Oficina de Seguimiento y Evaluación Estratégica

ALERTA ESCUELA:

Metodología para el cálculo del riesgo de deserción interanual en el Perú con *machine learning*.

Identificando riesgos de interrupción de estudios mediante métodos predictivos

ALERTA ESCUELA

PERÚ Ministerio de Educación

BICENTENARIO PERÚ 2024

ALERTA ESCUELA:
Metodología para el cálculo del riesgo de deserción interanual en el Perú con *machine learning*.

Identificando riesgos de interrupción de estudios mediante métodos predictivos

PERÚ Ministerio de Educación

ALERTA ESCUELA: METODOLOGÍA PARA EL CÁLCULO DEL RIESGO DE DESERCIÓN INTER-ANUAL EN EL PERÚ CON MACHINE LEARNING.

Identificando riesgos de interrupción de estudios mediante métodos predictivos

EDICIÓN
Ministerio de Educación
Oficina de Seguimiento y Evaluación Estratégica
Calle Del Comercio N.° 193, San Borja
Lima 15021, Perú
Teléfono: (511) 615-5800
https://www.gob.pe/minedu

Morgan Niccolo Quero Gaime
Ministro de Educación

Néstor Alfonso Supanta Velásquez
Secretario de Planificación Estratégica

Saraí Sirley Valdivia Zapana
Jefa de la Oficina de Seguimiento y Evaluación Estratégica

Claudia Paola Lisboa Vásquez
Jefa de la Unidad de Estadística

AUTORES
Erik Carl Candela Rojas
Cristian Dominico Centeno Guzmán
Severo Alfredo Aquino Baldeón

DISEÑO Y DIAGRAMACIÓN
Franco Martínez Monge

AGRADECIMIENTOS
Annie Chumpitaz Torres – Consultora del Banco Mundial
Ciro Avitabile – Economista senior del Banco Mundial
Mauricio Romero – Consultor del Banco Mundial
Pablo Augusto Lavado Padilla – Investigador del Centro de Investigación de la Universidad del Pacífico -CIUP
IPA Perú

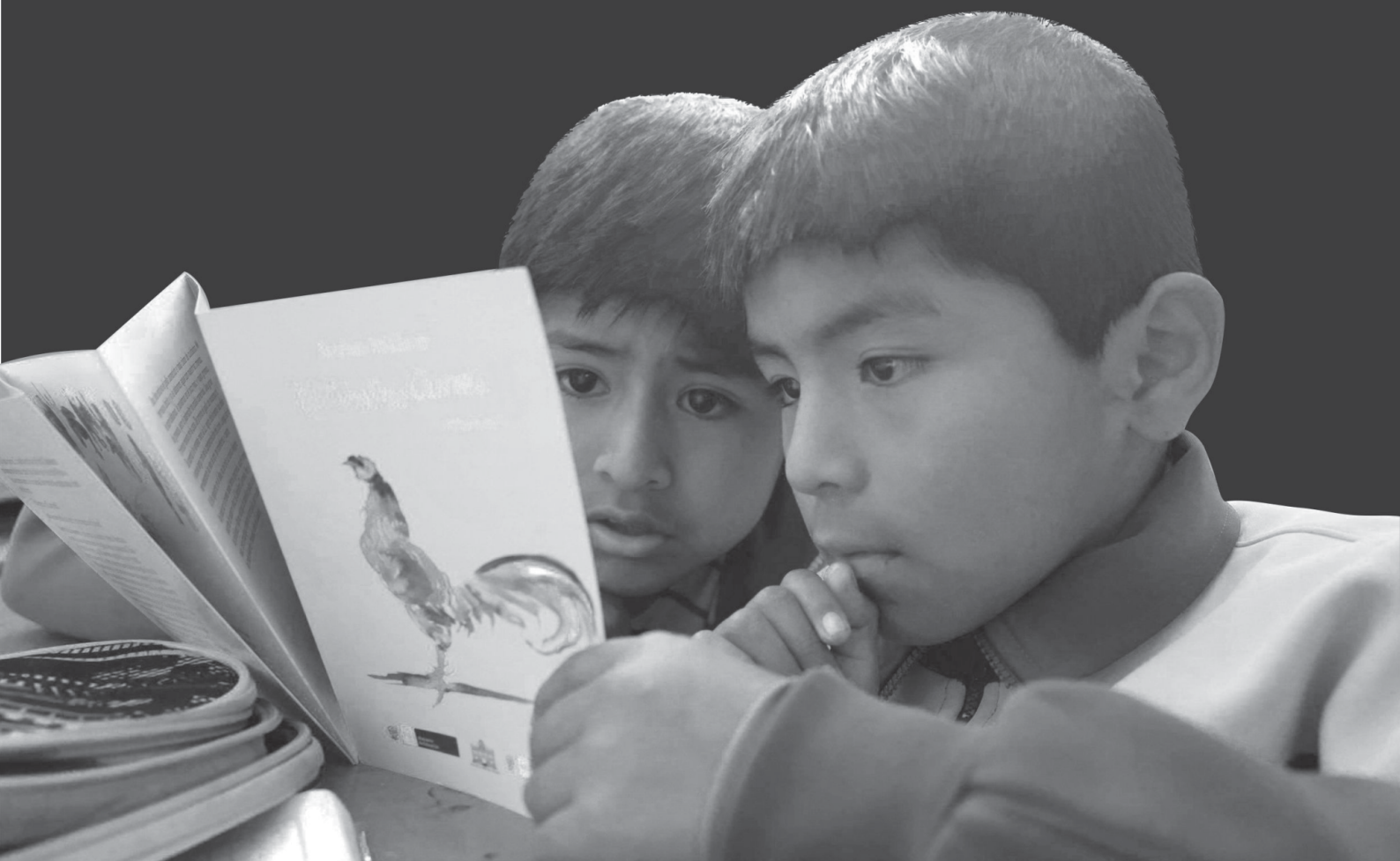
Primera edición digital, octubre 2024
Hecho el Depósito Legal en la Biblioteca Nacional del Perú N.° 2024-11568
ISBN: 978-9972-246-90-6

PERÚ Ministerio de Educación

ALERTA
ESCUELA

ALERTA ESCUELA:
METODOLOGÍA PARA EL CÁLCULO DEL RIESGO DE DESERCIÓN INTERANUAL EN EL PERÚ CON MACHINE LEARNING.

Identificando riesgos de interrupción de estudios mediante métodos predictivos.



ALERTA ESCUELA: Metodología para el cálculo del riesgo de deserción interanual en el Perú con *machine learning*.

Identificando riesgos de interrupción de estudios mediante métodos predictivos

Oficina de Seguimiento y Evaluación Estratégica[1]

1. Se extiende el presente agradecimiento a Ciro Avitabile, Mauricio Romero y Pablo Lavado que contribuyeron brindando asistencia técnica. Asimismo, se externa el agradecimiento al equipo de IPA Perú que contribuyó en las discusiones de revisión del diseño inicial del modelo y a Annie Chumpitaz por sus valiosos comentarios. .

RESUMEN

El presente documento describe la metodología utilizada para desarrollar un modelo basado en técnicas de Machine Learning (ML) que calcula el riesgo de deserción interanual que tienen los estudiantes matriculados en Educación Básica Regular (EBR) para un determinado año en el Perú. Para el desarrollo del modelo se empleó principalmente datos administrativos del Ministerio de Educación, los cuales evidenciaron su gran potencial para el desarrollo del modelo ML. De este modo se desarrolló un modelo ML que logra resultados satisfactorios en cuanto a la precisión y sensibilidad para los niveles de inicial, primaria y secundaria de EBR. Finalmente, se detalla cómo estos resultados se integran en la gestión educativa, a través del sistema «Alerta Escuela».

Palabras claves: deserción interanual, Machine Learning, Alerta Escuela

ABSTRACT

Inglés

This report describes the methodology used to develop a Machine Learning (ML) model that estimates the inter-annual dropout risk of students enrolled in Regular Basic Education (EBR) for a given school year in Peru. Administrative data from the Ministry of education was the main source of information used to estimate the model, which also highlights its potential for machine learning model development. In this way, an ML model was developed that achieves satisfactory results in terms of precision and sensitivity for the pre-primary, primary and secondary EBR levels. Finally, it details how these results are integrated into educational management, through the «Alerta Escuela» System.

Keywords: Dropout, Machine Learning, Alerta Escuela

ALERTA ESCUELA

ÍNDICE DE CONTENIDO

RESUMEN 6
ABSTRACT 7
ÍNDICE DE CONTENIDO 9
ÍNDICE DE TABLAS 11
ÍNDICE DE FIGURAS 12
INTRODUCCIÓN 13

CAPÍTULO I: COMPRENSIÓN DEL PROBLEMA 16
1.1 Objetivo 16
1.2 Evaluación de la situación inicial 16
1.3 Determinación de los objetivos de minería de datos 17
1.4 Revisión de la literatura 17

CAPÍTULO II: COMPRENSIÓN DE LOS DATOS 22
2.1 Recopilación de datos 22
2.2 Descripción de datos 23
2.3 Exploración de los datos 23
2.3.1 Evolución de la deserción interanual por nivel 24
2.4 Verificación de calidad de los datos 28
2.4.1 Balanceo de datos 28
2.4.2 Outliers 28

CAPÍTULO III: PREPARACIÓN DE LOS DATOS 29
3.1 Selección de variables 29
3.2 Limpieza de los datos 30
3.2.1 Eliminar observaciones duplicadas 30
3.2.2 Filtrar valores atípicos no deseados 31
3.2.3 Corregir errores estructurales 31
3.2.4 Corregir los datos faltantes 31
3.2.5 Validar datos 31
3.3 Construcción de nuevos datos derivados. 32
3.4 Integración de los datos 32
3.5 Formato de datos 32

ÍNDICE DE CONTENIDO

CAPÍTULO IV: MODELADO 33
4.1 Selección de técnicas de modelado 33
4.1.1 Tipo de *Machine Learning* 33
4.1.2 Algoritmos de *Machine Learning* 34
4.1.3 Métricas de desempeño 35
4.2 Generación de un diseño de comprobación 35
4.3 Generación de los modelos 36
4.3.1 Selección del algoritmo de *Machine Learning* 36
4.3.2 Configuración del Modelo: 37
4.3.3 Estimación del Modelo 38
4.4 Evaluación del modelo 47
4.4.1 Evaluación de resultados del nivel Inicial 48
4.4.2 Evaluación de resultados del nivel Primaria 50
4.4.3 Evaluación de resultados del nivel Secundaria 53

CAPÍTULO V: DESPLIEGUE 56

CONCLUSIONES 60

OPORTUNIDADES PARA LA MEJORA 61

BIBLIOGRAFÍA 63

ANEXOS 67
ANEXO 1: Machine Learning (ML) 67
ANEXO 2: Diccionario de datos 68
ANEXO 3: Macro regiones 70
ANEXO 4: Criterios de selección de métricas de desempeño 71
ANEXO 5: Cálculo de métricas 73
ANEXO 6: Criterios para la división de datos en entrenamiento y validación 75
ANEXO 7: División de datos en entrenamiento y validación 77
ANEXO 8: Hiperparámetros 80
ANEXO 9: Métricas por grado y macro región 83

ÍNDICE DE TABLAS

Tabla 1 Fuentes de información utilizada para la metodología 22
Tabla 2 Grupos de variables 23
Tabla 3 Niveles y Grados 25
Tabla 4 Macro regiones 26
Tabla 5 Tasa de Deserción Interanual (%) por grado y macro región -2022-2023 27
Tabla 6 Resultado de evaluación mediante validación cruzada con 10 iteraciones 36
Tabla 7 Total de Estudiantes por grado y macro región, 2022 37
Tabla 8 Validación cruzada con 10 iteraciones para cada nivel educativo 47
Tabla 9 Métricas con validación cruzada de 10 iteraciones – Nivel Inicial 48
Tabla 10 Métricas con validación cruzada de 10 iteraciones – Nivel Primaria 51
Tabla 11 Métricas con validación cruzada de 10 iteraciones – Nivel Secundaria..... 53
Tabla 12 Diccionario de datos 68
Tabla 13 Matriz de confusión 73
Tabla 14 Validación cruzada con 10 iteraciones para el Nivel Inicial 77
Tabla 15 Validación cruzada con 10 iteraciones para el Nivel Primaria 78
Tabla 16 Validación cruzada con 10 iteraciones para el Nivel Secundaria 79
Tabla 17 Hiperparámetros para configuración automática 80
Tabla 18 Hiperparámetros para configuración manual 80
Tabla 19 Métricas con VC de 10 iteraciones – Ciclo 2 del nivel Inicial y macro región 83
Tabla 20 Métricas con VC de 10 iteraciones – 1° grado de primaria y macro región 86
Tabla 21 Métricas con VC de 10 iteraciones – 2° grado de primaria y macro región 89
Tabla 22 Métricas con VC de 10 iteraciones – 3° grado de primaria y macro región 92
Tabla 23 Métricas con VC de 10 iteraciones – 4° grado de primaria y macro región 95
Tabla 24 Métricas con VC de 10 iteraciones – 5° grado de primaria y macro región 98
Tabla 25 Métricas con VC de 10 iteraciones – 6° grado de primaria y macro región 101
Tabla 26 Métricas con VC de 10 iteraciones – 1° grado de secundaria y macro región 104
Tabla 27 Métricas con VC de 10 iteraciones – 2° grado de secundaria y macro región 107
Tabla 28 Métricas con VC de 10 iteraciones – 3° grado de secundaria y macro región 110
Tabla 29 Métricas con VC de 10 iteraciones – 4° grado de secundaria y macro región 113
Tabla 30 Métricas con VC de 10 iteraciones – 5° grado de secundaria y macro región 116

ÍNDICE DE FIGURAS

Figura 1 Evolución de la deserción interanual 25
Figura 2 Evolución de la deserción interanual por grado 26
Figura 3 Tasa de la deserción interanual por grado y macro región – 2020-2021 27
Figura 4 Importancia de Variables – Nivel Inicial 39
Figura 5 Importancia de Variables – Primaria 42
Figura 6 Importancia de Variables – Secundaria 46
Figura 7 Curva ROC - Inicial (10 iteraciones) 49
Figura 8 Curva PR – Inicial (10 iteraciones) 50
Figura 9 Curva ROC – Primaria (10 iteraciones) 52
Figura 10 Curva PR - Primaria (10 iteraciones) 52
Figura 11 Curva ROC - Secundaria (10 iteraciones) 54
Figura 12 Curva PR - Secundaria (10 iteraciones) 54
Figura 13 Despliegue de los resultados en el sistema «Alerta Escuela» 57
Figura 14 Momento de envío de resultados hacia el sistema «Alerta Escuela» 58
Figura 15 Modelo de aprendizaje supervisado 67
Figura 16 Macro regiones 70
Figura 17 Curva Receiver Operating Characteristic (ROC) 74
Figura 18 Curva Precisión Recall (PR) 74
Figura 19 Validación cruzada de 10 iteraciones 76

INTRODUCCIÓN

La deserción escolar representa un desafío significativo para los sistemas educativos a nivel mundial, tal como señala el informe de la OCDE (2020). Los estudiantes que abandonan sus estudios enfrentan mayores barreras para ingresar y mantenerse en el mercado laboral, debido a la falta de conocimientos y habilidades que la educación proporciona para su desarrollo personal y participación cívica. Además, la interrupción de los estudios retrasa la graduación de los estudiantes, lo que a su vez demora su ingreso al mercado laboral y el inicio de su contribución económica a la sociedad.

Frente al desafío de la deserción escolar, los sistemas de alerta temprana se presentan como una respuesta efectiva para mitigar este problema. Instituciones multilaterales, como el Banco Mundial y el Banco Interamericano de Desarrollo (BID), han destacado la importancia de estas herramientas para identificar a los estudiantes en riesgo de abandonar sus estudios. Molina et al. (2024) mencionan que los sistemas de alerta temprana impulsados por inteligencia artificial permiten una intervención más rápida y precisa, lo que subraya la relevancia de utilizar técnicas avanzadas de *Machine Learning*[2] para analizar grandes volúmenes de datos y detectar patrones de riesgo de manera efectiva en el ámbito educativo. De manera similar, Giambruno et al. (2024), en su estudio sobre la educación en la región amazónica, hacen referencia al uso de sistemas de alerta temprana como un ejemplo exitoso de cómo la inteligencia artificial puede anticipar el riesgo de deserción escolar, ayudando a mitigar uno de los principales desafíos educativos en áreas remotas y con altos índices de abandono escolar. Ambos informes resaltan la necesidad de aprovechar el potencial de la inteligencia artificial para aumentar la retención de estudiantes.

Desde el Ministerio de Educación, se ha implementado desde el año 2020 el sistema de alerta temprana conocido como Alerta Escuela, como parte de las estrategias para combatir la deserción escolar. Este sistema emplea algoritmos de *Machine Learning* para calcular el riesgo de deserción interanual que tiene un estudiante matriculado en cierto año de no continuar sus estudios en el siguiente. Alerta Escuela permite generar alertas oportunas para identificar a posibles estudiantes en riesgo, facilitando la adopción de intervenciones personalizadas para garantizar su permanencia en el sistema educativo. Los niveles de riesgo estimados pueden ser puestos a disposición a los distintos actores del sistema educativo mediante los sistemas de Alerta Temprana, con el objetivo de establecer acciones y estrategias preventivas para evitar que el riesgo se materialice. (Arias Ortiz, et al., 2021)..

El presente documento tiene como objetivo detallar la metodología que se empleó para el cálculo del riesgo de deserción interanual de los estu-

2. Ver Anexo 1 «Machine Learning (ML)».

INTRODUCCIÓN

diantes de Educación Básica Regular (EBR) del Perú. En ese sentido, se describen las acciones empleadas para el desarrollo del modelo ML que calcula dicho riesgo, las métricas de rendimiento obtenidas y la integración de los riesgos en el sistema «Alerta Escuela».

El esquema de trabajo empleado para el desarrollo del modelo fue *Cross-Industry Standard Process for Data Mining* (CRISP-DM), el cual es una metodología ampliamente usada para el desarrollo de proyectos de minería de datos, ciencia de datos e inteligencia artificial. En ese sentido, la estructura del presente documento está basada en las fases que sigue esta metodología, que van desde la comprensión del problema hasta la puesta en marcha en la gestión.

El primer capítulo, titulado «Comprensión del problema», contiene los objetivos que busca la gestión educativa, así como el objetivo analítico de minería de datos, una evaluación de la situación actual y la revisión de la literatura vinculada al tema de investigación. En el segundo capítulo, «Comprensión de los datos», se describe el proceso de recopilación, descripción y exploración de los datos, así como la verificación de su calidad. Luego, en el tercer capítulo se aprecia la «Preparación de los datos», donde se detallan técnicamente las estrategias empleadas para la selección, limpieza, construcción, integración, formateo y generación de los datos analizados. Posteriormente, en el cuarto capítulo titulado «Modelado», se describe la selección de técnicas de modelado, la generación de un diseño de comprobación, así como la generación y evaluación de resultados del modelo estimado. En el quinto capítulo se describe la integración de los resultados en el sistema «Alerta Escuela». Por último, se presentan las conclusiones y recomendaciones.

ALERTA
ESCUELA

CAPÍTULO I: COMPRENSIÓN DEL PROBLEMA

1.1 Objetivo

El Ministerio de Educación busca promover la continuidad educativa de los estudiantes en la educación básica con énfasis en la población más vulnerable. Para ello, es importante poder identificar estudiantes con mayor riesgo de deserción interanual de forma preventiva con la finalidad de que los directores puedan prevenir que se retiren del sistema educativo.

1.2 Evaluación de la situación inicial

A inicios del 2020, el Ministerio de Educación no contaba con una herramienta implementada en la gestión que permita identificar estudiantes con mayor riesgo de deserción interanual a nivel nominal. Al ser el primer proyecto que emplearía técnicas de Machine Learning se procedió a evaluar la disponibilidad de los recursos de minería de datos.

Por un lado, se pudo identificar al Sistema de Información de apoyo a la Gestión de la Institución Educativa (SIAGIE)[3] como primera fuente de información para iniciar el proceso de análisis ya que cuenta con información demográfica y académica de los estudiantes de las instituciones de educación básica. Por otro lado, la OSEE contaba con expertos en técnicas de minería de datos para desarrollar el análisis correspondiente.

Un aspecto importante que se determinó desde el inicio fue que, para efectos del trabajo realizado, un estudiante de EBR -inicial, primaria o secundaria- que deserta de manera interanual será aquel que se encuentra matriculado en el año T y no se matriculará en el año T+1, excluyendo a aquellos que en el año T fallecieron o aprobaron el 5° grado de secundaria. Es decir, se utiliza la misma definición de deserción interanual empleada por el Ministerio de Educación (MINEDU, 2021).

Es relevante resaltar que la interrupción de estudios también puede ocurrir durante el año en curso T; sin embargo, escapa del alcance del modelado descrito en este documento.

3. El SIAGIE es administrado por la Unidad de Estadística (UE) de la Oficina de Seguimiento y Evaluación Estratégica (OSEE).

1.3 Determinación de los objetivos de minería de datos

El objetivo planteado es la creación de un modelo de ML utilizando datos administrativos del sector educación para calcular el riesgo de deserción interanual de cada estudiante de educación básica regular.

1.4 Revisión de la literatura

Existen diversas investigaciones que han analizado la interrupción de estudios de estudiantes de colegios públicos y privados. Sobre la revisión de la literatura realizada, se analizaron dos tipos de investigaciones, enfocadas a los factores asociados de la interrupción de estudios y a las metodologías para su predicción.

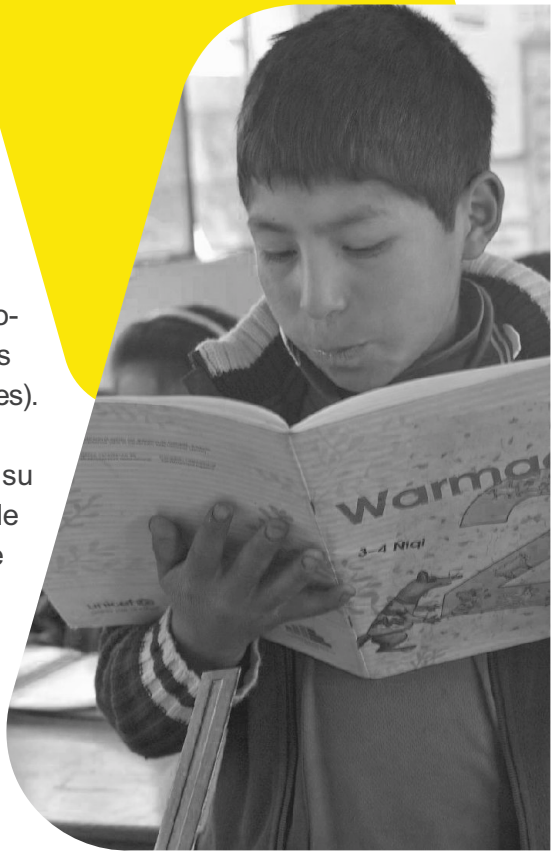
Entre las investigaciones orientadas a identificar los factores asociados con la interrupción de estudios, resaltan las siguientes:

- El primer grupo está conformado por las características del estudiante, tales como el rendimiento académico, el comportamiento del estudiante (compromiso con su aprendizaje, desviación social y situación laboral), sus actitudes (expectativas educativas y autopercepciones) y su background (datos demográficos, salud y experiencias pasadas).

- El segundo grupo corresponde a las características de la familia, tales como su estructura (si la familia está completa, su tamaño, la situación laboral de la madre, etc.), sus recursos (información económica y nivel académico de los padres) y las prácticas familiares (expectativas de los padres, prácticas de crianza, hermanos desertores).

- El tercer grupo está relacionado a la escuela, donde destaca su composición de estudiantes, su estructura (lugar, tamaño, tipo de gestión, etc.), sus recursos (ratio de alumnos-docentes y calidad de docentes) y sus prácticas escolares (relación de estudiantes y docentes).

- Por último, está el grupo de factores relacionados a la comunidad, como el porcentaje de desempleo, el porcentaje de pobreza, el ingreso promedio, las desventajas del lugar, familias encabezadas por mujeres, entre otros.



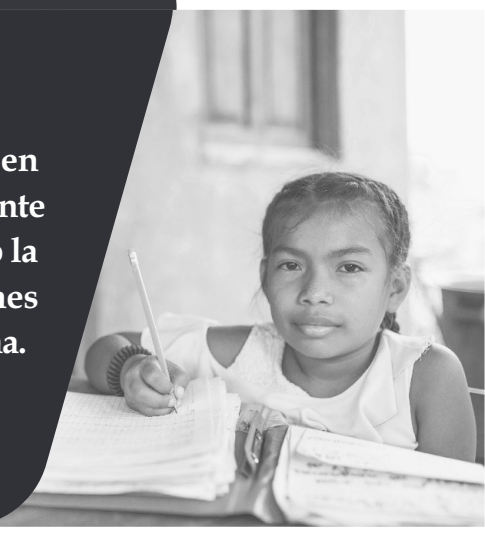
Sin embargo, es importante contextualizar los factores asociados en el ámbito nacional. Por este motivo se analizó el estudio de Jacoby (1994) sobre las restricciones crediticias y el progreso a través de la escuela en el Perú. Los resultados del autor sugieren que los estudiantes de nivel primaria que pertenecen a hogares con menores ingresos se retiran del colegio prematuramente.

En esa misma línea, Lavado y Gallegos (2005) analizaron la dinámica de la deserción escolar en el Perú a lo largo del ciclo escolar. Para dicho análisis, emplearon modelos de duración y tablas de supervivencia. Los investigadores también analizaron el efecto que tiene un programa de transferencia de dinero sobre la interrupción de estudios. Como parte de los resultados, se enfatiza la preponderancia del factor económico para la continuidad de los estudios en las zonas rurales de la sierra y la selva. También encontraron que la falta de la oferta educativa es un determinante del ausentismo y abandono en zonas rurales. Además, pudieron determinar que la mayor probabilidad de interrupción se encuentra en el primer año de secundaria. Por último, los autores señalan que el programa de transferencias puede tener un efecto positivo para combatir la interrupción de estudios.

Asimismo, Alcázar (2008) realizó un estudio sobre la asistencia y deserción en escuelas secundarias rurales del Perú. La autora pudo confirmar que los factores que originan la interrupción de estudios están asociados a la pobreza, el trabajo, la valorización de los estudios, la precariedad de las relaciones afectivas dentro del hogar, el historial educativo del estudiante (repeticiones previas o problemas de rendimiento), la percepción del estudiante sobre la calidad educativa, el costo de oportunidad de estudiar, las relaciones que existe entre los miembros del hogar y la maternidad temprana.

Igualmente, el estudio de Cueto et al. (2020) analiza los factores que están asociados a la deserción escolar en el Perú. Señalan que la necesidad de trabajar, la falta de interés en sus estudios, la lengua materna indígena, el bajo rendimiento y el haber repetido de grado son factores que están relacionados con la interrupción de estudios. Asimismo, mencionan que, si el abandono ocurre más temprano, mayor será el efecto en sus habilidades en matemáticas cuando cumplan 19 años.

> **La interrupción de estudios en escuelas rurales está fuertemente influenciada por factores como la pobreza, el trabajo, las relaciones familiares y maternidad temprana.**



Por otro lado, se cuenta con investigaciones enfocadas al desarrollo de técnicas o metodologías para predecir la deserción escolar. A continuación se describen las investigaciones internacionales[4] que se analizaron para el presente documento:

En primer lugar, se cuenta con el estudio realizado por Adelman et al. (2017) quienes estimaron modelos de alerta temprana sobre deserción escolar a partir de datos administrativos del Ministerio de Educación de Guatemala y la Secretaría de Educación de Honduras. Los investigadores emplearon variables a nivel de individuo, familia y escuela para estimar modelos de probabilidad lineal (MPL). Los autores señalan que los modelos estimados y descritos en su investigación pueden identificar el 80% de los estudiantes de 6to grado de primaria que van a desertar en su transición a educación secundaria.

Asimismo, Sansone (2017) realizó un estudio para predecir la deserción escolar empleando información del noveno grado. El investigador enfatiza que el uso de predictores que emplean técnicas de *Big Data* y *Machine Learning* mejora la detección de estudiantes que abandonarán la escuela. Los algoritmos ML evaluados fueron *Support Vector Machine(SVM), Boosted Regression y Post-LASSO*. Como parte de los resultados obtenidos, Sansone identifica que el GPA en noveno grado es el predictor que más influencia tiene para realizar la predicción.

Bianchi et al. (2019) desarrollaron un modelo ML que les permitió identificar posibles estudiantes que interrumpen sus estudios en las escuelas secundarias públicas de la provincia de Buenos Aires para el año 2018. Para ello, los autores emplearon la información de la carga inicial del sistema de información de educación «Mis Alumnos» para entrenar el algoritmo CatBoost. Los autores concluyeron que a través de ML se podían hacer predicciones razonables que podrían mejorar si se emplea mayor información.

4. A la fecha de presentación del presente documento, no se encontraron otros estudios que aborden el desarrollo de metodologías para calcular el riesgo de deserción escolar de educación básica regular a nivel de estudiante en el Perú.

Rodríguez et al. (2023) elaboraron un marco de trabajo para estimar predicciones de deserción en los sistemas escolares. En su estudio, desarrollaron un modelo de *Machine Learning* basado en datos administrativos del sistema educativo chileno, que utiliza variables a nivel individual, familiar y escolar, como el rendimiento académico, la asistencia y el estatus socioeconómico. Los resultados obtenidos indican que el modelo tiene una capacidad predictiva un 20% mayor en comparación con investigaciones previas, resaltando la importancia de factores no individuales en la predicción del abandono escolar.

