INFORME

Alerta Escuela: Machine Learning para el cálculo del riesgo de interrupción de estudios en el Perú.

Oficina de Seguimiento y Evaluación Estratégica (OSEE)

PERÚ Ministerio de Educación

Diciembre 2022

Jefe de la Oficina de Seguimiento y Evaluación Estratégica

Juan Manuel García Carpio

Elaboración de contenidos:

- Erik Carl Candela Rojas
- Cristian Doménico Centeno Guzmán

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-CIUP
- IPA Perú

.

MINISTERIO DE EDUCACIÓN

Oficina de Seguimiento y Evaluación Estratégico

Diciembre - 2022

Sede Central: Calle Del Comercio N° 193 Lima - Lima - San Borja - 15021 Perú

Teléfono: (01) 615-5800

https://www.gob.pe/minedu

ALERTA ESCUELA: MACHINE LEARNING PARA EL CÁLCULO DEL RIESGO DE INTERRUPCIÓN DE ESTUDIOS EN EL PERÚ.

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 contribuyo en las discusiones de revisión del diseño inicial del modelo y a Annie Chumpitaz por sus valiosos comentarios.

RESUMEN

El presente informe describe la metodología utilizada para desarrollar un modelo basado en técnicas de *Machine Learning* (ML) que calcula el riesgo de interrupción de estudios 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: Interrupción de Estudios, *Machine Learning*, Alerta Escuela

ABSTRACT

This report describes the methodology used to develop a Machine Learning (ML) model that estimate the dropout risk of students enrolled in Regular Basic Education (EBR) for a given school year in Perú. 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

ÍNDICE DE CONTENIDO

RESUMEN II
ABSTRACT III
ÍNDICE DE CONTENIDO IV
ÍNDICE DE TABLAS VI
ÍNDICE DE FIGURAS VIII
INTRODUCCIÓN 9
CAPÍTULO I: COMPRENSIÓN DEL PROBLEMA 11
1.1 Determinación del objetivo que busca la gestión 11
1.2 Evaluación de la situación inicial 11
1.3 Determinación de los objetivos de minería de datos 12
1.4 Revisión de la literatura 12
CAPÍTULO II: COMPRENSIÓN DE LOS DATOS 15
2.1 Recopilación de datos 15
2.2 Descripción de datos 16
2.3 Exploración de los datos 16
2.3.1 Evolución de la interrupción de estudios por nivel. 17
2.4 Verificación de calidad de los datos 21
2.4.1 Balanceo de datos 21
2.4.2 Outliers 21
CAPÍTULO III: PREPARACIÓN DE LOS DATOS 22
3.1 Selección de variables 22
3.2 Limpieza de los datos 23
3.2.1 Eliminar observaciones duplicadas 23
3.2.2 Filtrar valores atípicos no deseados 23
3.2.3 Corregir errores estructurales 24
3.2.4 Corregir los datos faltantes 24
3.2.5 Validar datos 24
3.3 Construcción de nuevos datos derivados. 24

3.4 Integración de los datos 25
3.5 Formato de datos 25
CAPÍTULO IV: MODELADO 26
4.1 Selección de técnicas de modelado 26
4.1.1 Tipo de *Machine Learning* 26
4.1.2 Algoritmos de *Machine Learning* 27
4.1.3 Métricas de desempeño 28
4.2 Generación de un diseño de comprobación 28
4.2.1 Selección del algoritmo de *Machine Learning* 29
4.2.2 Configuración del Modelo: 30
4.2.3 Estimación del Modelo 31
4.3 Evaluación del modelo 35
4.3.1 Evaluación de resultados del nivel inicial 36
4.3.2 Evaluación de resultados del nivel Primaria 38
4.3.3 Evaluación de resultados del nivel Secundaria 40
CAPÍTULO V: DESPLIEGUE 42
CONCLUSIONES 45
LÍNEAS DE MEJORA 46
BIBLIOGRAFÍA 48
ANEXOS 51
ANEXO 1: Machine Learning (ML) 51
ANEXO 2: Diccionario de datos 52
ANEXO 3: Macro regiones 54
ANEXO 4: Criterios de selección de métricas de desempeño 55
ANEXO 5: Cálculo de métricas 57
ANEXO 6: Criterios para la división de datos en entrenamiento y validación 59
ANEXO 7: División de datos en entrenamiento y validación 61
ANEXO 8: Hiperparámetros 64
ANEXO 9: Métricas por grado y macro región 67

ÍNDICE DE TABLAS

Tabla 1 Fuentes de información 15

Tabla 2 Grupos de variables..... 16

Tabla 3 Niveles y Grados..... 18

Tabla 4 Macro regiones 19

Tabla 5 Tasa de Interrupción de Estudios (%) por grado y macro región -2020-2021 20

Tabla 6 Resultado de evaluación mediante validación cruzada con 10 iteraciones 29

Tabla 7 Total de Estudiantes por grado y macro región, 2020 30

Tabla 8 Validación cruzada con 10 iteraciones para cada nivel educativo 35

Tabla 9 Métricas con validación cruzada de 10 iteraciones – Nivel Inicial..... 36

Tabla 10 Métricas con validación cruzada de 10 iteraciones – Nivel Primaria 38

Tabla 11 Métricas con validación cruzada de 10 iteraciones – Nivel Secundaria..... 40

Tabla 12 Diccionario de datos 52

Tabla 13 Matriz de confusión..... 57

Tabla 14 Validación cruzada con 10 iteraciones para el Nivel Inicial 61

Tabla 15 Validación cruzada con 10 iteraciones para el Nivel Primaria..... 62

Tabla 16 Validación cruzada con 10 iteraciones para el Nivel Secundaria..... 63

Tabla 17 Hiperparámetros para configuración automática 64

Tabla 18 Hiperparámetros para configuración manual 64

Tabla 19 Métricas con VC de 10 iteraciones – Ciclo 2 del nivel Inicial y macro región ... 67

Tabla 20 Métricas con VC de 10 iteraciones – 1° grado de primaria y macro región 69

Tabla 21 Métricas con VC de 10 iteraciones – 2° grado de primaria y macro región 71

Tabla 22 Métricas con VC de 10 iteraciones – 3° grado de primaria y macro región 73

Tabla 23 Métricas con VC de 10 iteraciones – 4° grado de primaria y macro región 75

Tabla 24 Métricas con VC de 10 iteraciones – 5° grado de primaria y macro región 77

Tabla 25 Métricas con VC de 10 iteraciones – 6° grado de primaria y macro región 79

Tabla 26 Métricas con VC de 10 iteraciones – 1° grado de secundaria y macro región . 81

Tabla 27 Métricas con VC de 10 iteraciones – 2° grado de secundaria y macro región . 83

Tabla 28 Métricas con VC de 10 iteraciones – 3° grado de secundaria y macro región . 85

Tabla 29 Métricas con VC de 10 iteraciones – 4° grado de secundaria y macro región . 87

Tabla 30 Métricas con VC de 10 iteraciones – 5° grado de secundaria y macro región . 89

ÍNDICE DE FIGURAS

Figura 1 Evolución de la interrupción de estudios 18

Figura 2 Evolución de la interrupción de estudios por grado 19

Figura 3 Tasa de Interrupción de Estudios por grado y macro región – 2020-2021 20

Figura 4 Importancia de Variables – Nivel Inicial 32

Figura 5 Importancia de Variables – Primaria 33

Figura 6 Importancia de Variables – Secundaria 34

Figura 7 Curva ROC - Inicial (10 iteraciones) 37

Figura 8 Curva PR – Inicial (10 iteraciones) 37

Figura 9 Curva ROC – Primaria (10 iteraciones) 39

Figura 10 Curva PR - Primaria (10 iteraciones) 39

Figura 11 Curva ROC - Secundaria (10 iteraciones) 41

Figura 12 Curva PR - Secundaria (10 iteraciones) 41

Figura 13 Despliegue de los resultados en el sistema «Alerta Escuela» 43

Figura 14 Momento de envió de resultados hacia el sistema «Alerta Escuela» 44

Figura 15 Modelo de aprendizaje supervisado 51

Figura 16 Macro regiones 54

Figura 17 Curva Receiver Operating Characteristic (ROC) 58

Figura 18 Curva Precisión Recall (PR) 58

Figura 19 Validación cruzada de 10 iteraciones 60

INTRODUCCIÓN

El problema de la interrupción de estudios reviste una gran importancia por el alto costo que tendrá en los estudiantes de educación básica que abandonan la escuela, ya que no podrán desempeñarse exitosamente en los mercados laborales y su vida en sociedad (CAF, 2018). Este problema se encuentra presente en el sistema educativo peruano, aunque el Perú ha logrado importantes avances[2] en los últimos años. Sin embargo, a inicios del 2020 se declaró la emergencia sanitaria por el COVID-19, el cual agravo la crisis económica y social del país. Producto de la emergencia, se previó que su impacto agudizaría el problema de interrupción de estudios.

Como respuesta a este nuevo escenario, el Ministerio de Educación (MINEDU) lanzó en septiembre del 2020 la campaña comunicacional «Movilización nacional contra la deserción escolar y la promoción del retorno al servicio educativo»[3], la cual tuvo la participación de representantes de otros sectores del Poder Ejecutivo, agencias internacionales de cooperación, academia y sociedad civil (MINEDU, 2020b). En dicha campaña, se hizo énfasis en la «deserción escolar» y se propuso cambiar el término por «interrupción de estudios»[4]. Incluso se resaltó que esta interrupción es la consecuencia de múltiples factores sociales que van más allá de la decisión individual de abandonar la escuela Para efectos del presente informe, se tomó en cuenta la propuesta de emplear el término «interrupción de estudios» en lugar de «deserción escolar».

En ese sentido, un aspecto importante es la identificación de los estudiantes con riesgo en interrumpir sus estudios mediante técnicas que hagan uso de la información disponible del sector educativo. Existen diversos estudios que muestran el empleo de técnicas de *Machine Learning*[5] (ML) para la identificación de estudiantes vulnerables con mayor riesgo de interrumpir sus estudios. Los niveles de riesgos 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 informe tiene como objetivo detallar la metodología que se empleó para el cálculo del riesgo de interrupción de estudios de los estudiantes de Educación Básica

[2] Durante los últimos años previos a la pandemia del COVID 19, la tasa de interrupción de estudios en el Perú ha presentado una tendencia a la baja (MINEDU, 2019).
[3] La campaña fue transmitida on-line y se encuentra disponible en el siguiente enlace: https://www.facebook.com/mineduperu/videos/648110396139204/
[4] El cambio se sustenta desde el minuto 49:50 de la transmisión y se emplea por primera vez en el minuto 56:33 (MINEDU, 2020b).
[5] Ver Anexo 1 «Machine Learning (ML)».

Regular (EBR) del Perú. En ese sentido, se describe 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 informe 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 detalla 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.

CAPÍTULO I: COMPRENSIÓN DEL PROBLEMA

1.1 Determinación del objetivo que busca la gestión

El Ministerio de Educación busca promover la continuidad educativa de los estudiantes en la educación básica con énfasis en la población más vulnerable y en el contexto de la emergencia sanitaria provocada por el COVID-19. Para ello, es importante poder identificar estudiantes con mayor riesgo de interrumpir sus estudios 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 interrumpir sus estudios 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)[6] 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 personal capacitado 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 interrumpe sus estudios será aquel que se encuentra matriculado en el año T y no se matriculara en el año T+1, excluyendo aquellos que en el año T fallecieron o aprobaron el 5° grado de secundaria. Es decir, se utiliza la misma definición de interrupción de estudios 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 informe.

[6] 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 interrupción de estudios de cada estudiante de educación básica regular.

1.4 Revisión de la literatura

Existen diversas investigaciones que han analizado la interrupción de estudios de los alumnos y alumnas 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 metodologías para su predicción.

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

En el ámbito internacional, se revisó la investigación realizada por Rumberger y Lim (2008), la cual es una extensa revisión de 25 años de estudios sobre deserción escolar. En esta revisión encontraron que la deserción escolar está relacionada con cuatro grupos de factores:

- 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, 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 con 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 relacionadas 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 interrupción de estudios 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 la zona 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 e interrupción de estudios 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 valorizacion 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 interrupción de estudios 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.

Por otro lado, se cuenta con investigaciones enfocadas al desarrollo de técnicas o metodologías para predecir la interrupción de estudios. A continuación se describen las investigaciones internacionales[7] que se analizaron para el presente informe:

En primer lugar, se cuenta con el estudio realizado por Adelman et al. (2017) quienes estimaron modelos de alerta temprana sobre interrupción de estudios a partir de datos administrativos del Ministerio de Educación de Guatemala y la Secretaria 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 puede identificar el 80% de los estudiantes de 6to grado de primaria que van a desertar en su transición a educación secundaria.

Asimismo, Sansone (2017) realizo un estudio para predecir la interrupción de estudios empleando información del noveno grado. El investigador enfatiza que el uso de predictores que emplean técnicas de *Big Data* y *Machine Learning* mejoran la detección de estudiantes que abandonaran 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.

Por último, 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.

[7] A la fecha de presentación del presente informe, no se encontraron estudios que aborden el desarrollo de metodologías para predecir la interrupción de estudios a nivel nacional.

CAPÍTULO II: COMPRENSIÓN DE LOS DATOS

2.1 Recopilación de datos

Se ha trabajado con 5 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 empleó las proyecciones de ingreso de los hogares de cada estudiante (información generada por la UE). Todas las fuentes de información se encuentran descritas en la tabla 1.

Tabla 1

Fuentes de información

Fuente	Descripción
SIAGIE	Datos de matrícula y evaluaciones de los estudiantes de EBR de los años 2014, 2015, 2016, 2017, 2018, 2019, 2020 y 2021.
ESCALE	Información de servicios educativos de los años 2019 y 2020
ECE	Evaluación Censal de estudiantes del año 2018
NEXUS	Información de control de plazas de los años 2015, 2016, 2017, 2018, 2019 y 2020.
JUNTOS	Verificación de cumplimiento de corresponsabilidad de los estudiantes de los años 2014, 2015, 2016, 2017, 2018, 2019 y 2020.
Ingresos	Proyecciones de ingresos económicos al 2021 de los estudiantes EBR matriculados en 2020. Los ingresos se proyectaron en función a la variación de los ingresos mensuales en 2009-2020 y la dinámica de recuperación del índice de actividad económica para 2021

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 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 del 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)

Exploración de los datos

La exploración de los datos se realizó en función a la variable que representa la interrupción de estudios. Como se vio en el punto 1.2, la interrupción de estudios 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 interrupción de estudios interanual.

Por consiguiente, el concepto de interrupción de estudios 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 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 interrupción de estudios bajo la siguiente notación:

$$Interrupción\ de\ Estudios = [M_t - (M_t \cap M_{t+1}) - A5S_t - F_t]$$

i. $M_t = Matriculados\ en\ el\ año\ t$
ii. $M_t \cap M_{t+1} = Matriculados\ en\ el\ año\ t\ y\ t+1$
iii. $A5S_t = Alumnos\ que\ aprobaron\ el\ quinto\ año\ de\ secundaria\ t\ y\ t+1$
iv. $F_t = Alumnos\ que\ fallecieron\ en\ el\ año\ t$

Y se define la tasa de interrupción de estudios como:

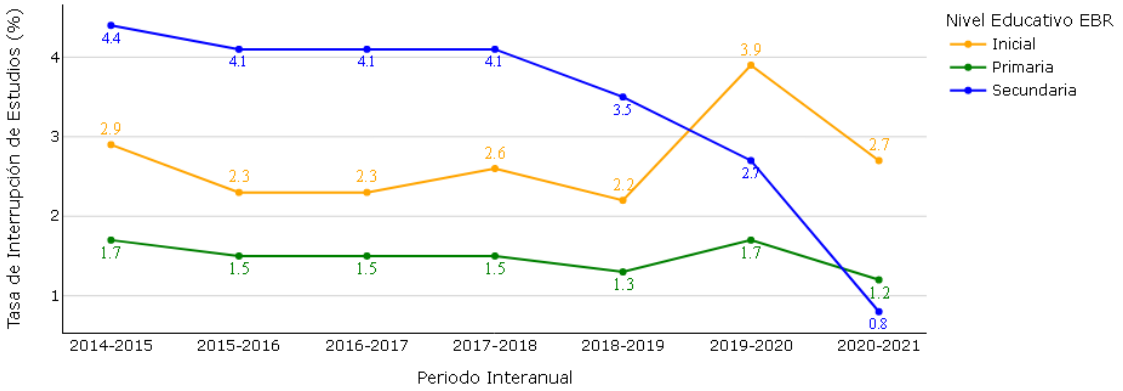
- $Tasa\ de\ interrupción\ de\ estudios = \frac{Interrupción\ de\ Estudios}{M_t}$

2.3.1 Evolución de la interrupción de estudios por nivel.

La figura 1 muestra las tendencias de interrupción de estudios por cada nivel educativo de EBR. Por un lado, se muestra una tendencia a la baja de la tasa de interrupción de estudios del nivel secundaria, mientras que la tasa de interrupción de estudios del nivel primaria no presenta una tendencia muy clara.

Figura 1

Evolución de la interrupción de estudios



Nota. El gráfico representa la evolución de la tasa de interrupción de estudios por nivel educativo EBR.

Un aspecto importante a resaltar de la figura 1 es que las tasas de interrupción de estudios de los niveles de inicial y primaria del periodo interanual 2019-2020 se incrementaron, muy probablemente por las consecuencias de la emergencia sanitaria del covid-19 (Cueto, Felipe, & León, 2020). A diferencia de la tasa de interrupción de estudios del nivel secundaria que continuo con su tendencia a la baja.

