase 1), lo que señala que el modelo es eficaz para clasificar correctamente ambas clases de estudiantes (ausentes crónicos y no ausentes crónicos).

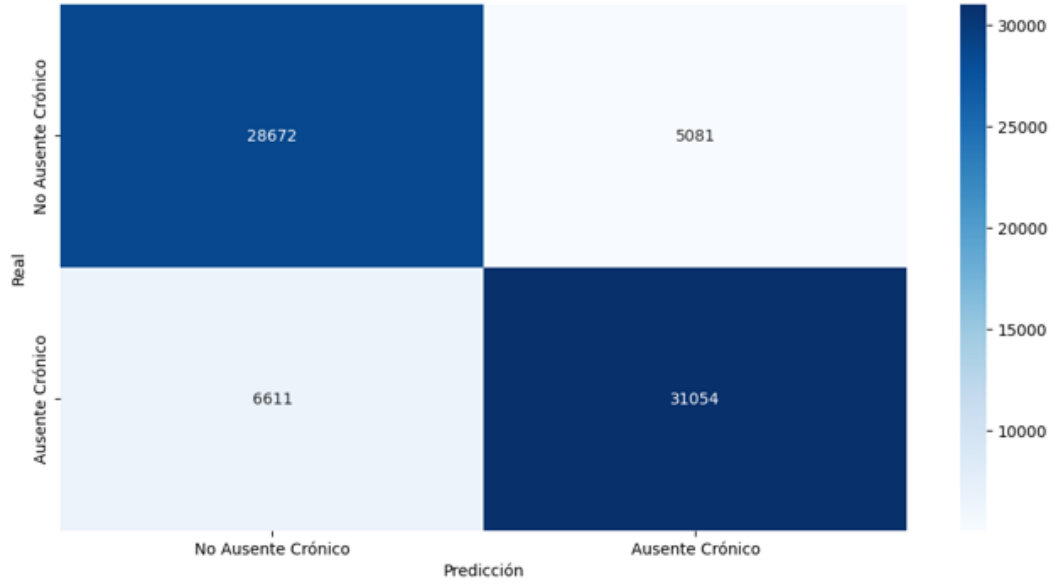
TABLA 6
REPORTE DE CLASIFICACIÓN DEL MODELO CATBOOST

Clases	Precision	Recall	F1-Score	Support
0 (No ausente crónico)	0,81	0,85	0,83	33.753
1 (Ausente crónico)	0,86	0,82	0,84	37.665

Nota: el modelo muestra un buen rendimiento en términos de precision, recall y F1-Score para ambas clases.

En función de estas medidas, el modelo CatBoost podría considerarse una buena herramienta de diagnóstico temprano del ausentismo crónico. En este sentido, cabe recordar que el modelo está diseñado para identificar, en el mes de junio, a los estudiantes con riesgo de convertirse en ausentes crónicos hacia el fin del año lectivo. Las métricas obtenidas aseguran que el porcentaje de ausentes crónicos captados es del 82%.

GRÁFICO 16
MATRIZ DE CONFUSIÓN DEL MODELO CATBOOST



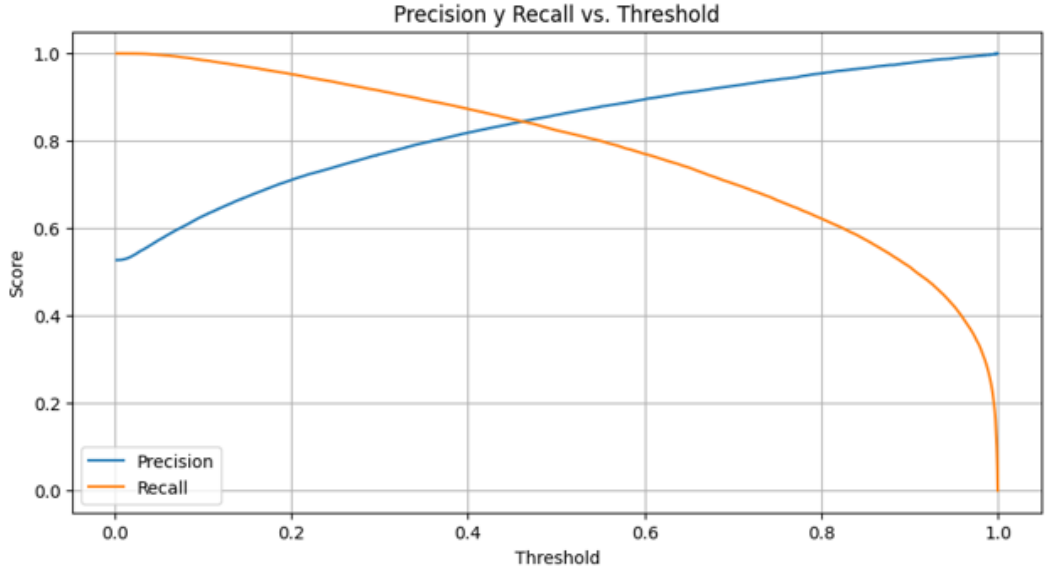
Nota: el modelo predice correctamente como ausentes crónicos a un total de 31.054 estudiantes (cuadrante abajo a la derecha).

Para esta clasificación, el umbral (*threshold)* óptimo para maximizar el F1-Score es aproximadamente 0,42 (gráfico 17) y el F1-Score máximo alcanzado con este umbral es 0,8455. El *threshold* es el valor de probabilidad a partir del cual el modelo clasifica una predicción como positiva (1). En los problemas de clasificación binaria, el *threshold* por defecto suele ser 0,5. Sin embargo, es posible ajustar este valor para generar otro tipo de equilibrio entre la *precisión* y el *recall*. En el modelo actual, el porcentaje de intervenciones innecesarias, calculado como la cantidad de falsos positivos con relación al total de casos negativos verdaderos (los ejemplos clasificados como *no ausentes*), es 15,1%.

Como mencionamos, el ausentismo crónico es un problema que, si no se aborda de manera temprana, puede convertirse en abandono escolar. Por lo tanto, proponemos ajustar el *threshold* (modificándolo a 0,30) para mejorar el *recall* y con esto capturar una mayor cantidad de ausentes crónicos. En el plano de la implementación de medidas para disminuir el ausentismo crónico, un ajuste del *threshold* redundaría en que la eventual intervención sobre los alumnos tendría una mayor cobertura, lo cual, como se discutirá más adelante, puede ser conveniente o no dependiendo de los recursos disponibles y del tipo de intervenciones que se realicen.