A su vez, Berniell et al. (2023) presentaron el diseño de un Sistema de Alerta Temprana (SAT) para identificar a los estudiantes en riesgo de abandonar la escuela. Utilizando datos del Sistema de Gestión Educativa de Mendoza (GEM), este SAT emplea variables como inasistencias, calificaciones, nivel socioeconómico y características familiares. A través de un modelo predictivo basado en inteligencia artificial (CatBoost), el sistema permite estimar el riesgo de deserción escolar. Los resultados iniciales muestran que este enfoque mejora la capacidad de identificar estudiantes en riesgo y permite focalizar las intervenciones para reducir el abandono escolar en los niveles más críticos.

Por último, Psyridou et al. (2024) abordaron la predicción de la deserción escolar en la educación secundaria superior. En su estudio, utilizaron datos longitudinales recolectados durante 13 años, desde el jardín de infantes hasta el noveno grado, para predecir el abandono escolar. Las variables analizadas incluyeron habilidades académicas y cognitivas de los estudiantes, su motivación, comportamiento, bienestar y contexto familiar. Aplicaron modelos de ML, como el Bosque Aleatorio Balanceado, para clasificar a los estudiantes en riesgo de desertar. Los resultados indicaron que con datos hasta el sexto grado se alcanzó un área bajo la curva (AUC) de 0.61, mientras que al incluir datos hasta el noveno grado, el AUC mejoró a 0.65, lo que demuestra el potencial de estos modelos para identificar a los estudiantes en riesgo desde etapas tempranas.


ALERTA
ESCUELA

CAPÍTULO II: COMPRENSIÓN DE LOS DATOS

2.1 Recopilación de datos

Se ha trabajado con 6 fuentes de información.

En primer lugar, se cuenta con la base de datos del Sistema de Información de Apoyo a la Gestión de la Institución Educativa (SIAGIE), el cual es la principal fuente de información de los estudiantes de EBR. En segundo lugar, se accedió a la base de datos del portal web de Estadística de la Calidad Educativa (ESCALE) que provee información de distintas variables relacionadas a servicios educativos. Asimismo, a través de la Evaluación Censal de Estudiantes (ECE), se pudo extraer información relacionada a la evaluación censal de estudiantes. Con la información del Sistema de Administración y Control de Plazas (NEXUS) se pudo extraer información relacionada a los docentes del servicio educativo. También se cuenta con información del Programa JUNTOS para conocer información de cumplimiento de la corresponsabilidad de matrícula y asistencia escolar. Finalmente, se emplearon diversos indicadores generados por la Unidad de Estadística (UE), como las proyecciones de ingresos de los hogares de cada estudiante, el gradiente de ruralidad, entre otros. Todas las fuentes de información se encuentran descritas en la tabla 1.

Tabla 1: Fuentes de información utilizada para la metodología

FUENTE	DESCRIPCIÓN
SIAGIE	Datos de matrícula y evaluaciones de los estudiantes de EBR de los años 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022 y 2023.
ESCALE	Información de servicios educativos de los años 2019 y 2022.
ECE	Evaluación Censal de estudiantes del año 2018
NEXUS	Información de control de plazas de los años 2015, 2016, 2017, 2018, 2019, 2020, 2021 y 2022.
JUNTOS	Verificación de cumplimiento de corresponsabilidad de los estudiantes de los años 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021 y 2022.
UE	Se incorporan diversos indicadores educativos generados por la Unidad de Estadística, como la proyección de ingresos de los hogares de los estudiantes, la clasificación de los servicios educativos según su gradiente de ruralidad, los tiempos de desplazamiento en minutos desde el servicio educativo hasta la capital departamental, provincial o distrital, así como los tiempos de traslado hacia la sede de la UGEL o DRE/GRE.

Nota. Esta tabla contiene el listado total de todas las fuentes de información empleadas para estimar el modelo de ML.

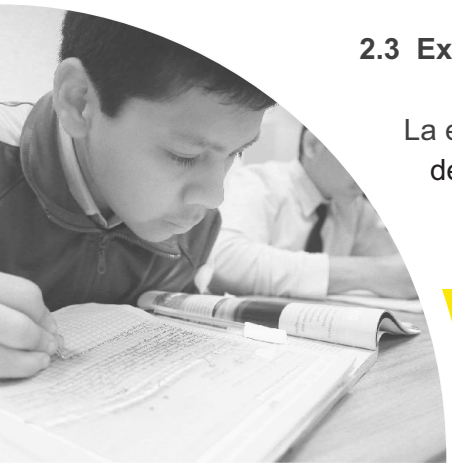
2.2 Descripción de datos

A partir de la recopilación de las fuentes de información se obtuvieron diversas variables, las cuales se clasificaron en 5 grupos. Cada uno de los grupos tienen una base teórica, las cuales se detallan en la tabla 2.

Tabla 2: Grupos de variables

GRUPO	DESCRIPCIÓN
Información propia del estudiante[a]	Dentro de este grupo se encuentran variables socio demográficas del estudiante, tales como: edad, sexo, lengua materna, nacionalidad, entre otros.
Información de contexto familiar[b]	Alberga variables sobre estructura y miembros del hogar del estudiante, tales como: grado de instrucción del apoderado, sexo del apoderado, entre otros.
Información contexto de servicio educativo[c]	Contiene variables que caracterizan al servicio educativo, tales como: tipo de gestión, ruralidad, entre otros.
Desempeño académico del estudiante[d]	Contiene variables sobre el rendimiento académico del estudiante, tales como: situación académica previa del estudiante, número de desviaciones estándares de la nota del estudiante en dichas áreas con respecto al promedio del aula, interrupciones previas de estudio, entre otros.
Información económica y de contexto[e]	Contiene variables de índole económico, tales como: proyección de ingresos del hogar del estudiante, participación en el Programa JUNTOS, costos previos de matrícula, entre otros.

Nota. La lista completa de todas las variables de cada grupo se encuentra disponibles en el Anexo 2 «Diccionario de datos».
a. Rumberger y Lim (2008), Cueto et al. (2020)
b. Rumberger y Lim (2008), Cueto et al. (2020)
c. Rumberger y Lim (2008), Adelman et al. (2017)
d. Rumberger y Lim (2008), Alcázar (2008), Sansone (2017) y Cueto et al. (2020)
e. Jacoby (1994), Lavado y Gallegos (2005), Rumberger y Lim (2008), Alcázar (2008)



2.3 Exploración de datos

La exploración de los datos se realizó en función a la variable que representa la deserción interanual. Como se vio en el punto 1.2, la deserción interanual se calcula empleando el mismo criterio utilizado por MINEDU (2021).

El MINEDU clasifica la interrupción de estudios en dos tipos: permanente e interanual. La presente investigación solo se ocupó de analizar la deserción interanual.

Por consiguiente, el concepto de deserción interanual será el siguiente:

> «… Alumnos matriculados en un determinado nivel de Educación Básica Regular (EBR) -inicial, primaria o secundaria- en el año t, que no volvieron a ser matriculados en EBR en el año t+1, excluyendo a aquellos que en el año t fallecieron o aprobaron el 5° grado de secundaria»
>
> (MINEDU, 2021).

Por lo anterior señalado, se representará el valor de la deserción interanual bajo la siguiente notación:

$$\textbf{Deserción interanual} = [\, M_t - (M_t \cap M_{t+1}) - A5S_t - F_t \,]$$

M_t = Matriculados en el año t

$M_t \cap M_{t+1}$ = Matriculados en el año t y t +1

$A5S_t$ = Alumnos que aprobaron el quinto año de secundaria t y t +1

F_t = Alumnos que fallecieron en el año t

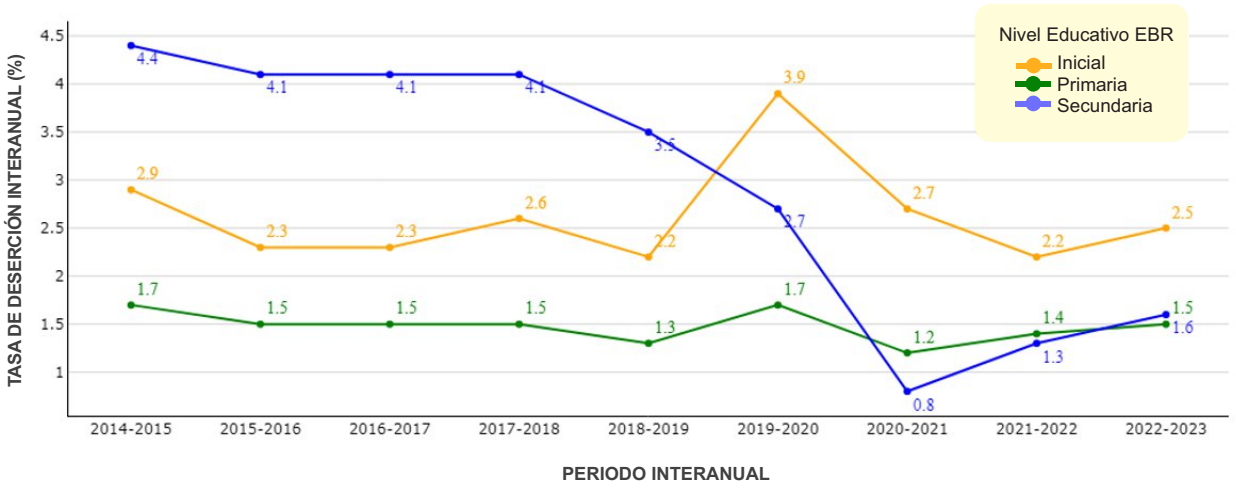
Y se define la tasa de deserción interanual como:

$$\textbf{Tasa de deserción interanual} = \frac{\textbf{Deserción interanual}}{M_t}$$

2.3.1 Evolución de la deserción interanual por nivel

La figura 1 muestra las tendencias de la tasa de deserción interanual por cada nivel educativo de EBR. Se puede observar cómo varía esta tendencia en cada período interanual «t - t+1», que corresponde a los estudiantes matriculados en el año t que no continuaron sus estudios en el año t+1.

Figura 1: Evolución de la tasa de deserción interanual



Nota. El gráfico representa la evolución de la tasa de deserción interanual por nivel educativo EBR.

Un aspecto importante a resaltar de la figura 1 es que las tasas de deserción interanual de los niveles de inicial y primaria del periodo interanual 2019-2020 se incrementaron, muy probablemente debido a las consecuencias de la emergencia sanitaria del covid-19 (Cueto, Felipe, & León, 2020).

Asimismo, el análisis también se realizó por cada uno de los grados de los niveles educativos de EBR detallados en la tabla 3, a excepción del ciclo 1 (0 a 2 años) del nivel inicial ya que los estudiantes en dicho grado cuentan con poca información histórica.

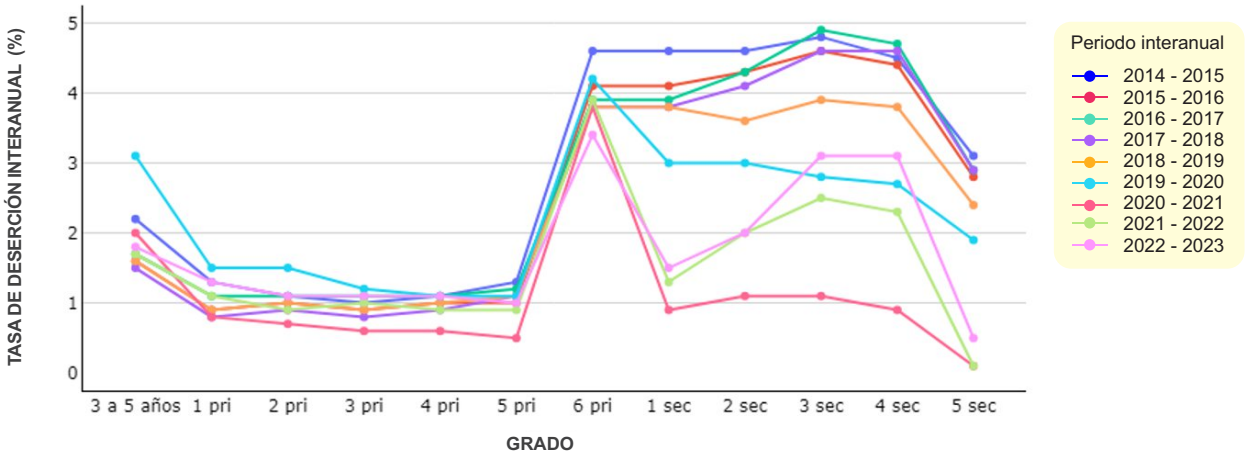
Tabla 3: Niveles y grados

NIVEL	GRADOS
Inicial[a]	Ciclo 2 (3 a 5 años)
Primaria	1ero, 2do, 3ro, 4to, 5to y 6to
Secundaria	1ero, 2do, 3ro, 4to y 5to

Nota. Esta tabla muestra los grados que conforman cada nivel educativo EBR.
a. No se incluye el ciclo 1 (0 a 2 años) del nivel inicial.

En base a la tabla 3, se analizó la tasa de deserción interanual en los distintos periodos interanuales comprendidos desde 2014-2015 hasta 2022-2023 como se muestra en la figura 2, donde se evidencia diferencias muy marcadas, principalmente en la transición del nivel de primaria al nivel de secundaria.

Figura 2: Evolución de la tasa de deserción interanual por grado



Nota. El gráfico representa la evolución histórica de la tasa de deserción interanual por cada grado de educación básica.

Entre los distintos periodos interanuales se tomó el periodo interanual 2022-2023[5] como referencia para analizar la deserción interanual a mayor detalle, por ser el último periodo con información de matrícula registrada al 100% en el SIAGIE. Además, para el análisis del periodo 2022-2023 se incorporó la variable de macro región, la cual es una agrupación de regiones del Perú como se detalla en la tabla 4.

Tabla 4: Macro regiones

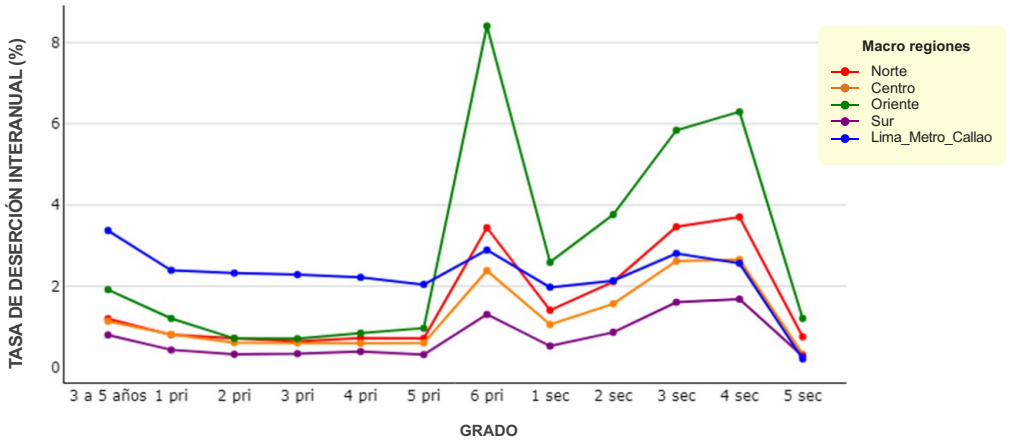
MACRO REGIÓN	REGIONES
Lima_Metro_Callao	Lima Metropolitana y Callao.
Norte	Cajamarca, La Libertad, Lambayeque, Piura y Tumbes.
Sur	Arequipa, Apurímac, Cusco, Madre de Dios, Moquegua, Puno y Tacna.
Centro	Áncash, Lima Provincias, Ayacucho, Huancavelica, Huánuco, Junín, Pasco y Ica.
Oriente	Amazonas, Loreto, San Martin y Ucayali.

Nota. Los criterios de agrupación se encuentran en el Anexo 3 «Macro regiones».

5. Estudiantes matriculados en 2022 que interrumpen sus estudios en 2023

Tomando en cuenta lo descrito en la tabla 3 y 4, se pudo desagregar la tasa de deserción interanual por grados y macro regiones para el periodo interanual 2022-2023, como se muestra en la figura 3 y tabla 5. Bajo este enfoque, se muestra que la macro región oriente ha tenido la mayor tasa en 6to grado de primaria y el nivel de secundaria, mientras que en el nivel primaria la macro región «lima_metro_callao» es la que resalta.

Figura 3: Tasa de deserción interanual por grado y macro región 2022-2023



Nota. El gráfico representa la tasa de deserción interanual por cada grado de EBR y macro región del periodo interanual 2022-2023.

Tabla 5: Tasa de Deserción Interanual (%) por grado y macro región 2022-2023

Grado	Norte	Centro	Oriente	Sur	Lima_metro y Callao
CICLO II	1.19	1.13	1.91	0.79	3.37
1 Primaria	0.80	0.81	1.20	0.43	2.39
2 Primaria	0.71	0.61	0.71	0.32	2.32
3 Primaria	0.64	0.59	0.70	0.33	2.28
4 Primaria	0.72	0.59	0.84	0.39	2.21
5 Primaria	0.71	0.60	0.96	0.31	2.04
6 Primaria	3.44	2.38	8.41	1.30	2.89
1 Secundaria	1.40	1.05	2.59	0.52	1.97
2 Secundaria	2.11	1.56	3.76	0.86	2.13
3 Secundaria	3.46	2.61	5.84	1.60	2.80
4 Secundaria	3.70	2.65	6.30	1.68	2.56
5 Secundaria	0.75	0.32	1.20	0.27	0.20

Nota. Esta tabla muestra la tasa de deserción interanual por cada grado de educación básica y macro región del periodo interanual 2022-2023.

Contar con datos de la tasa alta de deserción interanual contribuye con la identificación de tendencias y patrones.

2.4 Verificación de calidad de los datos

2.4.1 Balanceo de datos

En relación al punto 2.3, se puede evidenciar que la proporción de estudiantes que desertaron interanualmente es significativamente menor[6] al comparar el total de estudiantes matriculados en 2022 con aquellos que no se matricularon en 2023. Esto trae como consecuencia que existan pocos casos para analizar la deserción interanual.

Los datos desbalanceados se refieren a aquellos tipos de conjuntos de datos en los que existe una distribución desigual de las observaciones, es decir, una etiqueta de una clase tiene un número muy alto de observaciones y la otra tiene un número muy bajo de observaciones. Esta distribución desigual se encuentra presente en los estudiantes que desertan de manera interanual. Para resolver este desequilibrio existen métodos que consisten en generar nuevos registros de la clase con menor participación (oversampling) o disminuir registros de la clase con mayor participación (undersampling). Sin embargo, se ha preferido que dicho desequilibrio sea resuelto por el modelo a través de la calibración de hiperparámetros, lo cual está detallado en el Anexo 8 «Hiperparámetros».

2.4.2 Outliers

Un valor atípico es un punto individual de datos que está distante de otros puntos en el conjunto de datos. Los valores atípicos pueden sesgar las tendencias de los resultados de predicción de deserción interanual. Los métodos de detección de valores atípicos son una parte importante para la calidad del modelo de *Machine Learning.* Para este caso y como primera configuración se utilizó el *undersampling* para resolver el problema de los puntos atípicos.

6. La proporción de estudiantes que desertaron de manera interanual en el periodo 2022 y 2023 es de 2.5 en nivel Inicial, 1.6 en Primaria y 1.5 en Secundaria.

CAPÍTULO III: PREPARACIÓN DE LOS DATOS

3.1 Selección de variables

El proceso de selección de variables busca seleccionar un conjunto de variables predictoras óptimas para estimar un modelo parsimonioso[7]. Este proceso se compone de cuatro etapas: En una primera etapa se descartan aquellas variables que puedan contener información futura del estudiante; en la segunda etapa se retiran variables que presentan valores que no contribuyen con la predicción al contar con un único valor o ser redundantes; en la tercera etapa se seleccionan solo las variables que tienen una contribución integral diferente a cero en la predicción; y en una cuarta etapa se descartan variables que cuentan con una contribución espuria. A continuación, se detalla cada una de estas etapas:

La primera etapa del procedimiento de selección de variables consistió en verificar que las variables no cuenten con valores que describan el futuro del estudiante. Si se requiere calcular el riesgo de deserción interanual en el año T+1, es necesario contar con toda la información necesaria del año T, año T-1 y de años anteriores. Sin embargo, para el año escolar T, no se deberán considerar variables relacionadas al resultado final, las notas u otras variables que son obtenidas al final del periodo escolar de dicho año, ya que dichas variables no se encontrarían disponibles cuando se desee generar el listado con los riesgos respectivos de los estudiantes en meses iniciales del año T.

La segunda etapa consistió en verificar que las variables no tengan valores constantes; es decir, que los valores de las variables no se concentren en su totalidad en un solo valor. Si la variable tenía el mismo valor en todos los casos, se consideró como una constante y se eliminó de la base de datos. Luego se realizó una preselección de variables, evaluando las correlaciones bivariadas para cada pareja de variables candidatas, si entre ellas la correlación bivariada era alta, se seleccionaba únicamente a aquella de mayor correlación con la variable dependiente de deserción interanual.

7. La idea de un modelo parsimonioso hace referencia a la ley de la brevedad, el cual señala que no se debería usar más variables de las necesarias.

A través de la tercera etapa se estimó un modelo de *Machine Learning (LightGBM)* para poder determinar la contribución que tiene cada variable independiente (de forma individual y en conjunto con otras variables independientes) en la predicción de la variable dependiente. El objetivo es poder seleccionar variables que puedan contribuir de forma significativa con la predicción, evitando así la incorporación de variables innecesarias que puedan sobreajustar el modelo. Para obtener un conjunto robusto de variables significativas, se realizó un proceso de validación cruzada (CV)[8] con 10 iteraciones. Como indicador de contribución se empleó la importancia de tipo *gain*[9] al momento de estimar el modelo LightGBM .Solo se seleccionaron las variables con importancia diferente a cero.

Por último, la cuarta etapa consistió en retirar variables con importancia espuria; es decir, variables que tienen una importancia significativa solo por el hecho de ser de tipo continuas o discretas con alto nivel de cardinalidad, pero cuando son puestas a prueba con datos de validación, dicha contribución no existe. Para identificar dichas variables, se analizó la diferencia resultante entre los indicadores de robustez de entrenamiento y validación, para cada variable incorporada al modelo. Si la incorporación de dicha variable genera que la diferencia sea mayor a 0.15[10] y al mismo tiempo, los indicadores de robustez de validación no mejoran, entonces dicha variable será retirada.

3.2 Limpieza de los datos

A continuación, se describen las principales estrategias empleadas para la limpieza de los datos.

3.2.1 Eliminar observaciones duplicadas

Los datos duplicados ocurren con mayor frecuencia durante el proceso de recopilación de datos de distintas fuentes. Al momento de combinar varias fuentes o tablas de información se puede duplicar registros, por lo cual fue necesario realizar la siguiente validación:

Se verificó que la variable **«ID_PERSONA»** de la base de datos del SIAGIE no cuente con valores duplicados.

8. Ver Anexo 6 «Criterios para la división de datos en entrenamiento y validación».
9. «gain» es un criterio de importancia que hace referencia a la ganancia promedio de una variable cuando es incorporada en un modelo. Para mayor detalle ver el siguiente enlace: https://eli5.readthedocs.io/en/latest/libraries/lightgbm.html
10. Número referencial basado en el estudio de Adelman et al. (2017) sobre predicción de la deserción interanual, donde la diferencia del indicador de sensibilidad de entrenamiento y validación es de ~0.12.

3.2.2 Filtrar valores atípicos no deseados

Los valores atípicos son valores inusuales en el conjunto de datos (Aggarwal, 2017). Mantenerlos o removerlos es una decisión subjetiva que dependerá de la problemática que se está analizando. En la etapa de la preparación de los datos se pudo detectar que las variables de «Edad del Estudiante», «La Proyección de Ingresos del Hogar» y «Costos relacionado con la matricula, pensión y APAFA de la institución educativa» cuentan con valores atípicos. Sin embargo, en lugar de removerlas manualmente, se optó por transferir la gestión de estos valores no deseados al proceso de modelado el cual emplea principalmente algoritmo ML basados en árboles de decisión, los cuales no son sensibles a los valores atípicos (Kotu & Deshpande, 2019).

3.2.3 Corregir errores estructurales

Los errores estructurales son casos como convenciones de nomenclatura inusuales, errores tipográficos o uso incorrecto de mayúsculas, o registros inconsistentes que generan categorías mal etiquetadas. Por ejemplo, se encontraron fechas sin formato, valores vacíos etiquetados con None, entre otros.

3.2.4 Corregir los datos faltantes

Existen diversas técnicas para poder manejar los datos faltantes, entre las más resaltantes se encuentran: descarte de instancias, adquisición de nuevos datos, imputación, entre otros (Saar-Tsechansky & Provost, 2007). Sin embargo, se decidió principalmente que los datos faltantes sean gestionados en el mismo modelado. Para las variables imputadas, como la proyección de ingresos, se generó una variable adicional que indica si la variable original fue imputada o no.

3.2.5 Validar datos

Se realizaron las siguientes acciones de verificación para garantizar que los datos estén bien estructurados y listos para el entrenamiento.

- Los conjuntos de datos tienen al menos 10 mil registros como mínimo.
- Variables con más de 90% de datos faltantes fueron descartadas.
- No se incluyó información de identificación personal.
- Los datos con frases de texto largas fueron renombradas a texto corto.

3.3 Construcción de nuevos datos derivados

Para la construcción de nuevos datos se emplearon los distintos cortes anuales disponibles para poder generar múltiples ratios estadísticos (Totales, Medias, Desviaciones estándares, Mínimos, Máximos, entre otros) a nivel de estudiante y servicio educativo. Adicionalmente, todas las variables categóricas fueron transformadas a múltiples variables dicotómicas que representan a cada categoría.

3.4 Integración de los datos

Se procedió a integrar todas las variables en una única tabla, la cual contiene todas las variables independientes (predictores) y la variable dependiente (variable dicotómica que indica si el estudiante desertó sus estudios o no).

3.5 Formato de datos

Los modelos que hacen uso de algoritmos basados en árboles ensamblados no requieren que los datos de entrenamiento estén escalados o normalizados, ya que son invariables a estas transformaciones (Chen, 2014). Por este motivo, no se aplicó ninguna transformación general a todas las variables de la nueva base de datos integrada.

CAPÍTULO IV: MODELADO

En este capítulo se detallan todos los criterios que se consideraron y los resultados obtenidos en la fase de modelado. Para ello se emplearon los datos administrativos descritos en capítulos anteriores, los cuales permitieron la estimación de un modelo robusto que calcula el riesgo de un estudiante matriculado en el año T deserte sus estudios en el año T+1. El modelo obtenido es representado a través de la siguiente notación:

$$Y_{it} = E(\text{Deserción interanual}_{it+1} \mid P_{it}, F_{it}, S_{it-1}, A_{it-1}, C_{it})$$

Donde

Y: dummy igual a 1 si el estudiante deserta de manera interanual
P: Información propia del estudiante
F: Información del contexto familiar
S: Información de contexto del servicio educativo
A: Desempeño académico del estudiante

A partir de lo señalado en el punto 2.3, se estableció el valor de T = 2022. En ese sentido, se estimó un modelo ML para calcular la probabilidad que tiene un estudiante de educación básica regular, matriculado en el año 2022 de desertar sus estudios en 2023 (no matricularse en el año T+1).

4.1 Selección de técnicas de modelado

En este punto se abordarán todos los criterios que se tomaron en cuenta para una adecuada selección de técnicas de modelado.

4.1.1 Tipo de *Machine Learning*

En primer lugar, es importante resaltar que el objetivo que busca este modelado es identificar qué estudiantes van a desertar de manera interanual. Al tratarse de un caso de clasificación (entre deserta y no deserta), se procedió a emplear el aprendizaje supervisado[11] (*Supervised Learning*), el cual es un tipo de ML empleado típicamente para clasificar una etiqueta (Géron, 2019, pág. 8).

11. Ver anexo 1 para mayor detalle.

4.1.2 Algoritmos de *Machine Learning*

Se tomaron en cuenta los algoritmos de ML que fueron empleados por diversos estudios que calculan el riesgo de deserción interanual y que fueron descritos en el punto 1.4 “Revisión de la literatura”, tales como:

- **Regresión logística (lr):** Modelo de regresión empleado tradicionalmente para predecir variables categóricas, y el efecto de cambios en sus determinantes a partir de la función denominada logit (Berkson, 1944). Este algoritmo es empleado por Adelman et al. (2017) y en diversos estudios.

- **SVM – Linear Kernel (svm):** Algoritmo que recibe de entrada vectores de características y los asigna de forma no lineal a un espacio de características de muy alta dimensión, donde se construye una superficie de decisión lineal (Cortes & Vapnik, 1995). Este algoritmo fue uno de los empleados por Sansone (2017) en su investigación.

- **CatBoost Classifier (cat):** Algoritmo que implementa un boosting ordenado e incorpora un nuevo algoritmo de procesamiento de variables categóricas (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2019). Bianchi et al. (2019) lo emplearon exclusivamente en su investigación.

Asimismo, se vio conveniente incorporar otros algoritmos que emplean métodos de *boosting*[12] en lugar de las redes neuronales artificiales, por requerir un menor tiempo en el entrenamiento de los datos y la optimización de los hiperparámetros (Al daoud, 2019).

- **Extreme Gradient Boosting (xgb):** Algoritmo que emplea un sistema de boosting de árboles escalable de extremo a extremo (Chen & Guestrin, 2016).

- **Light Gradient Boosting Machine (lgb):** Algoritmo basado en árboles de decisión que tiene entre sus principales ventajas su alta tasa de velocidad para entrenar los datos con el menor consumo de memoria requerida (Microsoft, 2017).

Por último, se agregó un clasificador *Dummy* (dum) que realiza una clasificación aleatoria el cual servirá como línea base y permitirá evidenciar si los modelos evaluados cuentan con poder predictivo.

