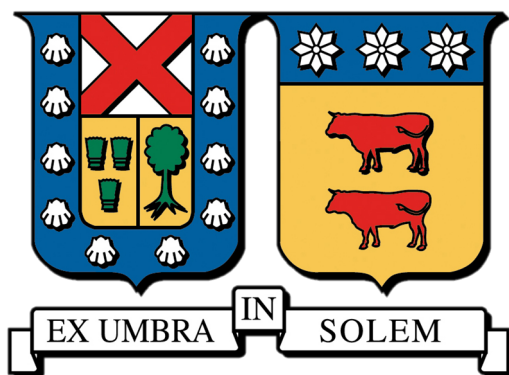




UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS

**Predicción e interpretación del riesgo de no continuidad escolar
en Chile: evidencia por tramos educativos y cohortes mediante
regresión logística y aprendizaje automático**

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL INDUSTRIAL

AUTOR: SOFIA CATALINA FERREIRA GÄDICKE

PROFESOR GUÍA: DR. BERNARDO PINCHEIRA S.

PROFESOR CORREFERENTE: MG. RODOLFO SALAZAR A.

ABRIL 2026



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Predicción e interpretación del riesgo de no continuidad escolar en Chile: evidencia por tramos educativos y cohortes mediante regresión logística y aprendizaje automático.

Nombre del candidato(a): Sofia Catalina Ferreira Gädicke

Carrera / Grado: Ingeniería Civil Industrial

Campus: Casa Central Valparaíso

Departamento: Departamento de Industrias

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Bernardo Pincheira Sarmiento, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 16 de junio 2026

Firma: 

Estudiante o Candidato(a):

Fecha: 16 de junio 2026

Firma: 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

TABLA DE CONTENIDOS

Índice de tablas	4
Índice de figuras	7
1. Problema de investigación	8
2. Objetivos	13
2.1. Objetivo General	13
2.2. Objetivos específicos	13
3. Marco teórico	15
3.1. Trayectorias educativas, continuidad escolar y derecho a la educación	17
3.2. No continuidad escolar, abandono, interrupción y revinculación	19
3.3. Factores asociados a la no continuidad escolar	20
3.4. Transiciones críticas, heterogeneidad temporal y cohortes	23
3.5. Enfoques predictivos, sistemas de alerta temprana y posicionamiento de la investigación	24
3.6. Síntesis del marco teórico	28
4. Datos	30
4.1. Fuentes de información	30
4.1.1. Registros administrativos de matrícula y rendimiento escolar	30
4.1.2. Evaluaciones SIMCE y cuestionarios de contexto	31
4.1.3. Información de los establecimientos	32
4.2. Definición de cohortes de estudio	33
4.3. Construcción del panel longitudinal	35
4.4. Etiqueta de no continuidad escolar	38
4.4.1. Continuidad escolar anual, observaciones elegibles y casos censurados	39
4.4.2. Clasificación en no continuidad temprana, intermedia y tardía	40
4.5. Variables de trayectoria	41
4.5.1. Edad y sobriedad	41
4.5.2. Movilidad entre establecimientos	42
4.5.3. Interrupciones previas y reingreso	43
4.5.4. Variables académicas rezagadas (enfoque mid-year)	44

4.6. Construcción de variables desde SIMCE	44
4.6.1. Puntajes académicos	44
4.6.2. Dimensiones psicosociales y de contexto escolar y familiar	44
4.6.3. Cobertura desigual y marcador de disponibilidad	47
4.7. Descripción de la muestra y cobertura de datos	47
4.7.1. Tamaño de la muestra modelable	48
4.7.2. Censura de la etiqueta	50
4.7.3. Prevalencia del evento por cohorte teórica	50
4.7.4. Prevalencia del evento por cohorte y tramo educativo	51
4.7.5. Disponibilidad de SIMCE y cobertura de variables seleccionadas	52
5. Metodología	56
5.1. Diseño predictivo y objetivo del estudio	56
5.2. Construcción de la etiqueta anual y tratamiento de censura	57
5.3. Escenario operativo y prevención de filtración de información	57
5.4. Niveles de información	58
5.5. Modelos comparados	59
5.6. Estrategias de partición y evaluación fuera de muestra	60
5.7. Métricas de desempeño y priorización operativa	61
5.8. Calibración del riesgo	63
5.9. Análisis por tramo educativo	64
5.10. Síntesis del flujo metodológico	64
6. Resultados	66
6.1. Comparación de ventanas y modelos	66
6.2. Generalización temporal entre cohortes y benchmark por estudiante	68
6.3. Desempeño global del modelo principal	69
6.4. Curvas ROC, PR y distribución de puntajes	70
6.5. Heterogeneidad del desempeño por tramo educativo	72
6.6. Calibración del riesgo	74
6.7. Heterogeneidad del desempeño entre cohortes	75
6.8. Estabilidad y variación de asociaciones entre cohortes	77
6.9. Relevancia de variables y lectura sustantiva del riesgo	78
6.10. Estabilidad y variación de asociaciones entre tramos educativos	80
6.11. Análisis de sensibilidad: incorporación exploratoria de dimensiones psicosociales	82