En un nuevo reporte de clasificación del modelo CatBoost, con un menor *threshold* (0,30 en lugar de 0,42), el *accuracy* disminuye en tres puntos porcentuales (del 84% a 81%); en cambio, el *recall* aumenta en 9 puntos porcentuales (del 82% al 91%).

GRÁFICO 17
PRECISION Y RECALL VS. THRESHOLD



Nota: el *threshold* óptimo del modelo es aproximadamente 0,42.

TABLA 7
REPORTE DE CLASIFICACIÓN DEL MODELO CATBOOST CON THRESHOLD = 0,30

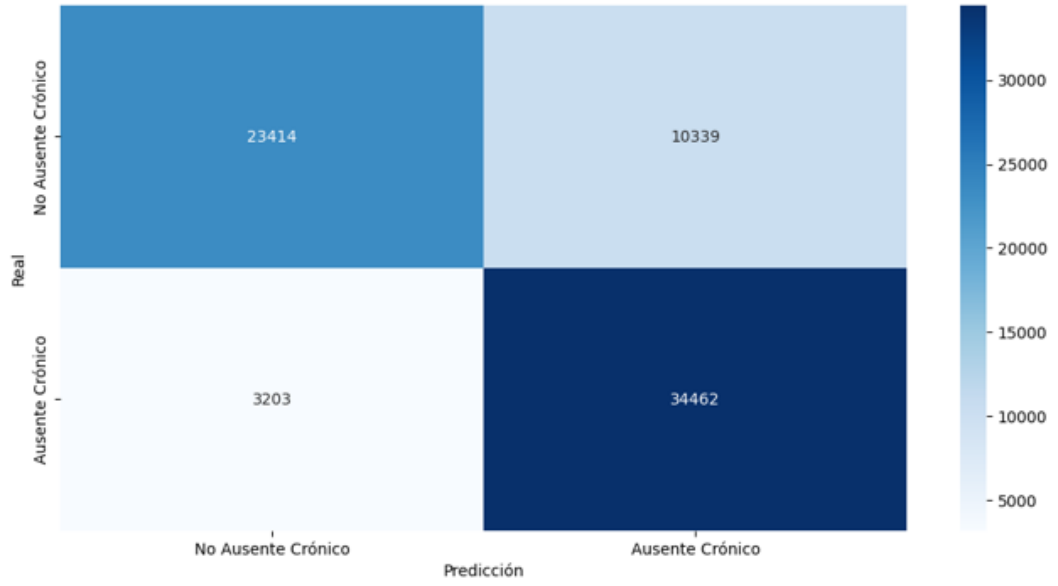
Clases	Precision	Recall	F1-Score	Support
0 (No ausente crónico)	0,88	0,69	0,78	33.753
1 (Ausente crónico)	0,77	0,91	0,84	37.665

Nota: el modelo ajustado mejora el mejor *recall* de la clase 1 en 9 puntos porcentuales.

Con la disminución del *threshold*, obtenemos un porcentaje de ausentes crónicos captados del 91,5% (casi 10 puntos de mejora en el *recall*); mientras que el porcentaje de intervenciones innecesarias se duplica (30,6%).

GRÁFICO 18

MATRIZ DE CONFUSIÓN DEL MODELO CATBOOST CON THRESHOLD = 0,30



Nota: el modelo predice correctamente como ausentes crónicos a un total de 34.462 estudiantes (cuadrante abajo a la derecha).

ANÁLISIS DE IMPORTANCIA DE CARACTERÍSTICAS

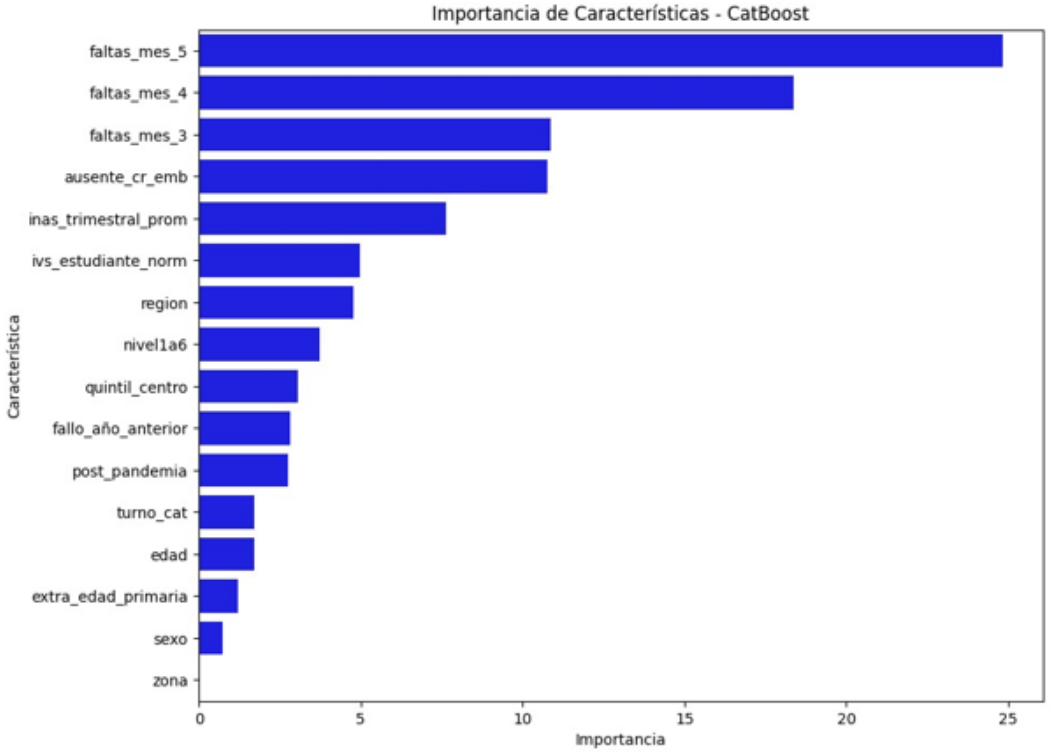
El análisis de importancia de las variables permite identificar aquellos factores que más contribuyen a predecir el riesgo de ausentismo crónico en estudiantes de educación secundaria básica pública en Uruguay. Este análisis resulta esencial, dado que proporciona información clave sobre las variables que deben ser monitoreadas y consideradas en las intervenciones preventivas. A continuación, se presenta un análisis detallado de las variables más relevantes de CatBoost.