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

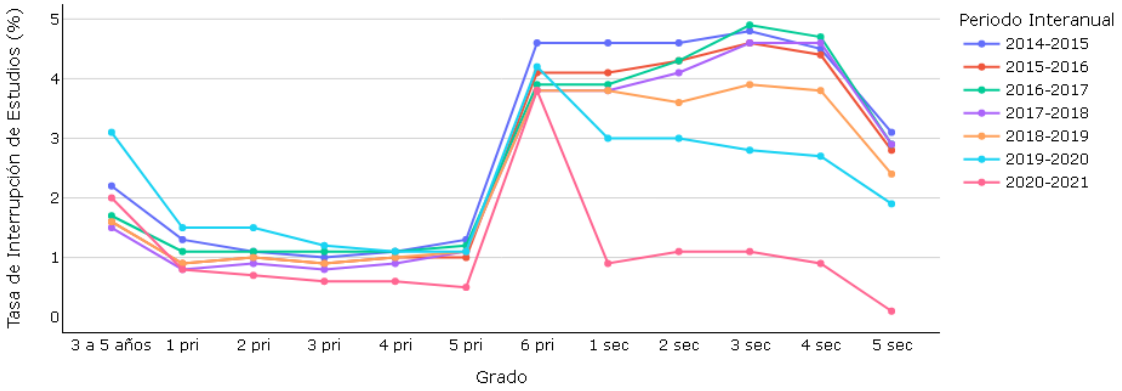
Nivel	Grados
Inicial[a]	Ciclo 2 (3 a 5 años)
Primaria	1ero, 2do, 3ro, 4to, 5to y 6to
Secundaria	1ero, 2do, 3ro, 4to y 5to

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 interrupción de estudios en los distintos periodos interanuales comprendidos desde 2014-2015 hasta 2020-2021 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 interrupción de estudios por grado



Nota. El gráfico representa la evolución histórica de la tasa de interrupción de estudios por cada grado de educación básica.

Entre los distintos periodos interanuales se tomó el periodo interanual 2020-2021[8] como referencia para analizar la interrupción de estudios a mayor detalle, por ser el último periodo con información de matrícula registrada al 100% en el SIAGIE. Además, para el análisis del periodo 2020-2021 se incorporó la variable de macro región, el cual es una agrupación de regiones del Perú como se detallada 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.

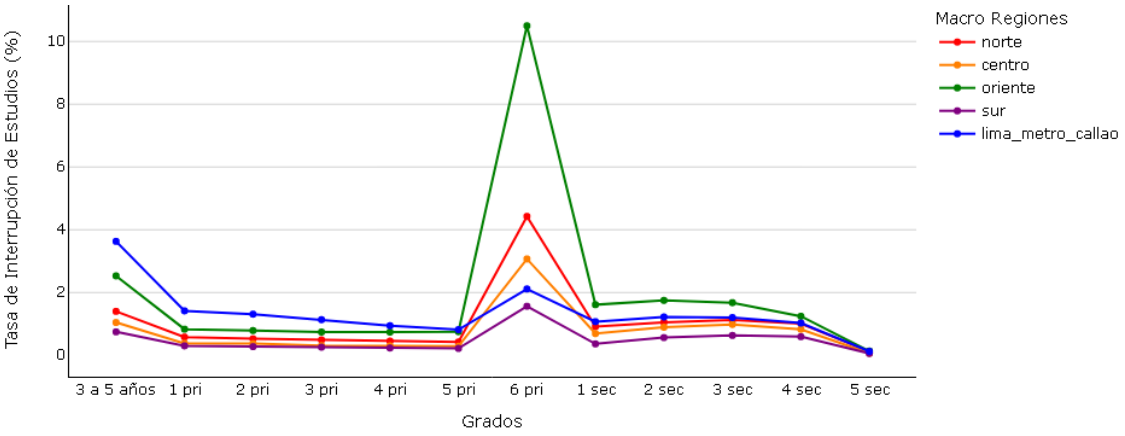
Nota. Los criterios de agrupación se encuentran en el Anexo 3 «Macro regiones».

Tomando en cuenta lo descrito en la tabla 3 y 4, se pudo desagregar la tasa de interrupción de estudios por grados y macro regiones para el periodo interanual 2020-2021, 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 el nivel secundaria, mientras que en el nivel primaria la macro región «lima_metro_callao» es la que resalta.

[8] Estudiantes matriculados en 2020 que interrumpen sus estudios en 2021

Figura 3

Tasa de Interrupción de Estudios por grado y macro región – 2020-2021



Nota. El gráfico representa la tasa de interrupción de estudios por cada grado de EBR y macro región del periodo interanual 2020-2021.

Tabla 5

Tasa de Interrupción de Estudios (%) por grado y macro región -2020-2021

Grado	Norte	Centro	Oriente	Sur	Lima_metro y Callao
Ciclo II	1.40	1.05	2.53	0.75	3.63
1 Primaria	0.58	0.38	0.83	0.30	1.42
2 Primaria	0.52	0.37	0.79	0.28	1.31
3 Primaria	0.49	0.31	0.74	0.26	1.13
4 Primaria	0.45	0.30	0.74	0.23	0.94
5 Primaria	0.42	0.29	0.75	0.22	0.82
6 Primaria	4.43	3.07	10.50	1.56	2.11
1 Secundaria	0.91	0.69	1.61	0.36	1.07
2 Secundaria	1.05	0.89	1.75	0.56	1.22
3 Secundaria	1.13	0.98	1.68	0.63	1.20
4 Secundaria	1.01	0.83	1.24	0.59	1.03
5 Secundaria	0.12	0.07	0.13	0.06	0.12

Nota. Esta tabla muestra la tasa de interrupción de estudios por cada grado de educación básica y macro región del periodo interanual 2020-2021.

Contar con datos de la tasa alta de interrupción de estudios 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 interrumpen sus estudios es muy menor (2.3% aproximadamente) en comparación con el total de estudiantes. Esto trae como consecuencia que exista pocos casos para analizar la interrupción de estudios.

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 interrumpen sus estudios. Para el periodo interanual de referencia, solo hay 2.3% de estudiantes matriculados en el 2020 que no se matricularan en el 2021. 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 esta 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 interrupción de estudios. 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.

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[9]. 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 un único valor o ser redundantes; en la tercera etapa se seleccionaron 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 interrupción de estudios 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 interrupción de estudios.

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

[9] 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.

variables significativas, se realizó un proceso de validación cruzada (CV)[10] con 10 iteraciones. Como indicador de contribución se empleó la importancia de tipo *gain*[11] 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 [12] 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 describe 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 verifico que la variable «ID_PERSONA» de la base de datos del SIAGIE no cuente con valores duplicados.

3.2.2 Filtrar valores atípicos no deseados

Los valores atípicos son valores inusuales en su 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

[10] Ver Anexo 6 «Criterios para la división de datos en entrenamiento y validación».
[11] «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
[12] Número referencial basado en el estudio de Adelman et al. (2017) sobre prediccón de la interrupción de estudios, donde la diferencia del indicador de sensibilidad de entrenamiento y validacion es de ~0.12.

«Costos relacionado a 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 arboles de decisiones, 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 extrañas, errores tipográficos o mayúsculas incorrectas o registros inconsistentes que generan categorías mal etiquetadas. Por ejemplo, se encontró 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 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 realizó 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 de 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 nuevo datos se empleó 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 variables independientes (predictores) y la variable dependiente (variable dicotómica que indica si el estudiante interrumpió sus estudios).

3.5 Formato de datos

Los modelos que hacen uso de algoritmos basados en arboles 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 una transformación general a todas las variables de la nueva base de datos integrada.

CAPÍTULO IV: MODELADO

En este capítulo se detalla todos los criterios que se consideraron y los resultados obtenidos en la fase de modelado. Para ello se empleó 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 en interrumpir sus estudios en el año T+1. El modelo obtenido es representado a través de la siguiente notación:

$$Y_{it} = E\ (\ Interrupción\ de\ Estudios\ _{it+1} \mid P_{it}, F_{it}, S_{it-1}, A_{it-1}, C_{it})$$

Donde:

i. $Y: dummy\ igual\ a\ 1\ si\ el\ estudiante\ interrumpe\ sus\ estudios$

ii. $P: Información\ propia\ del\ estudiante$

iii. $F: Información\ del\ contexto\ familiar$

iv. $S: Información\ de\ contexto\ del\ servicio\ educativo$

v. $A: Desempeño\ académico\ del\ estudiante$

vi. $C: Informacón\ económica\ y\ de\ contexto$

A partir de lo señalado en el punto 2.3, se estableció el valor de T = 2020. En ese sentido, se estimó un modelo ML que calcule la probabilidad que tiene un estudiante de educación básica regular, matriculado en el año 2020 de interrumpir sus estudios en 2021 (no matricularse en el año T+1).

4.1 Selección de técnicas de modelado

En este punto se abordará 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 interrumpir sus estudios. Al tratarse de un caso de clasificación (entre interrumpe y no interrumpe sus estudios), se procedió a emplear el aprendizaje supervisado [13](*Supervised Learning*), el cual es un tipo de ML empleado típicamente para clasificar una etiqueta (Géron, 2019, pág. 8).

[13] 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 interrupción de estudios 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 asignan 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*[14] 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 decisiones 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 ultimito, 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.

[14] 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 (interrumpe o no interrumpe 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 encuentra detallada 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, los cual es un indicador de un bajo sobreajuste u *overfitting*[15] 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 interrumpir los

[15] 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).

estudios de un estudiante matriculado en el año T es mayor al 50% entonces se estima que el estudiante interrumpirá sus estudios en el año T+1. Generación de los modelos

4.2.1 Selección del algoritmo de *Machine Learning*

Con respecto a los datos, se tomó una muestra de 79,966 estudiantes tomados, de forma estratificada[16], 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 [17] las métricas de rendimiento, tales como la Precisión, Sensibilidad, Especificidad, F1, PRAUC y ROCAUC, y se estableció que el algoritmo con mayor valor F1[18] será seleccionado para realizar la estimación final el 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 interrupción de estudios 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 interrupción de estudios.

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

4.2.2 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 dividió 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 estimara 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, 2020

Grado	Norte	Centro	Oriente	Sur	Lima_metro y Callao
3 a 5 años	374,584	345,338	184,824	247,998	465,376
1 Primaria	143,939	129,768	73,220	95,651	179,101
2 Primaria	150,131	132,271	84,811	96,347	184,568
3 Primaria	146,598	130,153	83,059	94,778	177,312
4 Primaria	144,683	129,629	79,190	94,787	174,084
5 Primaria	143,705	130,385	78,377	96,135	174,950
6 Primaria	136,684	128,114	71,788	92,144	163,964
1 Secundaria	127,068	118,859	64,340	86,218	159,326
2 Secundaria	125,112	116,932	62,054	86,354	156,138
3 Secundaria	117,613	114,374	54,591	85,177	152,551
4 Secundaria	116,518	114,576	51,355	88,391	155,614
5 Secundaria	105,443	104,430	43,973	86,414	145,252

Nota. Esta tabla muestra el total de estudiantes por cada grado de educación básica y macro región del año 2020. 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 evidencia 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 interrupción de estudios para las características de una determinada macro región y grado.

4.2.3 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 encuentra disponible 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 interrupción de estudios. 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[19], 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).

Asimismo, en lugar de tener 60 graficas 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 graficas 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, la cual se encuentra presente en todos los grados del nivel educativo y en la mayoría (mayor o igual a 3) de las macro regiones. A continuación, se describen las gráficas de importancia de cada nivel educativo:

[19] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.plot_importance.html

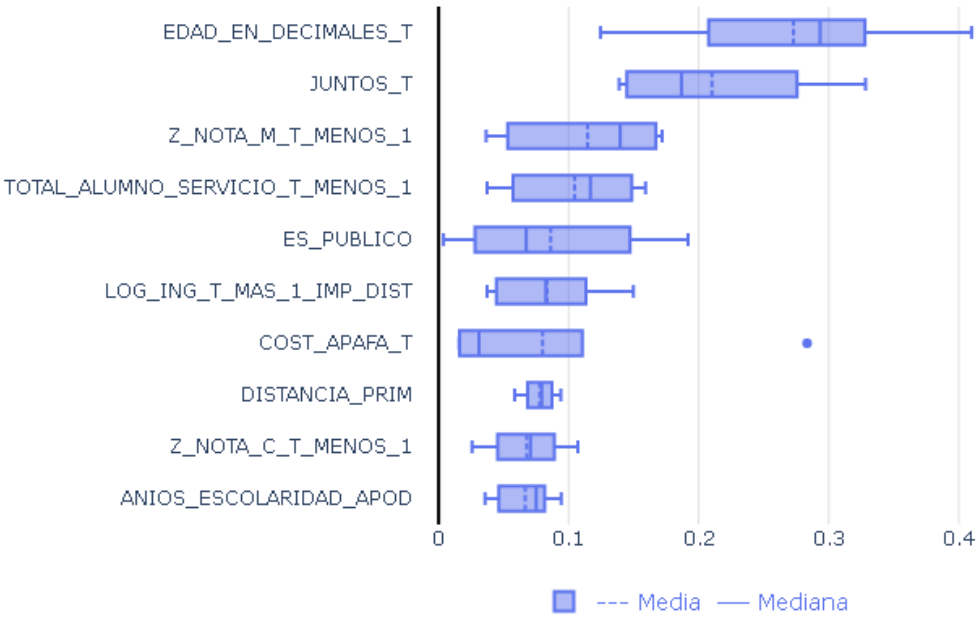
Nivel Inicial: Como muestra la figura 4, la extra-edad se posiciona como la variable que contribuye en mayor medida a explicar la interrupción de estudios en este nivel, seguido de la participación en el programa JUNTOS, el rendimiento académico en matemáticas obtenidos el año anterior y el total de estudiantes en el año anterior. Existe otras variables, tales como el tipo de gestión del servicio educativo, la proyección de ingresos, entre otros; sin embargo, su contribución es menos significativa.

Estos resultados están relacionados con los hallazgos de Rumberger y Lim (2008), quienes encuentran que el rendimiento académico está asociado con la interrupción de estudios. Con respecto a la variable JUNTOS, Lavado y Gallegos (2005) pudieron determinar que un programa de transferencia puede tener un efecto sobre la interrupción de estudios.

Esto sugiere que los estudiantes con mayor riesgo son aquellos con extra-edad, que no reciben el beneficio del programa JUNTOS, y que no han venido teniendo un buen desempeño académico en relación con el de sus compañeros.

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.

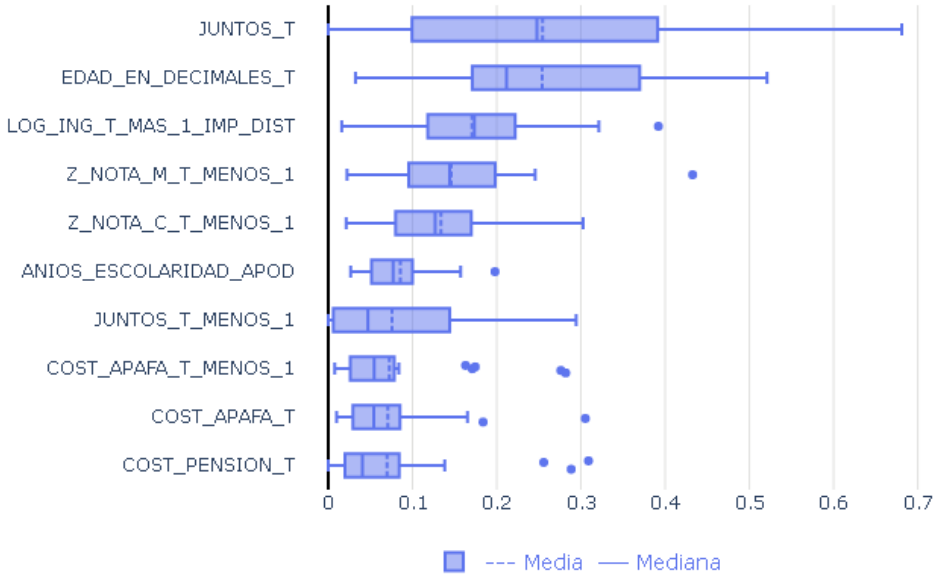
Nivel Primaria: Mediante la figura 5 se puede ver que la participación en el programa JUNTOS contribuye en mayor medida a explicar el abandono escolar en todos los grados de este nivel, seguido de la extra-edad, la proyección de ingresos del hogar y el rendimiento académico en comunicación y matemáticas obtenidos el año anterior. Otras variables como los años de escolaridad del apoderado, su participación en el programa JUNTOS en el año anterior, así como los costos de APAFA y pensión contribuyen en menor medida.

Estos resultados concuerdan con los hallazgos de Cueto y otros (2020), Lavado y Gallegos (2005) y Jacoby(1994), quienes encuentran que la situación económica y las limitaciones de los padres para obtener ingresos, son factores asociados a la interrupción de estudios. Con respecto a la variable JUNTOS, Lavado y Gallegos (2005) pudieron determinar que un programa de transferencia puede tener un efecto sobre la interrupción de estudios.

Esto sugiere que los estudiantes con mayor riesgo son aquellos que no reciben el beneficio del programa JUNTOS, tienen extra-edad, cuentan con menores proyecciones de ingresos en su hogar, y no han venido teniendo un buen desempeño académico serían aquellos con mayor riesgo de abandonar los estudios.

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.