12. El método de *boosting* es una técnica que permite entrenar secuencialmente varios modelos y combinarlos con el objetivo de tener una mejorar precisión.

4.1.3 Métricas de desempeño

La selección de una métrica de desempeño juega un papel fundamental al momento de estimar un modelo de ML. Sun et al. (2009) señalan que la selección de la métrica es importante porque puede orientar la estimación del modelo y permite realizar la evaluación de la estimación realizada. Además, para el caso de un modelo que busca realizar una clasificación binaria (deserta o no deserta sus estudios), los autores emplean el concepto de matriz de confusión para el cálculo de métricas rendimiento (Sun & Wong, 2009).

Los criterios empleados para la selección de las métricas de rendimiento del modelo ML se encuentran detallados en el Anexo 4 «Criterios de selección de métricas de desempeño», el cual señala que las métricas más adecuadas para evaluar el rendimiento del modelo son la precisión, sensibilidad, F1, curva ROC y curva PR. Adicionalmente se consideró las métricas de subcobertura y filtración para contextualizar los resultados bajo un enfoque de focalización.

Asimismo, las fórmulas de las métricas se encuentran detalladas en el Anexo 5 «Cálculo de métricas».

4.2 Generación de un diseño de comprobación

La comprobación se realizó mediante la predicción para un conjunto de datos de prueba o validación (test), los cuales no forman parte del proceso de creación del modelo. Se calcularon las métricas de Precisión, *Recall*, F1, ROC AUC, PR AUC, Subcobertura y filtración.

Se realizó una comparación entre las métricas calculadas a partir de los datos de validación y las obtenidas a partir de los datos de entrenamiento (training), donde se espera comprobar que la diferencia sea mínima, lo cual es un indicador de un bajo sobreajuste u *overfitting*[13] del modelo.

Para generalizar las métricas obtenidas, se empleó la metodología de validación cruzada o Cross Validation con 10 iteraciones en base a lo señalado en el Anexo 6 «Criterios para la división de datos en entrenamiento y validación».

Cabe precisar que para el cálculo de las métricas de rendimiento se empleó un punto de corte o umbral referencial de 0.5, es decir, si la probabilidad de desertar los estudios de un estudiante matriculado en el año T es mayor al 50 % entonces se estima que el estudiante desertará sus estudios en el año T+1.

13. Hawkins (2003) señala que el sobreajuste u *overfitting* ocurre cuando se incorpora más variables de las que son necesarias. También se da cuando se emplea enfoques más complejos de lo requerido, pudiendo afectar en las métricas de desempeño del modelo (Hawkins, 2003).

4.3 Generación de los modelos

En esta etapa, tras haber experimentado con diversos enfoques, se procede a la generación de los modelos seleccionados para su análisis y validación. Se implementan los algoritmos previamente identificados como los más prometedores, con el objetivo de evaluar su rendimiento y ajustarlos según sea necesario.

4.3.1 Selección del algoritmo de *Machine Learning*

Para la selección del algoritmo, se está tomando como referencia un análisis previo realizado en 2022 con los siguientes datos: se tomó una muestra de 79,966 estudiantes tomados, de forma estratificada[14], de los distintos grados de educación básica y macro regiones del año 2020. Los datos contienen 966 estudiantes que no se matricularon en 2021.

Además, se calcularon[15] las métricas de rendimiento, tales como la Precisión, Sensibilidad, Especificidad, F1, PRAUC y ROCAUC, y se estableció que el algoritmo con mayor valor F1[16] será seleccionado para realizar la estimación final la cual emplea el total de los datos de estudiantes. La tabla 6 contiene los resultados obtenidos.

Tabla 6: Resultado de evaluación mediante validación cruzada con 10 iteraciones

Modelo	Precisión	Sensibilidad	Especificidad	F1	PRAUC	ROCAUC
lgb	0.044	0.359	0.904	0.078	0.023	0.736
xgb	0.051	0.141	0.968	0.075	0.018	0.706
cat	0.400	0.005	0.999	0.011	0.017	0.781
lr	0.116	0.004	1.000	0.008	0.014	0.757
svm	0.051	0.011	0.988	0.004	0.012	0.000
dum	0.000	0.000	1.000	0.000	0.012	0.500

Nota. Esta tabla muestra los resultados de la evaluación, ordenados de mayor a menor valor en base a la métrica F1.

Como se puede evidenciar en la tabla 6, se concluye que el algoritmo *Light Gradient Boosting Machine* es el mejor para el cálculo del riesgo de deserción interanual de la educación básica. Es importante señalar que estas métricas NO representan el resultado final del estudio y su único propósito es proporcionar información para seleccionar el algoritmo idóneo para el cálculo del riesgo de deserción interanual.

14. Por cada combinación de los grados de educación básica y macro regiones, se tomaron un 1% de estudiantes, respetando la proporción de estudiantes que desertan o no desertan de manera interanual.
15. Para obtener resultados confiables se empleó la validación cruzada con 10 iteraciones en base a lo descrito en el Anexo 6 del presente documento. Asimismo, las fórmulas de cada una de las métricas se encuentran especificadas en el Anexo 5.
16. Se seleccionó la métrica F1 en base a lo señalado en el punto 4.1.3 (Métricas de desempeño) del documento.

4.3.2 Configuración del Modelo:

Configuración de Hiperparámetros: Uno de los aspectos más importantes al momento de estimar el modelo es la configuración de los hiperparámetros que recibe el modelo antes del entrenamiento de los datos. Muchos de estos valores pueden ser asignados de forma manual o a través de una rutina de optimización (Probst, Boulesteix, & Bischl, 2019). Los criterios empleados para el cálculo de los hiperparámetros se encuentran especificados en el Anexo 8 «Hiperparámetros».

Configuración de los datos de entrenamiento: En lugar de emplear el total de los datos para entrenar un único modelo, se dividieron los datos en 60 muestras a partir de la combinación de los grados de EBR y las macro regiones, como se muestra en la tabla 7. Cabe señalar que no se estimará modelos relacionados al ciclo 1 (0 a 2 años) del nivel inicial en relación al punto 2.3.1.

Tabla 7: Total de Estudiantes por grado y macro región, 2022

Grado	Norte	Centro	Oriente	Sur	Lima_metro y Callao
3 a 5 años	375,646	348,958	193,137	244,844	451,499
1 Primaria	139,227	127,534	74,965	93,368	177,747
2 Primaria	143,131	131,253	74,661	96,138	178,534
3 Primaria	146,272	131,360	75,474	96,370	178,157
4 Primaria	151,818	133,287	85,801	96,804	181,673
5 Primaria	147,840	131,050	83,566	95,115	175,356
6 Primaria	145,761	130,473	79,464	95,156	172,162
1 Secundaria	140,687	129,321	74,171	95,644	171,666
2 Secundaria	132,695	126,539	66,627	91,358	160,748
3 Secundaria	125,813	118,306	63,105	86,014	156,579
4 Secundaria	122,615	115,510	60,359	85,459	152,505
5 Secundaria	114,837	113,172	53,431	84,378	149,178

Nota. Esta tabla muestra el total de estudiantes por cada grado de educación básica y macro región del año 2022. Las celdas en amarillo contienen totales menores que 100 mil mientras que las celdas verdes cuentan con totales mayores a 100 mil estudiantes.

La decisión de dividir en 60 muestras se tomó a partir de evaluaciones previas donde se evidenció que un modelo especializado por macro región y grado cuenta con una mejor métrica F1 que un modelo general que incluye diversas macro regiones y grados. Esto podría deberse a que un modelo especializado podría ver mejor los efectos específicos de la deserción interanual para las características de una determinada macro región y grado.

4.3.3 Estimación del Modelo

A partir de las 60 muestras descritas en el punto 4.3.2, se estimaron 60 modelos cuyas métricas de rendimiento fueron calculadas a través de la validación cruzada con 10 iteraciones y se encuentran disponibles en el Anexo 9 «Métricas por grado y macro región».

Sin embargo, para la descripción de los modelos estimados se emplearon los predictores o variables más importantes que contribuyeron al cálculo del riesgo de deserción interanual. El *framework LightGBM* permite describir los resultados de los modelos de ML a partir de una gráfica que muestra la contribución global de cada variable en la estimación[17], sin embargo, no puede explicar la contribución individual de cada variable para una predicción en particular.

Por esta razón, se empleó *Shapley Additive Explanations* (SHAP) el cual permite describir la contribución global y local que tiene cada variable en la estimación en general y en una predicción particular, respectivamente (Lundberg & Lee, 2017).

Para describir la contribución global, se calculó el valor absoluto medio de la contribución local (valores SHAP) de cada variable en todas las observaciones del conjunto de datos de validación. Este valor absoluto medio refleja la contribución promedio de cada variable a las predicciones del modelo, sin considerar la dirección (positiva o negativa) de su efecto. Además, se describirá la dirección del efecto de las variables más importantes en cada nivel educativo.

Cabe precisar que los valores de contribución de cada variable, que pueden variar sin un rango fijo, son relativos entre sí, donde un número más alto indica un mayor impacto en las predicciones del modelo.

17. https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.plot_importance.html

Asimismo, en lugar de tener 60 gráficas que describen los distintos modelos por grado y macro región, se optó por agrupar por nivel educativo los valores SHAP de las variables de cada modelo, obteniendo así tres gráficas que describen de forma general los predictores más importantes del nivel inicial, primaria y secundaria.

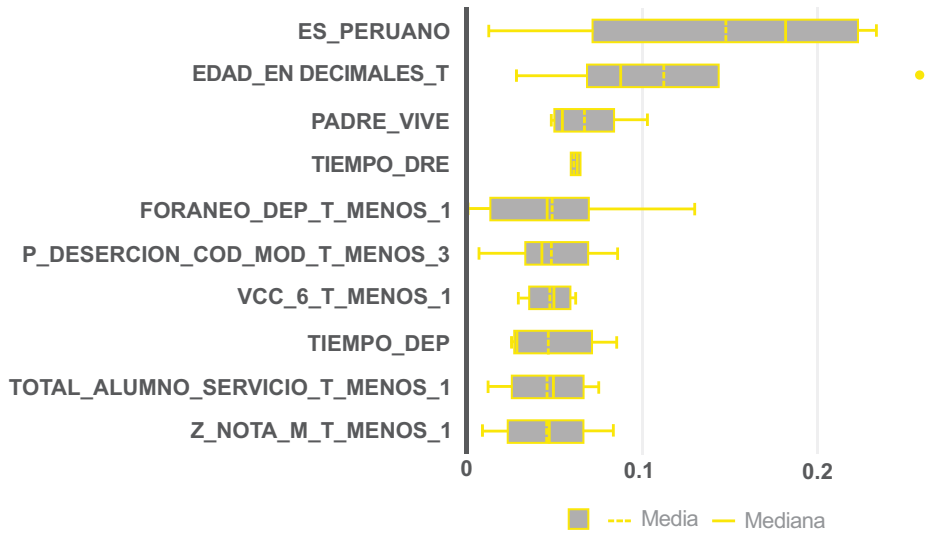
La importancia de cada variable será representada mediante un diagrama de cajas a partir de los valores SHAP. Se estableció que las variables con media y mediana mayor a 0.1 aportan significativamente.

Solo se consideraron las 10 variables más importantes por cada nivel educativo. A continuación, se describen las gráficas de importancia de cada nivel educativo:

Nivel Inicial: Como se observa en la Figura 4, la variable dicotómica que indica si un estudiante es peruano (ES_PERUANO) es la que el modelo otorga mayor peso en la predicción de la deserción interanual en el nivel educativo inicial. Le siguen la extra-edad (EDAD_EN_DECIMALES_T) y la condición de si el padre del estudiante está vivo (PADRE_VIVE).

El análisis reveló que cuando la variable ES_PERUANO toma el valor de cero, es decir, cuando el estudiante no es peruano, el modelo asigna una mayor probabilidad de deserción. Por otro lado, para la variable EDAD_EN_DECIMALES_T, el modelo otorga una mayor probabilidad de deserción a los estudiantes con mayor edad.

Figura 4: Importancia de Variables – Nivel Inicial



Nota. El gráfico representa el ranking de las 10 variables más importantes, ordenadas en forma decreciente, del nivel inicial.

En cuanto a la variable PADRE_VIVE, el modelo presenta un comportamiento mixto. En algunos casos, cuando el padre no está vivo, la probabilidad de deserción aumenta, mientras que en otros no muestra esta tendencia, lo que sugiere una relación no lineal condicionada por otras variables que emplea el modelo.

Otras variables que contribuyen al modelo son las siguientes:

- **TIEMPO_DRE y TIEMPO_DEP:** Estas variables miden el tiempo en minutos desde el centro educativo hasta la DRE/GRE y la capital departamental, respectivamente. Se encontró que, a menor tiempo de desplazamiento, mayor es la probabilidad de deserción interanual. Esto sugiere que los centros educativos más cercanos a zonas urbanas presentan un mayor riesgo de deserción en este nivel educativo. De hecho, los resultados de deserción interanual para 2022-2023 mostraron una tasa más alta en áreas urbanas (2.8%) en comparación con las áreas rurales (1.9%) para el nivel inicial.
- **FORANEO_DEP_T_MENOS_1:** Esta variable dicotómica indica si el departamento de nacimiento del estudiante es diferente al departamento donde se encontraba el centro educativo el año anterior. El modelo asigna una mayor probabilidad de deserción cuando el valor es 1, es decir, cuando el estudiante no proviene del mismo departamento en que está ubicado su servicio educativo.
- **P_DESERCION_COD_MOD_T_MENOS_3:** Esta variable representa el porcentaje de deserción interanual del servicio educativo del estudiante hace tres años. El modelo asigna una mayor probabilidad de deserción a los estudiantes que asisten a centros educativos con un mayor porcentaje de deserción interanual durante el periodo 2019-2020.
- **VCC_6_T_MENOS_1:** Esta variable dicotómica indica si el estudiante cumplió con la sexta corresponsabilidad del programa Juntos durante el año anterior. El modelo asigna una mayor probabilidad de deserción a los estudiantes que no cumplieron con dicha corresponsabilidad.
- **TOTAL_ALUMNO_SERVICIO_T_MENOS_1:** Esta variable refleja el total de alumnos en el servicio educativo durante el año anterior. El modelo asigna una mayor probabilidad de deserción a los estudiantes que pertenecen a servicios educativos con una menor cantidad de estudiantes.
- **Z_NOTA_M_T_MENOS_1:** Esta variable indica cuántas desviaciones estándar está la nota de matemáticas de un estudiante por encima o por debajo del promedio de su grado escolar en el año anterior. Aunque esta variable mejora la capacidad predictiva del modelo, su efecto en la deserción es mixto: en algunos casos, el modelo asigna una mayor probabilidad de deserción cuando el valor de la variable es menor, mientras que en otros casos esta probabilidad disminuye, lo que evidencia una relación no lineal posiblemente influenciada por otras variables.


ALERTA
ESCUELA
SONRIE. CRISTO
TE AMA

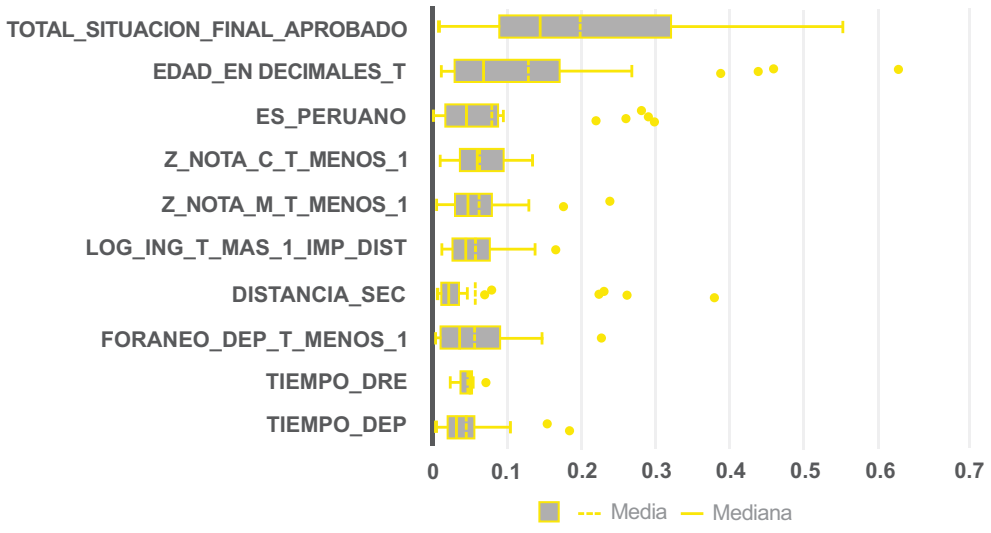
Nivel Primaria: La Figura 5 muestra que la variable con mayor peso en la predicción de la deserción interanual en el nivel primaria es TOTAL_SITUACION_FINAL_APROBADO, que indica el número total de años escolares que el estudiante ha aprobado. Le siguen, en importancia, la extra-edad (EDAD_EN_DECIMALES_T) y la variable dicotómica que indica si el estudiante es de nacionalidad peruana (ES_PERUANO).

El análisis reveló que el modelo asigna una mayor probabilidad de deserción cuando el valor de la variable TOTAL_SITUACION_FINAL_APROBADO es bajo.

Por otro lado, para la variable EDAD_EN_DECIMALES_T, el modelo asigna una mayor probabilidad de deserción a los estudiantes que tienen una mayor edad en comparación con la edad promedio de su grado escolar.

En cuanto a la variable ES_PERUANO, cuando toma el valor de cero, es decir, cuando el estudiante no es peruano, el modelo asigna una mayor probabilidad de deserción.

Figura 5: Importancia de Variables – Primaria



Nota. El gráfico representa el ranking de las 10 variables más importantes, ordenadas en forma decreciente, del nivel primaria.

Otras variables que contribuyen al modelo son las siguientes:

- **Z_NOTA_C_T_MENOS_1 y Z_NOTA_M_T_MENOS_1:** Indican cuántas desviaciones estándar está la nota de comunicación y matemáticas de un estudiante por encima o por debajo del promedio de su grado escolar en el año anterior. Ambas variables mejoran la capacidad predictiva del modelo, sin embargo, su efecto en la deserción es mixto: en algunos casos, el modelo asigna una mayor probabilidad de deserción cuando el valor de la variable es menor, mientras que en otros casos esta probabilidad disminuye, lo que sugiere una relación no lineal posiblemente influenciada por otras variables del modelo.

- **LOG_ING_T_MAS_1_IMP_DIST:** Esta variable representa la proyección de los ingresos del hogar del estudiante en una escala logarítmica. El modelo asigna una mayor probabilidad de deserción cuando el valor de la variable es bajo y no ha sido imputado con el promedio de los ingresos proyectados a nivel distrital en caso de que la variable esté nula.

- **DISTANCIA_SEC:** Esta variable representa la distancia euclidiana entre el centro educativo de nivel primaria que el estudiante asiste y el centro educativo de nivel secundaria más cercano. El modelo indica que los estudiantes que viven más lejos de la institución secundaria tienen una mayor probabilidad de deserción. Cabe destacar que este efecto es más relevante para los estudiantes de 6º grado de primaria.

- **FORANEO_DEP_T_MENOS_1:** Esta variable dicotómica indica si el departamento de nacimiento del estudiante es diferente al departamento en el que se ubicaba el servicio educativo el año anterior. El modelo asigna una mayor probabilidad de deserción cuando el valor es 1, es decir, cuando los departamentos son distintos.

- **TIEMPO_DRE y TIEMPO_DEP:** Estas variables miden el tiempo en minutos desde el servicio educativo hasta la DRE/GRE y la capital del departamento, respectivamente. El modelo muestra un comportamiento mixto en relación con estas variables: para ciertos grados escolares y macro regiones, un mayor tiempo de desplazamiento incrementa la probabilidad de deserción, mientras que en otros casos no se observa esta tendencia. Esto sugiere que el efecto de estas variables está condicionado tanto por el grado escolar como por la macro región en cuestión.

ALERTA
ESCUELA

Nivel Secundaria: Como se muestra en la Figura 6, la extra-edad (EDAD_EN_DECIMALES_T) es la variable que el modelo le da mayor peso para la predicción de la deserción interanual en el nivel educativo secundaria. Esta variable es seguida por Z_NOTA_M_T_MENOS_1, el cual indica cuántas desviaciones estándar está la nota de matemáticas de un estudiante por encima o por debajo del promedio de su grado escolar en el año anterior y TOTAL_SITUACION_FINAL_APROBADO el cual indica el número total de años escolares que el estudiante ha aprobado.

El análisis mostró que, para la variable EDAD_EN_DECIMALES_T, el modelo asigna una mayor probabilidad de deserción a los estudiantes que tienen una mayor edad en comparación con la edad promedio de su grado escolar.

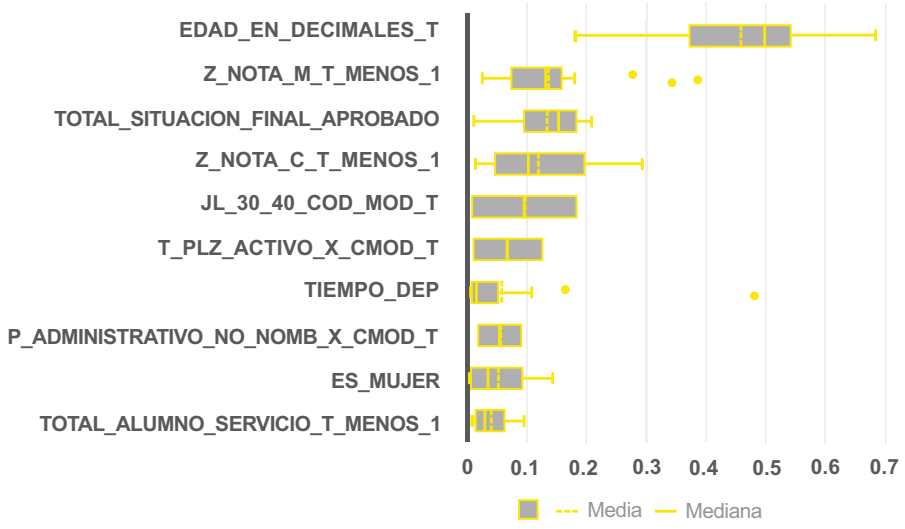
En el caso de la variable Z_NOTA_M_T_MENOS_1, el modelo exhibe un comportamiento mixto. Asigna una mayor probabilidad de deserción en diversos escenarios tanto cuando el valor de la variable es alto como cuando es bajo, lo que sugiere una relación no lineal influenciada por otras variables consideradas en el modelo.

En cuanto a la variable TOTAL_SITUACION_FINAL_APROBADO, cuando tenía un menor valor, el modelo otorga una mayor probabilidad de deserción interanual.

Otras variables que contribuyen al modelo son las siguientes:

- **Z_NOTA_C_T_MENOS_1:** Tiene un comportamiento similar al de la variable Z_NOTA_M_T_MENOS_1, pero se centra en la nota de comunicaciones.
- **JL_30_40_COD_MOD_T:** Representa la cantidad de personal en el servicio educativo que trabaja entre 30 y 40 horas a la semana. El modelo asigna una mayor probabilidad de deserción a los estudiantes cuyos servicios educativos cuentan con menos personal en esta jornada laboral.
- **T_PLZ_ACTIVO_X_CMOD_T:** Esta variable representa la cantidad de plazas activas en el servicio educativo. El modelo asigna una mayor probabilidad de deserción interanual a los estudiantes cuyos servicios educativos cuentan con un menor número de plazas activas.
- **P_ADMINISTRATIVO_NO_NOMB_X_CMOD_T:** Esta variable indica el porcentaje de personal administrativo no nombrado en el servicio educativo del estudiante. El modelo asigna una mayor probabilidad de deserción cuando este porcentaje es bajo.
- **ES_MUJER:** Esta variable representa el género del estudiante (masculino o femenino). El modelo asigna una mayor probabilidad de deserción interanual cuando el estudiante es femenino.
- **TOTAL_ALUMNO_SERVICIO_T_MENOS_1:** Esta variable refleja el total de alumnos en el servicio educativo durante el año anterior. El modelo asigna una mayor probabilidad de deserción a los estudiantes que pertenecen a servicios educativos con una menor cantidad de estudiantes.

Figura 6: Importancia de Variables – Secundaria



Nota. El gráfico representa el ranking de las 10 variables más importantes, ordenadas en forma decreciente, del nivel secundaria.

4.4 Evaluación del modelo

Como se mencionó en el punto 4.3.3, las métricas de los 60 modelos se encuentran disponibles en el Anexo 9 «Métricas por grado y macro región», las cuales fueron calculadas mediante la validación cruzada con 10 iteraciones. Sin embargo, para una representación más general de las métricas de rendimiento por nivel educativo, se optó por emplear las muestras descritas en el Anexo 7, agrupando los E_i y V_i de cada iteración de los grados y macro regiones de un nivel educativo.

De esta manera se formó los resultados de validación cruzada con 10 iteraciones por cada nivel de inicial, primaria y secundaria de la EBR, como se muestra en la tabla 8.

Tabla 8: Validación cruzada con 10 iteraciones para cada nivel educativo

NIVEL	TOTAL	P	E_i^{nivel}	$E_i^{nivel}P$	V_i^{nivel}	$V_i^{nivel}P$
Inicial	1614084	29287	1452674	26358	161410	2929
Primaria	3769517	55621	3392551	50055	376966	5566
Secundaria	2790717	56661	2511635	50995	279082	5666

Nota. Esta tabla muestra el total de estudiantes para el entrenamiento y validación del modelo de deserción interanual.

Donde E_i^{nivel} representa el total de datos de entrenamiento de la iteración i de un nivel educativo, mientras que V_i^{nivel} son los datos para la validación del modelo en la iteración i para el respectivo nivel educativo. Asimismo $E_i^{nivel}P$ representa el total de registros de la clase positiva de los datos de entrenamiento de la iteración i, mientras que $V_i^{nivel}P$ conforma el total de registros de la clase positiva de los datos de validación de la iteración i.

Además, se graficaron las curvas ROC y PR por nivel educativo. Cada una de estas gráficas cuentan con una línea base el cual representa un modelo que no cuenta con poder predictivo (aleatorio).

Cabe precisar que las tablas de métricas contienen el valor mínimo, máximo, promedio y la desviación estándar de cada métrica con el objetivo de medir la robustez del modelo.

4.4.1 Evaluación de resultados del nivel Inicial

La tabla 9 presenta la evaluación del nivel inicial, donde se evidencia que el modelo está identificando un 37% de los estudiantes que desertaron de manera interanual (subcobertura del 63%). Asimismo, la clasificación que realizó el modelo tiene una precisión del 23% (filtración del 77%).

Asimismo, se pudo observar que el modelo cuenta con poco sobreajuste[18]. Esto se comprobó al analizar la desviación estándar de cada una de las métricas, cuyo valor es muy cercano a cero. De la misma forma, se encontró que la diferencia máxima entre las métricas de entrenamiento y de validación es sólo 0.03.

Tabla 9: Métricas con validación cruzada de 10 iteraciones – Nivel Inicial

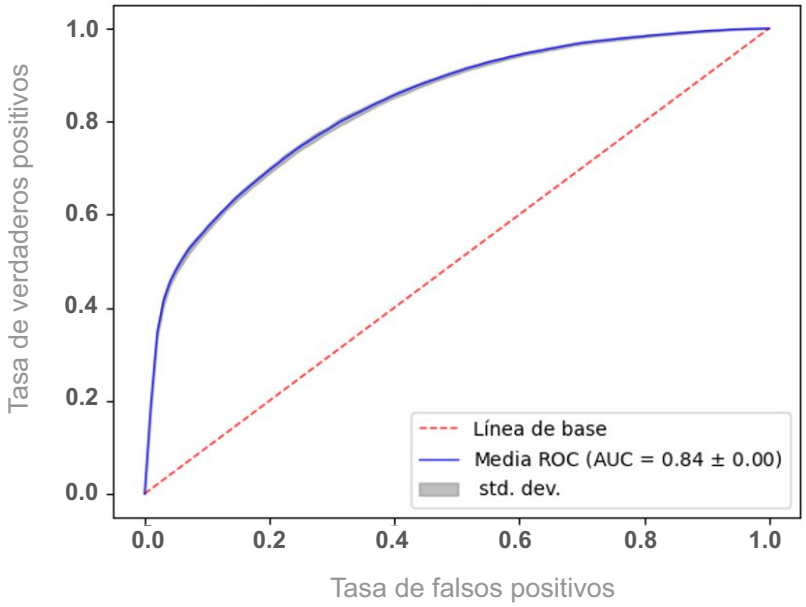
MÉTRICA	DATOS	PROMEDIO	DE	Min.	Max.
Precisión	Validación	0.23	0.01	0.22	0.23
Precisión	Entrenamiento	0.23	0.00	0.23	0.23
Sensibilidad	Validación	0.37	0.01	0.36	0.39
Sensibilidad	Entrenamiento	0.38	0.00	0.38	0.38
Especificidad	Validación	0.98	0.00	0.98	0.98
Especificidad	Entrenamiento	0.98	0.00	0.98	0.98
F1	Validación	0.28	0.01	0.27	0.29
F1	Entrenamiento	0.29	0.00	0.29	0.29
PR AUC	Validación	0.15	0.00	0.14	0.16
PR AUC	Entrenamiento	0.16	0.00	0.16	0.16
ROC AUC	Validación	0.84	0.00	0.83	0.84
ROC AUC	Entrenamiento	0.87	0.00	0.87	0.87
Filtración	Validación	0.77	0.01	0.77	0.78
Filtración	Entrenamiento	0.77	0.00	0.77	0.77
Subcobertura	Validación	0.63	0.01	0.61	0.64
Subcobertura	Entrenamiento	0.62	0.00	0.62	0.62

Nota. Esta tabla muestra las métricas obtenidas del nivel inicial mediante la validación cruzada de 10 iteraciones.