Las ausencias tempranas de los estudiantes en el año lectivo, específicamente en marzo, abril y mayo, son las variables con mayor importancia en el modelo. Esto apoya la hipótesis del proyecto de que el patrón de inasistencias tempranas es un fuerte predictor del riesgo de ausentismo crónico hacia fin de año. El hallazgo también coincide con lo que plantea London (2016), quien sostiene que los estudiantes que presentan inasistencias desde el inicio del año lectivo tienden a consolidar patrones de ausentismo, lo cual incrementa su riesgo de desvinculación del sistema educativo. En el análisis de las variables logramos identificar que los ausentes crónicos alcanzan el umbral del 10% al mes y medio de iniciar las clases, y continúan este patrón hasta doblar el porcentaje promedio hacia fin de año (gráfico 8). Así, la inclusión de estas ausencias tempranas como variables predictivas en el modelo resulta particularmente relevante.

Por otro lado, la variable ausente_cr_emb, que indica si el estudiante ha tenido ausentismo crónico en años previos, se posiciona como un factor predictivo de relevancia en el modelo. Este resultado subraya que se trata de un fenómeno más estructural y no es solamente un evento aislado, reflejando problemas más profundos y persistentes en la trayectoria

educativa de los estudiantes. El análisis exploratorio ya sugería el fuerte vínculo entre ausente crónico y esta variable.

GRÁFICO 19
IMPORTANCIA DE CARACTERÍSTICAS DEL MODELO CATBOOST



La importancia de las variables analizadas hasta el momento sugiere que el correcto registro de las inasistencias de los estudiantes es de crítica importancia para predecir (y prevenir, en el mejor de los casos) el ausentismo crónico. En Uruguay, la reciente implementación de la transformación educativa establece cambios en los criterios de pasaje de grado. A partir del 2023, los estudiantes que acumulan más de 30 inasistencias fictas durante el año lectivo y tengan hasta tres unidades curriculares con escaso o mínimo nivel de avance quedarán con un fallo de acreditación total o parcial[5]. Teniendo en cuenta que cada falta justificada pesa 0,5, un alumno podría faltar 60 días de forma justificada y promover. Además de ser un límite excesivamente elevado, la flexibilidad que introduce este cambio con respecto a las inasistencias genera preocupación con respecto a la calidad de los registros de las faltas para adelante. Concretamente, si los docentes y centros educativos no priorizan el monitoreo de la asistencia debido a estos esquemas de promoción más flexibles que habilita la transformación educativa, la fiabilidad de esta variable podría disminuir en el futuro.

[5] El artículo 74 del Reglamento de Evaluación del Estudiante aprobado en mayo de 2024 establece: "Al finalizar los cursos en el mes de diciembre, aquellos estudiantes que hayan acumulado más de 30 inasistencias fictas en el grado y tengan: a) hasta tres unidades curriculares con escaso o mínimo nivel de avance, según la referencia de evaluación del artículo 40, quedarán con un fallo de acreditación total o parcial; b) entre cuatro a seis unidades curriculares con escaso o mínimo nivel de avance, según la referencia de evaluación del artículo 40, quedarán con un fallo de acreditación diferida a la evaluación de los APE hasta el mes de febrero; c) más de seis unidades curriculares con escaso o mínimo nivel de avance, según la referencia de evaluación del artículo 40, la asamblea de Profesores podrá habilitar, por mayoría simple, la posibilidad de realizar los APE de febrero. En caso de empate en la votación, define quien presida la reunión final. En caso negativo, recursa el grado por desvinculación. Si el estudiante no asiste a ninguno de los APE que le correspondiere se emitirá el fallo de recursa por desvinculación de acuerdo al artículo 73" (ANEP, 2024).

Este escenario, por un lado, limita la capacidad de reflejar patrones de comportamiento reales y, por otro, compromete la posibilidad de predecir el ausentismo crónico.

La variable inas_trimestral_prom representa el promedio de inasistencias del grupo durante el primer trimestre. Su relevancia en el modelo indica el impacto del entorno de pares en el comportamiento de la asistencia individual. Este resultado es consistente con la teoría ecológica del desarrollo de Bronfenbrenner (1979), que sostiene que el contexto social influye en la conducta individual. En este caso, pertenecer a un grupo con altos niveles de ausentismo puede normalizar o incentivar las inasistencias entre sus miembros. Dräger et al. (2023) plantean que los estudiantes son influenciados por las normas grupales, de manera que un entorno con ausentismo elevado puede aumentar el riesgo de que se ausenten de forma similar. Al igual que en el caso anterior, el monitoreo de las inasistencias grupales podría verse afectado a partir de los cambios introducidos en el nuevo plan educativo. Si los cambios en el reglamento de evaluación y promoción afectan el relevamiento de las inasistencias, los promedios de grupo se verían afectados impactando la capacidad del modelo para identificar la influencia del contexto del grupo.

El IVS del estudiante también se destaca como factor de gran relevancia en el modelo, respaldando los resultados consistentes en la literatura sobre la relación entre la situación socioeconómica de los estudiantes y su propensión al ausentismo. En el contexto uruguayo, varios estudios han resaltado que los estudiantes en condiciones de mayor vulnerabilidad enfrentan barreras adicionales que dificultan su asistencia regular, tales como la necesidad de salir al mercado laboral desde edades tempranas o la falta de recursos familiares que apoyen su educación (ANEP, 2023b; de Melo et al., 2015). Además, Liu y Lee (2022) también enfatizan que la inseguridad económica es un factor crítico que incrementa el riesgo de inasistencia crónica, destacando que los estudiantes de contextos más desfavorecidos suelen priorizar otras necesidades urgentes, lo que afecta su continuidad educativa. La importancia que tiene la inclusión de esta variable en el modelo evidencia que el ausentismo debe abordarse desde una perspectiva integral que contemple las condiciones socioeconómicas de los estudiantes.