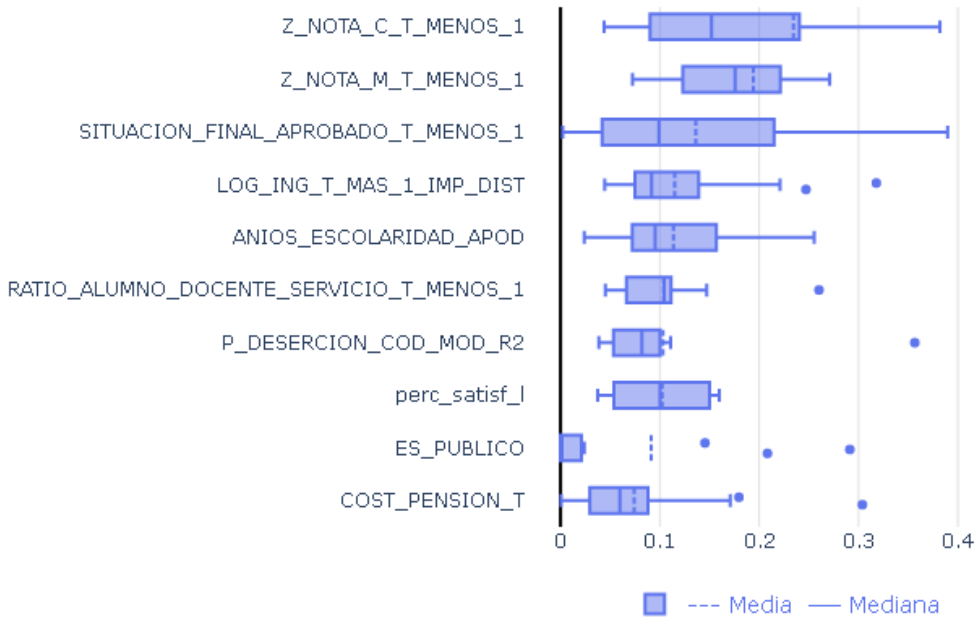
Nivel Secundaria: A través de la figura 6 se evidencia que rendimiento académico en comunicación y matemáticas obtenidos el año anterior se posicionan como las variables que contribuye en mayor medida a explicar el abandono escolar en todos los grados de este nivel, seguida de su situación final distinto de aprobado, la proyección de ingresos del hogar y los años de escolaridad del apoderado. Asimismo, ratios de alumnos/docentes, ratios históricas de interrupción de estudios por servicio educativo, porcentaje de estudiantes que obtuvieron un nivel de logro satisfactorio en comprensión lectora, entre otros, contribuyen, pero en menor medida.

Al respecto, Alcázar (2008) encuentra que aspectos relacionados con el historial educativo de los jóvenes (tales como repeticiones previas o problemas de rendimiento) son determinantes importantes en la interrupción de estudios.

Esto sugiere que estudiantes que no han venido teniendo un buen desempeño académico, que no han tenido una situación final aprobatoria previa, con menores proyecciones de ingresos en su hogar y menores años de escolaridad del apoderado, serían aquellos con mayor riesgo de abandonar los estudios secundarios.

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.

4.3 Evaluación del modelo

Como se mencionó en el punto 4.3.3, las métricas de los 60 modelos se encuentran disponible 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 el Anexo 7, agrupando los $\boldsymbol{E_i}$ y $\boldsymbol{V_i}$ de cada iteración i 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 EBR, como se muestra en la tabla 8.

Tabla 8

Validación cruzada con 10 iteraciones para cada nivel educativo

Nivel	Total	$\boldsymbol{P}$	$\boldsymbol{E_i^{nivel}}$	$\boldsymbol{E_i^{nivel}P}$	$\boldsymbol{V_i^{nivel}}$	$\boldsymbol{V_i^{nivel}P}$
Inicial	1618120	32323	1456306	29091	161814	3232
Primaria	3740326	42642	3366280	38379	374046	4263
Secundaria	2638673	21994	2374793	19793	263880	2201

Nota. Esta tabla muestra el total de estudiantes para el entrenamiento y validación del modelo de interrupción de estudios.

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

Además, se graficó las curvas ROC y PR por nivel educativo. Cada una de estas graficas 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.

4.3.1 Evaluación de resultados del nivel inicial

La tabla 9 presenta la evaluación del nivel inicial, donde se evidencia que el modelo está identificando un 6% de los estudiantes que interrumpieron sus estudios (subcobertura del 94%). Asimismo, la clasificación que realizó el modelo tiene una precisión del 19% (filtración del 79%).

Tabla 9

Métricas con validación cruzada de 10 iteraciones – Nivel Inicial

Métrica	Datos	Promedio	DE	Mín.	Máx.
Precisión	Validación	0.19	0.01	0.17	0.22
Precisión	Entrenamiento	0.21	0.01	0.20	0.23
Sensibilidad	Validación	0.06	0.00	0.05	0.06
Sensibilidad	Entrenamiento	0.06	0.00	0.06	0.07
Especificidad	Validación	1.00	0.00	0.99	1.00
Especificidad	Entrenamiento	1.00	0.00	0.99	1.00
F1	Validación	0.09	0.01	0.08	0.10
F1	Entrenamiento	0.10	0.00	0.09	0.10
PR AUC	Validación	0.10	0.00	0.09	0.10
PR AUC	Entrenamiento	0.11	0.00	0.10	0.11
ROC AUC	Validación	0.82	0.00	0.82	0.83
ROC AUC	Entrenamiento	0.85	0.00	0.84	0.85
Filtración	Validación	0.81	0.01	0.78	0.83
Filtración	Entrenamiento	0.79	0.01	0.77	0.80
Subcobertura	Validación	0.94	0.00	0.94	0.95
Subcobertura	Entrenamiento	0.94	0.00	0.93	0.94

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

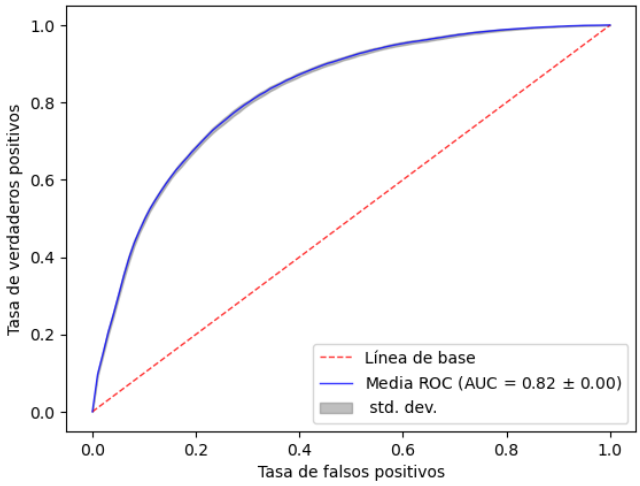
Asimismo, se pudo observar que el modelo cuenta con poco sobreajuste[20] . Esto se comprobó al analizar la desviación estándar de cada una de las métricas, cuyo valor es muy cercano a cero. De la misma forma, se encontró que la diferencia entre las métricas de entrenamiento y de validación es sólo 0.01.

Adicionalmente, la figura 7 muestra la curva ROC con un AUC de 82%, el cual sugeriría que se cuenta con una capacidad predictiva muy significativa.

[20] (Hawkins, 2003)

Figura 7

Curva ROC - Inicial (10 iteraciones)

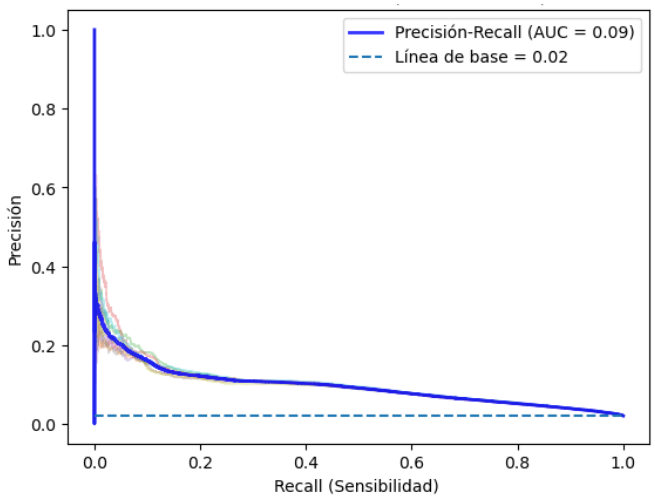


Nota. El grafico representa la curva ROC resultante a partir de la validación cruzada de 10 iteraciones del nivel inicial.

Sin embargo, la figura 8 muestra la curva PR con un AUC del 9%, principalmente por la clase positiva que se encuentra desbalanceada. Aunque el PR AUC no parece un buen resultado, es 4.5 veces mejor que su línea base.

Figura 8

Curva PR – Inicial (10 iteraciones)



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

4.3.2 Evaluación de resultados del nivel Primaria

La tabla 10 presenta la evaluación del nivel primaria, donde se evidencia que el modelo está identificando un 41% de los estudiantes que interrumpieron sus estudios (subcobertura del 59%). Asimismo, la clasificación que realizó el modelo tiene una precisión del 19% (filtración del 81%).

Tabla 10

Métricas con validación cruzada de 10 iteraciones – Nivel Primaria

Métrica	Datos	Promedio	DE	Mín.	Máx.
Precisión	Validación	0.19	0.00	0.19	0.20
Precisión	Entrenamiento	0.23	0.00	0.22	0.23
Sensibilidad	Validación	0.41	0.01	0.40	0.42
Sensibilidad	Entrenamiento	0.49	0.00	0.49	0.49
Especificidad	Validación	0.98	0.00	0.98	0.98
Especificidad	Entrenamiento	0.98	0.00	0.98	0.98
F1	Validación	0.26	0.00	0.26	0.27
F1	Entrenamiento	0.31	0.00	0.31	0.31
PR AUC	Validación	0.18	0.00	0.17	0.19
PR AUC	Entrenamiento	0.24	0.00	0.24	0.25
ROC AUC	Validación	0.91	0.00	0.91	0.91
ROC AUC	Entrenamiento	0.96	0.00	0.96	0.96
Filtración	Validación	0.81	0.00	0.80	0.81
Filtración	Entrenamiento	0.77	0.00	0.77	0.78
Subcobertura	Validación	0.59	0.01	0.58	0.60
Subcobertura	Entrenamiento	0.51	0.00	0.51	0.51

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

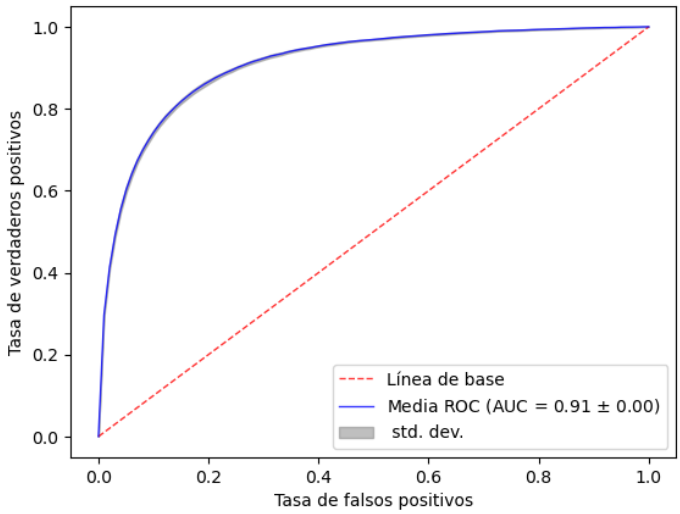
Por otro lado, se pudo observar que el modelo cuenta con poco sobreajuste[21]. Esto se comprobó al analizar la desviación estándar de cada uno de las métricas, cuyo valor es muy cercano a cero. De la misma forma, se encontró que la diferencia entre las métricas de entrenamiento y validación es menor a 0.01.

Adicionalmente, la figura 9 muestra la curva ROC con un AUC de 91%, el cual sugeriría que se cuenta con una capacidad predictiva muy significativa.

[21] (Hawkins, 2003)

Figura 9

Curva ROC – Primaria (10 iteraciones)

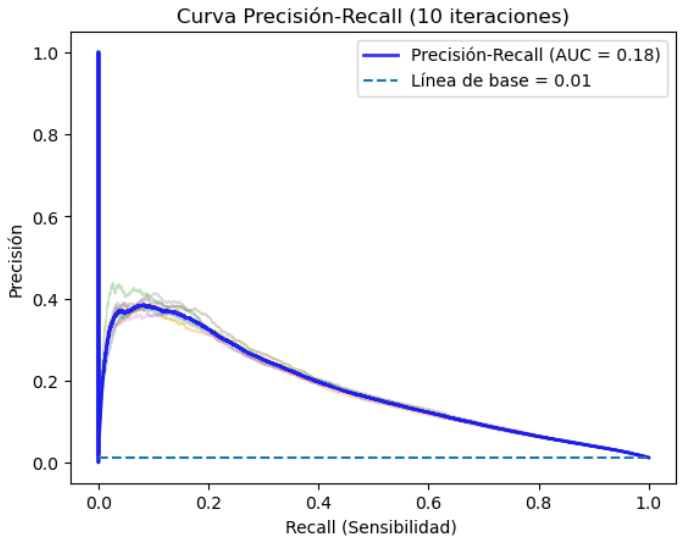


Nota. El grafico representa la curva ROC resultante a partir de la validación cruzada de 10 iteraciones del nivel primaria.

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 10

Curva PR - Primaria (10 iteraciones)



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

4.3.3 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 29% de los estudiantes que interrumpieron sus estudios (subcobertura del 71%). Asimismo, la clasificación que realizó el modelo tiene una precisión del 9% (filtración del 91%).

Tabla 11

Métricas con validación cruzada de 10 iteraciones – Nivel Secundaria

Métrica	Datos	Promedio	DE	Mín.	Máx.
Precisión	Validación	0.09	0.00	0.09	0.10
Precisión	Entrenamiento	0.12	0.00	0.12	0.13
Sensibilidad	Validación	0.29	0.01	0.27	0.30
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	0.14	0.00	0.13	0.15
F1	Entrenamiento	0.19	0.00	0.18	0.19
PR AUC	Validación	0.06	0.00	0.06	0.06
PR AUC	Entrenamiento	0.10	0.00	0.10	0.11
ROC AUC	Validación	0.88	0.00	0.88	0.89
ROC AUC	Entrenamiento	0.95	0.00	0.95	0.95
Filtración	Validación	0.91	0.00	0.90	0.91
Filtración	Entrenamiento	0.88	0.00	0.87	0.88
Subcobertura	Validación	0.71	0.01	0.70	0.73
Subcobertura	Entrenamiento	0.62	0.00	0.62	0.62

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

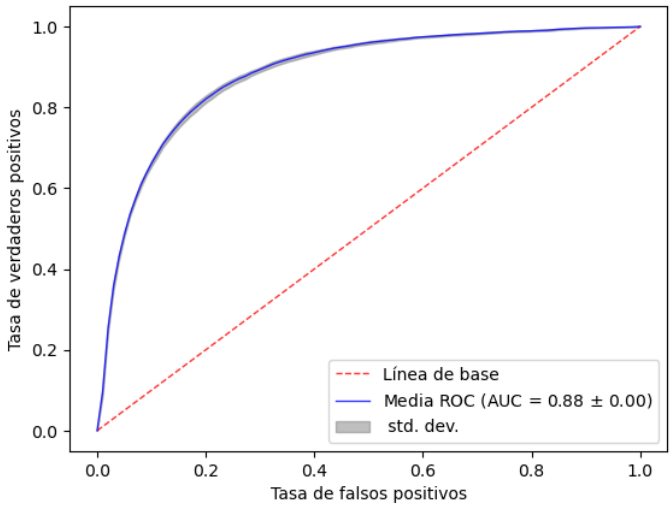
Por otro lado, se pudo observar que el modelo cuenta con poco sobreajuste22. Esto se comprobó al analizar la desviación estándar de cada uno de las métricas, cuyo valor es muy cercano a cero. De la misma forma, se encontró que la diferencia entre las métricas de entrenamiento y validación es 0.01.

Adicionalmente, la figura 11 muestra la curva ROC con un AUC de 88%, el cual sugeriría que se cuenta con una capacidad predictiva muy significativa.

[22] (Hawkins, 2003)

Figura 11

Curva ROC - Secundaria (10 iteraciones)

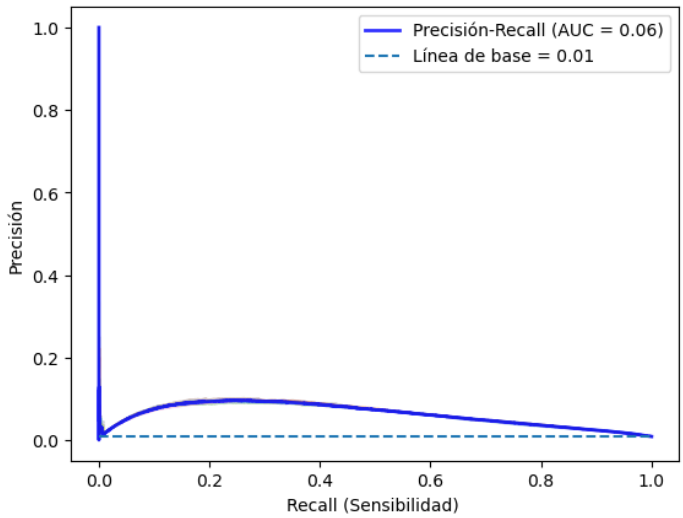


Nota. El grafico representa la curva ROC resultante a partir de la validación cruzada de 10 iteraciones del nivel primaria.

Sin embargo, la figura 12 muestra la curva PR con un AUC del 6%, principalmente por la clase positiva que se encuentra desbalanceada. Aunque el PR AUC no parece un buen resultado, es 6 veces mejor que su línea base.

Figura 12

Curva PR - Secundaria (10 iteraciones)



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

CAPÍTULO V: DESPLIEGUE

Los resultados fueron incorporados dentro del sistema de alerta temprana de interrupción de la trayectoria escolar[23] denominado «Alerta Escuela»[24], el cual se implementa a través de un módulo del SIAGIE que está a cargo de la Unidad 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. Esto facilita el poner a disposición el sistema «Alerta Escuela» a los directores, todo esto en el marco de «Movilización nacional contra la deserción escolar y la promoción del retorno al servicio educativo».