18. (Hawkins, 2003)

Adicionalmente, la figura 7 muestra la curva ROC con un AUC de 84%, el cual sugeriría que se cuenta con una capacidad predictiva muy significativa.

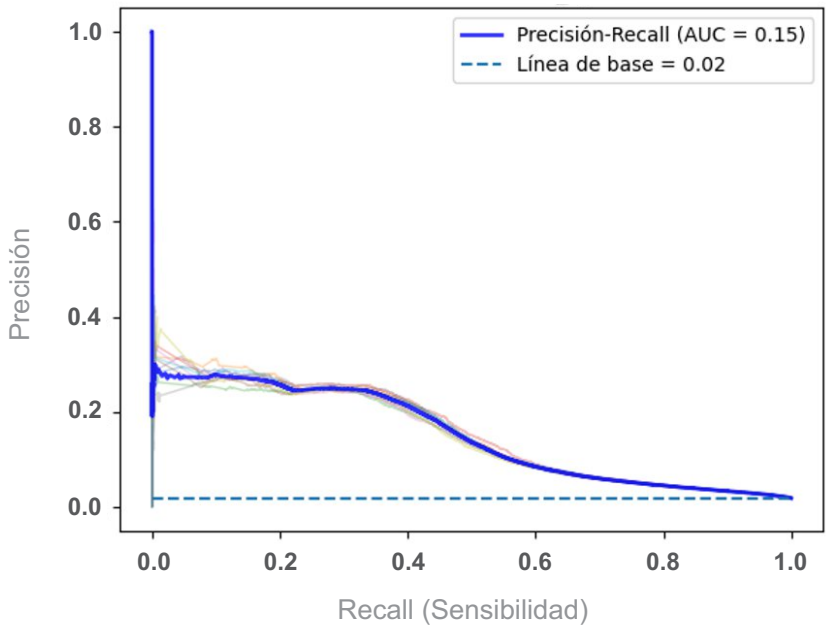
Figura 7: Curva ROC - Inicial (10 iteraciones)



Nota. El gráfico representa la curva ROC resultante a partir de la validación cruzada de 10 iteraciones del nivel inicial.

Sin embargo, la figura 8 muestra la curva PR con un AUC del 15%, principalmente por la clase positiva que se encuentra desbalanceada. Aunque el PR AUC no parece un buen resultado, es 7.5 veces mejor que su línea base.

Figura 8: Curva PR – Inicial (10 iteraciones)



Nota. El gráfico representa la curva PR resultante a partir de la validación cruzada de 10 iteraciones del nivel inicial.

4.4.2 Evaluación de resultados del nivel Primaria

La tabla 10 presenta la evaluación del nivel primaria, donde se evidencia que el modelo está identificando un 43% de los estudiantes que desertaron de manera interanual (subcobertura del 57%). Asimismo, la clasificación que realizó el modelo tiene una precisión del 22% (filtración del 78%).

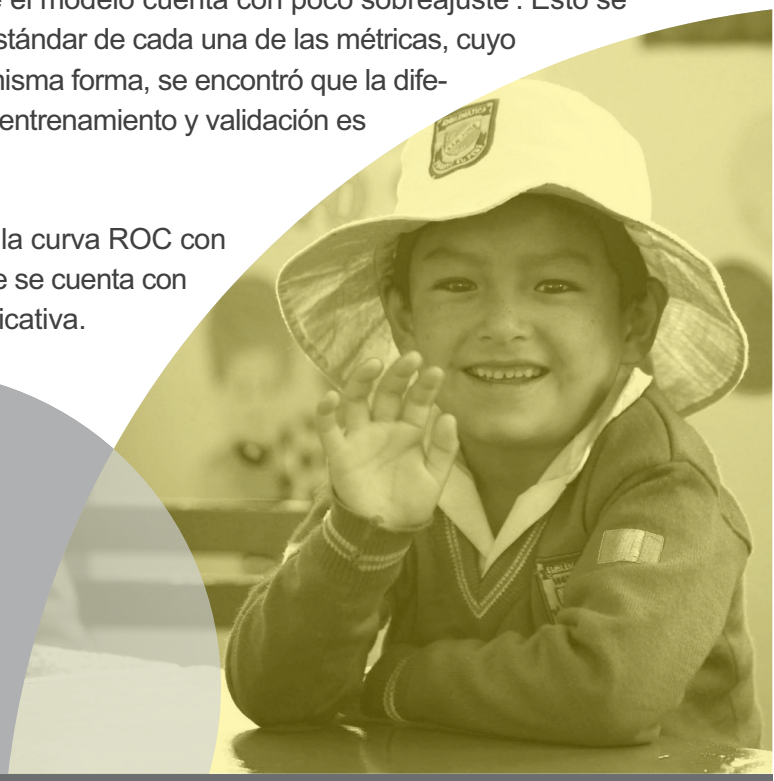
Tabla 10: Métricas con validación cruzada de 10 iteraciones – Nivel Primaria

MÉTRICA	DATOS	PROMEDIO	DE	Min.	Max.
Precisión	Validación	0.22	0.00	0.21	0.22
Precisión	Entrenamiento	0.23	0.00	0.23	0.23
Sensibilidad	Validación	0.43	0.00	0.43	0.45
Sensibilidad	Entrenamiento	0.47	0.00	0.47	0.47
Especificidad	Validación	0.98	0.00	0.98	0.98
Especificidad	Entrenamiento	0.98	0.00	0.98	0.98
F1	Validación	0.29	0.00	0.29	0.30
F1	Entrenamiento	0.31	0.00	0.31	0.31
PR AUC	Validación	0.18	0.00	0.18	0.19
PR AUC	Entrenamiento	0.22	0.00	0.22	0.22
ROC AUC	Validación	0.89	0.00	0.89	0.89
ROC AUC	Entrenamiento	0.94	0.00	0.94	0.94
Filtración	Validación	0.78	0.00	0.78	0.79
Filtración	Entrenamiento	0.77	0.00	0.77	0.77
Subcobertura	Validación	0.57	0.00	0.55	0.57
Subcobertura	Entrenamiento	0.53	0.00	0.53	0.53

Nota. Esta tabla muestra las métricas obtenidas del nivel primaria mediante la validación cruzada de 10 iteraciones.

Por otro lado, se pudo observar que el modelo cuenta con poco sobreajuste[19]. Esto se comprobó al analizar la desviación estándar de cada una de las métricas, cuyo valor es muy cercano a cero. De la misma forma, se encontró que la diferencia máxima entre las métricas de entrenamiento y validación es menor a 0.05.

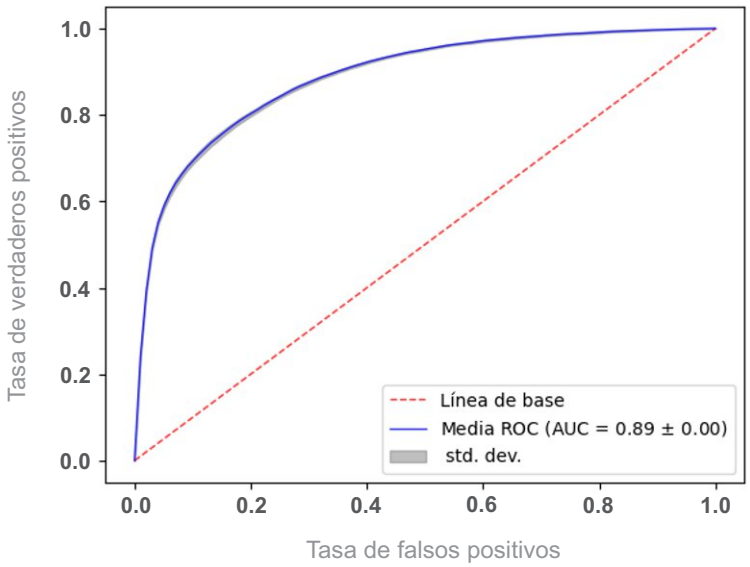
Adicionalmente, la figura 9 muestra la curva ROC con un AUC de 89%, el cual sugeriría que se cuenta con una capacidad predictiva muy significativa.



19. (Hawkins, 2003)

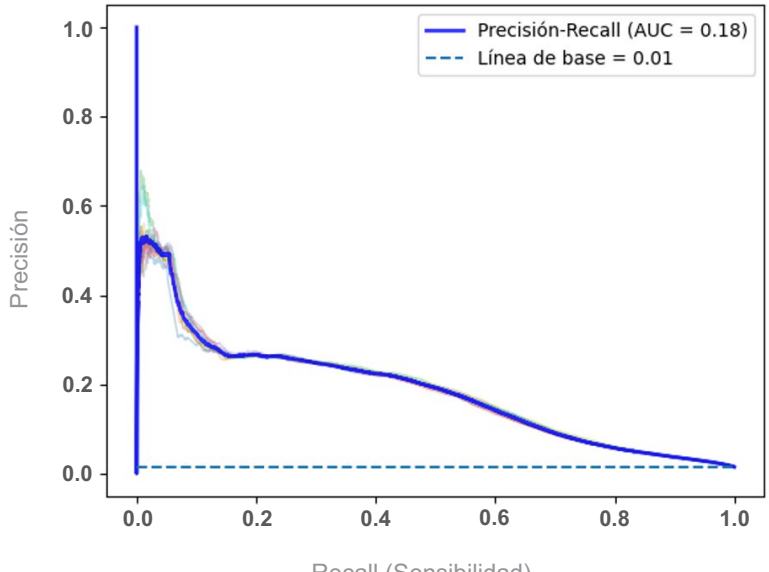
Sin embargo, la figura 10 muestra la curva PR con un AUC es de 18%, principalmente por la clase positiva que se encuentra desbalanceada. Aunque el PR AUC no parece un buen resultado, es 18 veces mejor que su línea base.

Figura 9: Curva ROC – Primaria (10 iteraciones)

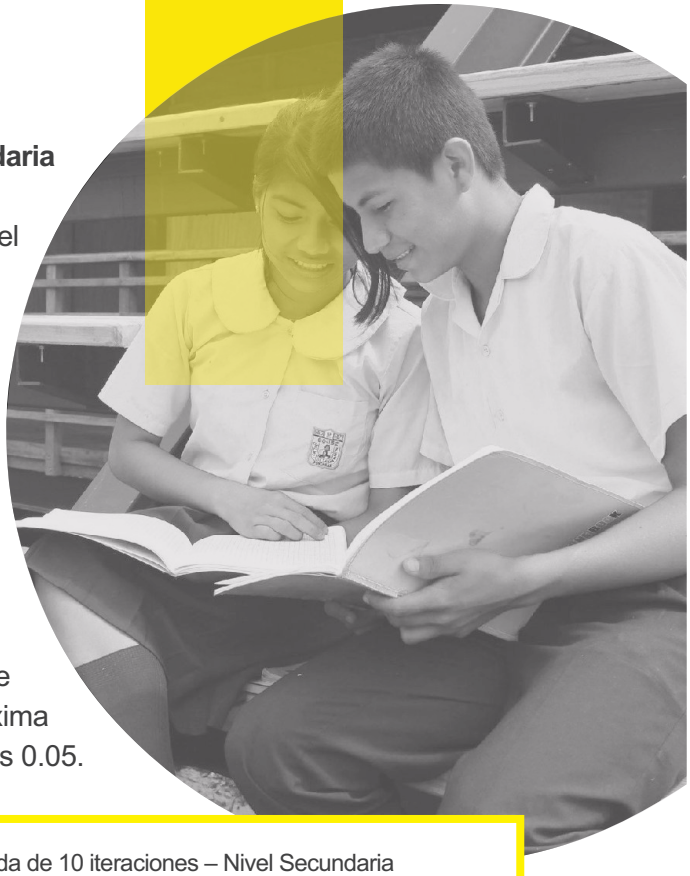


Nota. El gráfico representa la curva ROC resultante a partir de la validación cruzada de 10 iteraciones del nivel primaria.

Figura 10: Curva PR - Primaria (10 iteraciones)



Nota. El gráfico representa la curva PR resultante a partir de la validación cruzada de 10 iteraciones del nivel primaria.



4.4.3 Evaluación de resultados del nivel Secundaria

Por último, la tabla 11 presenta la evaluación del nivel secundaria, donde se evidencia que el modelo está identificando un 36% de los estudiantes que desertaron de manera interanual (subcobertura del 64%). Asimismo, la clasificación que realizó el modelo tiene una precisión del 19% (filtración del 81%).

Por otro lado, se pudo observar que el modelo cuenta con poco sobreajuste.[20] Esto se comprobó al analizar la desviación estándar de cada una de las métricas, cuyo valor es muy cercano a cero. De la misma forma, se encontró que la diferencia máxima entre las métricas de entrenamiento y validación es 0.05.

Tabla 11: Métricas con validación cruzada de 10 iteraciones – Nivel Secundaria

MÉTRICA	DATOS	PROMEDIO	DE	Min.	Max.
Precisión	Validación	0.19	0.00	0.19	0.20
Precisión	Entrenamiento	0.21	0.00	0.21	0.22
Sensibilidad	Validación	0.36	0.01	0.35	0.37
Sensibilidad	Entrenamiento	0.40	0.00	0.40	0.40
Especificidad	Validación	0.97	0.00	0.97	0.97
Especificidad	Entrenamiento	0.97	0.00	0.97	0.97
F1	Validación	0.25	0.00	0.24	0.26
F1	Entrenamiento	0.28	0.00	0.28	0.28
PR AUC	Validación	0.15	0.00	0.15	0.16
PR AUC	Entrenamiento	0.19	0.00	0.18	0.19
ROC AUC	Validación	0.87	0.00	0.87	0.88
ROC AUC	Entrenamiento	0.92	0.00	0.92	0.92
Filtración	Validación	0.81	0.00	0.80	0.81
Filtración	Entrenamiento	0.79	0.00	0.78	0.79
Subcobertura	Validación	0.64	0.01	0.63	0.65
Subcobertura	Entrenamiento	0.60	0.00	0.62	0.60

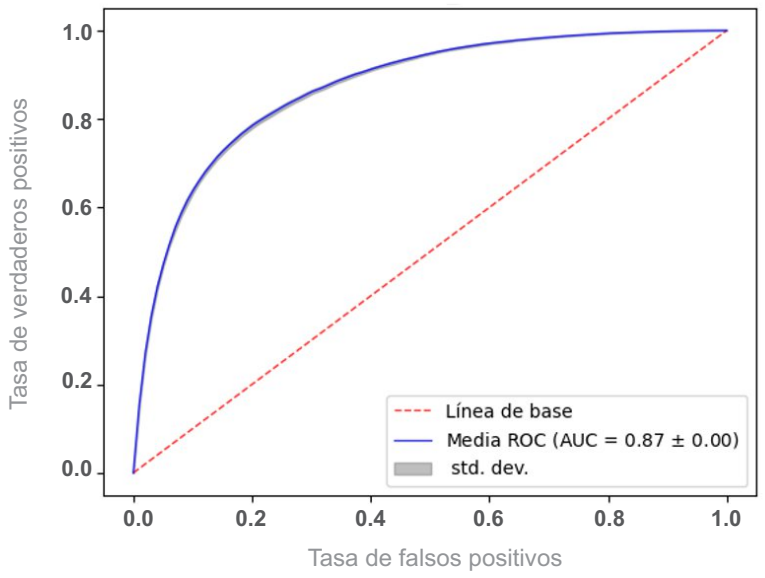
Nota. Esta tabla muestra las métricas obtenidas del nivel secundaria mediante la validación cruzada de 10 iteraciones. .

20. (Hawkins, 2003)

Adicionalmente, la figura 11 muestra la curva ROC con un AUC de 87%, el cual sugeriría que se cuenta con una capacidad predictiva muy significativa.

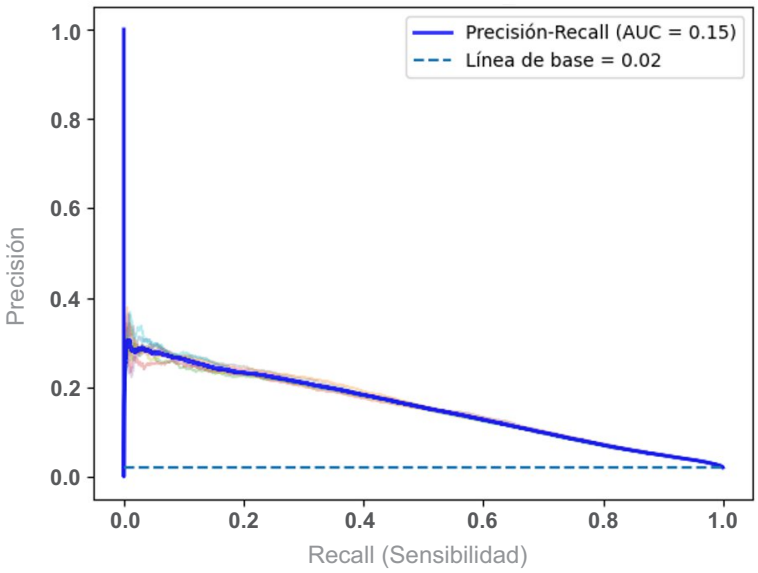
Sin embargo, la figura 12 muestra la curva PR con un AUC del 15%, principalmente por la clase positiva que se encuentra desbalanceada. Aunque el PR AUC no parece un buen resultado, es 7.5 veces mejor que su línea base.

Figura 11: Curva ROC – Secundaria (10 iteraciones)



Nota. El gráfico representa la curva ROC resultante a partir de la validación cruzada de 10 iteraciones del nivel primaria.

Figura 12: Curva PR - Secundaria (10 iteraciones)



Nota. El gráfico representa la curva PR resultante a partir de la validación cruzada de 10 iteraciones del nivel primaria.

ALERTA
ESCUELA

CAPÍTULO V: DESPLIEGUE

Los resultados fueron incorporados dentro del sistema de alerta temprana de deserción escolar Interanual[21] denominado «Alerta Escuela»[22], el cual se implementa a través de un módulo del SIAGIE que está a cargo de la Unidad de Estadística (UE) de la Oficina de Seguimiento y Evaluación Estratégica (OSEE).

Es importante resaltar que el SIAGIE, al ser un sistema ya institucionalizado, es empleado por los directores de las instituciones educativas de educación básica regular. Esto facilita el poner a disposición el sistema «Alerta Escuela» a los directores.

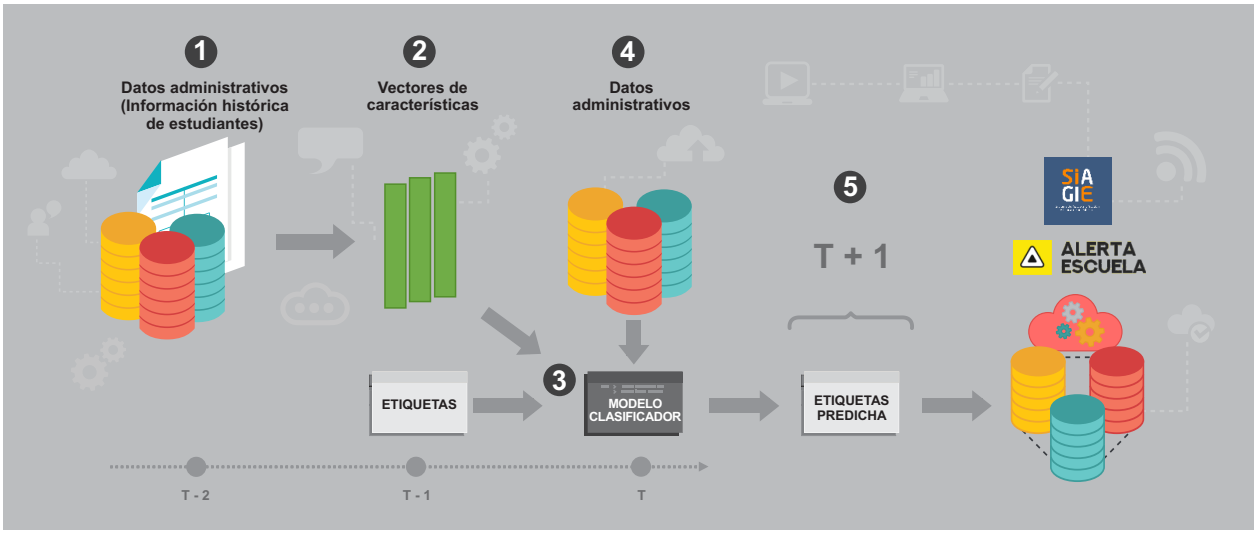
Para cada nuevo año escolar, el modelo es reentrenado con la información histórica más actualizada posible. La figura 13 muestra un flujo de trabajo sobre el despliegue de los resultados de riesgos de deserción interanual en el sistema «Alerta Escuela», el cual se detalla a continuación:

1. En el año T-2 se realiza un corte para obtener los datos de estudiantes matriculados en dicho año.
2. Para poder identificar que estudiantes desertaron interanualmente, se emplea la información del año T-1 y se aplican los criterios descritos en el punto 2.3 para poder determinar los estudiantes que interrumpen sus estudios.
3. Se emplea la información del año T-2 y T-1 para estimar el modelo que calcula el riesgo de deserción interanual.
4. El modelo es empleado para calcular el riesgo que tienen los estudiantes matriculados en el año T en interrumpir sus estudios en el año T+1.
5. El riesgo de deserción interanual de cada estudiante es exportado en una base de datos para su incorporación en el sistema «Alerta Escuela» del SIAGIE.

21. La literatura reciente ha enfatizado el rol de las alertas tempranas para disminuir la deserción interanual como un componente principal de los sistemas de protección de las trayectorias educativas, junto con las intervenciones de remediación y acompañamiento oportunas (Arias et al., 2021). (CAF, 2022) sintetiza las experiencias en uso de datos y de modelos de detección temprana de riesgo de deserción en Estados Unidos (Wisconsin), Australia (Victoria), y Argentina (Provincia de Buenos Aires, PBA). Cabe indicar que en este último se utiliza un sistema basado en el método CATBoost (Bianchi et al., 2019).

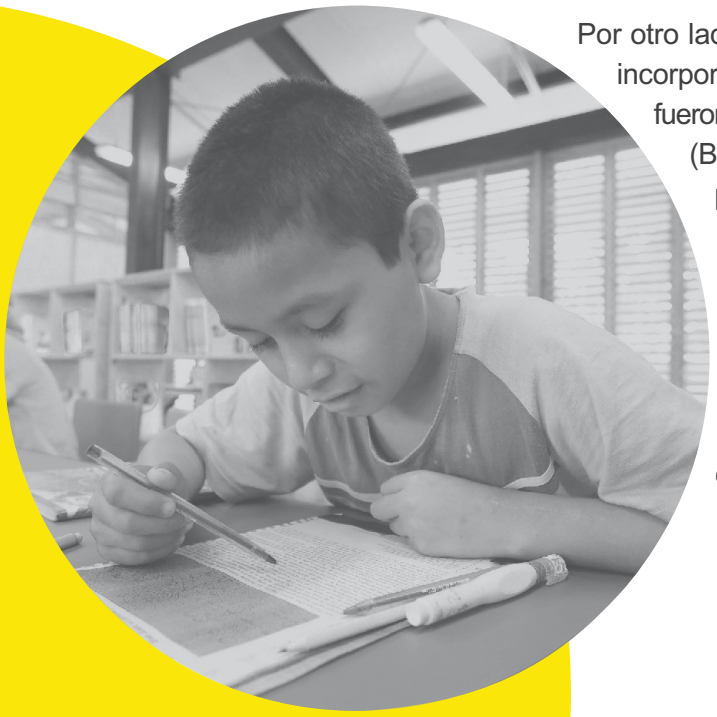
22. La información sobre este sistema está disponible en el siguiente enlace: https://alertaescuela.minedu.gob.pe/

Figura 13: Despliegue de los resultados en el sistema «Alerta Escuela»



Nota. El gráfico representa el flujo de trabajo necesario para poner a disposición los resultados de riesgo de deserción interanual en el sistema «Alerta Escuela».

Como se muestra en la figura 13, para calcular el riesgo que tienen los estudiantes matriculados en el año T en no matricularse en año T+1 fue necesario emplear el rango interanual «T-2; T-1», es decir, se debe estimar un modelo que haga uso de información del año T-2 hacia atrás y emplear la información del año T-1 para determinar si el estudiante desertó interanualmente sus estudios. No se puede emplear el periodo interanual «T-1; T» para estimar el modelo, ya que en los primeros meses del año T no se cuenta con el 100% de los datos registrados en el SIAGIE, el cual es un requisito para poder determinar si un estudiante interrumpe sus estudios en el año T.



Por otro lado, es importante señalar que los riesgos incorporados en el sistema de «Alerta Escuela», fueron categorizados previamente en tres grupos (Bajo, Medio y Alto) con el objetivo de simplificar la interpretación de la alerta por parte de los actores del sistema educativo. A diferencia del umbral 0.5 que se empleó para calcular las métricas de rendimiento, se tomaron en cuenta los siguientes criterios para seleccionar los nuevos umbrales para seleccionar los riesgos en Bajo, Medio y Alto:

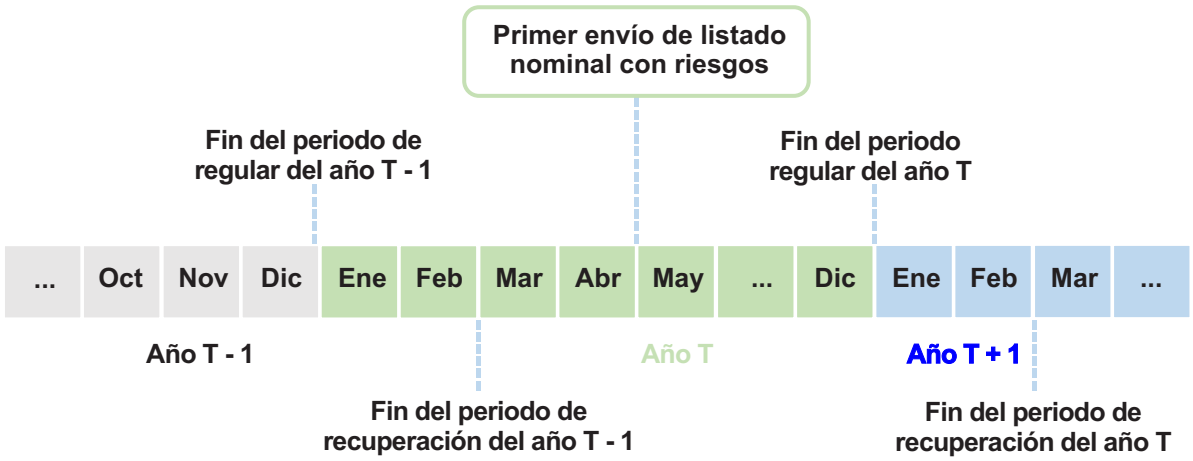
- **Riesgo Bajo:** Los estudiantes con un riesgo de deserción interanual menor a 0.5 se clasificarán como riesgo bajo.

- **Riesgo medio:** Todos los estudiantes que tengan un riesgo de deserción interanual superior a 0.5 y menor que un umbral intermedio previamente establecido serán considerados en esta categoría. La determinación exacta de este umbral se realiza en coordinación con la unidad estadística.

- **Riesgo Alto:** Este grupo incluye a todos los estudiantes cuyo riesgo supera el umbral intermedio.

Es importante señalar que la información que se registra en el SIAGIE es progresiva durante el transcurso del año escolar T, por ende, se recomienda generar las alertas de deserción interanual a partir del mes de mayo como se muestra en la figura 14. De esta manera se contará con más del 90%[23] de estudiantes registrados en el SIAGIE. Los riesgos de los estudiantes registrados después de la fecha podrán ser calculados en los meses restantes.

Figura 14: Momento de envío de resultados hacia el sistema «Alerta Escuela»



Nota. El gráfico representa el momento sugerido para el envío de información al sistema de «Alerta Escuela».

23. Para fines del mes de abril del 2022 se contaba con el ~92%de estudiantes registrados en el SIAGIE (Reporte de avance de matrícula al 24 de abril del 2022).


ALERTA
ESCUELA

CONCLUSIONES

Mediante el presente documento se describe la metodología empleada para calcular el riesgo de deserción interanual a través del uso de datos administrativos del sector educación y la aplicación de técnicas de *Machine Learning* (ML). Se tomó como marco de referencia la metodología CRISP-DM para desarrollar los distintos capítulos que contiene el presente documento.

En el capítulo I se señaló la importancia de poder identificar estudiantes vulnerables de desertar sus estudios. Para ello se creó un modelo que puede calcular el riesgo de deserción interanual. El riesgo es calculado a nivel de estudiante de EBR y puede ser empleado para el desarrollo de acciones preventivas a fin de mitigarlo.