La variable región muestra una relevancia en el modelo comparable a la del IVS, lo cual indica que el contexto geográfico es un factor significativo en la predicción del ausentismo crónico. Esto sugiere que las diferencias regionales en el acceso a recursos educativos, infraestructura y otras condiciones tienen un impacto en la frecuencia de ausencias de los estudiantes. En Uruguay, el contexto regional puede influir en aspectos como los tiempos de traslado, la disponibilidad de transporte, las inclemencias climáticas y las condiciones socioeconómicas, factores que pueden afectar la asistencia regular de los estudiantes. Así, la importancia de esta variable evidencia que el ausentismo crónico no solo es una cuestión individual, sino que también depende del contexto estructural y regional en el que habita el estudiante. Este resultado también está respaldado en la teoría ecológica del desarrollo, que resalta cómo el contexto ambiental influye en el comportamiento del individuo.

Salvo la variable zona, el resto de las variables incluidas en el modelo parecen ser relevantes para la predicción del ausentismo crónico, aunque con una menor importancia en comparación con las ausencias mensuales o el IVS. Sin embargo, estas variables contribuyen al modelo

al capturar factores contextuales y estructurales que influyen en el riesgo de ausentismo. Por ejemplo, el nivel educativo refleja cómo el grado puede afectar de forma diferencial el riesgo de ausentismo, mientras que el quintil del centro aporta información sobre el contexto social del entorno escolar. Las variables fallo en el año anterior y pospandemia brindan información sobre el compromiso académico previo y las circunstancias específicas recientes que han afectado la asistencia de los estudiantes y que pueden estar incidiendo en nuestra predicción. Asimismo, variables como edad, extraedad y sexo también añaden valor al modelo. La edad y extraedad permiten identificar a estudiantes que no están en la cohorte de edad esperada, lo cual se ha asociado a una mayor probabilidad de inasistencia y de trayectorias educativas irregulares. La variable sexo, aunque con menor peso en el modelo, aporta un matiz adicional que puede reflejar diferencias de género en los patrones de asistencia, asociadas a factores sociales y culturales. En conjunto, aunque estas variables son menos determinantes de manera individual, enriquecen la capacidad del modelo para captar una variedad de factores asociados al ausentismo crónico, permitiendo una visión más integral y contextualizada del fenómeno.

En conclusión, el análisis de importancia de características en el modelo CatBoost destaca una combinación de factores individuales, sociales y contextuales que inciden en el riesgo de ausentismo crónico en estudiantes de educación secundaria básica en Uruguay. Las variables más relevantes, como las inasistencias en los primeros meses del año, el historial de ausentismo crónico y el IVS, subrayan la importancia de patrones tempranos de inasistencia y factores socioeconómicos en la predicción del ausentismo. Además, variables como la región geográfica y el entorno grupal refuerzan el rol de los contextos estructurales y regionales, destacando la influencia del entorno en el comportamiento individual.

Al comparar estos resultados con los obtenidos en los modelos XGBoost, LightGBM y regresión logística, se observa una coherencia en las variables más relevantes para la predicción del ausentismo crónico (ver Anexo). Las ausencias en los primeros meses del año, el historial de ausentismo previo y el IVS destacan en todos los modelos como factores clave, lo que destaca la importancia de estas variables. Aunque la magnitud de la importancia relativa puede variar entre modelos, existe un consenso en que estos factores son predictores robustos del riesgo de ausentismo crónico. Adicionalmente, variables contextuales como la región geográfica y el promedio de inasistencias del grupo también aparecen con relevancia en varios modelos, reforzando la idea de que el entorno influye significativamente en el comportamiento de asistencia. Esta consistencia entre modelos aumenta la confiabilidad de los resultados, brindando una visión integral y validada de los factores críticos que deben considerarse en intervenciones preventivas contra el ausentismo crónico.

DISCUSIÓN

Este proyecto se orienta a la predicción temprana del ausentismo crónico en estudiantes que cursan educación secundaria básica pública en Uruguay, identificando a aquellos con mayor riesgo para implementar intervenciones preventivas oportunas que mejoren sus trayectorias educativas. La implementación de un modelo predictivo de ausentismo crónico es innovador en el ámbito educativo nacional, dado que permite anticipar problemas de asistencia antes de que se traduzca en una mayor probabilidad de fracaso académico o abandono escolar. En particular, esta herramienta responde a una necesidad urgente dentro del sistema educativo uruguayo, especialmente en los sectores más vulnerables, donde el ausentismo crónico tiene un mayor impacto en la continuidad académica.

En Uruguay se han realizado avances en el uso de modelos predictivos en educación media para identificar estudiantes en riesgo de abandono. Proyectos como el elaborado por Queiroga et al. (2022) en conjunto con la ANEP han implementado sistemas de alerta temprana para detectar el riesgo de deserción mediante modelos como *random forest* y redes neuronales. Estos modelos han alcanzado niveles altos de precisión, considerando a las faltas como factor de riesgo. Sin embargo, este proyecto representa un paso previo, al enfocarse en la predicción del ausentismo crónico específicamente, anticipando uno de los factores clave que aumentan el riesgo de abandono. Al identificar este riesgo de ausentismo crónico de forma temprana, este proyecto permite no solo identificar a los estudiantes para reducir la probabilidad de abandono, sino también optimizar las intervenciones preventivas sobre el ausentismo.

Uno de los principales hallazgos de este análisis es la importancia que cumplen las ausencias tempranas del estudiante en la predicción de riesgo de ausentismo crónico. El hecho de que las faltas de los primeros meses del año lectivo (marzo, abril y mayo) contribuyan significativamente al modelo respalda nuestro objetivo, dado que se busca predecir el ausentismo crónico basándonos en datos del primer trimestre. Este hallazgo sugiere que los patrones de inasistencia comienzan a establecerse temprano en el año, proporcionando una oportunidad para intervenir de manera preventiva y oportuna.