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 interrupción de estudios en el sistema «Alerta Escuela», el cual se detalla a continuación:

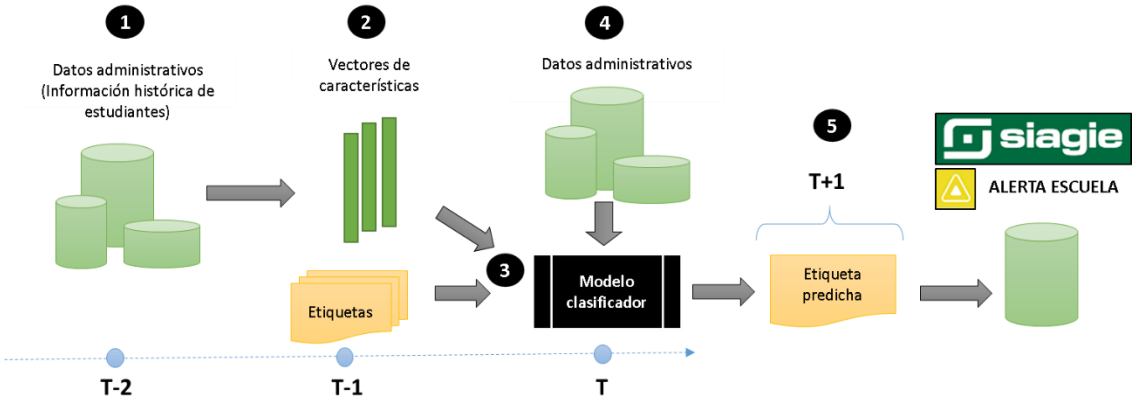
1. En el año T-2 se realiza un corte para obtener los datos de estudiantes matriculados en dicho año.
2. Para poder identificar que estudiantes interrumpen sus estudios, se emplea la información del año T-1 y se aplican los criterios descritos en el punto 2.3 para poder determinar los estudiantes que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

[23] La literatura reciente ha enfatizado el rol de las alertas tempranas para disminuir la interrupción de los estudios 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).

[24] 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 interrupción de estudios 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 matriculare en año T+1 fue necesario emplear el rango interanual «T-2; T-1», es decir, se debe estimar un modelo que haga uso información del año T-2 hacia atrás y emplear la información del año T-1 para determinar si el estudiante interrumpió 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 categorizaron 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: Se definió un nuevo punto de corte o umbral referencial U_{max} por cada modelo ML de macro región y grado escolar. El umbral U_{max} seleccionado es donde se maximiza la métrica F1. De este modo, los estudiantes con riesgo menor a U_{max} serán categorizados como riesgo Bajo.

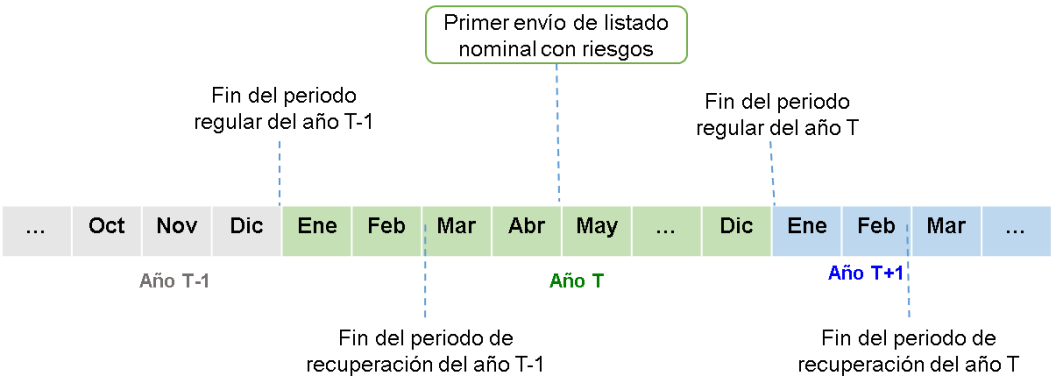
- Riesgo Medio: En este grupo se ubican los estudiantes que cuentan con un riesgo mayor a U_{max} y menor a 0.75. Se consideró el umbral 0.75 por ser un punto intermedio entre 0.5 (umbral de referencia) y 1.0 (máximo valor de umbral).

- Riesgo Alto: Para este grupo se considera todos los estudiantes cuyo riesgo es superior al umbral 0.75.

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 interrupción de estudios a partir del mes de mayo como se muestra en la figura 14. De esta manera se contará con más del 90%[25] 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 envió de resultados hacia el sistema «Alerta Escuela»



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

[25] 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).

CONCLUSIONES

Mediante el presente informe se describe la metodología empleada para calcular el riesgo de interrupción de estudios 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 informe.

En el capítulo I se señaló la importancia de poder identificar estudiantes vulnerables de interrumpir sus estudios. Para ello se creó un modelo que pude calcular el riesgo de interrupción de estudios. 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 interrupción de estudios, las cuales se han identificado a partir de la revisión de la literatura al respecto.

A través del diseño de comprobación propuesto en este informe, se verificó que los modelos para inicial, primaria y secundaria cuentan con poder predictivo para poder calcular el riesgo de interrupción de estudios. Asimismo, se evidenció un bajo nivel de sobre ajuste.

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 interrupción de estudios 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 interrupción de estudios 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 interrupción de estudios 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.

LÍNEAS DE MEJORA

Como futura mejora a la metodología se tiene contemplado emplear ML no supervisado para clusterizar 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 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 puede visibilizar mejor los distintos factores asociados a la interrupción de estudios de forma más contextualizada y específica, sin embargo, esta estrategia omite posibles tendencias generales de interrupción de estudios 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 interrupción de estudios, 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 normatividad actual para el registro de la asistencia genera que no haya 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 cambios en contexto que se consideró en este informe. 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.

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-xgboost-lightgbm-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 Bolíbar.

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

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-del-abandono-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

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 Univesidad 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/541161-094-2020-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 internanual: http://escale.minedu.gob.pe/tendencias-2016-portlet/servlet/tendencias/archivo?idCuadro=321&tipo=meta

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

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

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.*

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 informe emplea el aprendizaje supervisado el cual es utilizado típicamente para clasificar una etiqueta (Géron, 2019, pág. 8). La figura 15 muestra la representación graficas 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.

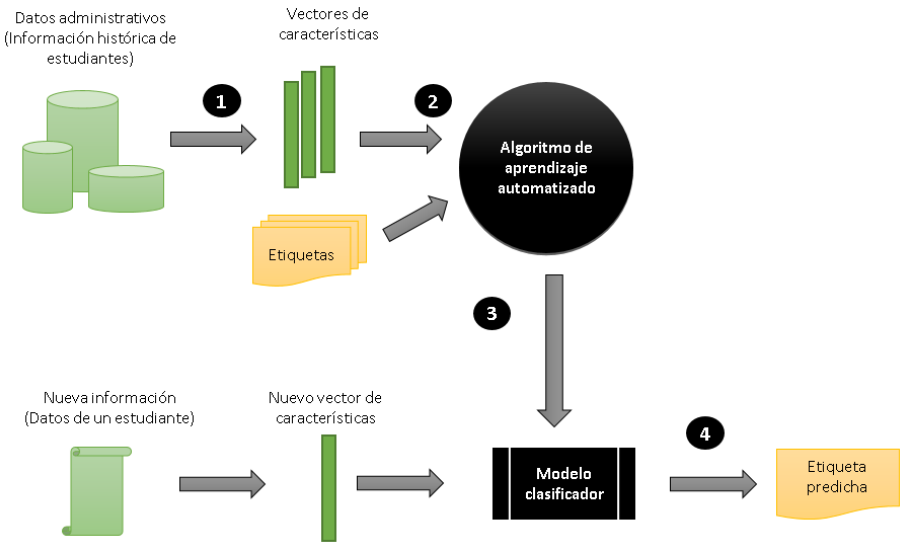
2) A cada uno de los vectores de características se le asigna una etiqueta (interrumpe o no interrumpe sus estudios), formando así el principal insumo para entrenar el algoritmo de aprendizaje supervisado.

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 cercana.	ESCALE
	Distancia euclidiana entre el servicio de nivel primaria del estudiante al servicio de nivel secundaria más cercana.	ESCALE
	Ratios de docentes, por modalidad de contrato (Nombrados, Contratados, Otros), a nivel de servicio educativo.	NEXUS

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 2009-2020 y la dinámica de recuperación del índice de actividad económica para 2021.	Ingresos
	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

Nota. Esta tabla muestra la descripción de cada variable empleada para calcular el riesgo de interrupción de estudios.

ANEXO 3: Macro regiones

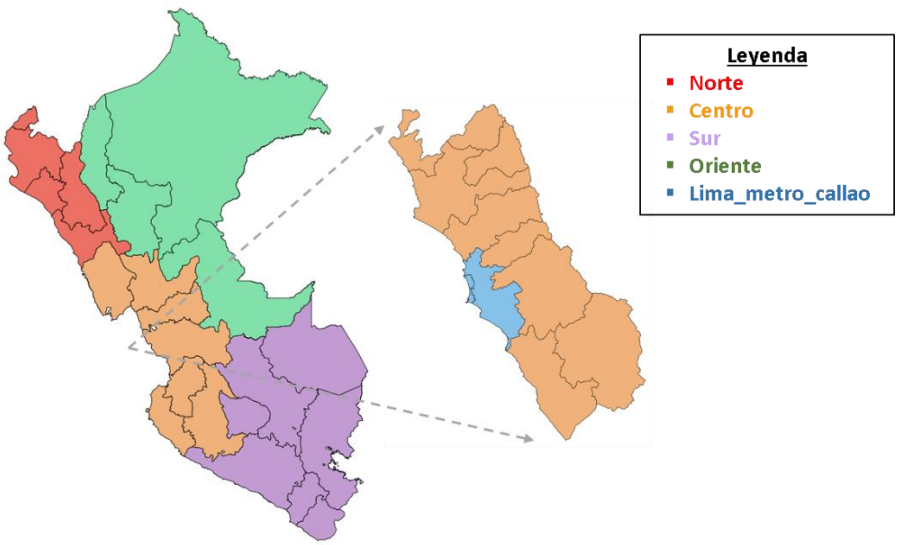
Se establecieron 5 macro regiones (Norte, Sur, Centro, Oriente y Lima metro_callao), como se muestra en la figura 16. A continuación se describen los criterios empleados para la conformación de cada macro región.

Para la macro región Norte, se consideró las regiones que conforman a la macro región Nor Oriente Del Perú[26], con excepción de las regiones de Amazonas, Loreto y San Martin, 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[27] .

Por otro lado, para definir la macro región Centro, se tomó como base inicial las regiones que conforman la Mancomunidad Regional Pacífico Centro Amazónica[28], 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 Martin y Ucayali.

Figura 16

Macro regiones



[26] 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.
[27] Se aprueba con Ordenanza N° 343-AREQUIPA, el cual incluye a los departamentos de Arequipa, Apurímac, Cusco, Madre de Dios, Moquegua, Puno, Tacna.
[28] 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 interrumpen 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 asigno 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 podria 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 apropiada 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 relacionados 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 interrumpen sus estudios, entre el total de estudiantes clasificados como que interrumpen sus estudios. Asimismo, la subcobertura está definida como el ratio de estudiantes que no fueron clasificados como que interrumpen sus estudios, entre el total de estudiantes que realmente interrumpen 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 interrumpen 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	
		Interrumpe sus estudios	No interrumpe sus estudios
Predicción	Interrumpe sus estudios	A (Verdadero Positivo)	B (Falso positivo)
	No interrumpe sus estudios	C (Falso Negativo)	D (Verdadero Negativo)

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

MÉTRICAS DE RENDIMIENTO

- $Exactitud\ (Accuracy) = (A + D)/(A + B + C + D)$
- $Sensibilidad\ (Recall/TPR) = A/(A + C)$
- $Especificidad = D/(B + D)$
- $Presicion = A/(A + B)$
- $F1 = 2 * Presicion * Recall/(Presicion + Recall)$
- $Ratio\ de\ falsos\ positivos\ (FPR) = 1 -\ Especificidad = B/(B + D)$
- $Subcobertura\ (error\ de\ exclusión) = C/(A + C)$
- $Filtración\ (error\ de\ inclusión\ ) = B/(A + B)$
- $Average\ Precision = \sum_n (Recall_n - Recall_{n-1}) * Presicion_n$,

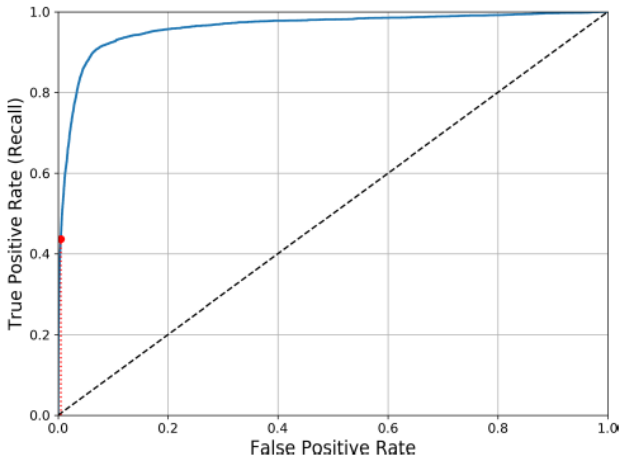
  $donde\ Recall_n\ y\ \ Presicion_n\ es\ la\ sensibilidad\ y\ la\ precision\ para\ un\ umbral\ n.$

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 significara una tasa de fasos 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)

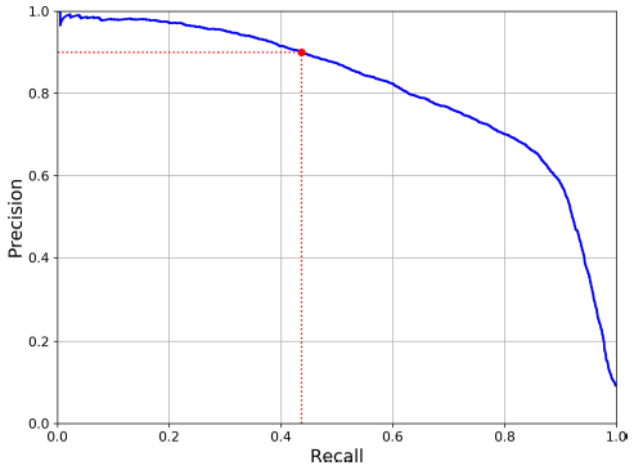


Nota. Adaptado de *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* (p.98), por A. Géron, 2019, O'REALLY.

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 este método es que solo asigna una única proporción de los datos para el entrenamiento (Kohavi, 1995).

En esa misma línea, existe otro grupo[29] de métodos que emplean los datos originales para generan 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 el mejor para medir el rendimiento de un modelo.

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

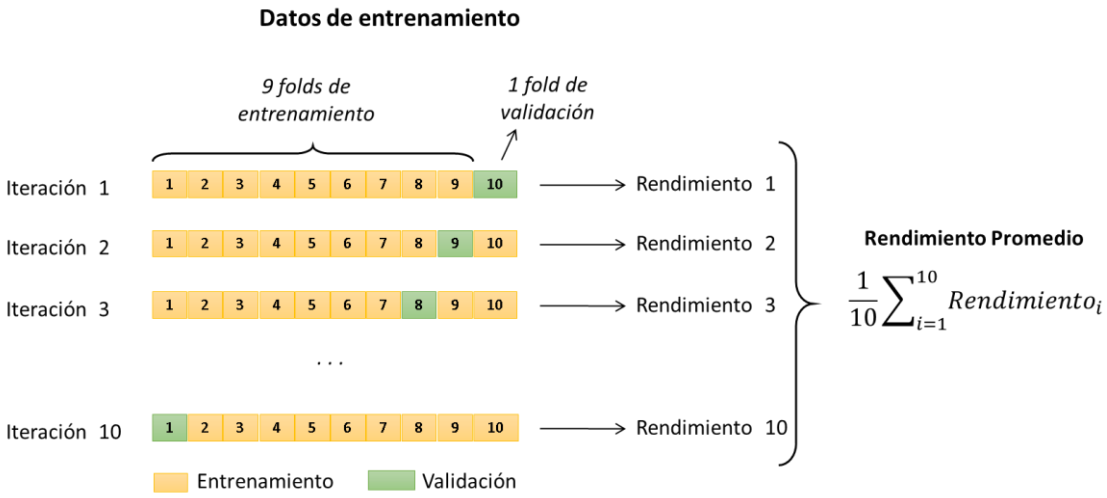
Adicionalmente, se espera que los datos de entrenamiento y validación cuenten con la misma proporción de registros de la clase positiva (estudiantes que interrumpen sus estudios) y clase negativa (estudiantes que no interrumpen sus estudios), es decir, que estén estratificados.

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

[29] 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[30] 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 submuetras 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	465376	16929	418838	15236	46538	1693
ciclo_2	norte	374584	5248	337125	4723	37459	525
ciclo_2	sur	247998	1853	223198	1668	24800	185
ciclo_2	centro	345338	3616	310804	3255	34534	361
ciclo_2	oriente	184824	4677	166341	4209	18483	468

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.