6.12. Robustez: particiones alternativas y estabilidad estadística	83
6.13. Alcance conceptual de la etiqueta anual	84
6.14. Síntesis de resultados	84
7. Conclusiones	86
7.1. Conclusiones generales	86
7.2. Hallazgos principales del estudio	87
7.3. Aporte de la investigación	89
7.4. Implicancias sustantivas y operativas	90
7.5. Limitaciones del estudio	91
7.6. Líneas futuras de investigación	93
7.7. Cierre	94
8. Anexos	101
8.1. Mapeo de dimensiones psicosociales	101
8.2. Screening exploratorio de dimensiones psicosociales sobre W2	103
8.3. Resultados complementarios de modelos logísticos por tramo	105
8.4. Resultados complementarios de modelos logísticos por cohorte	107
8.5. Especificación operativa de variables por ventana de información	108
8.6. Superposición entre no continuidad anual y desvinculación persistente	108

Índice de tablas

4.1. Años de aplicación SIMCE observados para cada cohorte de estudio escogida. . .	34
4.2. Ejemplos de elegibilidad y tratamiento de la etiqueta anual	40
4.3. Cobertura temporal y tamaño del panel total versus muestra modelable	49
4.4. Muestra modelable y tasa de no continuidad anual por tramo educativo	49
4.5. Censura de la etiqueta anual: motivos y solapamientos	50
4.6. Muestra modelable y tasa de no continuidad anual por cohorte teórica de egreso	51
4.7. Tasa de no continuidad anual por cohorte teórica y tramo educativo	52
4.8. Proporción de observaciones con al menos un antecedente SIMCE disponible . .	53
4.9. Cobertura de variables seleccionadas por tramo	54
6.1. Desempeño fuera de muestra por ventana y modelo en el escenario principal E1 (entrenamiento en cohortes 2017 y 2021; prueba en cohorte 2024).	67
6.2. Desempeño bajo benchmark de partición 80/20 por estudiante (holdout por MRUN).	68
6.3. Desempeño global del modelo principal seleccionado (E1, W2, HGB).	69
6.4. Desempeño por tramo educativo del modelo principal (E1, W2, HGB).	72
6.5. Métricas complementarias al 10 % por tramo educativo del modelo principal (E1, W2, HGB).	73

6.6. Calibración por deciles de riesgo del modelo principal (E1, W2, HGB).	74
6.7. Desempeño del modelo principal (W2, HGB) en ejercicios de transferencia temporal entre cohortes.	76
6.8. Capacidad predictiva por cohorte manteniendo fijo el modelo y el conjunto de variables (W2, HGB).	76
6.9. Odds ratios de variables clave en modelos logísticos parsimoniosos por cohorte. .	77
6.10. Variables más relevantes del modelo principal según importancia por permutación.	80
6.11. Odds ratios de variables numéricas clave en modelos logísticos parsimoniosos por tramo educativo.	81
6.12. Intervalos bootstrap para métricas seleccionadas del modelo principal (E1, W2, HGB).	83
8.1. Resumen del mapeo de dimensiones psicosociales y de contexto construidas a partir de cuestionarios SIMCE.	101
8.2. Screening exploratorio de dimensiones psicosociales sobre la ventana W2 en el escenario principal E1.	103
8.3. Resumen de ajuste de modelos logísticos parsimoniosos por tramo educativo. . .	105
8.4. Resultados completos de modelos logísticos parsimoniosos por tramo educativo. .	106
8.5. Resumen de ajuste de modelos logísticos parsimoniosos por cohorte teórica. . . .	107
8.6. Especificación operativa de variables por ventana de información.	108

8.7. Superposición entre no continuidad anual y desvinculación persistente observada a nivel estudiante.	109
--	-----

Índice de figuras

6.1. Curva ROC del modelo principal (E1, W2, HGB)	71
6.2. Curva precisión–recall del modelo principal (E1, W2, HGB)	71
6.3. Distribución de puntajes predichos del modelo principal (E1, W2, HGB)	72
6.4. Diagrama de confiabilidad del modelo principal (E1, W2, HGB)	75

1. Problema de investigación

La continuidad de las trayectorias educativas constituye un desafío persistente para los sistemas escolares, en tanto se relaciona directamente con el ejercicio efectivo del derecho a la educación, la inclusión social y las oportunidades futuras de niños, niñas y adolescentes. A nivel global, el problema continúa siendo de gran magnitud: UNESCO reportó que en 2024 había 273 millones de niños, adolescentes y jóvenes fuera de la escuela, equivalente a uno de cada seis en edad escolar, y advirtió que el avance en acceso y finalización se ha desacelerado en la última década (UNESCO, 2026). En este escenario, la permanencia escolar ya no puede entenderse solo como una cuestión de cobertura o matrícula, sino como la capacidad de los sistemas educativos para sostener trayectorias continuas, pertinentes y completas a lo largo del tiempo.