Es razonable considerar que estas inasistencias al inicio del año reflejan la influencia de factores subyacentes como el contexto familiar y el entorno escolar y comunitario, los cuales también contribuyen al ausentismo crónico. Aunque el análisis exploratorio no reveló altas correlaciones que sugieran multicolinealidad, estos factores contextuales y personales forman parte de una red más compleja de influencias que se reflejan tanto en las ausencias tempranas como en el ausentismo crónico al final del año, en sintonía con la literatura (Kearney y Childs, 2023; Liu y Lee, 2022), que también ha señalado la influencia

del entorno socioeconómico y escolar en la asistencia regular. Este estudio intenta predecir el ausentismo crónico, además de analizar los factores que inciden en él, por lo que la incorporación de estas variables resulta de gran utilidad para la precisión del modelo. Estos hallazgos no solo refuerzan el enfoque adoptado para la selección de variables en el proyecto, sino que también demuestran la capacidad del modelo de aprendizaje automático para captar la complejidad de estas interrelaciones, proporcionando un mayor entendimiento para el diseño de intervenciones.

En función de la literatura consultada y las variables disponibles, sería interesante que futuras investigaciones consideren la incorporación de variables que reflejen de manera más completa el entorno familiar y el bienestar socioemocional del estudiante, principalmente en relación con su percepción y experiencia en el entorno educativo, del cual no se incorpora información en este análisis. La literatura destaca que estos factores son determinantes en el comportamiento de ausentismo, dado que el apoyo familiar, el sentido de pertenencia con el centro educativo y la seguridad percibida en el entorno escolar influyen de manera significativa en la asistencia regular (Gilmore y Newcomer, 2022; Liu y Lee, 2022; Torino, 2023). Aunque actualmente esta información está disponible en el sistema educativo para algunas muestras de estudiantes, como los que participan en pruebas estandarizadas (por ejemplo, Aristas), buscar mecanismos para extender esta información al resto de la población estudiantil sería sumamente beneficioso. Incorporar esta información al algoritmo podría no solo mejorar la precisión predictiva del modelo, sino enriquecer la comprensión de los factores subyacentes al ausentismo crónico, permitiendo intervenciones más personalizadas y eficaces para mejorar la asistencia.

Otra limitación de este estudio es que se centra exclusivamente en el subsistema de estudiantes de liceos públicos, decisión motivada tanto por la complejidad de los datos de la Dirección General de Educación Técnico Profesional (DGETP) como por el gran volumen de información que implicaba considerar ambos subsistemas.

La implementación del modelo enfrentó desafíos significativos en la gestión de estos datos, debido a que los registros administrativos utilizados no fueron diseñados para análisis estadístico, lo cual demandó un proceso intensivo de limpieza y transformación, incluyendo la creación de variables adicionales como el IVS. Esta experiencia remarca la necesidad de contar con estándares uniformes en el registro de la información para mejorar la calidad de los datos en el sistema educativo uruguayo. Expandir el análisis a toda la educación media sería necesario, pero requeriría mayores recursos de infraestructura y tiempo, dada la heterogeneidad en los registros y en los planes educativos entre los subsistemas.

Es importante considerar el impacto que puede tener una posible interrupción en el monitoreo de las inasistencias debido a los cambios incorporados a partir de la transformación educativa. La mayor flexibilización en el pasaje de grado, donde las inasistencias ya no definen por sí solas la repetición, ha generado que algunos docentes dejen de pasar lista regularmente. Esto afecta tanto el seguimiento de la asistencia como la precisión del algoritmo que necesita de estos datos para ser efectivo. Es por esto que resulta fundamental que desde las autoridades educativas se exija el monitoreo riguroso y continuo de la asistencia.

Por otro lado, el análisis de la sensibilidad del modelo ante el ajuste del *threshold* de clasificación revela que si el objetivo es captar la mayor cantidad posible de estudiantes en riesgo (*recall*), un *threshold* más bajo incrementaría los verdaderos positivos, aunque también aumentaría los falsos positivos. Este ajuste puede ser apropiado para intervenciones preventivas de menor costo, como la intervención que se encuentra realizando Ceibal con el envío de cartas a las familias. Sin embargo, en el caso de intervenciones más intensivas, como el acompañamiento por parte de equipos de seguimiento, que requieren recursos significativos, mayores valores de *precision* (exactitud con un *threshold* más alto) podría ser preferible para optimizar el uso de los recursos y asegurar que las intervenciones lleguen a estudiantes con riesgo claro de ausentismo.

CONCLUSIONES

El problema del ausentismo crónico en la educación media es un fenómeno extendido y preocupante, que abarca aproximadamente al 53% de las cohortes analizadas en este estudio. Este dato sugiere que más de la mitad de los estudiantes de educación secundaria básica en el sector público presentan una frecuencia de inasistencias significativa, lo que implica una ausencia equivalente a al menos al 10% de los días lectivos durante el año. Este tipo de ausentismo se considera particularmente crítico porque no solo afecta la continuidad del aprendizaje, sino que también está directamente relacionado con un mayor riesgo de fracaso académico y abandono escolar.

El estudio revela que el ausentismo crónico es un fenómeno multidimensional, influenciado por factores sociodemográficos, académicos y contextuales. Los estudiantes pertenecientes a los sectores de mayor vulnerabilidad socioeconómica son los más afectados, ya que enfrentan desafíos adicionales que incrementan la probabilidad de ausentarse del centro educativo.

El ausentismo crónico no solo es un problema que afecta directamente la continuidad educativa de los estudiantes, sino que también refleja y acentúa profundas desigualdades en el acceso y las oportunidades de aprendizaje. Este fenómeno, al presentarse con mayor frecuencia en estudiantes provenientes de sectores de alta vulnerabilidad socioeconómica, subraya la disparidad existente en el sistema educativo. Los estudiantes con menos recursos enfrentan múltiples barreras que dificultan su asistencia regular y su participación activa en el proceso educativo, incluyendo la necesidad de contribuir económicamente a sus hogares, la falta de apoyo familiar y contextos escolares poco inclusivos o estimulantes. Como resultado, el ausentismo se convierte en un síntoma visible de desigualdades estructurales que perpetúan el ciclo de la exclusión social y educativa.

En este contexto, el uso de algoritmos de predicción, como los desarrollados en este trabajo, se presenta como una herramienta innovadora y valiosa para focalizar las políticas y las intervenciones educativas hacia aquellos estudiantes que más lo necesitan. Uruguay tiene la oportunidad de utilizar estos algoritmos para optimizar el uso de recursos limitados, mejorando así la eficiencia y el impacto de las estrategias de intervención. La educación pública enfrenta limitaciones presupuestales y de recursos humanos, lo que hace fundamental contar con herramientas que permitan priorizar de manera efectiva a los estudiantes con mayores riesgos. Los modelos predictivos desarrollados, al identificar tempranamente a quienes están en riesgo de ausentismo crónico, permiten a los actores educativos implementar intervenciones focalizadas, eficientes y basadas en evidencia.