[30] 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	179101	2550	161190	2295	17911	255
1_Prim	norte	143939	839	129545	755	14394	84
1_Prim	sur	95651	289	86085	260	9566	29
1_Prim	centro	129768	493	116791	444	12977	49
1_Prim	oriente	73220	617	65898	556	7322	61
2_Prim	lima_metro_callao	184568	2421	166111	2179	18457	242
2_Prim	norte	150131	791	135117	711	15014	80
2_Prim	sur	96347	266	86712	240	9635	26
2_Prim	centro	132271	494	119043	444	13228	50
2_Prim	oriente	84811	684	76329	615	8482	69
3_Prim	lima_metro_callao	177312	2022	159580	1819	17732	203
3_Prim	norte	146598	721	131938	649	14660	72
3_Prim	sur	94778	246	85300	222	9478	24
3_Prim	centro	130153	400	117137	360	13016	40
3_Prim	oriente	83059	623	74753	561	8306	62
4_Prim	lima_metro_callao	174084	1652	156675	1487	17409	165
4_Prim	norte	144683	657	130214	591	14469	66
4_Prim	sur	94787	220	85308	198	9479	22
4_Prim	centro	129629	389	116666	350	12963	39
4_Prim	oriente	79190	593	71271	534	7919	59
5_Prim	lima_metro_callao	174950	1436	157455	1293	17495	143
5_Prim	norte	143705	610	129334	549	14371	61
5_Prim	sur	96135	210	86521	189	9614	21
5_Prim	centro	130385	374	117346	337	13039	37
5_Prim	oriente	78377	600	70539	540	7838	60
6_Prim	lima_metro_callao	163964	3488	147567	3139	16397	349
6_Prim	norte	136684	6050	123015	5445	13669	605
6_Prim	sur	92144	1440	82929	1296	9215	144
6_Prim	centro	128114	3932	115302	3539	12812	393
6_Prim	oriente	71788	7535	64609	6782	7179	753

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	159326	1717	143393	1545	15933	172
1_Sec	norte	127068	1157	114361	1042	12707	115
1_Sec	sur	86218	310	77596	279	8622	31
1_Sec	centro	118859	820	106973	738	11886	82
1_Sec	oriente	64340	1028	57906	926	6434	102
2_Sec	lima_metro_callao	156138	1913	140524	1722	15614	191
2_Sec	norte	125112	1314	112600	1182	12512	132
2_Sec	sur	86354	487	77718	438	8636	49
2_Sec	centro	116932	1044	105238	939	11694	105
2_Sec	oriente	62054	1078	55848	970	6206	108
3_Sec	lima_metro_callao	152551	1854	137295	1668	15256	186
3_Sec	norte	117613	1327	105851	1194	11762	133
3_Sec	sur	85177	534	76659	481	8518	53
3_Sec	centro	114374	1123	102936	1011	11438	112
3_Sec	oriente	54591	908	49131	817	5460	91
4_Sec	lima_metro_callao	155614	1613	140052	1452	15562	161
4_Sec	norte	116518	1179	104866	1061	11652	118
4_Sec	sur	88391	517	79551	465	8840	52
4_Sec	centro	114576	946	103118	851	11458	95
4_Sec	oriente	51355	636	46219	572	5136	64
5_Sec	lima_metro_callao	145252	180	130726	162	14526	18
5_Sec	norte	105443	127	94898	114	10545	13
5_Sec	sur	86414	50	77772	45	8642	5
5_Sec	centro	104430	73	93987	66	10443	7
5_Sec	oriente	43973	59	39575	53	4398	6

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 detalla los criterios empleados para la configuración de hiperparámetros del modelo LightGBM.

Configuración automática: Se empleó Optuna[31] como framework para la búsqueda de valores idóneos para los hiperparámetros[32] 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.l

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 arboles) 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.

[31] Especificación de OPTUNA: https://optuna.org/
[32] 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 configuro *LightGBM* para que cada árbol en el *booting* 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 denoto las variables utilitarias para el cálculo.

   - N_n: $Número\ de\ estudiantes\ que\ no\ interrumpen\ sus\ estudios\ (casos\ negativos)$
   - N_p: $Número\ de\ estudiantes\ que\ si\ interrumpen\ sus\ estudios\ (casos\ positivos)$
   - $\lambda = \frac{N_p}{N_n}, siempre\ es\ menor\ que\ 1\ por\ el\ bajo\ número\ de\ casos\ positivos.$
   - $T_{mínimo}$: $Cantidad\ mínima\ de\ observaciones\ para\ realiza\ el\ entrenamiento.$ [33]

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

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

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 $= \lambda$ y pos_bagging_fraction $= 1$. Sin embargo, se debe cumplir que la $muestra > T_{mínimo}$, 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 arboles 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:

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

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

[33] Se emplea una cantidad referencial mínima de diez mil registros para reducir el sobreajuste del modelado con LightGBM

hiperparámetro scale_pos_weight. Es importante resaltar que el nuevo número de casos negativos estaría dato por $\alpha * N_n$ debido al muestreo descrito previamente, calculando el hiperparámetro de la siguiente manera:

- $\text{scale\_pos\_weight} = (\alpha * N_n)/N_p$

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	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.298	0.033	0.253	0.355
	Precisión	Entrenamiento	0.308	0.010	0.290	0.324
	Sensibilidad	Validación	0.014	0.003	0.009	0.021
	Sensibilidad	Entrenamiento	0.015	0.003	0.012	0.019
	Especificidad	Validación	0.999	0.000	0.998	0.999
	Especificidad	Entrenamiento	0.999	0.000	0.998	0.999
	F1	Validación	0.027	0.005	0.018	0.040
	F1	Entrenamiento	0.029	0.005	0.022	0.035
	PRAUC	Validación	0.128	0.002	0.125	0.133
	PRAUC	Entrenamiento	0.132	0.000	0.131	0.133
	ROCAUC	Validación	0.782	0.005	0.774	0.790
	ROCAUC	Entrenamiento	0.791	0.000	0.791	0.792
	Filtración	Validación	0.702	0.033	0.645	0.747
	Filtración	Entrenamiento	0.692	0.010	0.676	0.710
	Subcobertura	Validación	0.986	0.003	0.979	0.991
	Subcobertura	Entrenamiento	0.985	0.003	0.981	0.988
norte	Precisión	Validación	0.186	0.020	0.162	0.219
	Precisión	Entrenamiento	0.200	0.007	0.189	0.210
	Sensibilidad	Validación	0.118	0.017	0.095	0.152
	Sensibilidad	Entrenamiento	0.128	0.006	0.118	0.135
	Especificidad	Validación	0.993	0.001	0.992	0.994
	Especificidad	Entrenamiento	0.993	0.001	0.992	0.994
	F1	Validación	0.144	0.018	0.125	0.178
	F1	Entrenamiento	0.156	0.004	0.151	0.163
	PRAUC	Validación	0.100	0.013	0.085	0.125
	PRAUC	Entrenamiento	0.109	0.001	0.107	0.111
	ROCAUC	Validación	0.846	0.007	0.832	0.855
	ROCAUC	Entrenamiento	0.870	0.001	0.869	0.871
	Filtración	Validación	0.814	0.020	0.781	0.838
	Filtración	Entrenamiento	0.800	0.007	0.790	0.811
	Subcobertura	Validación	0.882	0.017	0.848	0.905
	Subcobertura	Entrenamiento	0.872	0.006	0.865	0.882
sur	Precisión	Validación	0.105	0.017	0.082	0.140
	Precisión	Entrenamiento	0.131	0.016	0.104	0.154
	Sensibilidad	Validación	0.094	0.020	0.065	0.130
	Sensibilidad	Entrenamiento	0.126	0.017	0.102	0.146
	Especificidad	Validación	0.994	0.002	0.992	0.996
	Especificidad	Entrenamiento	0.994	0.002	0.991	0.995
	F1	Validación	0.097	0.013	0.079	0.121
	F1	Entrenamiento	0.126	0.005	0.118	0.131

	PRAUC	Validación	0.054	0.012	0.042	0.085
	PRAUC	Entrenamiento	0.071	0.002	0.068	0.077
	ROCAUC	Validación	0.817	0.015	0.786	0.837
	ROCAUC	Entrenamiento	0.899	0.001	0.897	0.900
	Filtración	Validación	0.895	0.017	0.860	0.918
	Filtración	Entrenamiento	0.869	0.016	0.846	0.896
	Subcobertura	Validación	0.906	0.020	0.870	0.935
	Subcobertura	Entrenamiento	0.874	0.017	0.854	0.898
centro	Precisión	Validación	0.158	0.035	0.122	0.237
	Precisión	Entrenamiento	0.182	0.017	0.156	0.221
	Sensibilidad	Validación	0.071	0.011	0.055	0.094
	Sensibilidad	Entrenamiento	0.083	0.007	0.073	0.095
	Especificidad	Validación	0.996	0.001	0.995	0.997
	Especificidad	Entrenamiento	0.996	0.001	0.995	0.997
	F1	Validación	0.097	0.015	0.076	0.124
	F1	Entrenamiento	0.113	0.006	0.104	0.124
	PRAUC	Validación	0.070	0.010	0.055	0.093
	PRAUC	Entrenamiento	0.083	0.002	0.079	0.085
	ROCAUC	Validación	0.829	0.010	0.810	0.838
	ROCAUC	Entrenamiento	0.871	0.001	0.869	0.874
	Filtración	Validación	0.842	0.035	0.763	0.878
	Filtración	Entrenamiento	0.818	0.017	0.779	0.844
	Subcobertura	Validación	0.929	0.011	0.906	0.945
	Subcobertura	Entrenamiento	0.917	0.007	0.905	0.927
oriente	Precisión	Validación	0.293	0.024	0.253	0.333
	Precisión	Entrenamiento	0.305	0.009	0.291	0.318
	Sensibilidad	Validación	0.106	0.012	0.088	0.130
	Sensibilidad	Entrenamiento	0.112	0.010	0.094	0.124
	Especificidad	Validación	0.993	0.001	0.992	0.995
	Especificidad	Entrenamiento	0.993	0.001	0.992	0.995
	F1	Validación	0.155	0.015	0.138	0.187
	F1	Entrenamiento	0.163	0.010	0.144	0.174
	PRAUC	Validación	0.163	0.013	0.150	0.186
	PRAUC	Entrenamiento	0.178	0.002	0.176	0.181
	ROCAUC	Validación	0.854	0.009	0.844	0.873
	ROCAUC	Entrenamiento	0.876	0.001	0.875	0.877
	Filtración	Validación	0.707	0.024	0.667	0.747
	Filtración	Entrenamiento	0.695	0.009	0.682	0.709
	Subcobertura	Validación	0.894	0.012	0.870	0.912
	Subcobertura	Entrenamiento	0.888	0.010	0.876	0.906

Tabla 20

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.151	0.010	0.132	0.170
	Precisión	Entrenamiento	0.158	0.003	0.153	0.164
	Sensibilidad	Validación	0.246	0.028	0.208	0.302
	Sensibilidad	Entrenamiento	0.261	0.005	0.250	0.269
	Especificidad	Validación	0.980	0.001	0.979	0.981
	Especificidad	Entrenamiento	0.980	0.001	0.979	0.981
	F1	Validación	0.187	0.016	0.161	0.217
	F1	Entrenamiento	0.197	0.003	0.192	0.201
	PRAUC	Validación	0.112	0.012	0.095	0.128
	PRAUC	Entrenamiento	0.126	0.003	0.122	0.132
	ROCAUC	Validación	0.852	0.008	0.843	0.868
	ROCAUC	Entrenamiento	0.899	0.001	0.898	0.900
	Filtración	Validación	0.849	0.010	0.830	0.868
	Filtración	Entrenamiento	0.842	0.003	0.836	0.847
	Subcobertura	Validación	0.754	0.028	0.698	0.792
	Subcobertura	Entrenamiento	0.739	0.005	0.731	0.750
norte	Precisión	Validación	0.084	0.009	0.073	0.103
	Precisión	Entrenamiento	0.112	0.005	0.107	0.122
	Sensibilidad	Validación	0.411	0.035	0.373	0.476
	Sensibilidad	Entrenamiento	0.554	0.016	0.523	0.576
	Especificidad	Validación	0.974	0.002	0.971	0.977
	Especificidad	Entrenamiento	0.974	0.001	0.972	0.977
	F1	Validación	0.140	0.014	0.123	0.167
	F1	Entrenamiento	0.187	0.007	0.180	0.200
	PRAUC	Validación	0.078	0.010	0.065	0.097
	PRAUC	Entrenamiento	0.123	0.005	0.116	0.131
	ROCAUC	Validación	0.901	0.014	0.877	0.929
	ROCAUC	Entrenamiento	0.969	0.001	0.968	0.970
	Filtración	Validación	0.916	0.009	0.897	0.927
	Filtración	Entrenamiento	0.888	0.005	0.878	0.893
	Subcobertura	Validación	0.589	0.035	0.524	0.627
	Subcobertura	Entrenamiento	0.446	0.016	0.424	0.477
sur	Precisión	Validación	0.091	0.023	0.057	0.128
	Precisión	Entrenamiento	0.204	0.015	0.177	0.228
	Sensibilidad	Validación	0.301	0.059	0.172	0.379
	Sensibilidad	Entrenamiento	0.771	0.038	0.700	0.831
	Especificidad	Validación	0.991	0.001	0.989	0.993
	Especificidad	Entrenamiento	0.991	0.001	0.989	0.993
	F1	Validación	0.140	0.033	0.086	0.187
	F1	Entrenamiento	0.322	0.018	0.286	0.344
	PRAUC	Validación	0.081	0.026	0.039	0.129
	PRAUC	Entrenamiento	0.280	0.024	0.248	0.323

	ROCAUC	Validación	0.856	0.039	0.786	0.918
	ROCAUC	Entrenamiento	0.994	0.001	0.993	0.995
	Filtración	Validación	0.909	0.023	0.872	0.943
	Filtración	Entrenamiento	0.796	0.015	0.772	0.823
	Subcobertura	Validación	0.699	0.059	0.621	0.828
	Subcobertura	Entrenamiento	0.229	0.038	0.169	0.300
centro	Precisión	Validación	0.073	0.017	0.051	0.106
	Precisión	Entrenamiento	0.137	0.013	0.106	0.152
	Sensibilidad	Validación	0.322	0.078	0.224	0.460
	Sensibilidad	Entrenamiento	0.639	0.031	0.601	0.709
	Especificidad	Validación	0.984	0.002	0.980	0.988
	Especificidad	Entrenamiento	0.984	0.002	0.980	0.986
	F1	Validación	0.119	0.027	0.084	0.172
	F1	Entrenamiento	0.225	0.019	0.181	0.246
	PRAUC	Validación	0.060	0.013	0.034	0.078
	PRAUC	Entrenamiento	0.141	0.014	0.113	0.174
	ROCAUC	Validación	0.892	0.012	0.863	0.906
	ROCAUC	Entrenamiento	0.984	0.001	0.981	0.986
	Filtración	Validación	0.927	0.017	0.894	0.949
	Filtración	Entrenamiento	0.863	0.013	0.848	0.894
	Subcobertura	Validación	0.678	0.078	0.540	0.776
	Subcobertura	Entrenamiento	0.361	0.031	0.291	0.399
oriente	Precisión	Validación	0.102	0.013	0.087	0.126
	Precisión	Entrenamiento	0.133	0.004	0.125	0.138
	Sensibilidad	Validación	0.391	0.061	0.306	0.492
	Sensibilidad	Entrenamiento	0.522	0.025	0.483	0.559
	Especificidad	Validación	0.971	0.002	0.967	0.974
	Especificidad	Entrenamiento	0.971	0.001	0.969	0.973
	F1	Validación	0.162	0.021	0.136	0.201
	F1	Entrenamiento	0.212	0.007	0.201	0.222
	PRAUC	Validación	0.088	0.017	0.057	0.115
	PRAUC	Entrenamiento	0.152	0.013	0.132	0.172
	ROCAUC	Validación	0.876	0.021	0.839	0.904
	ROCAUC	Entrenamiento	0.970	0.001	0.968	0.970
	Filtración	Validación	0.898	0.013	0.874	0.913
	Filtración	Entrenamiento	0.867	0.004	0.862	0.875
	Subcobertura	Validación	0.609	0.061	0.508	0.694
	Subcobertura	Entrenamiento	0.478	0.025	0.441	0.517

Tabla 21

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.129	0.020	0.105	0.168
	Precisión	Entrenamiento	0.142	0.003	0.138	0.147
	Sensibilidad	Validación	0.188	0.024	0.161	0.236
	Sensibilidad	Entrenamiento	0.212	0.014	0.183	0.231
	Especificidad	Validación	0.983	0.002	0.980	0.985
	Especificidad	Entrenamiento	0.983	0.001	0.981	0.985
	F1	Validación	0.152	0.021	0.128	0.192
	F1	Entrenamiento	0.170	0.005	0.159	0.178
	PRAUC	Validación	0.092	0.008	0.082	0.103
	PRAUC	Entrenamiento	0.106	0.002	0.104	0.110
	ROCAUC	Validación	0.855	0.007	0.839	0.865
	ROCAUC	Entrenamiento	0.903	0.001	0.902	0.904
	Filtración	Validación	0.871	0.020	0.832	0.895
	Filtración	Entrenamiento	0.858	0.003	0.853	0.862
	Subcobertura	Validación	0.812	0.024	0.764	0.839
	Subcobertura	Entrenamiento	0.788	0.014	0.769	0.817
norte	Precisión	Validación	0.083	0.013	0.056	0.103
	Precisión	Entrenamiento	0.111	0.007	0.097	0.119
	Sensibilidad	Validación	0.376	0.069	0.266	0.506
	Sensibilidad	Entrenamiento	0.509	0.030	0.459	0.546
	Especificidad	Validación	0.978	0.002	0.973	0.980
	Especificidad	Entrenamiento	0.978	0.001	0.976	0.981
	F1	Validación	0.135	0.021	0.093	0.168
	F1	Entrenamiento	0.182	0.010	0.161	0.195
	PRAUC	Validación	0.076	0.018	0.050	0.099
	PRAUC	Entrenamiento	0.116	0.006	0.108	0.130
	ROCAUC	Validación	0.887	0.020	0.849	0.915
	ROCAUC	Entrenamiento	0.969	0.001	0.967	0.970
	Filtración	Validación	0.917	0.013	0.897	0.944
	Filtración	Entrenamiento	0.889	0.007	0.881	0.903
	Subcobertura	Validación	0.624	0.069	0.494	0.734
	Subcobertura	Entrenamiento	0.491	0.030	0.454	0.541
sur	Precisión	Validación	0.091	0.020	0.056	0.119
	Precisión	Entrenamiento	0.216	0.019	0.178	0.241
	Sensibilidad	Validación	0.335	0.102	0.185	0.462
	Sensibilidad	Entrenamiento	0.877	0.017	0.850	0.904
	Especificidad	Validación	0.991	0.001	0.988	0.992
	Especificidad	Entrenamiento	0.991	0.001	0.989	0.992
	F1	Validación	0.142	0.034	0.086	0.188
	F1	Entrenamiento	0.346	0.025	0.296	0.380
	PRAUC	Validación	0.062	0.019	0.037	0.089
	PRAUC	Entrenamiento	0.306	0.036	0.235	0.362
	ROCAUC	Validación	0.886	0.032	0.824	0.925