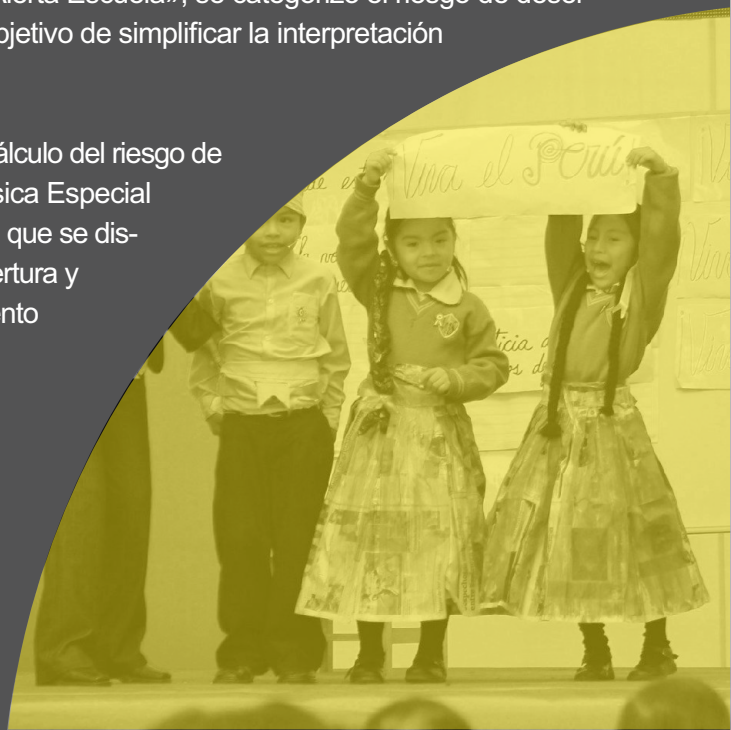
Los modelos fueron presentados por niveles Inicial, Primaria y Secundaria, describiendo las 10 principales variables más importantes por cada nivel que contribuyen en el cálculo del riesgo de deserción interanual, las cuales se han identificado a partir de la revisión de la literatura al respecto.

A través del diseño de comprobación propuesto en este documento, se verificó que los modelos para inicial, primaria y secundaria cuentan con poder predictivo para poder calcular el riesgo de deserción interanual. Asimismo, se evidenció un bajo nivel de sobreajuste.

En general, la evaluación de los resultados obtenidos por la metodología permite confirmar que la aplicación de técnicas de ML en datos administrativos del MINEDU hace posible el cálculo del riesgo de deserción interanual con niveles de precisión y de sensibilidad que varían según el grado escolar y macro región donde pertenece el estudiante. En ese sentido, se pone en evidencia el potencial que tienen los datos administrativos del MINEDU para la generación de alertas tempranas.

Para la incorporación de los resultados en el sistema de «Alerta Escuela», se categorizó el riesgo de deserción interanual en tres grupos (Alto, Medio y Bajo) con el objetivo de simplificar la interpretación de la alerta por parte de los actores del sistema educativo.

Finalmente, esta metodología puede ser empleada para el cálculo del riesgo de deserción interanual de las modalidades de Educación Básica Especial (EBE) y Educación Básica Alternativa (EBA) en la medida en que se disponga de cortes históricos de información con 100% de cobertura y una cantidad mínima de diez mil registros para el entrenamiento de modelo respectivo.

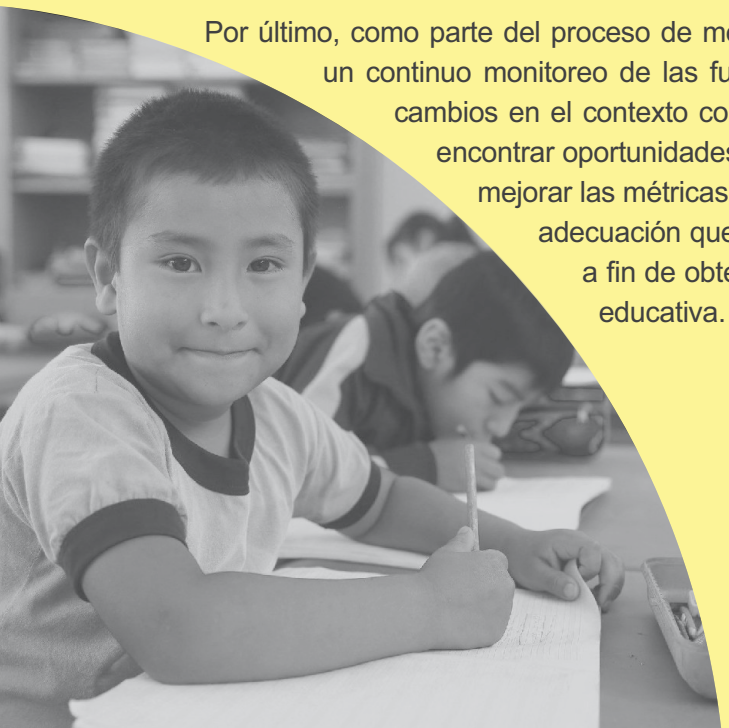


OPORTUNIDADES PARA LA MEJORA

Como futura mejora de la metodología se tiene contemplado emplear ML no supervisado para *clusterizar* a los estudiantes vulnerables identificados por el modelo a fin de poder diferenciar grupos que requieren tratamientos especializados. La literatura sugiere que uno de los beneficios sería que los formuladores de políticas podrían diseñar intervenciones especializadas por grupo. Asimismo, se podría evaluar como una política y podría afectar de forma diferente a los distintos grupos (Sansone, 2017).

Otra oportunidad de mejora está relacionada con el criterio de establecer múltiples modelos especializados por macro región y grado escolar. Se empleó este criterio para que la estimación del modelo de ML pueda visibilizar mejor los distintos factores asociados a la deserción interanual de forma más contextualizada y específica, sin embargo, esta estrategia omite posibles tendencias generales de deserción interanual que se evidenciarían al entrenar un único modelo general con todos los datos disponibles. Existe evidencia que propone la combinación de un modelo general y de múltiples modelos especializados para obtener un mejor resultado (Hinton, Vinyals, & Dean, 2015).

Por otro lado, se tiene contemplado incorporar a futuro la variable de asistencia del estudiante. Sobre la base de la literatura revisada, la asistencia del estudiante podría ser un predictor significativo para el cálculo del riesgo de deserción interanual, ya que requiere de costos implícitos para poder realizarlo (Cueto, Felipe, & León, 2020). Si bien los datos de la asistencia son registrados a través del SIAGIE y son declarativos bajo responsabilidad del director, no fueron incorporados en el análisis actual por los siguientes motivos: 1) La normativa vigente para el registro de la asistencia no establece un criterio uniforme para su registro en el SIAGIE. 2) No se cuenta con elementos para poder corroborar la veracidad del registro de la asistencia. 3) Los registros de asistencia podrían estar sesgados a favor de los estudiantes que se encuentran afiliados al Programa JUNTOS, realizados con el fin de no afectar su transferencia económica que reciben por el cumplimiento de la corresponsabilidad en educación.



Por último, como parte del proceso de mejora continua de la metodología, se recomienda realizar un continuo monitoreo de las fuentes de información, las métricas de rendimiento y los cambios en el contexto considerados en este documento. Este monitoreo permitirá encontrar oportunidades para identificar mejoras en la información a ser utilizada, mejorar las métricas de rendimiento obtenidas, así como alertar ante cualquier adecuación que requiera el modelo ML para un determinado año escolar a fin de obtener una mejor respuesta a las necesidades de la gestión educativa.

ALERTA
ESCUELA

BIBLIOGRAFÍA

Arias Ortiz, E., Giambruno, C., González Alarcón, N., Pérez Alfaro, M., Pombo, C., & Sánchez Ávalos, R. (2021). Camino hacia la inclusión educativa: 4 pasos para la construcción de sistemas de protección de trayectorias. BID.

Adelman, M., Haimovich, F., Ham, A., & Vázquez, E. (2017). Predicting School Dropout with Administrative Data. Education Global Practice Group - World Bank.

Aggarwal, C. C. (2017). An Introduction to Outlier Analysis. Outlier Analysis. doi: https://doi.org/10.1007/978-3-319-47578-3_1

Al daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. Obtenido de https://publications.waset.org/10009954/comparison-between-xgboostlightgbm-and-catboost-using-a-home-credit-dataset

Alcázar, L. (2008). Asistencia y deserción en escuelas secundarias rurales del Perú. GRADE.

Amaya, K., & Barrientos, E. H. (2014). Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos. Barranquilla: Universidad Simón Bolívar.

Berkson, J. (1944). Application of the logistic function to bio-assay. Journal of the American Statistical Association, pp. 357-365. doi:https://doi.org/10.2307/2280041

Berniell, L. L. (2023). Alertas tempranas para prevenir el abandono escolar: el caso de la provincia de Mendoza.

Bianchi, B., Pietto, M. L., & Kamienkowski, J. E. (2019). Estimación de la interrupción de las trayectorias escolares en escuelas secundarias públicas de la provincia de Buenos Aires. Programa Manos en la DATA. Universidad de Buenos Aires.

Bliss, J. (1993). The Cry-Wolf Phenomenon and its Effect on Alarm Responses. University of Central Florida.

Borra, S., & Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Computational Statistics and Data Analysis.

Breiman, L. (2001). Random Forests. Machine Learning. doi:https://doi.org/10.1023/A:1010933404324

CAF. (2018). El alto costo del abandono escolar en América Latina. Obtenido de https://www.caf.com/es/conocimiento/visiones/2018/08/el-alto-costo-delabandono-escolar-en-america-latina/

CAF. (2022). Uso estratégico de datos e inteligencia artificial en la educación. Policy Brief #5. América Latina y el Caribe. Obtenido de https://cafscioteca.azurewebsites.net/handle/123456789/1944

Cecilia Giambruno, J. C.-A. (2024). Education in the Amazon Region. Banco Interamericano de Desarrollo (BID).

Chen, T. (2014). Introduction to boosted trees. University of Washington Computer Science, 14-40. Obtenido de University of Washington Computer Science.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. ACM. doi: https://doi.org/10.1145/2939672.2939785

Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. sage journals.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning. doi: https://doi.org/10.1007/BF00994018

Cueto, S., Felipe, C., & León, J. (2020). Predictores de la deserción escolar en el Perú. Grupo de Análisis para el Desarrollo (GRADE).

Efron, B., & Tibshirani, R. (1993). An Introduction to the Bootstrap. CHAPMAN & HALL/CRC.

Elbir, A., Gündüz, E., & Diri, B. (2018). Estimating the School Dropout Trend by Using Data Mining Methods. IEEE.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow. O'REILLY.

Hawkins, D. (2003). The Problem of Overfitting. Minneapolis: School of Statistics,University of Minnesota.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arxiv.

Jacoby, H. (1994). Borrowing Constraints and Progress Through School: Evidence from Peru. The Review of Economics and Statistics.

Japkowicz, N. (2013). Assessment metrics for imbalanced learning.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Articial Intelligence . Stanford University.

Kotu, V., & Deshpande, B. (2019). Chapter 4 - Classification. En Data Science, Concepts and Practice. sciencedirect.

Lavado, P., & Gallegos, J. (2005). La dinámica de la deserción escolar en el Perú: un enfoque usando modelos de duración. Lima: Centro de Investigación de la Universidad del Pacífico.

Lundberg, S. (2020). From local explanations to global understanding with explainable AI for trees. nature machine learning.

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Long Beach, CA, USA: Conference on Neural Information Processing Systems.

Microsoft. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Microsoft Research.

MINEDU. (2019). Tendencias . Obtenido de ESCALE – Estadísticas de la Calidad Educativa.: https://escale.minedu.gob.pe/ueetendencias2016

MINEDU. (26 de abril de 2020a). Resolución Viceministerial N° 094-2020-MINEDU. Obtenido de https://www.gob.pe/institucion/minedu/normas-legales/541161094-2020-minedu

MINEDU. (10 de septiembre de 2020b). Únete a esta movilización nacional contra la deserción escolar para que peruanas y peruanos puedan cumplir sus sueños [video]. Obtenido de https://www.facebook.com/mineduperu/videos/648110396139204/

MINEDU. (9 de Diciembre de 2021). Tasa de deserción interanual. Obtenido de Tasa de deserción inter-anual: http://escale.minedu.gob.pe/tendencias-2016portlet/servlet/tendencias/archivo?idCuadro=321&tipo=meta

Molina, E. C. (2024). AI revolution in education: What you need to know. Washington, DC: International Bank for Reconstruction and Development / The World Bank.

OECD. (2020). Education at a Glance 2020: OECD Indicators. París, Francia: OECD Publishing.

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. Journal of Machine Learning Research.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). CatBoost: unbiased boosting with categorical features. ACM. doi: https://dl.acm.org/doi/abs/10.5555/3327757.3327770

Psyridou, M. P. (2024). Machine learning predicts upper secondary education dropout as early as the end of primary school. *Scientific Reports,*.

R, A. (3 de 10 de 2020). Machine Learning Explanation: Supervised Learning & Unsupervised Learning. Obtenido de medium: https://arifromadhan19.medium.com/machine-learning-explanation-supervised learning-unsupervised-learning-6d4c7f2bebb2

Rodríguez, P. V. (2023). A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. *Education and Information Technologies,* 10103-10149.

Rumberger, R., & Lim, S. (2008). Why Students Drop Out of School: A Review of 25 Years of Research. California Dropout Research Project Report.

Saar-Tsechansky, M., & Provost, F. (2007). Handling Missing Values when Applying Classification Models. Journal of Machine Learning Research.

Sansone, D. (2017). Beyond Early Warning Indicators: High School Dropout and Machine Learning. ssrn.

Sun, Y., Wong, A., & S. Kamel, M. (2009). Classification of imbalanced data: a review.

Vaarma, M. &. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. Technology in Society.

Van Domelen, J. (2007). Reaching the Poor and Vulnerable: Targeting Strategies for Social Funds and other Community-Driven Programs. World Bank.

Webb, G. (2011). Naïve Bayes. Boston: Encyclopedia of Machine Learning. Springer. doi: https://doi.org/10.1007/978-0-387-30164-8_576

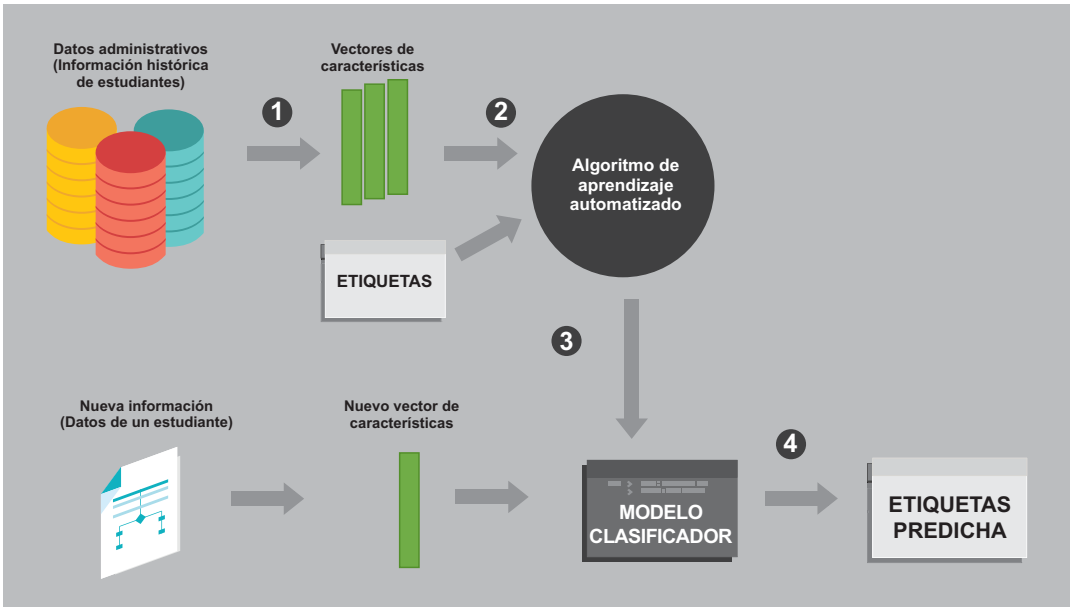
ANEXOS

ANEXO 1: Machine Learning (ML)

Machine Learning (ML) o aprendizaje automático es la ciencia (y arte) de programar las computadoras para que puedan aprender de los datos. Entre los distintos tipos de ML, el presente documento emplea el aprendizaje supervisado el cual es utilizado típicamente para clasificar una etiqueta (Géron, 2019, pág. 8). La figura 15 muestra la representación gráfica de los pasos requeridos para desarrollar un modelo supervisado:

1. Se procesa los datos administrativos para generar los vectores de características, donde cada vector representa el conjunto de características útiles que pueden explicar las etiquetas de interés.
2. A cada uno de los vectores de características se le asigna una etiqueta (deserta o no deserta sus estudios), formando así el principal insumo para entrenar el algoritmo de aprendizaje supervisado.
3. Se realiza el proceso de entrenamiento del algoritmo, dando como salida un modelo clasificador.
4. Se emplea el modelo clasificador para predecir, con cierto grado de precisión, la etiqueta de un nuevo vector de características.

Figura 15: Modelo de aprendizaje supervisado



Nota. Adaptado de Supervised Machine Learning Model [Figura], por Arif R, (2020), Medium (https://cutt.ly/zMJOk2c) .

ANEXO 2: Diccionario de datos

Tabla 12: Diccionario de datos

GRUPO	VARIABLE	FUENTE
Información propia del estudiante	Edad del estudiante.	SIAGIE
	Sexo del estudiante.	SIAGIE
	Lengua materna.	SIAGIE
	Si el estudiante tiene discapacidad.	SIAGIE
	Si el estudiante actualmente se encuentra trabajando.	SIAGIE
	Nacionalidad del estudiante.	SIAGIE
	Tipo de cercanía del lugar de nacimiento del estudiante con respecto al UBIGEO del servicio educativo: Comparación entre el lugar de nacimiento y la ubicación del servicio educativo a nivel de distrito, provincia y departamento.	SIAGIE
Información contexto familiar	Grado de instrucción del apoderado.	SIAGIE
	Años de escolaridad del apoderado.	SIAGIE
	Sexo del apoderado.	SIAGIE
	Parentesco que tiene el estudiante con el apoderado.	SIAGIE
	Si el padre o la madre viven.	SIAGIE
Información contexto del servicio educativo	Tipo de gestión del servicio educativo: Pública, Privada o Pública de gestión privada.	ESCALE
	Si el servicio educativo se encuentra en una zona rural o urbana.	ESCALE
	Si los estudiantes del servicio educativo son solo hombres, mujeres o mixto.	ESCALE
	Total de estudiantes hombres, estudiantes mujeres, secciones y docentes por servicio educativo.	ESCALE
	Cantidad promedio de Alumnos por Sección: Alumnos/Secciones.	ESCALE
	Distancia euclidiana entre el servicio de nivel inicial del estudiante al servicio de nivel primaria más cercano.	ESCALE
	Distancia euclidiana entre el servicio de nivel primaria del estudiante al servicio de nivel secundaria más cercano.	ESCALE

	Ratios de docentes, por modalidad de contrato (Nombrados, Contratados, Otros), a nivel de servicio educativo.	NEXUS
	Tiempo en minutos del servicio educativo a la capital departamental, provincial y distrital.	UE
	Tiempo en minutos del servicio educativo a la sede de UGEL y DRE/GRE.	UE
Desempeño académico del estudiante	Situación académica previa del estudiante: aprobado, desaprobado, requiere recuperación.	SIAGIE
	Desempeño previo del estudiante en matemáticas, comunicación y otras áreas: Número de desviaciones estándares de la nota del estudiante en dichas áreas con respecto al promedio del aula.	SIAGIE
	Porcentaje de estudiantes, a nivel de servicio educativo, que obtuvieron el nivel satisfactorio en matemáticas y comunicación en la prueba ECE.	ECE
	Deserción previa del estudiante: si el estudiante ha desertado en años previos.	SIAGIE
Información económica y de contexto	Proyección de ingresos del hogar: se proyectaron en función a la variación de los ingresos mensuales en 2009-2022.	UE
	Hogar focalizado por el programa JUNTOS actualmente y en periodos previos.	Juntos
	Costo previo de la matrícula de la institución educativa a la que asiste el estudiante.	SIAGIE
	Gradiente de ruralidad, que identifica niveles diferenciados al interior del ámbito rural, identificados como Rural 1, Rural 2 y Rural 3; donde Rural 1 identifica el ámbito más rural y Rural 3 identifica el ámbito menos rural.	UE

Nota. Esta tabla muestra la descripción de cada variable empleada para calcular el riesgo de deserción interanual.

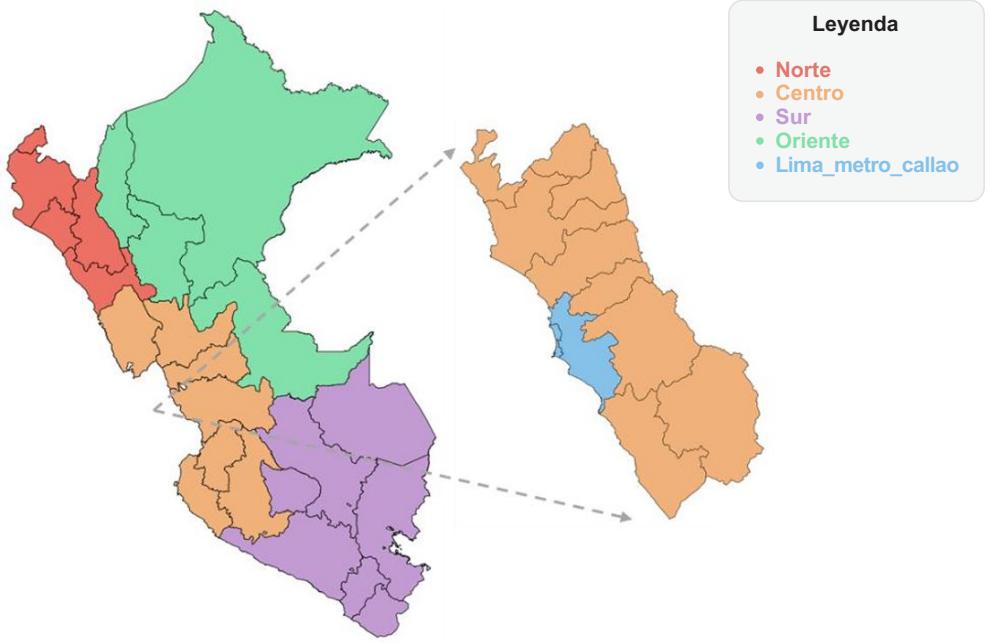
ANEXO 3: Macro regiones

Se establecieron 5 macro regiones (Norte, Sur, Centro, Oriente y Lima metro_callao), como se muestra en la figura 16. A continuación se describen los criterios empleados para la conformación de cada macro región.

Para la macro región Norte, se consideró las regiones que conforman a la macro región Nor Oriente Del Perú[24], con excepción de las regiones de Amazonas, Loreto y San Martín, por ubicarse en el lado oriente del Perú. Asimismo, para la macro región sur se tomó en cuenta todas las regiones que conforman a la Mancomunidad Regional Macrorregión Sur[25].

Por otro lado, para definir la macro región Centro, se tomó como base inicial las regiones que conforman la Mancomunidad Regional Pacífico Centro Amazónica[26], a excepción de la región de Ucayali por ubicarse en el extremo oriente. Adicionalmente, se incorporó la región de Áncash, Ayacucho e Ica por la cercanía a las regiones de la macro región Centro. No obstante, se excluyó la provincia de Lima y el Callo los cuales formaron el grupo Lima_metro_callao. Por último, para la macro región Oriente se tomó en cuenta las regiones de Amazonas, Loreto, San Martín y Ucayali.

Figura 16: Macro regiones



24. Se aprueba con Ordenanza Regional N° 006-2017-GRLL/CR, el cual incluye los departamentos de Tumbes, Amazonas, Cajamarca, Lambayeque, La Libertad, Loreto, Piura, y San Martín.
25. Se aprueba con Ordenanza N° 343-AREQUIPA, el cual incluye a los departamentos de Arequipa, Apurímac, Cusco, Madre de Dios, Moquegua, Puno, Tacna.
26. Se aprueba con Ordenanza Regional N° 229-GRJ/CR, el cual incluye a los departamentos de Lima, Huancavelica, Huánuco, Junín, Pasco, Ucayali.

ANEXO 4: Criterios de selección de métricas de desempeño

***Exactitud (Accuracy)*:** Si bien la exactitud es la métrica más empleada, Sun, Wong, & S. Kamel (2009) advierten que NO es apropiado emplearla cuando los datos de las clases están desbalanceados, ya que daría un valor muy alto principalmente por la clase predominante . De esta forma, se sustenta descartar este indicador para medir el rendimiento del modelo debido a su resultado que podría ser muy optimista. Por tal motivo, se ha optado por seleccionar otras métricas que tomen mayor consideración a la clase con menor predominancia, la cual es representada por los estudiantes que desertan sus estudios (clase positiva).

Sensibilidad y Especificidad: Inicialmente, se consideró la Sensibilidad (tasa de verdaderos positivos) y la Especificidad (tasa de verdaderos negativos). Japkowicz (2013) señala que, a diferencia de la métrica de exactitud, estas métricas pueden ser empleadas para evaluar, de forma independiente, a la clase con menor y mayor predominancia. Sin embargo, la sensibilidad y la especificidad pasan por alto los aciertos que tiene la clasificación de las observaciones que realiza el modelo hacia una determinada clase.

Precisión y *Recall*: Para abordar el aspecto faltante de la sensibilidad y especificidad, se incorporó en el análisis a la métrica de Precisión, el cual es la proporción de observaciones a los que el modelo les asignó una clasificación positiva y resultó ser verdaderamente positiva. Cabe mencionar que la métrica de precisión es comúnmente usada junto con la sensibilidad, la cual es llamada *Recall* cuando es empleada con la precisión.

Curva ROC: Géron (2019) señala que a partir de la curva característica operativa del receptor (curva ROC) se puede visualizar todos los valores que tiene la tasa de verdaderos positivos (TPR, también llamado *Recall*) y la tasa de falsos positivos (FPR) para todos los posibles umbrales de clasificación determinados por un valor de la probabilidad estimada. Asimismo, se emplea el área bajo la curva (*AUC*) para medir el rendimiento de la curva, cuyo valor está comprendido entre 0.5 y 1, donde 1 significa una predicción perfecta y 0.5 señala que el modelo no cuenta con capacidad para discriminar entre una clase positiva y negativa. Sin embargo, Cook & Ramadas (2020) advierten que en un escenario de datos desbalanceados, la figura de la curva ROC podría mostrar una vista muy optimista. A pesar de esta última advertencia, se optó por incorporar esta métrica porque resume de forma sencilla los distintos valores de la sensibilidad bajo diferentes umbrales de clasificación y además ha sido empleado en distintos estudios sobre interrupción de estudios. (Adelman, Haimovich, Ham, & Vázquez, 2017; Sansone, 2017)

Curva Precisión *Recall* (PR): Por otro lado, Cook & Ramadas (2020) señalan que la curva PR, al igual que la curva ROC, permite resumir de forma muy sencilla los distintos valores de la Precisión y Recall (Sensibilidad) bajo diferentes umbrales. Además, señalan que es más apropiado emplear esta curva cuando las clases están desbalanceadas. También se emplea el AUC para medir el rendimiento de la

curva PR, cuyo valor se le conoce como *Average Precision*. Por esta razón se optó por graficar la curva PR y se calculó su respectivo AUC (*average precision*) para conocer, de forma resumida, los distintos valores de la sensibilidad y la precisión para distintos umbrales.

Subcobertura y Filtración: En este punto es importante señalar que estas métricas están relacionadas a indicadores que se emplean para cuantificar la imperfección de una focalización, tales como la filtración o error de inclusión y subcobertura o error de exclusión. Van Domelen (2007) señala que la subcobertura hace referencia a hogares pobres que son excluidos del beneficio de algún programa, mientras que la filtración se refiere a los hogares no pobres que se benefician del programa. En ese sentido, se definió la métrica de filtración como el ratio de estudiantes que fueron clasificados incorrectamente como que desertan sus estudios, entre el total de estudiantes clasificados como que desertan sus estudios. Asimismo, la subcobertura está definida como el ratio de estudiantes que no fueron clasificados como que desertan sus estudios, entre el total de estudiantes que realmente desertan sus estudios.

Efecto *Cry Wolf*: Un aspecto adicional en tener en cuenta para la selección de la métrica de desempeño es la incorporación de los riesgos de interrupción en un sistema de alertas tempranas (Alerta Escueta). Los sistemas de alerta temprana están asociados al modismo inglés «llorar lobo» (*cry wolf*) que hace referencia a las falsas alarmas. Existe estudios que miden el efecto de «*cry wolf*» en las alertas, el cual evidencia que la presencia de una gran cantidad de falsos positivos (falsas alarmas) puede disminuir la confianza en las alertas reportadas por estos sistemas, reduciendo así el uso de estas herramientas (Bliss, 1993).

F1: Esta métrica representa la media armónica entre la Precisión y Sensibilidad (Japkowicz, 2013). Esta métrica es importante porque permite orientar la estimación del modelo con el objetivo que busque siempre maximizar su valor. De esta forma se optimiza la sensibilidad y precisión al mismo tiempo, identificando así la mayor cantidad de estudiantes que desertan sus estudios sin descuidar los falsos positivos que puede ser contraproducente debido al efecto *Cry Wolf*.

ANEXO 5: Cálculo de métricas

Tabla 13: Matriz de confusión

MATRIZ DE CONFUSIÓN		Evento Real	
		Deserta interanualmente	No deserta interanualmente
Predicción	Deserta interanualmente	A (Verdadero Positivo)	B (Falso Negativo)
	No deserta interanualmente	C (Falso Negativo)	D (Verdadero Positivo)

Nota. Esta tabla muestra la matriz de confusión de referencia para el cálculo de las métricas de rendimiento.