Durante el desarrollo del proyecto se probaron varios modelos de clasificación, como la regresión logística, XGBoost, LightGBM y CatBoost, siendo estos últimos los que lograron el mejor desempeño predictivo. El modelo CatBoost alcanzó un *accuracy* notable del 84%, con un AUC-ROC del 0,92, lo cual evidencia su capacidad para discriminar correctamente entre estudiantes en riesgo y aquellos que no lo están. El modelo es una herramienta de gran impacto para tomar decisiones informadas que conduzcan a intervenciones tempranas y efectivas. Aunque los modelos de *boosting* obtuvieron los mejores rendimientos, los otros algoritmos más tradicionales también obtuvieron métricas comparables, lo cual le da robustez a los resultados obtenidos. La consistencia entre los modelos garantiza que las predicciones no dependan de un único enfoque o conjunto de características, sino que, en cambio, se trata de una capacidad predictiva robusta que está bien fundamentada en los datos.

La limpieza y preparación de los datos utilizados para entrenar los modelos resultaron ser etapas cruciales para asegurar la calidad de los resultados. Los datos administrativos utilizados en este proyecto presentaban inicialmente varios problemas, como inconsistencias y faltantes, que debieron ser abordados antes de implementar los modelos de aprendizaje automático. Esto puso de manifiesto la importancia de mejorar los sistemas de información y la calidad de los registros educativos, dado que el éxito de cualquier estrategia basada en datos depende en gran medida de la calidad de la información con la que se trabaja. Es necesario seguir avanzando en la integración y sistematización de los datos administrativos para garantizar que las futuras aplicaciones de estos modelos puedan ser igual de efectivas y precisas.

Una de las principales contribuciones del proyecto es su capacidad para generar un sistema de alerta temprana de ausentismo crónico. Los resultados muestran que el modelo CatBoost puede identificar con un alto *recall* (91,5% luego de ajustar el *threshold*) a los estudiantes que tienen riesgo de convertirse en ausentes crónicos durante el primer trimestre del año lectivo. Esta capacidad de predicción temprana es fundamental porque permite actuar antes de que el problema se consolide y se torne más difícil de revertir. Detectar a estos estudiantes en las etapas iniciales del ciclo lectivo ofrece una ventana de oportunidad para la implementación de estrategias de apoyo y acompañamiento que pueden prevenir el abandono escolar y mejorar las trayectorias educativas.

El algoritmo desarrollado ofrece la posibilidad de capturar una población potencialmente en riesgo de ausentismo crónico, pero el éxito de la intervención no solo depende de la capacidad predictiva del modelo, sino también de los tipos de intervención que se implementen y los recursos disponibles para cada caso. Es decir, mientras que los algoritmos permiten identificar a los estudiantes en riesgo, la efectividad real dependerá de cómo se implementen las intervenciones. En función de los recursos disponibles, los umbrales de riesgo pueden ajustarse. Por ejemplo, para intervenciones de bajo costo y amplio alcance (como el envío de notificaciones a los hogares), se puede usar un umbral que maximice la sensibilidad del modelo. Para intervenciones más intensivas y costosas se debe optar por un umbral que priorice la precisión.

Este proyecto representa un avance significativo para el sistema educativo, ofreciendo una herramienta basada en datos para mejorar la eficiencia y efectividad de las intervenciones dirigidas a combatir el ausentismo crónico en estudiantes de liceos de educación media básica pública, que puede extenderse al resto de los subsistemas. A través del desarrollo de modelos predictivos robustos y la implementación de un sistema de alerta temprana, se pueden crear oportunidades para reducir la desigualdad en el acceso a la educación y promover trayectorias educativas más inclusivas y exitosas. No obstante, la efectividad de estas herramientas depende de la mejora continua de los sistemas de registro de información, así como de un compromiso país para priorizar recursos e intervenciones que respondan a las necesidades reales de los estudiantes en mayor situación de vulnerabilidad.