	ROCAUC	Entrenamiento	0.996	0.000	0.995	0.996
	Filtración	Validación	0.909	0.020	0.881	0.944
	Filtración	Entrenamiento	0.784	0.019	0.759	0.822
	Subcobertura	Validación	0.665	0.102	0.538	0.815
	Subcobertura	Entrenamiento	0.123	0.017	0.096	0.150
centro	Precisión	Validación	0.056	0.011	0.041	0.073
	Precisión	Entrenamiento	0.107	0.009	0.088	0.122
	Sensibilidad	Validación	0.306	0.080	0.200	0.469
	Sensibilidad	Entrenamiento	0.603	0.038	0.526	0.656
	Especificidad	Validación	0.981	0.002	0.978	0.985
	Especificidad	Entrenamiento	0.981	0.002	0.976	0.985
	F1	Validación	0.095	0.019	0.070	0.127
	F1	Entrenamiento	0.181	0.013	0.153	0.201
	PRAUC	Validación	0.047	0.013	0.027	0.072
	PRAUC	Entrenamiento	0.112	0.012	0.090	0.136
	ROCAUC	Validación	0.866	0.027	0.825	0.920
	ROCAUC	Entrenamiento	0.980	0.002	0.976	0.982
	Filtración	Validación	0.944	0.011	0.927	0.959
	Filtración	Entrenamiento	0.893	0.009	0.878	0.912
	Subcobertura	Validación	0.694	0.080	0.531	0.800
	Subcobertura	Entrenamiento	0.397	0.038	0.344	0.474
oriente	Precisión	Validación	0.092	0.017	0.069	0.134
	Precisión	Entrenamiento	0.134	0.012	0.121	0.157
	Sensibilidad	Validación	0.256	0.044	0.174	0.348
	Sensibilidad	Entrenamiento	0.392	0.036	0.317	0.439
	Especificidad	Validación	0.979	0.004	0.974	0.987
	Especificidad	Entrenamiento	0.979	0.003	0.976	0.984
	F1	Validación	0.134	0.021	0.106	0.174
	F1	Entrenamiento	0.199	0.015	0.186	0.231
	PRAUC	Validación	0.072	0.019	0.051	0.120
	PRAUC	Entrenamiento	0.138	0.014	0.120	0.171
	ROCAUC	Validación	0.872	0.020	0.832	0.901
	ROCAUC	Entrenamiento	0.964	0.001	0.962	0.966
	Filtración	Validación	0.908	0.017	0.866	0.931
	Filtración	Entrenamiento	0.866	0.012	0.843	0.879
	Subcobertura	Validación	0.744	0.044	0.652	0.826
	Subcobertura	Entrenamiento	0.608	0.036	0.561	0.683

Tabla 22

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.113	0.008	0.095	0.123
	Precisión	Entrenamiento	0.127	0.004	0.119	0.132
	Sensibilidad	Validación	0.250	0.019	0.228	0.287
	Sensibilidad	Entrenamiento	0.278	0.007	0.269	0.293
	Especificidad	Validación	0.977	0.002	0.975	0.980
	Especificidad	Entrenamiento	0.978	0.001	0.976	0.979
	F1	Validación	0.156	0.010	0.134	0.169
	F1	Entrenamiento	0.174	0.004	0.167	0.179
	PRAUC	Validación	0.085	0.006	0.076	0.095
	PRAUC	Entrenamiento	0.103	0.001	0.101	0.106
	ROCAUC	Validación	0.859	0.010	0.846	0.874
	ROCAUC	Entrenamiento	0.917	0.001	0.916	0.918
	Filtración	Validación	0.887	0.008	0.877	0.905
	Filtración	Entrenamiento	0.873	0.004	0.868	0.881
	Subcobertura	Validación	0.750	0.019	0.713	0.772
	Subcobertura	Entrenamiento	0.722	0.007	0.707	0.731
norte	Precisión	Validación	0.083	0.014	0.050	0.097
	Precisión	Entrenamiento	0.120	0.006	0.110	0.129
	Sensibilidad	Validación	0.370	0.061	0.222	0.431
	Sensibilidad	Entrenamiento	0.548	0.027	0.505	0.607
	Especificidad	Validación	0.980	0.001	0.978	0.982
	Especificidad	Entrenamiento	0.980	0.001	0.978	0.982
	F1	Validación	0.136	0.023	0.082	0.157
	F1	Entrenamiento	0.196	0.009	0.182	0.210
	PRAUC	Validación	0.074	0.014	0.048	0.094
	PRAUC	Entrenamiento	0.133	0.007	0.123	0.142
	ROCAUC	Validación	0.897	0.014	0.877	0.916
	ROCAUC	Entrenamiento	0.975	0.001	0.974	0.977
	Filtración	Validación	0.917	0.014	0.903	0.950
	Filtración	Entrenamiento	0.880	0.006	0.871	0.890
	Subcobertura	Validación	0.630	0.061	0.569	0.778
	Subcobertura	Entrenamiento	0.452	0.027	0.393	0.495
sur	Precisión	Validación	0.090	0.028	0.053	0.149
	Precisión	Entrenamiento	0.261	0.013	0.237	0.280
	Sensibilidad	Validación	0.252	0.077	0.160	0.440
	Sensibilidad	Entrenamiento	0.923	0.017	0.896	0.950
	Especificidad	Validación	0.993	0.001	0.992	0.995
	Especificidad	Entrenamiento	0.993	0.000	0.992	0.994
	F1	Validación	0.133	0.040	0.079	0.222
	F1	Entrenamiento	0.406	0.016	0.377	0.431
	PRAUC	Validación	0.063	0.016	0.037	0.088
	PRAUC	Entrenamiento	0.418	0.039	0.352	0.497
	ROCAUC	Validación	0.877	0.047	0.804	0.961

	ROCAUC	Entrenamiento	0.998	0.000	0.997	0.998
	Filtración	Validación	0.910	0.028	0.851	0.947
	Filtración	Entrenamiento	0.739	0.013	0.720	0.763
	Subcobertura	Validación	0.748	0.077	0.560	0.840
	Subcobertura	Entrenamiento	0.077	0.017	0.050	0.104
centro	Precisión	Validación	0.051	0.013	0.032	0.074
	Precisión	Entrenamiento	0.134	0.020	0.101	0.169
	Sensibilidad	Validación	0.283	0.070	0.150	0.375
	Sensibilidad	Entrenamiento	0.768	0.058	0.667	0.861
	Especificidad	Validación	0.984	0.002	0.980	0.987
	Especificidad	Entrenamiento	0.984	0.002	0.982	0.988
	F1	Validación	0.087	0.022	0.052	0.123
	F1	Entrenamiento	0.228	0.031	0.176	0.280
	PRAUC	Validación	0.043	0.013	0.026	0.068
	PRAUC	Entrenamiento	0.176	0.019	0.135	0.198
	ROCAUC	Validación	0.891	0.020	0.858	0.930
	ROCAUC	Entrenamiento	0.989	0.001	0.987	0.991
	Filtración	Validación	0.949	0.013	0.926	0.968
	Filtración	Entrenamiento	0.866	0.020	0.831	0.899
	Subcobertura	Validación	0.718	0.070	0.625	0.850
	Subcobertura	Entrenamiento	0.232	0.058	0.139	0.333
oriente	Precisión	Validación	0.110	0.025	0.079	0.163
	Precisión	Entrenamiento	0.163	0.011	0.141	0.176
	Sensibilidad	Validación	0.258	0.062	0.177	0.397
	Sensibilidad	Entrenamiento	0.399	0.030	0.335	0.440
	Especificidad	Validación	0.984	0.001	0.981	0.986
	Especificidad	Entrenamiento	0.984	0.001	0.982	0.986
	F1	Validación	0.154	0.035	0.109	0.231
	F1	Entrenamiento	0.231	0.015	0.202	0.246
	PRAUC	Validación	0.082	0.028	0.052	0.135
	PRAUC	Entrenamiento	0.162	0.011	0.143	0.173
	ROCAUC	Validación	0.867	0.023	0.824	0.908
	ROCAUC	Entrenamiento	0.971	0.001	0.969	0.974
	Filtración	Validación	0.890	0.025	0.837	0.921
	Filtración	Entrenamiento	0.837	0.011	0.824	0.859
	Subcobertura	Validación	0.742	0.062	0.603	0.823
	Subcobertura	Entrenamiento	0.601	0.030	0.560	0.665

Tabla 23

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.097	0.010	0.083	0.112
	Precisión	Entrenamiento	0.112	0.006	0.103	0.123
	Sensibilidad	Validación	0.258	0.032	0.200	0.291
	Sensibilidad	Entrenamiento	0.305	0.028	0.255	0.352
	Especificidad	Validación	0.977	0.003	0.971	0.983
	Especificidad	Entrenamiento	0.977	0.003	0.971	0.982
	F1	Validación	0.140	0.013	0.121	0.160
	F1	Entrenamiento	0.164	0.005	0.156	0.171
	PRAUC	Validación	0.079	0.006	0.065	0.088
	PRAUC	Entrenamiento	0.096	0.002	0.092	0.099
	ROCAUC	Validación	0.861	0.009	0.844	0.872
	ROCAUC	Entrenamiento	0.923	0.001	0.922	0.924
	Filtración	Validación	0.903	0.010	0.888	0.917
	Filtración	Entrenamiento	0.888	0.006	0.877	0.897
	Subcobertura	Validación	0.742	0.032	0.709	0.800
	Subcobertura	Entrenamiento	0.695	0.028	0.648	0.745
norte	Precisión	Validación	0.093	0.012	0.070	0.112
	Precisión	Entrenamiento	0.146	0.006	0.140	0.158
	Sensibilidad	Validación	0.420	0.049	0.318	0.485
	Sensibilidad	Entrenamiento	0.679	0.020	0.649	0.716
	Especificidad	Validación	0.981	0.001	0.980	0.985
	Especificidad	Entrenamiento	0.982	0.001	0.980	0.984
	F1	Validación	0.153	0.019	0.115	0.177
	F1	Entrenamiento	0.241	0.007	0.232	0.254
	PRAUC	Validación	0.091	0.017	0.062	0.115
	PRAUC	Entrenamiento	0.168	0.010	0.154	0.186
	ROCAUC	Validación	0.907	0.020	0.862	0.941
	ROCAUC	Entrenamiento	0.982	0.000	0.981	0.982
	Filtración	Validación	0.907	0.012	0.888	0.930
	Filtración	Entrenamiento	0.854	0.006	0.842	0.860
	Subcobertura	Validación	0.580	0.049	0.515	0.682
	Subcobertura	Entrenamiento	0.321	0.020	0.284	0.351
sur	Precisión	Validación	0.085	0.024	0.059	0.143
	Precisión	Entrenamiento	0.207	0.014	0.187	0.235
	Sensibilidad	Validación	0.341	0.074	0.227	0.455
	Sensibilidad	Entrenamiento	0.923	0.011	0.909	0.949
	Especificidad	Validación	0.991	0.001	0.989	0.994
	Especificidad	Entrenamiento	0.992	0.001	0.991	0.993
	F1	Validación	0.136	0.036	0.093	0.217
	F1	Entrenamiento	0.337	0.019	0.311	0.376
	PRAUC	Validación	0.080	0.040	0.032	0.156
	PRAUC	Entrenamiento	0.337	0.037	0.296	0.402
	ROCAUC	Validación	0.875	0.026	0.835	0.915

	ROCAUC	Entrenamiento	0.997	0.000	0.996	0.997
	Filtración	Validación	0.915	0.024	0.857	0.941
	Filtración	Entrenamiento	0.793	0.014	0.765	0.813
	Subcobertura	Validación	0.659	0.074	0.545	0.773
	Subcobertura	Entrenamiento	0.077	0.011	0.051	0.091
centro	Precisión	Validación	0.057	0.010	0.040	0.071
	Precisión	Entrenamiento	0.131	0.006	0.121	0.143
	Sensibilidad	Validación	0.295	0.047	0.231	0.359
	Sensibilidad	Entrenamiento	0.710	0.026	0.671	0.744
	Especificidad	Validación	0.985	0.001	0.983	0.987
	Especificidad	Entrenamiento	0.986	0.001	0.984	0.988
	F1	Validación	0.096	0.016	0.068	0.118
	F1	Entrenamiento	0.221	0.009	0.208	0.235
	PRAUC	Validación	0.042	0.010	0.029	0.058
	PRAUC	Entrenamiento	0.171	0.007	0.161	0.182
	ROCAUC	Validación	0.868	0.024	0.826	0.912
	ROCAUC	Entrenamiento	0.989	0.000	0.989	0.990
	Filtración	Validación	0.943	0.010	0.929	0.960
	Filtración	Entrenamiento	0.869	0.006	0.857	0.879
	Subcobertura	Validación	0.705	0.047	0.641	0.769
	Subcobertura	Entrenamiento	0.290	0.026	0.256	0.329
oriente	Precisión	Validación	0.095	0.020	0.071	0.139
	Precisión	Entrenamiento	0.165	0.012	0.150	0.183
	Sensibilidad	Validación	0.212	0.034	0.153	0.267
	Sensibilidad	Entrenamiento	0.402	0.026	0.371	0.463
	Especificidad	Validación	0.984	0.002	0.981	0.987
	Especificidad	Entrenamiento	0.985	0.001	0.983	0.987
	F1	Validación	0.131	0.025	0.101	0.183
	F1	Entrenamiento	0.234	0.015	0.215	0.262
	PRAUC	Validación	0.081	0.022	0.057	0.126
	PRAUC	Entrenamiento	0.178	0.013	0.157	0.199
	ROCAUC	Validación	0.870	0.023	0.843	0.917
	ROCAUC	Entrenamiento	0.975	0.001	0.973	0.978
	Filtración	Validación	0.905	0.020	0.861	0.929
	Filtración	Entrenamiento	0.835	0.012	0.817	0.850
	Subcobertura	Validación	0.788	0.034	0.733	0.847
	Subcobertura	Entrenamiento	0.598	0.026	0.537	0.629

Tabla 24

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.084	0.008	0.073	0.097
	Precisión	Entrenamiento	0.099	0.006	0.090	0.109
	Sensibilidad	Validación	0.300	0.037	0.236	0.371
	Sensibilidad	Entrenamiento	0.356	0.028	0.280	0.389
	Especificidad	Validación	0.973	0.003	0.970	0.980
	Especificidad	Entrenamiento	0.973	0.003	0.969	0.981
	F1	Validación	0.131	0.011	0.111	0.149
	F1	Entrenamiento	0.155	0.005	0.144	0.163
	PRAUC	Validación	0.070	0.008	0.056	0.082
	PRAUC	Entrenamiento	0.091	0.002	0.088	0.094
	ROCAUC	Validación	0.861	0.018	0.823	0.890
	ROCAUC	Entrenamiento	0.933	0.001	0.932	0.935
	Filtración	Validación	0.916	0.008	0.903	0.927
	Filtración	Entrenamiento	0.901	0.006	0.891	0.910
	Subcobertura	Validación	0.700	0.037	0.629	0.764
	Subcobertura	Entrenamiento	0.644	0.028	0.611	0.720
norte	Precisión	Validación	0.080	0.020	0.052	0.120
	Precisión	Entrenamiento	0.134	0.010	0.118	0.150
	Sensibilidad	Validación	0.397	0.084	0.262	0.525
	Sensibilidad	Entrenamiento	0.691	0.026	0.643	0.723
	Especificidad	Validación	0.980	0.002	0.977	0.984
	Especificidad	Entrenamiento	0.981	0.002	0.977	0.983
	F1	Validación	0.132	0.033	0.087	0.194
	F1	Entrenamiento	0.224	0.014	0.202	0.247
	PRAUC	Validación	0.081	0.027	0.045	0.133
	PRAUC	Entrenamiento	0.157	0.010	0.138	0.177
	ROCAUC	Validación	0.910	0.017	0.885	0.932
	ROCAUC	Entrenamiento	0.982	0.001	0.980	0.983
	Filtración	Validación	0.920	0.020	0.880	0.948
	Filtración	Entrenamiento	0.866	0.010	0.850	0.882
	Subcobertura	Validación	0.603	0.084	0.475	0.738
	Subcobertura	Entrenamiento	0.309	0.026	0.277	0.357
sur	Precisión	Validación	0.064	0.021	0.037	0.106
	Precisión	Entrenamiento	0.188	0.018	0.152	0.218
	Sensibilidad	Validación	0.290	0.110	0.190	0.476
	Sensibilidad	Entrenamiento	0.913	0.018	0.884	0.942
	Especificidad	Validación	0.991	0.002	0.986	0.994
	Especificidad	Entrenamiento	0.991	0.001	0.989	0.993
	F1	Validación	0.104	0.034	0.062	0.161
	F1	Entrenamiento	0.311	0.025	0.261	0.350
	PRAUC	Validación	0.054	0.017	0.031	0.086
	PRAUC	Entrenamiento	0.341	0.032	0.271	0.387
	ROCAUC	Validación	0.850	0.039	0.778	0.917