MÉTRICAS DE RENDIMIENTO

Exactitud (Accuracy) = (A + D) / (A + B + C + D)

Sensibilidad (Recall / TPR) = A / (A + C)

Especificidad = D / (B + D)

Precisión = A / (A + B)

F1 = 2 * Precisión * Recall / (Precisión + Recall)

Ratio de falsos positivos (FPR) = 1 − Especificidad = B / (B + D)

Subcobertura (error de exclusión) = C / (A + C)

Filtración (error de inclusión) = B / (A + B)

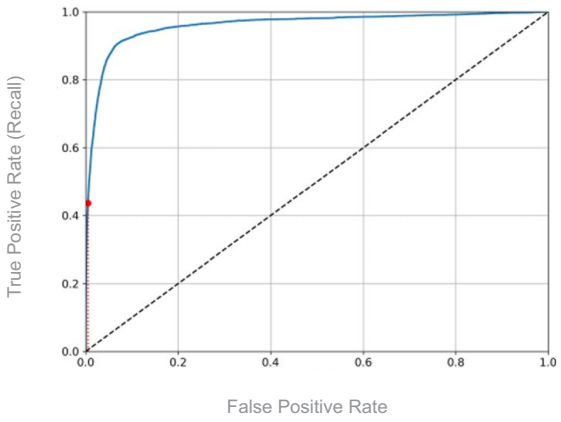
Average Precisión = $\sum_n (\text{Recall} - \text{Recall}_{n-1}) * \text{Precisión}_n$,

donde Recall_n y Precisión_n es la sensibilidad y la precisión para un umbral n.

CURVAS ROC Y PR

La figura 17 muestra la curva ROC, el cual se muestra al trazar la tasa de falsos positivos y la tasa de verdaderos positivos para todos los posibles umbrales de clasificación. Una mayor área debajo de la curva significará una tasa de falsos positivos cercanos a cero y una tasa de verdaderos positivos cercanos a 1. El punto rojo representa el ratio de verdaderos positivos y falsos positivos para un determinado umbral (Géron, 2019).

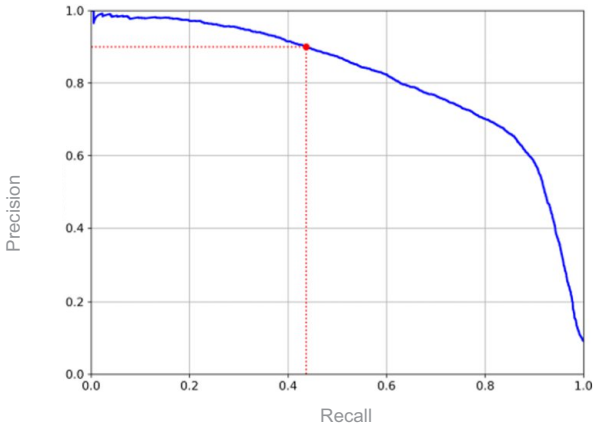
Figura 17: Curva Receiver Operating Characteristic (ROC)



Nota. Adaptado de Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow (p.98), por A. Géron, 2019, O'REALLY.

La figura 18 muestra los diferentes umbrales entre la precisión y recall (sensibilidad). Una mayor área debajo de la curva representa un mayor valor para la precisión y sensibilidad. El punto rojo representa la precisión y sensibilidad para un determinado umbral (Géron, 2019).

Figura 18: Curva Precisión Recall (PR)



Nota. Adaptado de Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow (p.96), por A. Géron, 2019, O'REALLY.

ANEXO 6: Criterios para la división de datos en entrenamiento y validación

Existen diversos métodos que dividen el conjunto de datos en entrenamiento y validación. Por un lado, se encuentra el método de retención o *holdout* el cual asigna 2/3 de los datos como entrenamiento y 1/3 de los datos como validación. Sin embargo, la desventaja de este método es que solo asigna una única proporción de los datos para el entrenamiento (Kohavi, 1995).

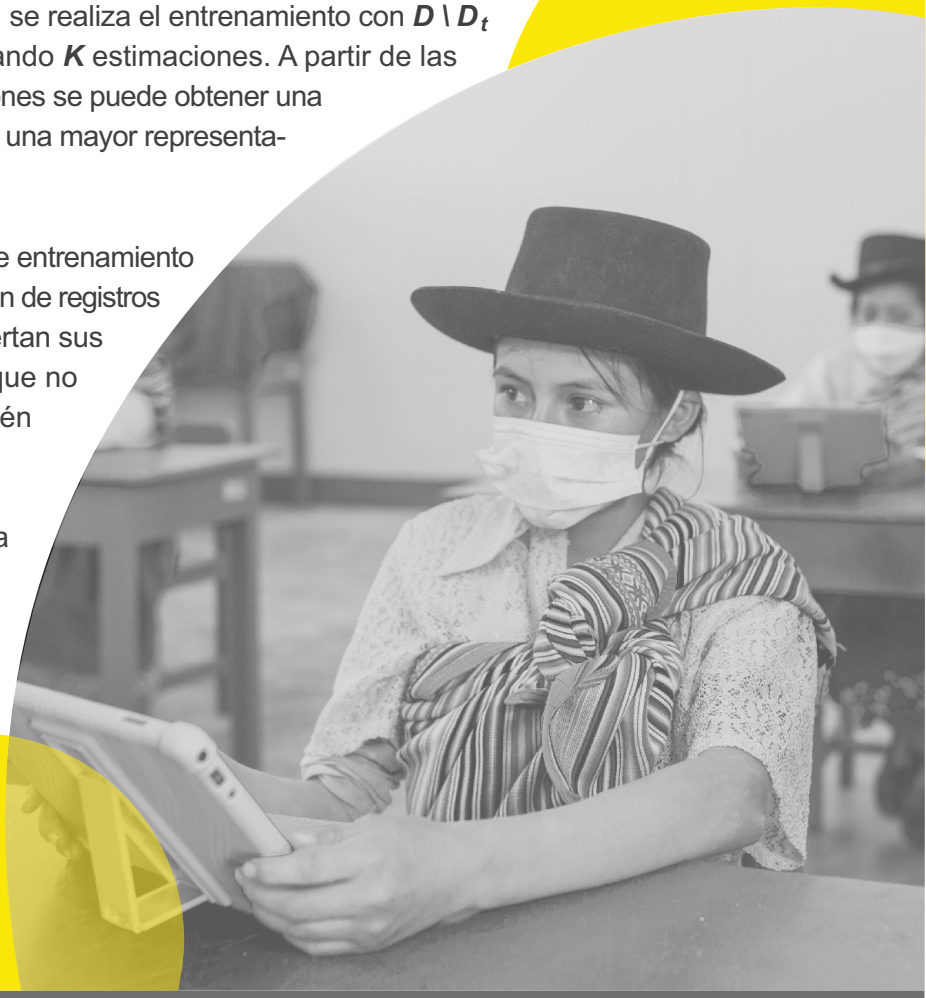
En esa misma línea, existe otro grupo[27] de métodos que emplean los datos originales para generar múltiples muestras de datos de entrenamiento y validación con el objetivo de poder promediar las distintas métricas de desempeño de los estimadores de cada una de las muestras, obteniendo así un resultado más representativo (Kohavi, 1995).

Para poder seleccionar el método más adecuado se revisaron los estudios de Kohavi (1995) y Borra & Di Ciaccio (2010), cuyos resultados sugieren que la técnica de *k-fold cross-validation* (validación cruzada de k iteraciones) es la mejor para medir el rendimiento de un modelo.

En la validación cruzada (CV) de k iteraciones, el conjunto de datos $\boldsymbol{D}$ es dividido en $\boldsymbol{K}$ submuestras $\boldsymbol{D_1}$, $\boldsymbol{D_2}$, $\boldsymbol{D_3}$,…, $\boldsymbol{D_k}$ de igual tamaño aproximadamente. Se entrena y valida $\boldsymbol{K}$ veces, donde cada vez $\boldsymbol{t} \in \{ 1,2,\dots k \}$ se realiza el entrenamiento con $\boldsymbol{D} \setminus \boldsymbol{D_t}$ y se emplea $\boldsymbol{D}$ para la validación, generando $\boldsymbol{K}$ estimaciones. A partir de las métricas de rendimiento de las $\boldsymbol{K}$ estimaciones se puede obtener una métrica de rendimiento promedio que tiene una mayor representatividad (Kohavi, 1995).

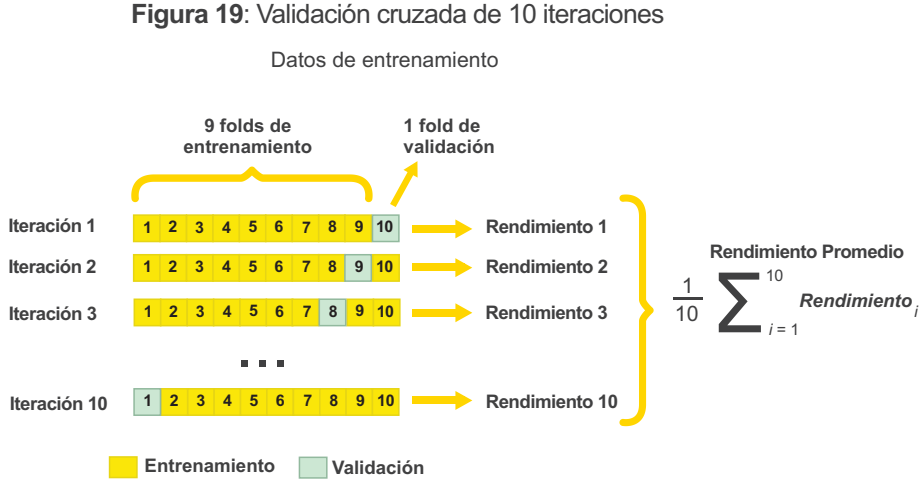
Adicionalmente, se espera que los datos de entrenamiento y validación cuenten con la misma proporción de registros de la clase positiva (estudiantes que desertan sus estudios) y clase negativa (estudiantes que no desertan sus estudios), es decir, que estén estratificados.

Por lo anterior señalado, se tomó en cuenta la recomendación realizada por Kohavi (1995, pág. 7), la cual consiste en realizar una validación cruzada de 10 iteraciones, como se muestra en la figura 19.



27. Kohavi (1995) señala los siguientes métodos: Bootstrap, k-fold cross-validation y Leave one out cross validation (LOOCV).

Figura 19: Validación cruzada de 10 iteraciones



Nota. El gráfico representa el esquema de trabajo para la validación cruzada de 10 iteraciones.

ANEXO 7: División de datos en entrenamiento y validación

Para la división de los datos en entrenamiento y validación se realizó la validación cruzada con 10 iteraciones[28] para cada muestra de grado escolar y macro región.

Para la validación cruzada de 10 iteraciones se procedió a dividir el conjunto de datos, el cual contiene al sub conjunto de registros de la clase positiva ***P***, en 10 submuestras iguales donde 9 submuestras se emplean para el entrenamiento y 1 submuestra para la validación.

El procedimiento se iteró 10 veces, rotando la submuestra de validación. De esta manera, cada iteración contó con sus respectivos conjuntos de datos de entrenamiento E_i y validación V_i. Asimismo, E_i cuenta con un sub conjunto de datos de entrenamiento de la clase positiva E_iP y V_i cuenta con un sub conjunto de datos de validación de la clase positiva V_iP. Es importante resaltar que E_iP y V_iP se encuentran distribuidos proporcionalmente (estratificados) en base al total de registros en E_i y V_i respectivamente.

Este proceso trajo como resultado distintas cantidades de registros de entrenamiento y validación como se muestran en las tablas 14,15 y 16.

Tabla 14: Validación cruzada con 10 iteraciones para el Nivel Inicial

Grado	Macro región	Total	*P*	E_i	E_iP	V_i	V_iP
ciclo_2	lima_metro_callao	451499	15207	406349	13687	45150	1520
ciclo_2	norte	375646	4486	338081	4037	37565	449
ciclo_2	sur	244844	1945	220359	1750	24485	195
ciclo_2	centro	348958	3959	314062	3563	34896	396
ciclo_2	oriente	193137	3690	173823	3321	19314	369

Nota. Esta tabla muestra el total de estudiantes para el entrenamiento y validación del modelo que corresponde a cada iteración de la validación cruzada para el nivel inicial.

28. Ver Anexo 6 «Criterios para la división de datos en entrenamiento y validación».

Tabla 15: Validación cruzada con 10 iteraciones para el Nivel Primaria

Grado	Macro región	Total	P	E_i	E_iP	V_i	V_iP
1 Prim	lima_metro_callao	177747	4244	159972	3820	17775	424
1 Prim	norte	139227	1111	125304	1000	13923	111
1 Prim	sur	93368	399	84031	359	9337	40
1 Prim	centro	127534	1034	114780	930	12754	104
1 Prim	oriente	74965	902	67468	812	7497	90
2 Prim	lima_metro_callao	178534	4139	160680	3725	17854	414
2 Prim	norte	143131	1015	128817	913	14314	102
2 Prim	sur	96138	309	86524	278	9614	31
2 Prim	centro	131253	797	118127	717	13126	80
2 Prim	oriente	74661	528	67194	475	7467	53
3 Prim	lima_metro_callao	178157	4062	160341	3656	17816	406
3 Prim	norte	146272	932	131644	838	14628	94
3 Prim	sur	96370	320	86733	288	9637	32
3 Prim	centro	131360	780	118224	702	13136	78
3 Prim	oriente	75474	532	67926	479	7548	53
4 Prim	lima_metro_callao	181673	4024	163505	3621	18168	403
4 Prim	norte	151818	1095	136636	986	15182	109
4 Prim	sur	96804	374	87123	336	9681	38
4 Prim	centro	133287	786	119958	708	13329	78
4 Prim	oriente	85801	722	77220	649	8581	73
5 Prim	lima_metro_callao	175356	3572	157820	3215	17536	357
5 Prim	norte	147840	1048	133056	944	14784	104
5 Prim	sur	95115	295	85603	265	9512	30
5 Prim	centro	131050	781	117945	703	13105	78
5 Prim	oriente	83566	804	75209	724	8357	80
6 Prim	lima_metro_callao	172162	4975	154945	4477	17217	498
6 Prim	norte	145761	5013	131184	4511	14577	502
6 Prim	sur	95156	1238	85640	1114	9516	124
6 Prim	centro	130473	3107	117425	2796	13048	311
6 Prim	oriente	79464	6683	71517	6014	7947	669

Nota. Esta tabla muestra el total de estudiantes para el entrenamiento y validación del modelo que corresponde a cada iteración de la validación cruzada para el nivel primaria.

Tabla 16: Validación cruzada con 10 iteraciones para el Nivel Secundaria

Grado	Macro región	Total	P	E_i	E_iP	V_i	V_iP
1 Sec	lima_metro_callao	171666	3380	154499	3042	17167	338
1 Sec	norte	140687	1973	126618	1776	14069	197
1 Sec	sur	95644	497	86079	447	9565	50
1 Sec	centro	129321	1359	116388	1223	12933	136
1 Sec	oriente	74171	1919	66753	1727	7418	192
2 Sec	lima_metro_callao	160748	3429	144673	3086	16075	343
2 Sec	norte	132695	2795	119425	2515	13270	280
2 Sec	sur	91358	784	82222	706	9136	78
2 Sec	centro	126539	1975	113885	1778	12654	197
2 Sec	oriente	66627	2506	59964	2256	6663	250
3 Sec	lima_metro_callao	156579	4391	140921	3952	15658	439
3 Sec	norte	125813	4357	113231	3921	12582	436
3 Sec	sur	86014	1377	77412	1239	8602	138
3 Sec	centro	118306	3093	106475	2784	11831	309
3 Sec	oriente	63105	3686	56794	3317	6311	369
4 Sec	lima_metro_callao	152505	3902	137254	3512	15251	390
4 Sec	norte	122615	4541	110353	4087	12262	454
4 Sec	sur	85459	1434	76913	1291	8546	143
4 Sec	centro	115510	3060	103959	2754	11551	306
4 Sec	oriente	60359	3801	54323	3421	6036	380
5 Sec	lima_metro_callao	149178	305	134260	275	14918	30
5 Sec	norte	114837	861	103353	775	11484	86
5 Sec	sur	84378	231	75940	208	8438	23
5 Sec	centro	113172	362	101854	325	11318	37
5 Sec	oriente	53431	643	48087	578	5344	65

Nota. Esta tabla muestra el total de estudiantes para el entrenamiento y validación del modelo que corresponde a cada iteración de la validación cruzada para el nivel secundaria.

ANEXO 8: Hiperparámetros

A continuación, se detallan los criterios empleados para la configuración de hiperparámetros del modelo *LightGBM*.

Configuración automática: Se empleó Optuna[29] como framework para la búsqueda de valores idóneos para los hiperparámetros[30] descritos en la tabla 17.

Tabla 17: Hiperparámetros para configuración automática

Hiperparámetro	Descripción
num_leaves	Número máximo de hojas de un árbol. Permite controlar el nivel de complejidad del modelo.
learning_rate	Tasa de aprendizaje
max_bin	Número máximo de rangos de números empleados para discretizar variables continuas.

Nota. Esta tabla muestra los hiperparámetros para la configuración automática.

Configuración manual: La configuración estuvo enfocado en calcular los valores de algunos hiperparámetros de tal forma que puedan reducir los datos desbalanceados y el sobreajuste. La tabla 18 describe los hiperparámetros empleados.

Tabla 18: Hiperparámetros para configuración manual

Hiperparámetro	Descripción
pos_bagging_fraction	Proporción de la clase positiva que se emplea en el *Bagging*.
neg_ bagging_fraction	Proporción de la clase negativa que se emplea en el *Bagging*.
bagging_freq	Indica con qué frecuencia (cada cuantos árboles) se debe hacer un muestreo (o si se debe usar la base de datos completa original). Su valor debe ser mayor a cero para habilitar *pos_bagging_fraction* y *neg_bagging_fraction*.
scale_pos_weight	Peso de las etiquetas de la clase positiva (N_n / N_p).
early_stopping	Permite detener el entrenamiento si la métrica de desempeño especificada para los datos de validación no mejora en los *early_stopping* iteraciones.
num_boost_round	Número de árboles o iteraciones empleados para el *boosting*.

Nota. Esta tabla muestra los hiperparámetros para la configuración manual.

29. Especificación de OPTUNA: https://optuna.org/
30. Especificación de los hiperparámetros: https://lightgbm.readthedocs.io/en/latest/Parameters.html

En primer lugar, para la gestión de los datos desbalanceados, se tomó en cuenta que *LightGBM* es un modelo que soporta *Bagging,* el cual es un proceso que permite realizar el entrenamiento basado sobre múltiples muestras aleatorias sin reemplazo. De esta manera, se configuró *LightGBM* para que cada árbol en el *boosting* sea entrenado con una muestra de los casos positivos y los casos negativos. Para el cálculo de la muestra, se siguieron los siguientes pasos:

1) Se denotó las variables utilitarias para el cálculo.

   N_n **:** Número de estudiantes que no interrumpen sus estudios (casos negativos)

   N_p **:** Número de estudiantes que sí interrumpen sus estudios (casos positivos)

   $\lambda \frac{N_p}{N_n}$, siempre es menor que 1 por el bajo número de casos positivos

   $T_{mínimo}$**:** cantidad mínima de observaciones para realizar el entrenamiento.[31]

2) Se formuló la muestra de cada árbol con la siguiente expresión:

   $$\textbf{muestra} = neg\_bagging\_fraction * N_n + pos\_bagging\_fraction * N_p$$

3) Se fijó *neg_bagging_fraction* y *pos_bagging_fraction* de tal modo que los casos positivos y casos negativos estén balanceados para entrenar cada árbol, es decir *neg_bagging_fraction* = λ y *pos_bagging_fraction* = *1* . Sin embargo, se debe cumplir que la $\textbf{muestra} > T_{mínimo}$, ese decir, se debe encontrar un nuevo valor λ que permita cumplir dicha condición. Este nuevo valor será denotado con α, el cual será calculado de la siguiente forma:

$$\alpha * N_n + N_p = T_{mínimo}$$

$$\alpha = \frac{T_{mínimo} - N_p}{N_n}$$

4) La idea es que todos los árboles usen una base más balanceada que la original, por lo que se estableció *bagging_freq* = 1. En resumen, los parámetros de *bagging* configurados quedan de la siguiente manera:

$$pos\_bagging\_fraction = 1$$
$$neg\_bagging\_fraction = \alpha$$
$$bagging\_freq = 1$$

31. Se emplea una cantidad referencial mínima de diez mil registros para reducir el sobreajuste del modelado con LightGBM

5) No obstante, **α** no asegura que el peso de los casos positivos y negativos sean igual. Para ello, se realizó el reajuste de los pesos de la clase positiva a través del hiperparámetro *scale_pos_weight*. Es importante resaltar que el nuevo número de casos negativos estaría dato por **α** * N_n debido al muestreo descrito previamente, calculando el hiperparámetro de la siguiente manera:

$$scale\_pos\_weight = (\alpha * N_n) / N_p$$

Gestión de sobreajuste: Se configuró el hiperparámetro ***early_stopping*** con un valor de 200. Asimismo, se estableció ***num_boost_round*** como el número máximo de interacciones que contribuyen con la métrica de desempeño previo a la interrupción del entrenamiento del modelo por ***early_stopping***.

ANEXO 9: Métricas por grado y macro región

Tabla 19: Métricas con VC de 10 iteraciones – Ciclo 2 del nivel Inicial y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.247	0.009	0.230	0.258
	Precisión	Entrenamiento	0.247	0.001	0.245	0.249
	Sensibilidad	Validación	0.369	0.014	0.343	0.387
	Sensibilidad	Entrenamiento	0.370	0.002	0.367	0.374
	Especificidad	Validación	0.961	0.001	0.959	0.962
	Especificidad	Entrenamiento	0.961	0.000	0.960	0.961
	F1	Validación	0.296	0.011	0.275	0.308
	F1	Entrenamiento	0.296	0.001	0.294	0.299
	PRAUC	Validación	0.162	0.008	0.148	0.174
	PRAUC	Entrenamiento	0.169	0.002	0.165	0.172
	ROCAUC	Validación	0.766	0.006	0.755	0.776
	ROCAUC	Entrenamiento	0.791	0.001	0.790	0.792
	Filtración	Validación	0.753	0.009	0.742	0.770
	Filtración	Entrenamiento	0.753	0.001	0.751	0.755
	Subcobertura	Validación	0.631	0.014	0.613	0.657
	Subcobertura	Entrenamiento	0.630	0.002	0.626	0.633
Norte	Precisión	Validación	0.264	0.012	0.245	0.294
	Precisión	Entrenamiento	0.266	0.002	0.263	0.269
	Sensibilidad	Validación	0.409	0.014	0.392	0.438
	Sensibilidad	Entrenamiento	0.415	0.005	0.407	0.425
	Especificidad	Validación	0.986	0.001	0.985	0.988
	Especificidad	Entrenamiento	0.986	0.000	0.986	0.986
	F1	Validación	0.321	0.012	0.305	0.347
	F1	Entrenamiento	0.324	0.002	0.321	0.327
	PRAUC	Validación	0.177	0.010	0.159	0.192
	PRAUC	Entrenamiento	0.189	0.003	0.182	0.193
	ROCAUC	Validación	0.868	0.006	0.856	0.874
	ROCAUC	Entrenamiento	0.908	0.001	0.907	0.910
	Filtración	Validación	0.736	0.012	0.706	0.755
	Filtración	Entrenamiento	0.734	0.002	0.731	0.737
	Subcobertura	Validación	0.591	0.014	0.563	0.608
	Subcobertura	Entrenamiento	0.585	0.005	0.575	0.593

Sur	Precisión	Validación	0.244	0.022	0.210	0.277
	Precisión	Entrenamiento	0.253	0.004	0.246	0.261
	Sensibilidad	Validación	0.308	0.023	0.263	0.338
	Sensibilidad	Entrenamiento	0.325	0.005	0.317	0.336
	Especificidad	Validación	0.992	0.001	0.990	0.993
	Especificidad	Entrenamiento	0.992	0.000	0.992	0.992
	F1	Validación	0.272	0.021	0.236	0.299
	F1	Entrenamiento	0.284	0.004	0.277	0.292
	PRAUC	Validación	0.137	0.017	0.102	0.159
	PRAUC	Entrenamiento	0.167	0.005	0.159	0.175
	ROCAUC	Validación	0.820	0.012	0.804	0.839
	ROCAUC	Entrenamiento	0.931	0.001	0.930	0.932
	Filtración	Validación	0.756	0.022	0.723	0.790
	Filtración	Entrenamiento	0.747	0.004	0.739	0.754
	Subcobertura	Validación	0.692	0.023	0.662	0.737
	Subcobertura	Entrenamiento	0.675	0.005	0.664	0.683
Centro	Precisión	Validación	0.229	0.017	0.195	0.260
	Precisión	Entrenamiento	0.233	0.005	0.224	0.245
	Sensibilidad	Validación	0.313	0.022	0.265	0.348
	Sensibilidad	Entrenamiento	0.319	0.006	0.307	0.327
	Especificidad	Validación	0.988	0.000	0.987	0.989
	Especificidad	Entrenamiento	0.988	0.000	0.987	0.989
	F1	Validación	0.265	0.019	0.225	0.298
	F1	Entrenamiento	0.269	0.003	0.263	0.273
	PRAUC	Validación	0.133	0.014	0.112	0.157
	PRAUC	Entrenamiento	0.144	0.004	0.140	0.153
	ROCAUC	Validación	0.830	0.009	0.815	0.842
	ROCAUC	Entrenamiento	0.888	0.001	0.886	0.890
	Filtración	Validación	0.771	0.017	0.740	0.805
	Filtración	Entrenamiento	0.767	0.005	0.755	0.776
	Subcobertura	Validación	0.687	0.022	0.652	0.735
	Subcobertura	Entrenamiento	0.681	0.006	0.673	0.693

Oriente	Precisión	Validación	0.154	0.010	0.130	0.167
	Precisión	Entrenamiento	0.165	0.001	0.163	0.168
	Sensibilidad	Validación	0.456	0.026	0.409	0.509
	Sensibilidad	Entrenamiento	0.491	0.004	0.486	0.498
	Especificidad	Validación	0.951	0.002	0.947	0.953
	Especificidad	Entrenamiento	0.952	0.001	0.951	0.953
	F1	Validación	0.230	0.015	0.197	0.252
	F1	Entrenamiento	0.247	0.002	0.244	0.250
	PRAUC	Validación	0.150	0.012	0.126	0.164
	PRAUC	Entrenamiento	0.167	0.004	0.161	0.176
	ROCAUC	Validación	0.864	0.009	0.848	0.880
	ROCAUC	Entrenamiento	0.899	0.001	0.898	0.901
	Filtración	Validación	0.846	0.010	0.833	0.870
	Filtración	Entrenamiento	0.835	0.001	0.832	0.837
	Subcobertura	Validación	0.544	0.026	0.491	0.591
	Subcobertura	Entrenamiento	0.509	0.004	0.502	0.514

Tabla 20: Métricas con VC de 10 iteraciones – 1° grado de primaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.177	0.009	0.161	0.189
	Precisión	Entrenamiento	0.181	0.002	0.178	0.183
	Sensibilidad	Validación	0.365	0.023	0.327	0.408
	Sensibilidad	Entrenamiento	0.373	0.005	0.362	0.380
	Especificidad	Validación	0.958	0.001	0.955	0.960
	Especificidad	Entrenamiento	0.959	0.001	0.958	0.960
	F1	Validación	0.238	0.013	0.216	0.257
	F1	Entrenamiento	0.244	0.002	0.240	0.245
	PRAUC	Validación	0.128	0.012	0.107	0.147
	PRAUC	Entrenamiento	0.140	0.003	0.135	0.145
	ROCAUC	Validación	0.800	0.013	0.783	0.826
	ROCAUC	Entrenamiento	0.862	0.001	0.860	0.864
	Filtración	Validación	0.823	0.009	0.811	0.839
	Filtración	Entrenamiento	0.819	0.002	0.817	0.822
	Subcobertura	Validación	0.635	0.023	0.592	0.673
	Subcobertura	Entrenamiento	0.627	0.005	0.620	0.638
Norte	Precisión	Validación	0.219	0.023	0.188	0.254
	Precisión	Entrenamiento	0.224	0.004	0.219	0.232
	Sensibilidad	Validación	0.455	0.060	0.351	0.550
	Sensibilidad	Entrenamiento	0.466	0.015	0.427	0.486
	Especificidad	Validación	0.987	0.001	0.986	0.988
	Especificidad	Entrenamiento	0.987	0.001	0.986	0.988
	F1	Validación	0.296	0.033	0.245	0.344
	F1	Entrenamiento	0.302	0.005	0.295	0.310
	PRAUC	Validación	0.152	0.021	0.114	0.193
	PRAUC	Entrenamiento	0.185	0.005	0.179	0.194
	ROCAUC	Validación	0.901	0.018	0.874	0.930
	ROCAUC	Entrenamiento	0.975	0.001	0.975	0.976
	Filtración	Validación	0.781	0.023	0.746	0.813
	Filtración	Entrenamiento	0.776	0.004	0.768	0.781
	Subcobertura	Validación	0.545	0.060	0.450	0.649
	Subcobertura	Entrenamiento	0.534	0.015	0.514	0.573

Sur	Precisión	Validación	0.180	0.035	0.125	0.250
	Precisión	Entrenamiento	0.300	0.021	0.269	0.340
	Sensibilidad	Validación	0.278	0.067	0.125	0.400
	Sensibilidad	Entrenamiento	0.538	0.012	0.519	0.557
	Especificidad	Validación	0.995	0.001	0.992	0.996
	Especificidad	Entrenamiento	0.995	0.001	0.994	0.995
	F1	Validación	0.217	0.044	0.125	0.286
	F1	Entrenamiento	0.385	0.019	0.358	0.418
	PRAUC	Validación	0.103	0.022	0.070	0.153
	PRAUC	Entrenamiento	0.228	0.015	0.205	0.251
	ROCAUC	Validación	0.865	0.020	0.826	0.900
	ROCAUC	Entrenamiento	0.991	0.000	0.990	0.992
	Filtración	Validación	0.820	0.035	0.750	0.875
	Filtración	Entrenamiento	0.700	0.021	0.660	0.731
	Subcobertura	Validación	0.722	0.067	0.600	0.875
	Subcobertura	Entrenamiento	0.462	0.012	0.443	0.481
Centro	Precisión	Validación	0.184	0.017	0.061	0.216
	Precisión	Entrenamiento	0.194	0.005	0.186	0.203
	Sensibilidad	Validación	0.488	0.048	0.423	0.573
	Sensibilidad	Entrenamiento	0.524	0.015	0.489	0.542
	Especificidad	Validación	0.982	0.002	0.980	0.985
	Especificidad	Entrenamiento	0.982	0.001	0.981	0.984
	F1	Validación	0.267	0.022	0.238	0.307
	F1	Entrenamiento	0.283	0.006	0.276	0.296
	PRAUC	Validación	0.149	0.019	0.124	0.183
	PRAUC	Entrenamiento	0.179	0.004	0.170	0.185
	ROCAUC	Validación	0.897	0.015	0.873	0.930
	ROCAUC	Entrenamiento	0.974	0.001	0.973	0.975
	Filtración	Validación	0.816	0.017	0.784	0.839
	Filtración	Entrenamiento	0.806	0.005	0.797	0.814
	Subcobertura	Validación	0.512	0.048	0.427	0.577
	Subcobertura	Entrenamiento	0.476	0.015	0.458	0.511