ANEXO

IMPORTANCIA DE LAS CARACTERÍSTICAS DE LAS VARIABLES

GRÁFICO A.1
IMPORTANCIA DE CARACTERÍSTICAS DEL MODELO XGBOOST

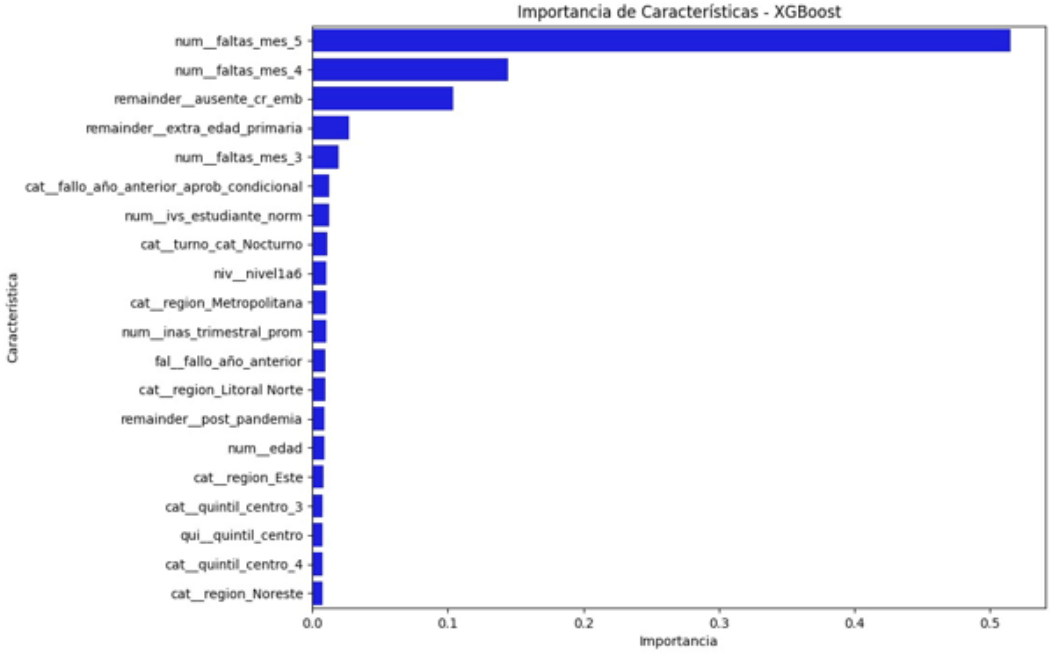


GRÁFICO A.2

IMPORTANCIA DE CARACTERÍSTICAS DEL MODELO LIGHTGBM

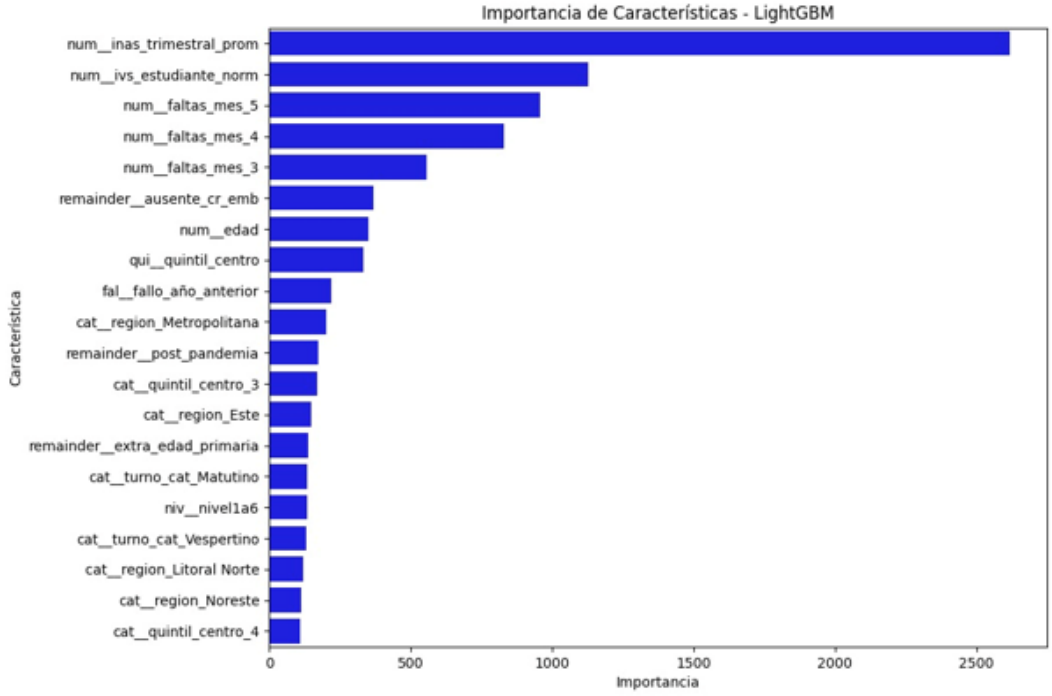
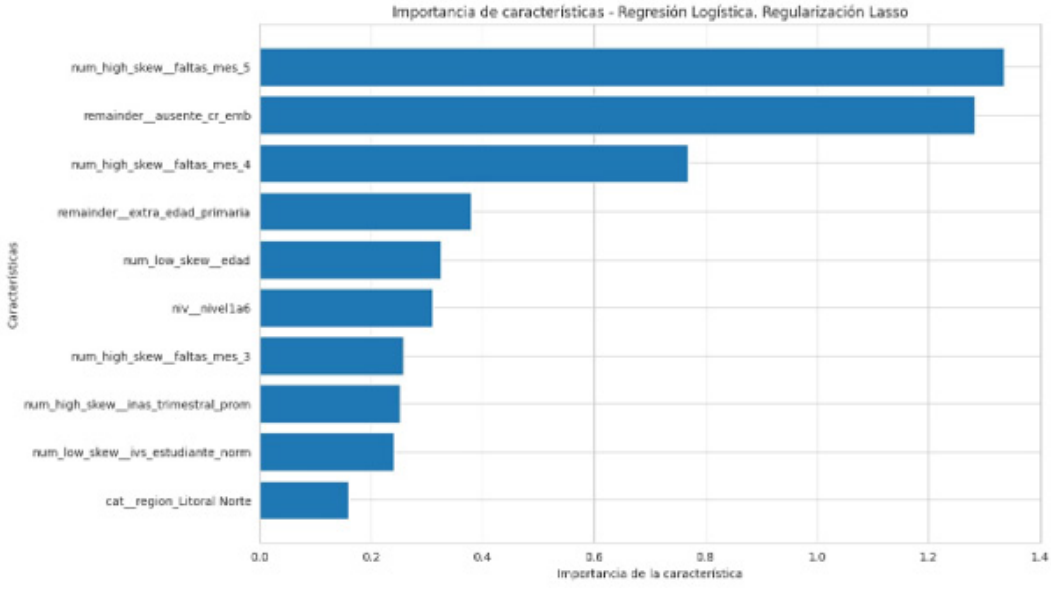


GRÁFICO A.3

IMPORTANCIA DE CARACTERÍSTICAS DEL MODELO LASSO



BIBLIOGRAFÍA

Alvez Legelén, M. (2024). *Estudio panel sobre los factores asociados a las trayectorias educativas en Educación Media ¿Es posible construir un sistema de alertas tempranas en Uruguay?* (Udelar). Recuperado de https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/44706

ANEP. (2021). *Rendición de Cuentas 2020. Tomo 5. Educación en tiempos de pandemia. Acción 2020.* Recuperado de https://www.anep.edu.uy/sites/default/files/images/2021/noticias/julio/20210701/TOMO 5 EDUCACIÓN EN TIEMPOS DE PANDEMIA - ACCIÓN 2020 Rendición de Cuentas 2020.pdf

ANEP. (2022). *Índice de Vulnerabilidad Socioeconómica en Enseñanza Media. Diciembre 2022.* Recuperado de https://observatorio.anep.edu.uy/sites/default/files/arch/IVSEducMediaANEP_InformeMetodologico202303.pdf