	ROCAUC	Entrenamiento	0.997	0.000	0.996	0.997
	Filtración	Validación	0.936	0.021	0.894	0.963
	Filtración	Entrenamiento	0.812	0.018	0.782	0.848
	Subcobertura	Validación	0.710	0.110	0.524	0.810
	Subcobertura	Entrenamiento	0.087	0.018	0.058	0.116
centro	Precisión	Validación	0.061	0.020	0.023	0.082
	Precisión	Entrenamiento	0.155	0.014	0.129	0.176
	Sensibilidad	Validación	0.275	0.102	0.108	0.405
	Sensibilidad	Entrenamiento	0.765	0.029	0.703	0.813
	Especificidad	Validación	0.988	0.002	0.985	0.991
	Especificidad	Entrenamiento	0.988	0.001	0.985	0.990
	F1	Validación	0.099	0.033	0.038	0.137
	F1	Entrenamiento	0.257	0.020	0.220	0.287
	PRAUC	Validación	0.049	0.019	0.022	0.077
	PRAUC	Entrenamiento	0.204	0.017	0.177	0.232
	ROCAUC	Validación	0.877	0.029	0.818	0.915
	ROCAUC	Entrenamiento	0.991	0.001	0.990	0.992
	Filtración	Validación	0.939	0.020	0.918	0.977
	Filtración	Entrenamiento	0.845	0.014	0.824	0.871
	Subcobertura	Validación	0.725	0.102	0.595	0.892
	Subcobertura	Entrenamiento	0.235	0.029	0.187	0.297
oriente	Precisión	Validación	0.110	0.028	0.068	0.165
	Precisión	Entrenamiento	0.182	0.009	0.167	0.192
	Sensibilidad	Validación	0.235	0.050	0.133	0.300
	Sensibilidad	Entrenamiento	0.430	0.040	0.337	0.474
	Especificidad	Validación	0.985	0.002	0.982	0.989
	Especificidad	Entrenamiento	0.985	0.001	0.983	0.987
	F1	Validación	0.150	0.034	0.090	0.209
	F1	Entrenamiento	0.255	0.013	0.224	0.271
	PRAUC	Validación	0.089	0.024	0.058	0.130
	PRAUC	Entrenamiento	0.170	0.007	0.162	0.185
	ROCAUC	Validación	0.873	0.020	0.829	0.906
	ROCAUC	Entrenamiento	0.974	0.001	0.973	0.976
	Filtración	Validación	0.890	0.028	0.835	0.932
	Filtración	Entrenamiento	0.818	0.009	0.808	0.833
	Subcobertura	Validación	0.765	0.050	0.700	0.867
	Subcobertura	Entrenamiento	0.570	0.040	0.526	0.663

Tabla 25

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.248	0.013	0.231	0.265
	Precisión	Entrenamiento	0.258	0.004	0.252	0.267
	Sensibilidad	Validación	0.340	0.017	0.321	0.384
	Sensibilidad	Entrenamiento	0.354	0.008	0.340	0.365
	Especificidad	Validación	0.978	0.001	0.976	0.979
	Especificidad	Entrenamiento	0.978	0.001	0.977	0.980
	F1	Validación	0.286	0.013	0.270	0.313
	F1	Entrenamiento	0.298	0.002	0.295	0.301
	PRAUC	Validación	0.222	0.012	0.207	0.246
	PRAUC	Entrenamiento	0.242	0.004	0.236	0.247
	ROCAUC	Validación	0.871	0.008	0.856	0.882
	ROCAUC	Entrenamiento	0.904	0.001	0.903	0.905
	Filtración	Validación	0.752	0.013	0.735	0.769
	Filtración	Entrenamiento	0.742	0.004	0.733	0.748
	Subcobertura	Validación	0.660	0.017	0.616	0.679
	Subcobertura	Entrenamiento	0.646	0.008	0.635	0.660
norte	Precisión	Validación	0.433	0.016	0.409	0.463
	Precisión	Entrenamiento	0.449	0.003	0.445	0.454
	Sensibilidad	Validación	0.479	0.018	0.431	0.499
	Sensibilidad	Entrenamiento	0.502	0.005	0.495	0.511
	Especificidad	Validación	0.971	0.001	0.968	0.973
	Especificidad	Entrenamiento	0.972	0.001	0.971	0.972
	F1	Validación	0.455	0.016	0.420	0.481
	F1	Entrenamiento	0.474	0.002	0.471	0.478
	PRAUC	Validación	0.431	0.020	0.380	0.455
	PRAUC	Entrenamiento	0.451	0.004	0.447	0.456
	ROCAUC	Validación	0.912	0.004	0.906	0.918
	ROCAUC	Entrenamiento	0.924	0.001	0.923	0.925
	Filtración	Validación	0.567	0.016	0.537	0.591
	Filtración	Entrenamiento	0.551	0.003	0.546	0.555
	Subcobertura	Validación	0.521	0.018	0.501	0.569
	Subcobertura	Entrenamiento	0.498	0.005	0.489	0.505
sur	Precisión	Validación	0.264	0.017	0.239	0.301
	Precisión	Entrenamiento	0.292	0.006	0.283	0.302
	Sensibilidad	Validación	0.556	0.035	0.507	0.625
	Sensibilidad	Entrenamiento	0.623	0.004	0.617	0.632
	Especificidad	Validación	0.975	0.002	0.973	0.978
	Especificidad	Entrenamiento	0.976	0.001	0.975	0.977
	F1	Validación	0.358	0.022	0.327	0.406
	F1	Entrenamiento	0.397	0.005	0.389	0.406
	PRAUC	Validación	0.335	0.029	0.286	0.382
	PRAUC	Entrenamiento	0.385	0.009	0.368	0.396
	ROCAUC	Validación	0.922	0.010	0.897	0.934

	ROCAUC	Entrenamiento	0.969	0.000	0.968	0.969
	Filtración	Validación	0.736	0.017	0.699	0.761
	Filtración	Entrenamiento	0.708	0.006	0.698	0.717
	Subcobertura	Validación	0.444	0.035	0.375	0.493
	Subcobertura	Entrenamiento	0.377	0.004	0.368	0.383
centro	Precisión	Validación	0.402	0.027	0.356	0.431
	Precisión	Entrenamiento	0.418	0.004	0.413	0.428
	Sensibilidad	Validación	0.523	0.028	0.487	0.570
	Sensibilidad	Entrenamiento	0.553	0.005	0.544	0.561
	Especificidad	Validación	0.975	0.002	0.971	0.979
	Especificidad	Entrenamiento	0.976	0.001	0.975	0.977
	F1	Validación	0.454	0.025	0.412	0.487
	F1	Entrenamiento	0.476	0.002	0.471	0.481
	PRAUC	Validación	0.422	0.024	0.379	0.444
	PRAUC	Entrenamiento	0.446	0.003	0.442	0.450
	ROCAUC	Validación	0.925	0.006	0.913	0.934
	ROCAUC	Entrenamiento	0.943	0.001	0.942	0.944
	Filtración	Validación	0.598	0.027	0.569	0.644
	Filtración	Entrenamiento	0.582	0.004	0.572	0.587
	Subcobertura	Validación	0.477	0.028	0.430	0.513
	Subcobertura	Entrenamiento	0.447	0.005	0.439	0.456
oriente	Precisión	Validación	0.583	0.015	0.562	0.603
	Precisión	Entrenamiento	0.596	0.002	0.592	0.599
	Sensibilidad	Validación	0.624	0.015	0.602	0.651
	Sensibilidad	Entrenamiento	0.639	0.003	0.635	0.645
	Especificidad	Validación	0.948	0.003	0.942	0.953
	Especificidad	Entrenamiento	0.949	0.000	0.948	0.950
	F1	Validación	0.603	0.010	0.585	0.620
	F1	Entrenamiento	0.617	0.002	0.614	0.621
	PRAUC	Validación	0.617	0.016	0.587	0.639
	PRAUC	Entrenamiento	0.635	0.002	0.631	0.639
	ROCAUC	Validación	0.923	0.002	0.919	0.926
	ROCAUC	Entrenamiento	0.930	0.000	0.929	0.930
	Filtración	Validación	0.417	0.015	0.397	0.438
	Filtración	Entrenamiento	0.404	0.002	0.401	0.408
	Subcobertura	Validación	0.376	0.015	0.349	0.398
	Subcobertura	Entrenamiento	0.361	0.003	0.355	0.365

Tabla 26

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.136	0.011	0.121	0.158
	Precisión	Entrenamiento	0.155	0.008	0.143	0.171
	Sensibilidad	Validación	0.263	0.036	0.216	0.326
	Sensibilidad	Entrenamiento	0.301	0.014	0.281	0.326
	Especificidad	Validación	0.982	0.002	0.978	0.985
	Especificidad	Entrenamiento	0.982	0.002	0.980	0.985
	F1	Validación	0.179	0.016	0.156	0.213
	F1	Entrenamiento	0.205	0.007	0.196	0.218
	PRAUC	Validación	0.101	0.015	0.077	0.131
	PRAUC	Entrenamiento	0.123	0.002	0.120	0.127
	ROCAUC	Validación	0.870	0.012	0.852	0.892
	ROCAUC	Entrenamiento	0.928	0.001	0.926	0.931
	Filtración	Validación	0.864	0.011	0.842	0.879
	Filtración	Entrenamiento	0.845	0.008	0.829	0.857
	Subcobertura	Validación	0.737	0.036	0.674	0.784
	Subcobertura	Entrenamiento	0.699	0.014	0.674	0.719
norte	Precisión	Validación	0.119	0.010	0.109	0.142
	Precisión	Entrenamiento	0.147	0.006	0.137	0.160
	Sensibilidad	Validación	0.299	0.030	0.250	0.353
	Sensibilidad	Entrenamiento	0.384	0.013	0.357	0.405
	Especificidad	Validación	0.980	0.002	0.977	0.983
	Especificidad	Entrenamiento	0.980	0.001	0.978	0.981
	F1	Validación	0.170	0.014	0.155	0.202
	F1	Entrenamiento	0.213	0.007	0.201	0.226
	PRAUC	Validación	0.103	0.017	0.086	0.143
	PRAUC	Entrenamiento	0.137	0.004	0.130	0.145
	ROCAUC	Validación	0.887	0.016	0.867	0.912
	ROCAUC	Entrenamiento	0.955	0.001	0.953	0.957
	Filtración	Validación	0.881	0.010	0.858	0.891
	Filtración	Entrenamiento	0.853	0.006	0.840	0.863
	Subcobertura	Validación	0.701	0.030	0.647	0.750
	Subcobertura	Entrenamiento	0.616	0.013	0.595	0.643
sur	Precisión	Validación	0.084	0.018	0.059	0.107
	Precisión	Entrenamiento	0.180	0.010	0.164	0.192
	Sensibilidad	Validación	0.335	0.075	0.226	0.452
	Sensibilidad	Entrenamiento	0.806	0.034	0.749	0.871
	Especificidad	Validación	0.987	0.002	0.984	0.990
	Especificidad	Entrenamiento	0.987	0.001	0.985	0.989
	F1	Validación	0.134	0.028	0.094	0.173
	F1	Entrenamiento	0.294	0.013	0.272	0.312
	PRAUC	Validación	0.080	0.021	0.052	0.113
	PRAUC	Entrenamiento	0.231	0.027	0.196	0.291
	ROCAUC	Validación	0.885	0.023	0.847	0.915

	ROCAUC	Entrenamiento	0.992	0.001	0.992	0.994
	Filtración	Validación	0.916	0.018	0.893	0.941
	Filtración	Entrenamiento	0.820	0.010	0.808	0.836
	Subcobertura	Validación	0.665	0.075	0.548	0.774
	Subcobertura	Entrenamiento	0.194	0.034	0.129	0.251
centro	Precisión	Validación	0.087	0.009	0.075	0.100
	Precisión	Entrenamiento	0.112	0.004	0.106	0.119
	Sensibilidad	Validación	0.360	0.024	0.317	0.402
	Sensibilidad	Entrenamiento	0.472	0.019	0.442	0.501
	Especificidad	Validación	0.974	0.003	0.970	0.977
	Especificidad	Entrenamiento	0.974	0.002	0.971	0.976
	F1	Validación	0.140	0.013	0.122	0.157
	F1	Entrenamiento	0.180	0.005	0.175	0.191
	PRAUC	Validación	0.079	0.012	0.054	0.100
	PRAUC	Entrenamiento	0.123	0.007	0.111	0.136
	ROCAUC	Validación	0.887	0.018	0.857	0.912
	ROCAUC	Entrenamiento	0.965	0.001	0.964	0.966
	Filtración	Validación	0.913	0.009	0.900	0.925
	Filtración	Entrenamiento	0.888	0.004	0.881	0.894
	Subcobertura	Validación	0.640	0.024	0.598	0.683
	Subcobertura	Entrenamiento	0.528	0.019	0.499	0.558
oriente	Precisión	Validación	0.170	0.017	0.139	0.195
	Precisión	Entrenamiento	0.191	0.006	0.182	0.201
	Sensibilidad	Validación	0.292	0.037	0.223	0.350
	Sensibilidad	Entrenamiento	0.331	0.018	0.299	0.363
	Especificidad	Validación	0.977	0.002	0.974	0.980
	Especificidad	Entrenamiento	0.977	0.001	0.975	0.980
	F1	Validación	0.215	0.022	0.171	0.250
	F1	Entrenamiento	0.242	0.007	0.234	0.251
	PRAUC	Validación	0.137	0.022	0.109	0.171
	PRAUC	Entrenamiento	0.179	0.006	0.171	0.189
	ROCAUC	Validación	0.885	0.016	0.858	0.924
	ROCAUC	Entrenamiento	0.950	0.001	0.949	0.951
	Filtración	Validación	0.830	0.017	0.805	0.861
	Filtración	Entrenamiento	0.809	0.006	0.799	0.818
	Subcobertura	Validación	0.708	0.037	0.650	0.777
	Subcobertura	Entrenamiento	0.669	0.018	0.637	0.701

Tabla 27

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.156	0.023	0.113	0.184
	Precisión	Entrenamiento	0.172	0.003	0.165	0.175
	Sensibilidad	Validación	0.274	0.031	0.220	0.319
	Sensibilidad	Entrenamiento	0.305	0.011	0.290	0.329
	Especificidad	Validación	0.982	0.002	0.979	0.984
	Especificidad	Entrenamiento	0.982	0.001	0.980	0.983
	F1	Validación	0.199	0.026	0.149	0.232
	F1	Entrenamiento	0.220	0.004	0.213	0.225
	PRAUC	Validación	0.111	0.019	0.095	0.148
	PRAUC	Entrenamiento	0.129	0.002	0.127	0.131
	ROCAUC	Validación	0.868	0.012	0.848	0.886
	ROCAUC	Entrenamiento	0.924	0.001	0.922	0.925
	Filtración	Validación	0.844	0.023	0.816	0.887
	Filtración	Entrenamiento	0.828	0.003	0.825	0.835
	Subcobertura	Validación	0.726	0.031	0.681	0.780
	Subcobertura	Entrenamiento	0.695	0.011	0.671	0.710
norte	Precisión	Validación	0.135	0.013	0.113	0.155
	Precisión	Entrenamiento	0.163	0.003	0.157	0.167
	Sensibilidad	Validación	0.288	0.042	0.191	0.333
	Sensibilidad	Entrenamiento	0.359	0.014	0.334	0.376
	Especificidad	Validación	0.980	0.002	0.978	0.985
	Especificidad	Entrenamiento	0.980	0.001	0.979	0.982
	F1	Validación	0.184	0.020	0.147	0.204
	F1	Entrenamiento	0.224	0.004	0.217	0.229
	PRAUC	Validación	0.106	0.019	0.084	0.146
	PRAUC	Entrenamiento	0.145	0.003	0.139	0.149
	ROCAUC	Validación	0.882	0.011	0.867	0.906
	ROCAUC	Entrenamiento	0.949	0.001	0.948	0.950
	Filtración	Validación	0.865	0.013	0.845	0.887
	Filtración	Entrenamiento	0.837	0.003	0.833	0.843
	Subcobertura	Validación	0.712	0.042	0.667	0.809
	Subcobertura	Entrenamiento	0.641	0.014	0.624	0.666
sur	Precisión	Validación	0.101	0.019	0.075	0.135
	Precisión	Entrenamiento	0.194	0.016	0.169	0.217
	Sensibilidad	Validación	0.292	0.054	0.204	0.367
	Sensibilidad	Entrenamiento	0.610	0.062	0.530	0.689
	Especificidad	Validación	0.985	0.002	0.981	0.987
	Especificidad	Entrenamiento	0.986	0.000	0.985	0.986
	F1	Validación	0.150	0.027	0.110	0.195
	F1	Entrenamiento	0.294	0.025	0.257	0.330
	PRAUC	Validación	0.088	0.031	0.051	0.157
	PRAUC	Entrenamiento	0.211	0.016	0.183	0.235
	ROCAUC	Validación	0.891	0.030	0.826	0.934

	ROCAUC	Entrenamiento	0.986	0.001	0.985	0.987
	Filtración	Validación	0.899	0.019	0.865	0.925
	Filtración	Entrenamiento	0.806	0.016	0.783	0.831
	Subcobertura	Validación	0.708	0.054	0.633	0.796
	Subcobertura	Entrenamiento	0.390	0.062	0.311	0.470
centro	Precisión	Validación	0.114	0.011	0.089	0.127
	Precisión	Entrenamiento	0.148	0.006	0.137	0.157
	Sensibilidad	Validación	0.299	0.035	0.219	0.346
	Sensibilidad	Entrenamiento	0.402	0.016	0.365	0.421
	Especificidad	Validación	0.979	0.001	0.976	0.980
	Especificidad	Entrenamiento	0.979	0.001	0.977	0.982
	F1	Validación	0.165	0.017	0.127	0.182
	F1	Entrenamiento	0.217	0.006	0.205	0.226
	PRAUC	Validación	0.096	0.011	0.078	0.112
	PRAUC	Entrenamiento	0.141	0.006	0.127	0.149
	ROCAUC	Validación	0.892	0.014	0.865	0.916
	ROCAUC	Entrenamiento	0.958	0.001	0.956	0.959
	Filtración	Validación	0.886	0.011	0.873	0.911
	Filtración	Entrenamiento	0.852	0.006	0.843	0.863
	Subcobertura	Validación	0.701	0.035	0.654	0.781
	Subcobertura	Entrenamiento	0.598	0.016	0.579	0.635
oriente	Precisión	Validación	0.150	0.012	0.132	0.168
	Precisión	Entrenamiento	0.178	0.007	0.167	0.187
	Sensibilidad	Validación	0.211	0.017	0.185	0.241
	Sensibilidad	Entrenamiento	0.253	0.023	0.213	0.289
	Especificidad	Validación	0.979	0.001	0.977	0.982
	Especificidad	Entrenamiento	0.979	0.002	0.976	0.983
	F1	Validación	0.175	0.013	0.154	0.198
	F1	Entrenamiento	0.209	0.009	0.194	0.224
	PRAUC	Validación	0.108	0.014	0.093	0.135
	PRAUC	Entrenamiento	0.156	0.005	0.147	0.163
	ROCAUC	Validación	0.857	0.012	0.840	0.874
	ROCAUC	Entrenamiento	0.938	0.001	0.936	0.941
	Filtración	Validación	0.850	0.012	0.832	0.868
	Filtración	Entrenamiento	0.822	0.007	0.813	0.833
	Subcobertura	Validación	0.789	0.017	0.759	0.815
	Subcobertura	Entrenamiento	0.747	0.023	0.711	0.787