Oriente	Precisión	Validación	0.150	0.024	0.095	0.181
	Precisión	Entrenamiento	0.179	0.010	0.164	0.197
	Sensibilidad	Validación	0.351	0.047	0.278	0.429
	Sensibilidad	Entrenamiento	0.425	0.047	0.341	0.497
	Especificidad	Validación	0.975	0.004	0.968	0.984
	Especificidad	Entrenamiento	0.976	0.003	0.971	0.982
	F1	Validación	0.210	0.028	0.142	0.243
	F1	Entrenamiento	0.251	0.012	0.234	0.266
	PRAUC	Validación	0.159	0.027	0.109	0.198
	PRAUC	Entrenamiento	0.197	0.009	0.180	0.209
	ROCAUC	Validación	0.900	0.021	0.851	0.926
	ROCAUC	Entrenamiento	0.961	0.001	0.959	0.962
	Filtración	Validación	0.850	0.024	0.819	0.905
	Filtración	Entrenamiento	0.821	0.010	0.803	0.836
	Subcobertura	Validación	0.649	0.047	0.571	0.722
	Subcobertura	Entrenamiento	0.575	0.047	0.503	0.659

Tabla 21: Métricas con VC de 10 iteraciones – 2° grado de primaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.176	0.010	0.154	0.189
	Precisión	Entrenamiento	0.178	0.001	0.176	0.180
	Sensibilidad	Validación	0.387	0.026	0.336	0.433
	Sensibilidad	Entrenamiento	0.392	0.004	0.385	0.396
	Especificidad	Validación	0.957	0.001	0.955	0.959
	Especificidad	Entrenamiento	0.957	0.000	0.956	0.958
	F1	Validación	0.242	0.014	0.211	0.263
	F1	Entrenamiento	0.245	0.001	0.242	0.246
	PRAUC	Validación	0.120	0.011	0.103	0.136
	PRAUC	Entrenamiento	0.137	0.003	0.134	0.143
	ROCAUC	Validación	0.788	0.014	0.767	0.810
	ROCAUC	Entrenamiento	0.859	0.001	0.857	0.861
	Filtración	Validación	0.824	0.000	0.811	0.846
	Filtración	Entrenamiento	0.822	0.001	0.820	0.824
	Subcobertura	Validación	0.613	0.026	0.567	0.664
	Subcobertura	Entrenamiento	0.608	0.004	0.604	0.615
Norte	Precisión	Validación	0.174	0.013	0.146	0.188
	Precisión	Entrenamiento	0.194	0.004	0.187	0.202
	Sensibilidad	Validación	0.527	0.029	0.455	0.564
	Sensibilidad	Entrenamiento	0.604	0.009	0.591	0.625
	Especificidad	Validación	0.982	0.002	0.978	0.984
	Especificidad	Entrenamiento	0.982	0.000	0.982	0.983
	F1	Validación	0.261	0.016	0.229	0.277
	F1	Entrenamiento	0.293	0.005	0.284	0.303
	PRAUC	Validación	0.147	0.017	0.116	0.184
	PRAUC	Entrenamiento	0.188	0.005	0.179	0.196
	ROCAUC	Validación	0.898	0.013	0.881	0.924
	ROCAUC	Entrenamiento	0.979	0.000	0.978	0.980
	Filtración	Validación	0.826	0.013	0.813	0.854
	Filtración	Entrenamiento	0.806	0.004	0.798	0.813
	Subcobertura	Validación	0.473	0.029	0.436	0.545
	Subcobertura	Entrenamiento	0.396	0.009	0.375	0.409

Sur	Precisión	Validación	0.116	0.027	0.087	0.177
	Precisión	Entrenamiento	0.232	0.011	0.212	0.251
	Sensibilidad	Validación	0.401	0.098	0.323	0.548
	Sensibilidad	Entrenamiento	0.896	0.025	0.849	0.950
	Especificidad	Validación	0.990	0.001	0.988	0.992
	Especificidad	Entrenamiento	0.990	0.000	0.990	0.991
	F1	Validación	0.180	0.042	0.137	0.268
	F1	Entrenamiento	0.369	0.016	0.340	0.393
	PRAUC	Validación	0.085	0.033	0.046	0.152
	PRAUC	Entrenamiento	0.280	0.025	0.246	0.324
	ROCAUC	Validación	0.867	0.035	0.815	0.950
	ROCAUC	Entrenamiento	0.995	0.000	0.995	0.996
	Filtración	Validación	0.884	0.027	0.823	0.913
	Filtración	Entrenamiento	0.768	0.011	0.749	0.788
	Subcobertura	Validación	0.559	0.098	0.452	0.677
	Subcobertura	Entrenamiento	0.104	0.025	0.050	0.151
Centro	Precisión	Validación	0.172	0.024	0.135	0.208
	Precisión	Entrenamiento	0.195	0.006	0.189	0.205
	Sensibilidad	Validación	0.369	0.047	0.304	0.438
	Sensibilidad	Entrenamiento	0.426	0.022	0.392	0.449
	Especificidad	Validación	0.989	0.001	0.986	0.991
	Especificidad	Entrenamiento	0.989	0.001	0.988	0.991
	F1	Validación	0.234	0.031	0.195	0.282
	F1	Entrenamiento	0.267	0.005	0.255	0.274
	PRAUC	Validación	0.110	0.016	0.090	0.133
	PRAUC	Entrenamiento	0.169	0.003	0.164	0.175
	ROCAUC	Validación	0.878	0.015	0.848	0.909
	ROCAUC	Entrenamiento	0.979	0.001	0.978	0.980
	Filtración	Validación	0.828	0.024	0.792	0.865
	Filtración	Entrenamiento	0.805	0.006	0.795	0.811
	Subcobertura	Validación	0.631	0.047	0.563	0.696
	Subcobertura	Entrenamiento	0.574	0.022	0.551	0.608

Oriente	Precisión	Validación	0.098	0.018	0.054	0.121
	Precisión	Entrenamiento	0.169	0.013	0.0137	0.185
	Sensibilidad	Validación	0.284	0.064	0.151	0.396
	Sensibilidad	Entrenamiento	0.542	0.038	0.459	0.608
	Especificidad	Validación	0.980	0.002	0.977	0.983
	Especificidad	Entrenamiento	0.981	0.002	0.976	0.984
	F1	Validación	0.138	0.028	0.080	0.182
	F1	Entrenamiento	0.257	0.018	0.219	0.279
	PRAUC	Validación	0.078	0.017	0.058	0.115
	PRAUC	Entrenamiento	0.176	0.008	0.164	0.189
	ROCAUC	Validación	0.855	0.026	0.808	0.904
	ROCAUC	Entrenamiento	0.978	0.001	0.976	0.979
	Filtración	Validación	0.908	0.018	0.879	0.946
	Filtración	Entrenamiento	0.831	0.013	0.815	0.863
	Subcobertura	Validación	0.716	0.064	0.604	0.849
	Subcobertura	Entrenamiento	0.458	0.038	0.392	0.541

Tabla 22: Métricas con VC de 10 iteraciones – 3° grado de primaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.180	0.006	0.171	0.192
	Precisión	Entrenamiento	0.183	0.001	0.181	0.185
	Sensibilidad	Validación	0.438	0.016	0.419	0.466
	Sensibilidad	Entrenamiento	0.447	0.003	0.438	0.451
	Especificidad	Validación	0.954	0.001	0.952	0.956
	Especificidad	Entrenamiento	0.954	0.001	0.953	0.955
	F1	Validación	0.255	0.009	0.243	0.272
	F1	Entrenamiento	0.260	0.001	0.258	0.262
	PRAUC	Validación	0.128	0.007	0.117	0.143
	PRAUC	Entrenamiento	0.146	0.002	0.142	0.148
	ROCAUC	Validación	0.799	0.013	0.776	0.825
	ROCAUC	Entrenamiento	0.869	0.001	0.866	0.870
	Filtración	Validación	0.820	0.006	0.808	0.829
	Filtración	Entrenamiento	0.817	0.001	0.815	0.819
	Subcobertura	Validación	0.562	0.016	0.534	0.581
	Subcobertura	Entrenamiento	0.553	0.003	0.549	0.562
Norte	Precisión	Validación	0.187	0.021	0.151	0.223
	Precisión	Entrenamiento	0.196	0.003	0.192	0.201
	Sensibilidad	Validación	0.407	0.047	0.333	0.468
	Sensibilidad	Entrenamiento	0.424	0.016	0.391	0.454
	Especificidad	Validación	0.989	0.001	0.987	0.990
	Especificidad	Entrenamiento	0.989	0.001	0.988	0.990
	F1	Validación	0.256	0.027	0.208	0.296
	F1	Entrenamiento	0.268	0.003	0.263	0.274
	PRAUC	Validación	0.134	0.021	0.102	0.170
	PRAUC	Entrenamiento	0.166	0.004	0.162	0.174
	ROCAUC	Validación	0.907	0.016	0.884	0.929
	ROCAUC	Entrenamiento	0.979	0.001	0.978	0.980
	Filtración	Validación	0.813	0.021	0.777	0.849
	Filtración	Entrenamiento	0.804	0.003	0.799	0.808
	Subcobertura	Validación	0.593	0.047	0.532	0.667
	Subcobertura	Entrenamiento	0.576	0.016	0.546	0.609

Sur	Precisión	Validación	0.192	0.073	0.069	0.317
	Precisión	Entrenamiento	0.368	0.035	0.309	0.427
	Sensibilidad	Validación	0.209	0.090	0.063	0.406
	Sensibilidad	Entrenamiento	0.461	0.029	0.410	0.497
	Especificidad	Validación	0.997	0.000	0.996	0.998
	Especificidad	Entrenamiento	0.997	0.000	0.997	0.998
	F1	Validación	0.200	0.079	0.066	0.356
	F1	Entrenamiento	0.408	0.027	0.352	0.443
	PRAUC	Validación	0.096	0.032	0.062	0.165
	PRAUC	Entrenamiento	0.258	0.018	0.237	0.291
	ROCAUC	Validación	0.827	0.041	0.778	0.905
	ROCAUC	Entrenamiento	0.994	0.000	0.993	0.995
	Filtración	Validación	0.808	0.073	0.683	0.931
	Filtración	Entrenamiento	0.632	0.035	0.573	0.691
	Subcobertura	Validación	0.791	0.090	0.594	0.938
	Subcobertura	Entrenamiento	0.539	0.029	0.503	0.590
Centro	Precisión	Validación	0.170	0.019	0.145	0.202
	Precisión	Entrenamiento	0.198	0.010	0.176	0.212
	Sensibilidad	Validación	0.405	0.061	0.295	0.513
	Sensibilidad	Entrenamiento	0.477	0.015	0.457	0.504
	Especificidad	Validación	0.988	0.001	0.987	0.990
	Especificidad	Entrenamiento	0.988	0.001	0.986	0.989
	F1	Validación	0.240	0.029	0.197	0.290
	F1	Entrenamiento	0.280	0.010	0.261	0.297
	PRAUC	Validación	0.112	0.017	0.092	0.138
	PRAUC	Entrenamiento	0.172	0.004	0.165	0.176
	ROCAUC	Validación	0.872	0.019	0.837	0.894
	ROCAUC	Entrenamiento	0.980	0.001	0.978	0.981
	Filtración	Validación	0.830	0.019	0.798	0.855
	Filtración	Entrenamiento	0.802	0.010	0.788	0.824
	Subcobertura	Validación	0.595	0.061	0.487	0.705
	Subcobertura	Entrenamiento	0.523	0.015	0.496	0.543

Oriente	Precisión	Validación	0.113	0.043	0.049	0.172
	Precisión	Entrenamiento	0.234	0.038	0.173	0.291
	Sensibilidad	Validación	0.118	0.081	0.038	0.264
	Sensibilidad	Entrenamiento	0.236	0.083	0.148	0.404
	Especificidad	Validación	0.994	0.004	0.986	0.997
	Especificidad	Entrenamiento	0.994	0.003	0.988	0.997
	F1	Validación	0.107	0.049	0.043	0.207
	F1	Entrenamiento	0.222	0.020	0.194	0.258
	PRAUC	Validación	0.074	0.019	0.048	0.121
	PRAUC	Entrenamiento	0.171	0.009	0.154	0.186
	ROCAUC	Validación	0.843	0.023	0.814	0.884
	ROCAUC	Entrenamiento	0.975	0.001	0.973	0.976
	Filtración	Validación	0.887	0.043	0.828	0.951
	Filtración	Entrenamiento	0.766	0.038	0.709	0.827
	Subcobertura	Validación	0.882	0.081	0.736	0.962
	Subcobertura	Entrenamiento	0.764	0.083	0.596	0.852

Tabla 23: Métricas con VC de 10 iteraciones – 4° grado de primaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.174	0.013	0.153	0.193
	Precisión	Entrenamiento	0.179	0.001	0.178	0.182
	Sensibilidad	Validación	0.425	0.029	0.387	0.474
	Sensibilidad	Entrenamiento	0.438	0.005	0.429	0.446
	Especificidad	Validación	0.954	0.002	0.951	0.956
	Especificidad	Entrenamiento	0.955	0.001	0.954	0.956
	F1	Validación	0.247	0.018	0.219	0.274
	F1	Entrenamiento	0.254	0.002	0.252	0.258
	PRAUC	Validación	0.126	0.013	0.104	0.143
	PRAUC	Entrenamiento	0.140	0.005	0.133	0.148
	ROCAUC	Validación	0.795	0.013	0.780	0.821
	ROCAUC	Entrenamiento	0.864	0.001	0.861	0.865
	Filtración	Validación	0.826	0.013	0.807	0.847
	Filtración	Entrenamiento	0.821	0.001	0.818	0.822
	Subcobertura	Validación	0.575	0.029	0.526	0.613
	Subcobertura	Entrenamiento	0.562	0.005	0.554	0.571
Norte	Precisión	Validación	0.199	0.016	0.175	0.235
	Precisión	Entrenamiento	0.208	0.007	0.197	0.218
	Sensibilidad	Validación	0.415	0.054	0.330	0.518
	Sensibilidad	Entrenamiento	0.444	0.032	0.391	0.478
	Especificidad	Validación	0.988	0.001	0.986	0.990
	Especificidad	Entrenamiento	0.988	0.001	0.986	0.990
	F1	Validación	0.268	0.023	0.238	0.323
	F1	Entrenamiento	0.283	0.004	0.278	0.289
	PRAUC	Validación	0.137	0.023	0.105	0.194
	PRAUC	Entrenamiento	0.170	0.003	0.168	0.176
	ROCAUC	Validación	0.897	0.014	0.877	0.923
	ROCAUC	Entrenamiento	0.973	0.001	0.971	0.974
	Filtración	Validación	0.801	0.016	0.765	0.825
	Filtración	Entrenamiento	0.792	0.007	0.782	0.803
	Subcobertura	Validación	0.585	0.054	0.482	0.670
	Subcobertura	Entrenamiento	0.556	0.032	0.522	0.609

Sur	Precisión	Validación	0.127	0.029	0.083	0.170
	Precisión	Entrenamiento	0.219	0.014	0.188	0.238
	Sensibilidad	Validación	0.413	0.104	0.237	0.568
	Sensibilidad	Entrenamiento	0.783	0.022	0.747	0.818
	Especificidad	Validación	0.989	0.001	0.987	0.991
	Especificidad	Entrenamiento	0.989	0.001	0.987	0.990
	F1	Validación	0.194	0.044	0.122	0.265
	F1	Entrenamiento	0.342	0.019	0.302	0.367
	PRAUC	Validación	0.107	0.041	0.051	0.205
	PRAUC	Entrenamiento	0.223	0.013	0.199	0.247
	ROCAUC	Validación	0.856	0.040	0.814	0.928
	ROCAUC	Entrenamiento	0.993	0.001	0.992	0.993
	Filtración	Validación	0.873	0.029	0.830	0.917
	Filtración	Entrenamiento	0.781	0.014	0.762	0.812
	Subcobertura	Validación	0.587	0.104	0.432	0.763
	Subcobertura	Entrenamiento	0.217	0.022	0.182	0.253
Centro	Precisión	Validación	0.171	0.017	0.146	0.195
	Precisión	Entrenamiento	0.185	0.006	0.176	0.194
	Sensibilidad	Validación	0.305	0.022	0.278	0.346
	Sensibilidad	Entrenamiento	0.336	0.021	0.307	0.373
	Especificidad	Validación	0.991	0.001	0.990	0.993
	Especificidad	Entrenamiento	0.991	0.001	0.990	0.992
	F1	Validación	0.219	0.017	0.196	0.240
	F1	Entrenamiento	0.239	0.003	0.235	0.245
	PRAUC	Validación	0.110	0.010	0.088	0.126
	PRAUC	Entrenamiento	0.153	0.003	0.148	0.157
	ROCAUC	Validación	0.869	0.016	0.839	0.898
	ROCAUC	Entrenamiento	0.976	0.001	0.975	0.997
	Filtración	Validación	0.829	0.017	0.805	0.854
	Filtración	Entrenamiento	0.815	0.006	0.806	0.824
	Subcobertura	Validación	0.695	0.022	0.654	0.722
	Subcobertura	Entrenamiento	0.664	0.021	0.627	0.693

Oriente	Precisión	Validación	0.106	0.012	0.089	0.129
	Precisión	Entrenamiento	0.153	0.008	0.138	0.168
	Sensibilidad	Validación	0.287	0.026	0.250	0.333
	Sensibilidad	Entrenamiento	0.426	0.027	0.367	0.460
	Especificidad	Validación	0.979	0.002	0.976	0.982
	Especificidad	Entrenamiento	0.980	0.002	0.977	0.983
	F1	Validación	0.155	0.015	0.133	0.184
	F1	Entrenamiento	0.225	0.009	0.208	0.240
	PRAUC	Validación	0.090	0.018	0.060	0.120
	PRAUC	Entrenamiento	0.152	0.007	0.143	0.171
	ROCAUC	Validación	0.862	0.016	0.837	0.894
	ROCAUC	Entrenamiento	0.967	0.001	0.966	0.968
	Filtración	Validación	0.894	0.012	0.871	0.911
	Filtración	Entrenamiento	0.847	0.008	0.832	0.862
	Subcobertura	Validación	0.713	0.026	0.667	0.750
	Subcobertura	Entrenamiento	0.574	0.027	0.540	0.633

Tabla 24: Métricas con VC de 10 iteraciones – 5° grado de primaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.165	0.012	0.145	0.184
	Precisión	Entrenamiento	0.170	0.002	0.168	0.172
	Sensibilidad	Validación	0.393	0.033	0.356	0.454
	Sensibilidad	Entrenamiento	0.406	0.004	0.399	0.414
	Especificidad	Validación	0.959	0.001	0.956	0.961
	Especificidad	Entrenamiento	0.959	0.001	0.958	0.960
	F1	Validación	0.232	0.017	0.206	0.258
	F1	Entrenamiento	0.240	0.002	0.237	0.243
	PRAUC	Validación	0.110	0.013	0.095	0.134
	PRAUC	Entrenamiento	0.129	0.006	0.124	0.143
	ROCAUC	Validación	0.800	0.019	0.770	0.833
	ROCAUC	Entrenamiento	0.871	0.001	0.870	0.873
	Filtración	Validación	0.835	0.012	0.816	0.855
	Filtración	Entrenamiento	0.830	0.002	0.828	0.832
	Subcobertura	Validación	0.607	0.033	0.546	0.644
	Subcobertura	Entrenamiento	0.594	0.004	0.586	0.601
Norte	Precisión	Validación	0.167	0.014	0.151	0.201
	Precisión	Entrenamiento	0.183	0.006	0.176	0.194
	Sensibilidad	Validación	0.370	0.025	0.324	0.400
	Sensibilidad	Entrenamiento	0.414	0.021	0.379	0.442
	Especificidad	Validación	0.987	0.002	0.984	0.990
	Especificidad	Entrenamiento	0.987	0.001	0.985	0.988
	F1	Validación	0.230	0.014	0.213	0.256
	F1	Entrenamiento	0.254	0.005	0.245	0.261
	PRAUC	Validación	0.124	0.019	0.103	0.153
	PRAUC	Entrenamiento	0.158	0.006	0.149	0.170
	ROCAUC	Validación	0.907	0.009	0.897	0.928
	ROCAUC	Entrenamiento	0.972	0.001	0.971	0.973
	Filtración	Validación	0.833	0.014	0.799	0.849
	Filtración	Entrenamiento	0.817	0.006	0.806	0.824
	Subcobertura	Validación	0.630	0.025	0.600	0.676
	Subcobertura	Entrenamiento	0.586	0.021	0.558	0.621

Sur	Precisión	Validación	0.115	0.028	0.082	0.164
	Precisión	Entrenamiento	0.238	0.009	0.215	0.248
	Sensibilidad	Validación	0.285	0.072	0.200	0.414
	Sensibilidad	Entrenamiento	0.661	0.026	0.634	0.698
	Especificidad	Validación	0.993	0.001	0.992	0.995
	Especificidad	Entrenamiento	0.993	0.000	0.992	0.994
	F1	Validación	0.163	0.039	0.117	0.229
	F1	Entrenamiento	0.350	0.011	0.324	0.366
	PRAUC	Validación	0.079	0.024	0.043	0.113
	PRAUC	Entrenamiento	0.277	0.024	0.235	0.311
	ROCAUC	Validación	0.838	0.052	0.766	0.936
	ROCAUC	Entrenamiento	0.995	0.000	0.994	0.995
	Filtración	Validación	0.885	0.028	0.836	0.918
	Filtración	Entrenamiento	0.762	0.009	0.752	0.785
	Subcobertura	Validación	0.715	0.072	0.586	0.800
	Subcobertura	Entrenamiento	0.339	0.026	0.302	0.366
Centro	Precisión	Validación	0.134	0.029	0.094	0.176
	Precisión	Entrenamiento	0.166	0.007	0.153	0.177
	Sensibilidad	Validación	0.351	0.064	0.253	0.462
	Sensibilidad	Entrenamiento	0.447	0.021	0.420	0.498
	Especificidad	Validación	0.986	0.001	0.984	0.989
	Especificidad	Entrenamiento	0.986	0.001	0.986	0.988
	F1	Validación	0.193	0.039	0.138	0.242
	F1	Entrenamiento	0.242	0.010	0.225	0.257
	PRAUC	Validación	0.093	0.026	0.059	0.135
	PRAUC	Entrenamiento	0.142	0.005	0.135	0.149
	ROCAUC	Validación	0.879	0.022	0.848	0.929
	ROCAUC	Entrenamiento	0.975	0.001	0.974	0.977
	Filtración	Validación	0.866	0.029	0.824	0.906
	Filtración	Entrenamiento	0.834	0.007	0.823	0.847
	Subcobertura	Validación	0.649	0.064	0.538	0.747
	Subcobertura	Entrenamiento	0.553	0.021	0.502	0.297

Oriente	Precisión	Validación	0.217	0.350	0.000	1.000
	Precisión	Entrenamiento	0.211	0.200	0.000	0.500
	Sensibilidad	Validación	0.007	0.015	0.000	0.050
	Sensibilidad	Entrenamiento	0.003	0.003	0.000	0.010
	Especificidad	Validación	1.000	0.000	1.000	1.000
	Especificidad	Entrenamiento	1.000	0.000	0.999	1.000
	F1	Validación	0.014	0.028	0.000	0.093
	F1	Entrenamiento	0.006	0.006	0.000	0.018
	PRAUC	Validación	0.104	0.024	0.074	0.156
	PRAUC	Entrenamiento	0.143	0.005	0.135	0.150
	ROCAUC	Validación	0.867	0.017	0.842	0.901
	ROCAUC	Entrenamiento	0.962	0.001	0.962	0.963
	Filtración	Validación	0.783	0.350	0.000	1.000
	Filtración	Entrenamiento	0.789	0.200	0.500	1.000
	Subcobertura	Validación	0.993	0.015	0.950	1.000
	Subcobertura	Entrenamiento	0.997	0.003	0.990	1.000

Tabla 25: Métricas con VC de 10 iteraciones – 6° grado de primaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.222	0.013	0.193	0.239
	Precisión	Entrenamiento	0.229	0.002	0.225	0.232
	Sensibilidad	Validación	0.377	0.023	0.315	0.400
	Sensibilidad	Entrenamiento	0.388	0.003	0.384	0.394
	Especificidad	Validación	0.961	0.001	0.958	0.963
	Especificidad	Entrenamiento	0.961	0.001	0.960	0.962
	F1	Validación	0.280	0.016	0.239	0.296
	F1	Entrenamiento	0.288	0.002	0.285	0.291
	PRAUC	Validación	0.192	0.015	0.152	0.216
	PRAUC	Entrenamiento	0.208	0.002	0.203	0.212
	ROCAUC	Validación	0.804	0.008	0.786	0.818
	ROCAUC	Entrenamiento	0.850	0.001	0.849	0.852
	Filtración	Validación	0.778	0.013	0.761	0.807
	Filtración	Entrenamiento	0.771	0.002	0.768	0.775
	Subcobertura	Validación	0.623	0.023	0.600	0.685
	Subcobertura	Entrenamiento	0.612	0.003	0.606	0.616
Norte	Precisión	Validación	0.334	0.011	0.319	0.347
	[1] Precisión	Entrenamiento	0.348	0.002	0.345	0.352
	Sensibilidad	Validación	0.529	0.020	0.509	0.567
	Sensibilidad	Entrenamiento	0.558	0.005	0.552	0.568
	Especificidad	Validación	0.962	0.002	0.959	0.965
	Especificidad	Entrenamiento	0.963	0.000	0.962	0.964
	F1	Validación	0.409	0.012	0.393	0.430
	F1	Entrenamiento	0.429	0.002	0.426	0.432
	PRAUC	Validación	0.370	0.023	0.341	0.416
	PRAUC	Entrenamiento	0.395	0.004	0.389	0.400
	ROCAUC	Validación	0.910	0.009	0.896	0.924
	ROCAUC	Entrenamiento	0.929	0.001	0.928	0.930
	Filtración	Validación	0.666	0.011	0.653	0.681
	Filtración	Entrenamiento	0.652	0.002	0.648	0.655
	Subcobertura	Validación	0.471	0.020	0.433	0.491
	Subcobertura	Entrenamiento	0.442	0.005	0.432	0.448

Sur	Precisión	Validación	0.253	0.017	0.229	0.290
	Precisión	Entrenamiento	0.281	0.005	0.272	0.290
	Sensibilidad	Validación	0.518	0.041	0.435	0.581
	Sensibilidad	Entrenamiento	0.597	0.009	0.583	0.613
	Especificidad	Validación	0.980	0.002	0.977	0.984
	Especificidad	Entrenamiento	0.980	0.001	0.979	0.981
	F1	Validación	0.339	0.018	0.300	0.365
	F1	Entrenamiento	0.382	0.004	0.375	0.389
	PRAUC	Validación	0.270	0.023	0.226	0.296
	PRAUC	Entrenamiento	0.340	0.016	0.313	0.370
	ROCAUC	Validación	0.907	0.011	0.887	0.926
	ROCAUC	Entrenamiento	0.973	0.001	0.972	0.974
	Filtración	Validación	0.747	0.017	0.710	0.771
	Filtración	Entrenamiento	0.719	0.005	0.710	0.728
	Subcobertura	Validación	0.482	0.041	0.419	0.565
	Subcobertura	Entrenamiento	0.403	0.009	0.387	0.417
Centro	Precisión	Validación	0.349	0.020	0.313	0.374
	Precisión	Entrenamiento	0.363	0.005	0.356	0.373
	Sensibilidad	Validación	0.455	0.027	0.399	0.505
	Sensibilidad	Entrenamiento	0.472	0.007	0.460	0.482
	Especificidad	Validación	0.979	0.001	0.977	0.981
	Especificidad	Entrenamiento	0.980	0.000	0.979	0.981
	F1	Validación	0.395	0.021	0.351	0.429
	F1	Entrenamiento	0.411	0.005	0.404	0.418
	PRAUC	Validación	0.342	0.028	0.307	0.410
	PRAUC	Entrenamiento	0.368	0.006	0.358	0.378
	ROCAUC	Validación	0.913	0.009	0.898	0.928
	ROCAUC	Entrenamiento	0.944	0.001	0.943	0.945
	Filtración	Validación	0.651	0.020	0.626	0.687
	Filtración	Entrenamiento	0.637	0.005	0.627	0.644
	Subcobertura	Validación	0.545	0.027	0.495	0.601
	Subcobertura	Entrenamiento	0.528	0.007	0.518	0.540