ANEP. (2023a). *Plan de acción para mejorar la asistencia en Educación Inicial y Primaria* (pp. 1-14). Recuperado de https://www.anep.edu.uy/sites/default/files/images/2023/noticias/agosto/230815/Plan ASISTE 2023.pdf

ANEP. (2023b). *Uruguay en PISA 2022. Volumen 1. Logros educativos, su evolución y contexto.* Recuperado de https://pisa.anep.edu.uy/sites/default/files/Recursos/Publicaciones/Informes/2022/Uruguay en PISA 2022_Volumen 1_Logros educativos, su evolución y contexto.pdf

ANEP. (2024). *Reglamento de Evaluación del Estudiante (REDE) de la Educación Básica Integrada* (pp. 1-26). Recuperado de https://transformacioneducativa.anep.edu.uy/sites/default/files/images/componentes/Curricular/documentos/ebi/REDE 2024.pdf

Arteaga Ramos, A. y Tapias López, N. (2024). *Modelo de recomendación de acciones para la prevención de la inasistencia escolar basado en un sistema predictivo de aprendizaje automático.* Recuperado de https://repositorio.unicordoba.edu.co/entities/publication/a0d9e45e-29d1-4119-bfb4-ff5c6c16e48a

Bergstra, J. y Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*, 281-305.

Bronfenbrenner, U. (1979). *The Ecology of Human Development. Experiments by Nature and Design.* https://doi.org/10.2307/j.ctv26071r6

Brown, S. (2021). Machine learning, explained. Recuperado de https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Brownlee, J. (2020). Feature Selection For Machine Learning in Python. Recuperado de https://machinelearningmastery.com/feature-selection-machine-learning-python/

Cardozo, S., Silveira, A. y Fonseca, B. (2022). Detección temprana del riesgo escolar. Predicción de trayectorias de rezago en la educación primaria en Uruguay mediante técnicas de machine learning. *Revista Latinoamericana de Estudios Educativos*, *52*(2), 297-326.

de Melo, G., Failache, E. y Machado, A. (2015). Adolescentes que no asisten a ciclo básico: caracterización de su trayectoria académica, condiciones de vida y decisión de abandono. *Páginas de Educación*, *8*(2), 66-88. Recuperado de http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S1688-74682015000200003

Dräger, J., Klein, M. y Sosu, E. (2023). *School absence trajectories and their consequences for achievemen* (p. 38). https://doi.org/10.35542/osf.io/ch4jq

Fernández-Casal, R., Costa, J. y Oviedo, M. (2024). *Métodos predictivos de aprendizaje estadístico*. https://doi.org/10.17979/spudc.9788497498937

Gilmore, D. y Newcomer, K. (2022).). Examining Chronic Absenteeism in Elementary Schools Among Minority Students Utilizing the Systemic Questioning Evaluative Framework. *Annals of Public Health & Epidemiology*, *1*. https://doi.org/10.33552/APHE.2022.01.000522

INEEd. (2020). *Aristas 2018. Informe de resultados de tercero de educación media*. Recuperado de https://www.ineed.edu.uy/images/Aristas/Publicaciones/Aristas2018/Aristas-2018-Informe-de-resultados.pdf

INEEd. (2023). *Aristas 2022. Informe de resultados de tercero de educación media*. Recuperado de https://www.ineed.edu.uy/images/Aristas/Publicaciones/Aristas2022/Aristas-2022-Informe-resultados-tercero-educacion-media.pdf

INEEd. (2024). *Egreso de media: nuevas cifras, viejas conclusiones*. Recuperado de https://www.ineed.edu.uy/images/boletines/2024/Egreso-media-nuevas-cifras-viejasconclusiones.pdf

Kashnitsky, Y. (2020). Topic 10. Gradient Boosting. Recuperado de https://www.kaggle.com/code/kashnitsky/topic-10-gradient-boosting

Kearney, C. A. y Childs, J. (2023). Translating sophisticated data analytic strategies regarding school attendance and absenteeism into targeted educational policy. *Improving Schools*, *26*(1), 5-22. https://doi.org/10.1177/13654802231174986

Lee, K., McMorris, B. J., Chi, C.-L., Looman, W. S., Burns, M. K. y Delaney, C. W. (2023). Using data-driven analytics and ecological systems theory to identify risk and protective factors for school absenteeism among secondary students. *Journal of School Psychology*, *98*, 148-180. https://doi.org/10.1016/j.jsp.2023.03.002

Liu, J. y Lee, M. (2022). *Beyond Chronic Absenteeism: The Dynamics and Disparities of Class Absences in Secondary School* (N.º 15.664). Bonn.

London, R. A., Sanchez, M. y Castrechini, S. (2016). The Dynamics of Chronic Absence and Student Achievement. *Education Policy Analysis Archives*, *24*(112). https://doi.org/10.14507/epaa.24.2741

Macarini et al., L. A. (2018). Using Data Mining Techniques to Follow Students Trajectories in Secondary Schools of Uruguay. *2018 XIII Latin American Conference on Learning Technologies (LACLO)*, 307-314. https://doi.org/10.1109/LACLO.2018.00061

Mukli, L. y Rista, A. (2022). Predicting and Analyzing Student Absenteeism Using Machine Learning Algorithm. *Integration of Education*, *26*, 216-228. https://doi.org/10.15507/1991-9468.107.026.202202.216-228

Queiroga, E. M., Batista Machado, M. F., Paragarino, V. R., Primo, T. T. y Cechinel, C. (2022). Early Prediction of At-Risk Students in Secondary Education: A Countrywide K-12 Learning Analytics Initiative in Uruguay. *Information*, *13*(9), 401.

Rojo-Echeburúa, A. (2024). What Is One Hot Encoding and How to Implement It in Python. Recuperado de https://www.datacamp.com/tutorial/one-hot-encoding-python-tutorial

Torino, M. T. (2023). *Deserción escolar en Argentina* (Universidad Torcuato Di Tella). Recuperado de https://repositorio.utdt.edu/items/6fe256a2-fc2c-48d9-8eae-3393bc680169

UNESCO. (2021). *Los sistemas de alerta temprana para prevenir el abandono escolar en América Latina y el Caribe*. Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000380354

Veneri, F. y Aguirre Imbriaco, E. (2018). *Modelización del desempeño educativo en la educación media mediante aprendizaje automático* (Udelar). Recuperado de https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/30571