Tabla 28

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.138	0.017	0.113	0.175
	Precisión	Entrenamiento	0.159	0.005	0.155	0.170
	Sensibilidad	Validación	0.258	0.026	0.210	0.292
	Sensibilidad	Entrenamiento	0.307	0.009	0.298	0.328
	Especificidad	Validación	0.980	0.002	0.977	0.983
	Especificidad	Entrenamiento	0.980	0.001	0.978	0.982
	F1	Validación	0.179	0.020	0.149	0.218
	F1	Entrenamiento	0.210	0.004	0.204	0.218
	PRAUC	Validación	0.102	0.015	0.074	0.124
	PRAUC	Entrenamiento	0.129	0.002	0.125	0.131
	ROCAUC	Validación	0.866	0.010	0.847	0.880
	ROCAUC	Entrenamiento	0.925	0.001	0.923	0.927
	Filtración	Validación	0.862	0.017	0.825	0.887
	Filtración	Entrenamiento	0.841	0.005	0.830	0.845
	Subcobertura	Validación	0.742	0.026	0.708	0.790
	Subcobertura	Entrenamiento	0.693	0.009	0.672	0.702
norte	Precisión	Validación	0.145	0.013	0.129	0.172
	Precisión	Entrenamiento	0.170	0.005	0.161	0.175
	Sensibilidad	Validación	0.305	0.036	0.248	0.361
	Sensibilidad	Entrenamiento	0.362	0.014	0.337	0.380
	Especificidad	Validación	0.979	0.002	0.976	0.982
	Especificidad	Entrenamiento	0.980	0.001	0.977	0.982
	F1	Validación	0.196	0.018	0.174	0.232
	F1	Entrenamiento	0.231	0.004	0.226	0.238
	PRAUC	Validación	0.111	0.018	0.086	0.145
	PRAUC	Entrenamiento	0.147	0.003	0.140	0.152
	ROCAUC	Validación	0.876	0.014	0.854	0.903
	ROCAUC	Entrenamiento	0.947	0.001	0.946	0.949
	Filtración	Validación	0.855	0.013	0.828	0.871
	Filtración	Entrenamiento	0.830	0.005	0.825	0.839
	Subcobertura	Validación	0.695	0.036	0.639	0.752
	Subcobertura	Entrenamiento	0.638	0.014	0.620	0.663
sur	Precisión	Validación	0.115	0.016	0.084	0.141
	Precisión	Entrenamiento	0.186	0.015	0.152	0.206
	Sensibilidad	Validación	0.292	0.042	0.241	0.358
	Sensibilidad	Entrenamiento	0.499	0.027	0.464	0.552
	Especificidad	Validación	0.986	0.002	0.983	0.988
	Especificidad	Entrenamiento	0.986	0.001	0.983	0.988
	F1	Validación	0.165	0.022	0.124	0.202
	F1	Entrenamiento	0.271	0.019	0.230	0.300
	PRAUC	Validación	0.082	0.016	0.058	0.107
	PRAUC	Entrenamiento	0.192	0.019	0.154	0.217
	ROCAUC	Validación	0.864	0.032	0.823	0.923

	ROCAUC	Entrenamiento	0.981	0.001	0.980	0.983
	Filtración	Validación	0.885	0.016	0.859	0.916
	Filtración	Entrenamiento	0.814	0.015	0.794	0.848
	Subcobertura	Validación	0.708	0.042	0.642	0.759
	Subcobertura	Entrenamiento	0.501	0.027	0.448	0.536
centro	Precisión	Validación	0.124	0.020	0.090	0.168
	Precisión	Entrenamiento	0.144	0.006	0.137	0.160
	Sensibilidad	Validación	0.361	0.055	0.239	0.469
	Sensibilidad	Entrenamiento	0.418	0.015	0.393	0.440
	Especificidad	Validación	0.975	0.002	0.972	0.977
	Especificidad	Entrenamiento	0.975	0.002	0.974	0.979
	F1	Validación	0.185	0.029	0.131	0.248
	F1	Entrenamiento	0.214	0.006	0.204	0.229
	PRAUC	Validación	0.103	0.014	0.084	0.135
	PRAUC	Entrenamiento	0.139	0.004	0.133	0.144
	ROCAUC	Validación	0.888	0.010	0.869	0.906
	ROCAUC	Entrenamiento	0.955	0.001	0.954	0.957
	Filtración	Validación	0.876	0.020	0.832	0.910
	Filtración	Entrenamiento	0.856	0.006	0.840	0.863
	Subcobertura	Validación	0.639	0.055	0.531	0.761
	Subcobertura	Entrenamiento	0.582	0.015	0.560	0.607
oriente	Precisión	Validación	0.156	0.020	0.127	0.187
	Precisión	Entrenamiento	0.186	0.008	0.177	0.201
	Sensibilidad	Validación	0.238	0.029	0.189	0.286
	Sensibilidad	Entrenamiento	0.286	0.006	0.274	0.294
	Especificidad	Validación	0.978	0.001	0.976	0.980
	Especificidad	Entrenamiento	0.979	0.001	0.977	0.980
	F1	Validación	0.188	0.023	0.152	0.226
	F1	Entrenamiento	0.226	0.007	0.216	0.238
	PRAUC	Validación	0.131	0.014	0.107	0.149
	PRAUC	Entrenamiento	0.193	0.005	0.186	0.202
	ROCAUC	Validación	0.861	0.015	0.837	0.881
	ROCAUC	Entrenamiento	0.953	0.001	0.952	0.955
	Filtración	Validación	0.844	0.020	0.813	0.873
	Filtración	Entrenamiento	0.814	0.008	0.799	0.823
	Subcobertura	Validación	0.762	0.029	0.714	0.811
	Subcobertura	Entrenamiento	0.714	0.006	0.706	0.726

Tabla 29

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.114	0.015	0.098	0.150
	Precisión	Entrenamiento	0.132	0.005	0.124	0.144
	Sensibilidad	Validación	0.280	0.034	0.222	0.342
	Sensibilidad	Entrenamiento	0.327	0.009	0.311	0.339
	Especificidad	Validación	0.977	0.001	0.975	0.980
	Especificidad	Entrenamiento	0.977	0.001	0.975	0.979
	F1	Validación	0.162	0.020	0.137	0.209
	F1	Entrenamiento	0.188	0.006	0.180	0.202
	PRAUC	Validación	0.088	0.013	0.075	0.117
	PRAUC	Entrenamiento	0.113	0.003	0.108	0.120
	ROCAUC	Validación	0.858	0.012	0.840	0.878
	ROCAUC	Entrenamiento	0.929	0.001	0.927	0.931
	Filtración	Validación	0.886	0.015	0.850	0.902
	Filtración	Entrenamiento	0.868	0.005	0.856	0.876
	Subcobertura	Validación	0.720	0.034	0.658	0.778
	Subcobertura	Entrenamiento	0.673	0.009	0.661	0.689
norte	Precisión	Validación	0.135	0.017	0.103	0.163
	Precisión	Entrenamiento	0.161	0.006	0.152	0.170
	Sensibilidad	Validación	0.311	0.038	0.254	0.373
	Sensibilidad	Entrenamiento	0.375	0.022	0.340	0.405
	Especificidad	Validación	0.979	0.002	0.977	0.982
	Especificidad	Entrenamiento	0.980	0.001	0.978	0.982
	F1	Validación	0.188	0.023	0.147	0.224
	F1	Entrenamiento	0.225	0.008	0.210	0.236
	PRAUC	Validación	0.098	0.016	0.069	0.128
	PRAUC	Entrenamiento	0.139	0.005	0.130	0.146
	ROCAUC	Validación	0.870	0.017	0.837	0.892
	ROCAUC	Entrenamiento	0.952	0.001	0.950	0.954
	Filtración	Validación	0.865	0.017	0.837	0.897
	Filtración	Entrenamiento	0.839	0.006	0.830	0.848
	Subcobertura	Validación	0.689	0.038	0.627	0.746
	Subcobertura	Entrenamiento	0.625	0.022	0.595	0.660
sur	Precisión	Validación	0.115	0.022	0.079	0.163
	Precisión	Entrenamiento	0.191	0.013	0.171	0.222
	Sensibilidad	Validación	0.309	0.060	0.196	0.412
	Sensibilidad	Entrenamiento	0.567	0.046	0.484	0.652
	Especificidad	Validación	0.986	0.001	0.984	0.988
	Especificidad	Entrenamiento	0.986	0.001	0.984	0.987
	F1	Validación	0.167	0.032	0.112	0.229
	F1	Entrenamiento	0.285	0.019	0.261	0.331
	PRAUC	Validación	0.086	0.019	0.064	0.130
	PRAUC	Entrenamiento	0.193	0.016	0.169	0.215
	ROCAUC	Validación	0.868	0.024	0.821	0.914

	ROCAUC	Entrenamiento	0.983	0.001	0.982	0.985
	Filtración	Validación	0.885	0.022	0.837	0.921
	Filtración	Entrenamiento	0.809	0.013	0.778	0.829
	Subcobertura	Validación	0.691	0.060	0.588	0.804
	Subcobertura	Entrenamiento	0.433	0.046	0.348	0.516
centro	Precisión	Validación	0.115	0.024	0.072	0.163
	Precisión	Entrenamiento	0.149	0.006	0.136	0.158
	Sensibilidad	Validación	0.313	0.071	0.181	0.436
	Sensibilidad	Entrenamiento	0.419	0.017	0.392	0.454
	Especificidad	Validación	0.980	0.002	0.976	0.982
	Especificidad	Entrenamiento	0.980	0.001	0.978	0.982
	F1	Validación	0.168	0.035	0.103	0.237
	F1	Entrenamiento	0.220	0.007	0.206	0.227
	PRAUC	Validación	0.095	0.023	0.060	0.144
	PRAUC	Entrenamiento	0.138	0.006	0.128	0.147
	ROCAUC	Validación	0.870	0.019	0.842	0.902
	ROCAUC	Entrenamiento	0.961	0.001	0.960	0.962
	Filtración	Validación	0.885	0.024	0.837	0.928
	Filtración	Entrenamiento	0.851	0.006	0.842	0.864
	Subcobertura	Validación	0.687	0.071	0.564	0.819
	Subcobertura	Entrenamiento	0.581	0.017	0.546	0.608
oriente	Precisión	Validación	0.125	0.032	0.075	0.164
	Precisión	Entrenamiento	0.178	0.010	0.168	0.195
	Sensibilidad	Validación	0.230	0.069	0.125	0.333
	Sensibilidad	Entrenamiento	0.336	0.020	0.295	0.364
	Especificidad	Validación	0.980	0.001	0.979	0.981
	Especificidad	Entrenamiento	0.980	0.001	0.978	0.982
	F1	Validación	0.162	0.044	0.094	0.220
	F1	Entrenamiento	0.232	0.012	0.214	0.254
	PRAUC	Validación	0.094	0.027	0.054	0.138
	PRAUC	Entrenamiento	0.179	0.006	0.168	0.188
	ROCAUC	Validación	0.862	0.027	0.812	0.889
	ROCAUC	Entrenamiento	0.969	0.001	0.967	0.970
	Filtración	Validación	0.875	0.032	0.836	0.925
	Filtración	Entrenamiento	0.822	0.010	0.805	0.832
	Subcobertura	Validación	0.770	0.069	0.667	0.875
	Subcobertura	Entrenamiento	0.664	0.020	0.636	0.705

Tabla 30

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

Macro región	Métrica	Datos	Promedio	DE	Mín.	Máx.
lima_metro_callao	Precisión	Validación	0.008	0.001	0.006	0.011
	Precisión	Entrenamiento	0.016	0.001	0.014	0.018
	Sensibilidad	Validación	0.478	0.071	0.333	0.611
	Sensibilidad	Entrenamiento	0.970	0.016	0.938	0.988
	Especificidad	Validación	0.923	0.007	0.911	0.933
	Especificidad	Entrenamiento	0.925	0.006	0.915	0.933
	F1	Validación	0.015	0.003	0.011	0.022
	F1	Entrenamiento	0.031	0.003	0.027	0.036
	PRAUC	Validación	0.021	0.018	0.007	0.071
	PRAUC	Entrenamiento	0.170	0.031	0.114	0.204
	ROCAUC	Validación	0.800	0.055	0.695	0.885
	ROCAUC	Entrenamiento	0.984	0.002	0.982	0.987
	Filtración	Validación	0.992	0.001	0.989	0.994
	Filtración	Entrenamiento	0.984	0.001	0.982	0.986
	Subcobertura	Validación	0.522	0.071	0.389	0.667
	Subcobertura	Entrenamiento	0.030	0.016	0.012	0.062
norte	Precisión	Validación	0.017	0.004	0.011	0.023
	Precisión	Entrenamiento	0.030	0.001	0.028	0.032
	Sensibilidad	Validación	0.559	0.121	0.385	0.769
	Sensibilidad	Entrenamiento	1.000	0.000	1.000	1.000
	Especificidad	Validación	0.960	0.003	0.953	0.964
	Especificidad	Entrenamiento	0.961	0.002	0.958	0.964
	F1	Validación	0.032	0.007	0.022	0.045
	F1	Entrenamiento	0.058	0.002	0.054	0.063
	PRAUC	Validación	0.059	0.034	0.020	0.107
	PRAUC	Entrenamiento	0.596	0.092	0.470	0.776
	ROCAUC	Validación	0.873	0.036	0.789	0.930
	ROCAUC	Entrenamiento	0.998	0.000	0.997	0.999
	Filtración	Validación	0.983	0.004	0.977	0.989
	Filtración	Entrenamiento	0.970	0.001	0.968	0.972
	Subcobertura	Validación	0.441	0.121	0.231	0.615
	Subcobertura	Entrenamiento	0.000	0.000	0.000	0.000
sur	Precisión	Validación	0.011	0.009	0.000	0.034
	Precisión	Entrenamiento	0.055	0.007	0.048	0.071
	Sensibilidad	Validación	0.220	0.166	0.000	0.600
	Sensibilidad	Entrenamiento	1.000	0.000	1.000	1.000
	Especificidad	Validación	0.988	0.001	0.986	0.991
	Especificidad	Entrenamiento	0.990	0.001	0.988	0.992
	F1	Validación	0.020	0.017	0.000	0.064
	F1	Entrenamiento	0.103	0.013	0.091	0.132
	PRAUC	Validación	0.045	0.063	0.000	0.208
	PRAUC	Entrenamiento	0.958	0.015	0.936	0.981
	ROCAUC	Validación	0.533	0.201	0.138	0.816

	ROCAUC	Entrenamiento	1.000	0.000	1.000	1.000
	Filtración	Validación	0.989	0.009	0.966	1.000
	Filtración	Entrenamiento	0.945	0.007	0.929	0.952
	Subcobertura	Validación	0.780	0.166	0.400	1.000
	Subcobertura	Entrenamiento	0.000	0.000	0.000	0.000
centro	Precisión	Validación	0.013	0.009	0.005	0.037
	Precisión	Entrenamiento	0.036	0.005	0.029	0.048
	Sensibilidad	Validación	0.396	0.236	0.125	0.875
	Sensibilidad	Entrenamiento	1.000	0.000	1.000	1.000
	Especificidad	Validación	0.980	0.003	0.975	0.986
	Especificidad	Entrenamiento	0.981	0.003	0.977	0.986
	F1	Validación	0.026	0.017	0.009	0.071
	F1	Entrenamiento	0.069	0.010	0.057	0.091
	PRAUC	Validación	0.170	0.160	0.006	0.476
	PRAUC	Entrenamiento	0.879	0.050	0.773	0.954
	ROCAUC	Validación	0.702	0.170	0.484	0.992
	ROCAUC	Entrenamiento	1.000	0.000	1.000	1.000
	Filtración	Validación	0.987	0.009	0.963	0.995
	Filtración	Entrenamiento	0.964	0.005	0.952	0.971
	Subcobertura	Validación	0.604	0.236	0.125	0.875
	Subcobertura	Entrenamiento	0.000	0.000	0.000	0.000
oriente	Precisión	Validación	0.064	0.027	0.026	0.111
	Precisión	Entrenamiento	0.163	0.012	0.148	0.187
	Sensibilidad	Validación	0.387	0.180	0.167	0.667
	Sensibilidad	Entrenamiento	1.000	0.000	1.000	1.000
	Especificidad	Validación	0.992	0.002	0.990	0.995
	Especificidad	Entrenamiento	0.993	0.001	0.992	0.994
	F1	Validación	0.109	0.046	0.047	0.182
	F1	Entrenamiento	0.280	0.017	0.259	0.315
	PRAUC	Validación	0.158	0.128	0.025	0.451
	PRAUC	Entrenamiento	0.926	0.053	0.789	0.986
	ROCAUC	Validación	0.870	0.117	0.545	0.981
	ROCAUC	Entrenamiento	1.000	0.000	1.000	1.000
	Filtración	Validación	0.936	0.027	0.889	0.974
	Filtración	Entrenamiento	0.837	0.012	0.813	0.852
	Subcobertura	Validación	0.613	0.180	0.333	0.833
	Subcobertura	Entrenamiento	0.000	0.000	0.000	0.000