Oriente	Precisión	Validación	0.438	0.013	0.422	0.464
	Precisión	Entrenamiento	0.452	0.002	0.447	0.455
	Sensibilidad	Validación	0.651	0.019	0.625	0.682
	Sensibilidad	Entrenamiento	0.675	0.005	0.665	0.680
	Especificidad	Validación	0.923	0.003	0.917	0.930
	Especificidad	Entrenamiento	0.925	0.001	0.923	0.927
	F1	Validación	0.524	0.014	0.507	0.546
	F1	Entrenamiento	0.541	0.001	0.540	0.545
	PRAUC	Validación	0.507	0.011	0.493	0.526
	PRAUC	Entrenamiento	0.529	0.001	0.526	0.531
	ROCAUC	Validación	0.906	0.005	0.899	0.912
	ROCAUC	Entrenamiento	0.918	0.000	0.917	0.918
	Filtración	Validación	0.562	0.013	0.536	0.578
	Filtración	Entrenamiento	0.548	0.002	0.545	0.553
	Subcobertura	Validación	0.349	0.019	0.318	0.375
	Subcobertura	Entrenamiento	0.325	0.005	0.320	0.335

Tabla 26: Métricas con VC de 10 iteraciones – 1° grado de secundaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.150	0.012	0.132	0.170
	Precisión	Entrenamiento	0.161	0.002	0.159	0.163
	Sensibilidad	Validación	0.304	0.023	0.260	0.340
	Sensibilidad	Entrenamiento	0.329	0.007	0.321	0.342
	Especificidad	Validación	0.965	0.002	0.962	0.968
	Especificidad	Entrenamiento	0.965	0.001	0.964	0.966
	F1	Validación	0.200	0.015	0.175	0.227
	F1	Entrenamiento	0.216	0.002	0.213	0.220
	PRAUC	Validación	0.121	0.019	0.098	0.160
	PRAUC	Entrenamiento	0.138	0.003	0.132	0.144
	ROCAUC	Validación	0.802	0.017	0.775	0.828
	ROCAUC	Entrenamiento	0.864	0.001	0.862	0.866
	Filtración	Validación	0.850	0.012	0.830	0.868
	Filtración	Entrenamiento	0.839	0.002	0.837	0.841
	Subcobertura	Validación	0.696	0.023	0.660	0.740
	Subcobertura	Entrenamiento	0.671	0.007	0.658	0.679
Norte	Precisión	Validación	0.210	0.018	0.184	0.250
	Precisión	Entrenamiento	0.221	0.004	0.214	0.230
	Sensibilidad	Validación	0.303	0.032	0.237	0.365
	Sensibilidad	Entrenamiento	0.328	0.012	0.310	0.349
	Especificidad	Validación	0.984	0.001	0.981	0.986
	Especificidad	Entrenamiento	0.984	0.001	0.982	0.985
	F1	Validación	0.247	0.022	0.212	0.297
	F1	Entrenamiento	0.264	0.005	0.255	0.274
	PRAUC	Validación	0.168	0.023	0.135	0.219
	PRAUC	Entrenamiento	0.191	0.005	0.184	0.202
	ROCAUC	Validación	0.880	0.009	0.868	0.894
	ROCAUC	Entrenamiento	0.939	0.000	0.938	0.940
	Filtración	Validación	0.790	0.018	0.750	0.816
	Filtración	Entrenamiento	0.779	0.004	0.770	0.786
	Subcobertura	Validación	0.697	0.032	0.635	0.763
	Subcobertura	Entrenamiento	0.672	0.012	0.651	0.690

Sur	Precisión	Validación	0.238	0.075	0.095	0.324
	Precisión	Entrenamiento	0.330	0.022	0.295	0.360
	Sensibilidad	Validación	0.195	0.063	0.080	0.280
	Sensibilidad	Entrenamiento	0.296	0.023	0.255	0.340
	Especificidad	Validación	0.997	0.000	0.996	0.998
	Especificidad	Entrenamiento	0.997	0.000	0.997	0.997
	F1	Validación	0.214	0.067	0.087	0.298
	F1	Entrenamiento	0.313	0.022	0.273	0.350
	PRAUC	Validación	0.136	0.039	0.059	0.175
	PRAUC	Entrenamiento	0.272	0.014	0.253	0.291
	ROCAUC	Validación	0.903	0.023	0.868	0.928
	ROCAUC	Entrenamiento	0.990	0.001	0.989	0.991
	Filtración	Validación	0.762	0.075	0.676	0.905
	Filtración	Entrenamiento	0.667	0.022	0.640	0.705
	Subcobertura	Validación	0.805	0.063	0.720	0.920
	Subcobertura	Entrenamiento	0.704	0.023	0.660	0.745
Centro	Precisión	Validación	0.136	0.016	0.012	0.165
	Precisión	Entrenamiento	0.150	0.002	0.146	0.152
	Sensibilidad	Validación	0.328	0.035	0.272	0.407
	Sensibilidad	Entrenamiento	0.368	0.014	0.341	0.383
	Especificidad	Validación	0.978	0.001	0.975	0.979
	Especificidad	Entrenamiento	0.978	0.001	0.976	0.979
	F1	Validación	0.192	0.022	0.159	0.235
	F1	Entrenamiento	0.213	0.003	0.206	0.218
	PRAUC	Validación	0.104	0.013	0.081	0.123
	PRAUC	Entrenamiento	0.131	0.003	0.124	0.134
	ROCAUC	Validación	0.894	0.010	0.878	0.907
	ROCAUC	Entrenamiento	0.952	0.001	0.951	0.953
	Filtración	Validación	0.864	0.016	0.835	0.888
	Filtración	Entrenamiento	0.850	0.002	0.848	0.854
	Subcobertura	Validación	0.672	0.035	0.593	0.728
	Subcobertura	Entrenamiento	0.632	0.014	0.617	0.659

Oriente	Precisión	Validación	0.226	0.024	0.179	0.272
	Precisión	Entrenamiento	0.253	0.010	0.240	0.271
	Sensibilidad	Validación	0.294	0.045	0.234	0.385
	Sensibilidad	Entrenamiento	0.326	0.008	0.308	0.334
	Especificidad	Validación	0.973	0.002	0.971	0.976
	Especificidad	Entrenamiento	0.974	0.001	0.973	0.977
	F1	Validación	0.256	0.032	0.203	0.319
	F1	Entrenamiento	0.285	0.008	0.272	0.296
	PRAUC	Validación	0.197	0.031	0.142	0.246
	PRAUC	Entrenamiento	0.228	0.008	0.214	0.243
	ROCAUC	Validación	0.874	0.011	0.858	0.894
	ROCAUC	Entrenamiento	0.924	0.001	0.922	0.925
	Filtración	Validación	0.774	0.024	0.728	0.821
	Filtración	Entrenamiento	0.747	0.010	0.729	0.760
	Subcobertura	Validación	0.706	0.045	0.615	0.766
	Subcobertura	Entrenamiento	0.674	0.008	0.666	0.692

Tabla 27: Métricas con VC de 10 iteraciones – 2° grado de secundaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.160	0.012	0.142	0.184
	Precisión	Entrenamiento	0.171	0.003	0.168	0.176
	Sensibilidad	Validación	0.366	0.027	0.300	0.394
	Sensibilidad	Entrenamiento	0.393	0.004	0.385	0.400
	Especificidad	Validación	0.958	0.003	0.954	0.963
	Especificidad	Entrenamiento	0.958	0.001	0.957	0.960
	F1	Validación	0.223	0.016	0.193	0.249
	F1	Entrenamiento	0.238	0.003	0.235	0.244
	PRAUC	Validación	0.135	0.016	0.112	0.160
	PRAUC	Entrenamiento	0.158	0.002	0.154	0.162
	ROCAUC	Validación	0.803	0.011	0.785	0.821
	ROCAUC	Entrenamiento	0.868	0.001	0.866	0.869
	Filtración	Validación	0.840	0.012	0.816	0.858
	Filtración	Entrenamiento	0.829	0.003	0.824	0.832
	Subcobertura	Validación	0.634	0.027	0.606	0.700
	Subcobertura	Entrenamiento	0.607	0.004	0.600	0.615
Norte	Precisión	Validación	0.221	0.017	0.199	0.247
	Precisión	Entrenamiento	0.236	0.004	0.229	0.241
	Sensibilidad	Validación	0.314	0.022	0.287	0.346
	Sensibilidad	Entrenamiento	0.339	0.007	0.328	0.351
	Especificidad	Validación	0.976	0.001	0.974	0.977
	Especificidad	Entrenamiento	0.976	0.001	0.975	0.977
	F1	Validación	0.260	0.019	0.238	0.289
	F1	Entrenamiento	0.278	0.004	0.273	0.286
	PRAUC	Validación	0.178	0.017	0.162	0.213
	PRAUC	Entrenamiento	0.200	0.004	0.195	0.209
	ROCAUC	Validación	0.868	0.011	0.841	0.882
	ROCAUC	Entrenamiento	0.914	0.001	0.913	0.915
	Filtración	Validación	0.779	0.017	0.753	0.801
	Filtración	Entrenamiento	0.764	0.004	0.759	0.771
	Subcobertura	Validación	0.689	0.022	0.654	0.713
	Subcobertura	Entrenamiento	0.661	0.007	0.649	0.672

Sur	Precisión	Validación	0.111	0.013	0.088	0.127
	Precisión	Entrenamiento	0.152	0.003	0.148	0.160
	Sensibilidad	Validación	0.455	0.044	0.367	0.526
	Sensibilidad	Entrenamiento	0.649	0.012	0.630	0.669
	Especificidad	Validación	0.968	0.002	0.965	0.972
	Especificidad	Entrenamiento	0.969	0.001	0.967	0.971
	F1	Validación	0.179	0.020	0.142	0.202
	F1	Entrenamiento	0.246	0.004	0.241	0.256
	PRAUC	Validación	0.129	0.017	0.101	0.157
	PRAUC	Entrenamiento	0.203	0.007	0.190	0.216
	ROCAUC	Validación	0.858	0.021	0.823	0.894
	ROCAUC	Entrenamiento	0.973	0.000	0.972	0.973
	Filtración	Validación	0.889	0.013	0.873	0.912
	Filtración	Entrenamiento	0.848	0.003	0.840	0.852
	Subcobertura	Validación	0.545	0.044	0.474	0.633
	Subcobertura	Entrenamiento	0.351	0.012	0.331	0.370
Centro	Precisión	Validación	0.160	0.021	0.130	0.185
	Precisión	Entrenamiento	0.179	0.003	0.174	0.182
	Sensibilidad	Validación	0.328	0.031	0.284	0.381
	Sensibilidad	Entrenamiento	0.373	0.022	0.326	0.393
	Especificidad	Validación	0.973	0.002	0.967	0.976
	Especificidad	Entrenamiento	0.973	0.002	0.971	0.976
	F1	Validación	0.215	0.025	0.178	0.247
	F1	Entrenamiento	0.241	0.005	0.228	0.246
	PRAUC	Validación	0.127	0.018	0.100	0.159
	PRAUC	Entrenamiento	0.154	0.003	0.148	0.158
	ROCAUC	Validación	0.866	0.014	0.843	0.891
	ROCAUC	Entrenamiento	0.927	0.001	0.925	0.929
	Filtración	Validación	0.840	0.021	0.815	0.870
	Filtración	Entrenamiento	0.821	0.003	0.818	0.826
	Subcobertura	Validación	0.672	0.031	0.619	0.716
	Subcobertura	Entrenamiento	0.627	0.022	0.607	0.674

Oriente	Precisión	Validación	0.223	0.020	0.194	0.253
	Precisión	Entrenamiento	0.245	0.005	0.238	0.252
	Sensibilidad	Validación	0.421	0.037	0.360	0.462
	Sensibilidad	Entrenamiento	0.467	0.006	0.460	0.477
	Especificidad	Validación	0.943	0.003	0.937	0.948
	Especificidad	Entrenamiento	0.944	0.002	0.941	0.947
	F1	Validación	0.291	0.026	0.255	0.324
	F1	Entrenamiento	0.321	0.005	0.314	0.326
	PRAUC	Validación	0.219	0.023	0.176	0.250
	PRAUC	Entrenamiento	0.255	0.005	0.248	0.265
	ROCAUC	Validación	0.847	0.009	0.829	0.858
	ROCAUC	Entrenamiento	0.899	0.001	0.897	0.900
	Filtración	Validación	0.777	0.020	0.747	0.806
	Filtración	Entrenamiento	0.755	0.005	0.748	0.762
	Subcobertura	Validación	0.579	0.037	0.538	0.640
	Subcobertura	Entrenamiento	0.533	0.006	0.523	0.540

Tabla 28: Métricas con VC de 10 iteraciones – 3° grado de secundaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.197	0.016	0.167	0.226
	Precisión	Entrenamiento	0.209	0.002	0.206	0.214
	Sensibilidad	Validación	0.343	0.016	0.312	0.371
	Sensibilidad	Entrenamiento	0.369	0.009	0.358	0.384
	Especificidad	Validación	0.959	0.002	0.955	0.963
	Especificidad	Entrenamiento	0.960	0.001	0.957	0.961
	F1	Validación	0.250	0.017	0.218	0.281
	F1	Entrenamiento	0.267	0.003	0.262	0.272
	PRAUC	Validación	0.168	0.013	0.145	0.191
	PRAUC	Entrenamiento	0.186	0.002	0.184	0.189
	ROCAUC	Validación	0.809	0.011	0.792	0.823
	ROCAUC	Entrenamiento	0.858	0.001	0.856	0.859
	Filtración	Validación	0.803	0.016	0.774	0.833
	Filtración	Entrenamiento	0.791	0.002	0.786	0.794
	Subcobertura	Validación	0.657	0.016	0.629	0.688
	Subcobertura	Entrenamiento	0.631	0.009	0.616	0.642
Norte	Precisión	Validación	0.269	0.024	0.235	0.319
	Precisión	Entrenamiento	0.279	0.003	0.273	0.282
	Sensibilidad	Validación	0.326	0.024	0.292	0.376
	Sensibilidad	Entrenamiento	0.341	0.005	0.333	0.349
	Especificidad	Validación	0.968	0.002	0.964	0.973
	Especificidad	Entrenamiento	0.968	0.001	0.967	0.970
	F1	Validación	0.295	0.023	0.260	0.332
	F1	Entrenamiento	0.307	0.003	0.301	0.310
	PRAUC	Validación	0.216	0.015	0.192	0.246
	PRAUC	Entrenamiento	0.237	0.003	0.232	0.244
	ROCAUC	Validación	0.857	0.008	0.843	0.868
	ROCAUC	Entrenamiento	0.889	0.001	0.888	0.890
	Filtración	Validación	0.731	0.024	0.681	0.765
	Filtración	Entrenamiento	0.721	0.003	0.718	0.727
	Subcobertura	Validación	0.674	0.024	0.624	0.708
	Subcobertura	Entrenamiento	0.659	0.005	0.651	0.667

Sur	Precisión	Validación	0.227	0.025	0.185	0.269
	Precisión	Entrenamiento	0.247	0.006	0.240	0.260
	Sensibilidad	Validación	0.289	0.030	0.255	0.336
	Sensibilidad	Entrenamiento	0.317	0.010	0.303	0.334
	Especificidad	Validación	0.984	0.002	0.981	0.988
	Especificidad	Entrenamiento	0.984	0.001	0.983	0.985
	F1	Validación	0.254	0.025	0.219	0.298
	F1	Entrenamiento	0.277	0.005	0.268	0.284
	PRAUC	Validación	0.163	0.025	0.128	0.222
	PRAUC	Entrenamiento	0.205	0.004	0.198	0.212
	ROCAUC	Validación	0.857	0.014	0.829	0.878
	ROCAUC	Entrenamiento	0.943	0.001	0.941	0.944
	Filtración	Validación	0.773	0.025	0.731	0.815
	Filtración	Entrenamiento	0.753	0.006	0.740	0.760
	Subcobertura	Validación	0.711	0.030	0.664	0.745
	Subcobertura	Entrenamiento	0.683	0.010	0.666	0.697
Centro	Precisión	Validación	0.192	0.010	0.176	0.212
	Precisión	Entrenamiento	0.204	0.002	0.200	0.206
	Sensibilidad	Validación	0.411	0.027	0.350	0.448
	Sensibilidad	Entrenamiento	0.439	0.007	0.430	0.447
	Especificidad	Validación	0.954	0.002	0.951	0.956
	Especificidad	Entrenamiento	0.954	0.001	0.952	0.955
	F1	Validación	0.262	0.014	0.235	0.287
	F1	Entrenamiento	0.278	0.002	0.274	0.281
	PRAUC	Validación	0.161	0.013	0.145	0.189
	PRAUC	Entrenamiento	0.181	0.003	0.176	0.188
	ROCAUC	Validación	0.853	0.010	0.832	0.871
	ROCAUC	Entrenamiento	0.900	0.001	0.898	0.902
	Filtración	Validación	0.808	0.010	0.788	0.824
	Filtración	Entrenamiento	0.796	0.002	0.794	0.800
	Subcobertura	Validación	0.589	0.027	0.552	0.650
	Subcobertura	Entrenamiento	0.561	0.007	0.553	0.570

Oriente	Precisión	Validación	0.324	0.012	0.307	0.345
	Precisión	Entrenamiento	0.348	0.006	0.339	0.359
	Sensibilidad	Validación	0.356	0.016	0.328	0.380
	Sensibilidad	Entrenamiento	0.391	0.004	0.380	0.397
	Especificidad	Validación	0.954	0.003	0.948	0.959
	Especificidad	Entrenamiento	0.954	0.001	0.953	0.957
	F1	Validación	0.339	0.012	0.321	0.362
	F1	Entrenamiento	0.368	0.003	0.363	0.374
	PRAUC	Validación	0.298	0.013	0.279	0.320
	PRAUC	Entrenamiento	0.332	0.006	0.322	0.343
	ROCAUC	Validación	0.844	0.009	0.827	0.853
	ROCAUC	Entrenamiento	0.879	0.001	0.878	0.881
	Filtración	Validación	0.676	0.012	0.655	0.693
	Filtración	Entrenamiento	0.652	0.006	0.641	0.661
	Subcobertura	Validación	0.644	0.016	0.620	0.672
	Subcobertura	Entrenamiento	0.609	0.004	0.603	0.620

Tabla 29: Métricas con VC de 10 iteraciones – 4° grado de secundaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.142	0.007	0.133	0.159
	Precisión	Entrenamiento	0.154	0.001	0.153	0.157
	Sensibilidad	Validación	0.389	0.023	0.356	0.425
	Sensibilidad	Entrenamiento	0.424	0.003	0.417	0.430
	Especificidad	Validación	0.938	0.002	0.934	0.941
	Especificidad	Entrenamiento	0.939	0.001	0.937	0.940
	F1	Validación	0.208	0.011	0.194	0.231
	F1	Entrenamiento	0.226	0.001	0.225	0.229
	PRAUC	Validación	0.133	0.009	0.117	0.143
	PRAUC	Entrenamiento	0.153	0.001	0.151	0.155
	ROCAUC	Validación	0.803	0.007	0.795	0.818
	ROCAUC	Entrenamiento	0.857	0.001	0.855	0.859
	Filtración	Validación	0.858	0.007	0.841	0.867
	Filtración	Entrenamiento	0.846	0.001	0.843	0.847
	Subcobertura	Validación	0.611	0.023	0.575	0.644
	Subcobertura	Entrenamiento	0.576	0.003	0.570	0.583
Norte	Precisión	Validación	0.216	0.012	0.190	0.237
	Precisión	Entrenamiento	0.232	0.003	0.228	0.237
	Sensibilidad	Validación	0.411	0.022	0.377	0.449
	Sensibilidad	Entrenamiento	0.441	0.006	0.433	0.453
	Especificidad	Validación	0.943	0.003	0.938	0.948
	Especificidad	Entrenamiento	0.944	0.001	0.943	0.945
	F1	Validación	0.283	0.015	0.253	0.310
	F1	Entrenamiento	0.304	0.004	0.299	0.309
	PRAUC	Validación	0.205	0.015	0.174	0.231
	PRAUC	Entrenamiento	0.223	0.003	0.218	0.229
	ROCAUC	Validación	0.852	0.008	0.840	0.867
	ROCAUC	Entrenamiento	0.881	0.001	0.880	0.882
	Filtración	Validación	0.784	0.012	0.763	0.810
	Filtración	Entrenamiento	0.768	0.003	0.763	0.772
	Subcobertura	Validación	0.589	0.022	0.551	0.623
	Subcobertura	Entrenamiento	0.559	0.006	0.547	0.567

Sur	Precisión	Validación	0.197	0.026	0.156	0.238
	Precisión	Entrenamiento	0.210	0.004	0.201	0.216
	Sensibilidad	Validación	0.269	0.044	0.203	0.350
	Sensibilidad	Entrenamiento	0.289	0.012	0.272	0.312
	Especificidad	Validación	0.981	0.001	0.979	0.984
	Especificidad	Entrenamiento	0.981	0.001	0.979	0.983
	F1	Validación	0.227	0.032	0.176	0.283
	F1	Entrenamiento	0.243	0.004	0.238	0.251
	PRAUC	Validación	0.137	0.022	0.105	0.175
	PRAUC	Entrenamiento	0.175	0.004	0.168	0.181
	ROCAUC	Validación	0.843	0.012	0.824	0.866
	ROCAUC	Entrenamiento	0.933	0.000	0.933	0.934
	Filtración	Validación	0.803	0.026	0.762	0.844
	Filtración	Entrenamiento	0.790	0.004	0.784	0.799
	Subcobertura	Validación	0.731	0.044	0.650	0.797
	Subcobertura	Entrenamiento	0.711	0.012	0.688	0.728
Centro	Precisión	Validación	0.160	0.008	0.148	0.173
	Precisión	Entrenamiento	0.173	0.002	0.171	0.179
	Sensibilidad	Validación	0.384	0.024	0.350	0.422
	Sensibilidad	Entrenamiento	0.418	0.005	0.409	0.430
	Especificidad	Validación	0.945	0.002	0.943	0.950
	Especificidad	Entrenamiento	0.946	0.001	0.943	0.948
	F1	Validación	0.226	0.011	0.208	0.245
	F1	Entrenamiento	0.245	0.002	0.243	0.250
	PRAUC	Validación	0.136	0.008	0.126	0.155
	PRAUC	Entrenamiento	0.157	0.003	0.152	0.163
	ROCAUC	Validación	0.840	0.008	0.829	0.857
	ROCAUC	Entrenamiento	0.887	0.000	0.886	0.888
	Filtración	Validación	0.840	0.008	0.827	0.852
	Filtración	Entrenamiento	0.827	0.002	0.821	0.829
	Subcobertura	Validación	0.616	0.024	0.578	0.650
	Subcobertura	Entrenamiento	0.582	0.005	0.570	0.591

Oriente	Precisión	Validación	0.291	0.014	0.270	0.321
	Precisión	Entrenamiento	0.315	0.004	0.306	0.321
	Sensibilidad	Validación	0.451	0.028	0.413	0.511
	Sensibilidad	Entrenamiento	0.493	0.006	0.485	0.506
	Especificidad	Validación	0.926	0.004	0.920	0.934
	Especificidad	Entrenamiento	0.928	0.001	0.925	0.931
	F1	Validación	0.353	0.017	0.329	0.379
	F1	Entrenamiento	0.384	0.003	0.377	0.390
	PRAUC	Validación	0.306	0.020	0.280	0.340
	PRAUC	Entrenamiento	0.346	0.004	0.339	0.352
	ROCAUC	Validación	0.838	0.009	0.821	0.853
	ROCAUC	Entrenamiento	0.874	0.001	0.873	0.875
	Filtración	Validación	0.709	0.014	0.679	0.730
	Filtración	Entrenamiento	0.685	0.004	0.679	0.694
	Subcobertura	Validación	0.549	0.028	0.489	0.587
	Subcobertura	Entrenamiento	0.507	0.006	0.494	0.515

Tabla 30: Métricas con VC de 10 iteraciones – 5° grado de secundaria y macro región

Macro región	Métrica	Datos	Promedio	DE	Min.	Max.
lima_metro_callao	Precisión	Validación	0.067	0.021	0.039	0.117
	Precisión	Entrenamiento	0.124	0.020	0.092	0.156
	Sensibilidad	Validación	0.292	0.081	0.167	0.400
	Sensibilidad	Entrenamiento	0.557	0.028	0.513	0.596
	Especificidad	Validación	0.992	0.001	0.990	0.994
	Especificidad	Entrenamiento	0.992	0.001	0.989	0.994
	F1	Validación	0.109	0.032	0.065	0.180
	F1	Entrenamiento	0.202	0.028	0.156	0.247
	PRAUC	Validación	0.046	0.019	0.018	0.077
	PRAUC	Entrenamiento	0.105	0.008	0.087	0.115
	ROCAUC	Validación	0.809	0.033	0.750	0.850
	ROCAUC	Entrenamiento	0.986	0.001	0.984	0.988
	Filtración	Validación	0.933	0.021	0.883	0.961
	Filtración	Entrenamiento	0.876	0.020	0.844	0.908
	Subcobertura	Validación	0.708	0.081	0.600	0.833
	Subcobertura	Entrenamiento	0.443	0.028	0.404	0.487
Norte	Precisión	Validación	0.186	0.051	0.108	0.263
	Precisión	Entrenamiento	0.210	0.015	0.193	0.240
	Sensibilidad	Validación	0.266	0.050	0.198	0.337
	Sensibilidad	Entrenamiento	0.319	0.021	0.280	0.366
	Especificidad	Validación	0.991	0.002	0.988	0.994
	Especificidad	Entrenamiento	0.991	0.001	0.990	0.993
	F1	Validación	0.218	0.050	0.140	0.287
	F1	Entrenamiento	0.253	0.013	0.233	0.280
	PRAUC	Validación	0.131	0.035	0.097	0.224
	PRAUC	Entrenamiento	0.191	0.012	0.175	0.222
	ROCAUC	Validación	0.926	0.013	0.902	0.953
	ROCAUC	Entrenamiento	0.977	0.001	0.976	0.978
	Filtración	Validación	0.814	0.051	0.737	0.892
	Filtración	Entrenamiento	0.790	0.015	0.760	0.807
	Subcobertura	Validación	0.734	0.050	0.663	0.802
	Subcobertura	Entrenamiento	0.842	0.024	0.806	0.881

Sur	Precisión	Validación	0.248	0.066	0.175	0.391
	Precisión	Entrenamiento	0.471	0.035	0.394	0.526
	Sensibilidad	Validación	0.286	0.086	0.174	0.435
	Sensibilidad	Entrenamiento	0.719	0.037	0.649	0.774
	Especificidad	Validación	0.998	0.001	0.996	0.998
	Especificidad	Entrenamiento	0.998	0.000	0.997	0.998
	F1	Validación	0.263	0.070	0.186	0.391
	F1	Entrenamiento	0.568	0.033	0.490	0.615
	PRAUC	Validación	0.168	0.052	0.096	0.269
	PRAUC	Entrenamiento	0.556	0.049	0.483	0.633
	ROCAUC	Validación	0.903	0.031	0.846	0.946
	ROCAUC	Entrenamiento	0.998	0.000	0.998	0.999
	Filtración	Validación	0.752	0.066	0.609	0.825
	Filtración	Entrenamiento	0.529	0.035	0.474	0.606
	Subcobertura	Validación	0.714	0.086	0.565	0.826
	Subcobertura	Entrenamiento	0.281	0.037	0.226	0.351
Centro	Precisión	Validación	0.222	0.033	0.167	0.273
	Precisión	Entrenamiento	0.377	0.053	0.252	0.456
	Sensibilidad	Validación	0.293	0.045	0.222	0.361
	Sensibilidad	Entrenamiento	0.591	0.030	0.546	0.646
	Especificidad	Validación	0.997	0.001	0.995	0.998
	Especificidad	Entrenamiento	0.997	0.001	0.994	0.998
	F1	Validación	0.251	0.032	0.190	0.296
	F1	Entrenamiento	0.458	0.046	0.353	0.520
	PRAUC	Validación	0.131	0.039	0.093	0.229
	PRAUC	Entrenamiento	0.320	0.045	0.211	0.398
	ROCAUC	Validación	0.918	0.020	0.891	0.958
	ROCAUC	Entrenamiento	0.995	0.001	0.992	0.996
	Filtración	Validación	0.778	0.033	0. 727	0.833
	Filtración	Entrenamiento	0.623	0.053	0.544	0.748
	Subcobertura	Validación	0.707	0.045	0.639	0.778
	Subcobertura	Entrenamiento	0.409	0.030	0.354	0.454

Oriente	Precisión	Validación	0.311	0.046	0.258	0.403
	Precisión	Entrenamiento	0.436	0.023	0.405	0.476
	Sensibilidad	Validación	0.392	0.054	0.313	0.484
	Sensibilidad	Entrenamiento	0.648	0.014	0.627	0.665
	Especificidad	Validación	0.989	0.002	0.986	0.992
	Especificidad	Entrenamiento	0.990	0.001	0.989	0.991
	F1	Validación	0.345	0.044	0.301	0.440
	F1	Entrenamiento	0.521	0.017	0.496	0.554
	PRAUC	Validación	0.272	0.058	0.173	0.379
	PRAUC	Entrenamiento	0.486	0.029	0.446	0.540
	ROCAUC	Validación	0.926	0.015	0.902	0.951
	ROCAUC	Entrenamiento	0.989	0.001	0.988	0.990
	Filtración	Validación	0.689	0.046	0.597	0.742
	Filtración	Entrenamiento	0.564	0.023	0.524	0.595
	Subcobertura	Validación	0.608	0.054	0.516	0.688
	Subcobertura	Entrenamiento	0.352	0.014	0.335	0.373

MINISTERIO DE EDUCACIÓN
Oficina de Seguimiento y Evaluación Estratégico
Octubre - 2024
Sede Central: Calle Del Comercio N° 193 Lima - Lima - San Borja - 15021 Perú
Teléfono: (511) 615-5800
https://www.gob.pe/minedu

ISBN: 978-9972-246-90-6