En Chile, esta preocupación también ha adquirido mayor densidad durante los últimos años. Aunque las cifras más recientes muestran una mejora relevante -la tasa de desvinculación escolar bajó de 1,7% en 2022 a 1,3% en 2024, y en 2025 la tasa de revinculación alcanzó 40,8%, equivalente a 18.442 estudiantes reincorporados al sistema- (Ministerio de Educación, 2025b; Ministerio de Educación, Centro de Estudios, 2025), ello no implica que el problema haya dejado de ser relevante. Más bien, sugiere que las trayectorias escolares deben pensarse de manera más fina: no solo en términos de acceso o egreso, sino también de interrupción, persistencia, retorno y vulnerabilidad diferencial entre grupos y territorios (Ministerio de Educación, 2025a). En otras palabras, incluso en un escenario de mejora agregada, sigue siendo necesario comprender qué factores configuran el riesgo de no continuidad escolar, para quiénes se expresa con mayor intensidad y en qué momentos del recorrido educativo se vuelve más probable.

Esta necesidad se refuerza por la complejidad conceptual del fenómeno. La literatura chilena

y la documentación técnica han mostrado que nociones como deserción escolar, desvinculación, exclusión educativa y trayectorias interrumpidas no son completamente equivalentes. Mientras algunas aproximaciones se centran en el abandono como no culminación de un nivel educativo, otras enfatizan la interrupción observable de la trayectoria, la salida del sistema regular o los procesos más amplios de exclusión. Esta distinción no es meramente terminológica: condiciona la forma de medir el problema, la lectura de sus causas y el tipo de respuestas que pueden proponerse. En investigaciones basadas en registros administrativos longitudinales, como la presente, esta discusión adquiere especial importancia, porque obliga a desplazar la mirada desde un evento terminal único hacia la observación de trayectorias que pueden presentar continuidad, interrupciones temporales o quiebres más persistentes (Romo & Cumsille, 2020).

La evidencia disponible coincide, además, en que la no continuidad escolar es un fenómeno multicausal. Los estudios realizados en Chile han asociado el riesgo de abandono o interrupción con factores individuales y sociodemográficos, como sexo, pobreza, ruralidad o rezago escolar; con antecedentes académicos y de trayectoria, como bajo rendimiento, asistencia deficiente, sobreedad o repitencia; y también con dimensiones relacionales e institucionales, como el clima escolar, las relaciones entre estudiantes y docentes y la capacidad de la escuela para sostener vínculos y apoyos significativos. En particular, la investigación longitudinal de Contreras et al. (2022) mostró que la percepción de un clima relacional negativo incrementa la probabilidad de abandono escolar, incluso controlando por otras características del estudiante y del establecimiento. Estos hallazgos refuerzan la idea de que la interrupción de la trayectoria educativa no puede atribuirse únicamente a decisiones individuales, sino que se configura en la intersección entre biografía escolar, condiciones sociales, entorno familiar, experiencia institucional y contexto territorial (Ministerio de Educación, Centro de Estudios, 2020).

En ese marco, el problema involucra a múltiples agentes. En primer lugar, a los propios estudiantes, cuyas trayectorias se ven afectadas por condiciones académicas, emocionales y materiales diversas. En segundo lugar, a las familias y redes cercanas, que cumplen un papel clave en las expectativas educacionales, el apoyo cotidiano y la capacidad de sostener la escolaridad. En tercer lugar, a los establecimientos educacionales y sus equipos, cuya organización, clima, prácticas pedagógicas y dispositivos de acompañamiento pueden contribuir a prevenir o profundizar trayectorias vulnerables. Finalmente, el problema compromete también a sostenedores y organismos públicos, en tanto participan en la definición de políticas, sistemas de seguimiento, asignación de recursos y estrategias de prevención o revinculación. Por ello, estudiar la no continuidad escolar exige una mirada que integre niveles de análisis distintos y que no reduzca la explicación del fenómeno a un único actor o dimensión (Gómez Vera et al., 2021).

A lo anterior se suma que en Chile ya existe una línea de trabajo predictivo y de alerta temprana sobre el problema, tanto desde la política pública como desde la investigación académica. El estudio de Rodríguez et al. (2023) mostró que es posible desarrollar modelos de aprendizaje automático para predecir abandono escolar a partir de trayectorias individuales y factores de distintos niveles, mientras que la evolución reciente de sistemas de alerta y protección de trayectorias confirma la relevancia práctica de anticipar riesgos antes de que la desvinculación se consolide. Sin embargo, la existencia de estas herramientas no agota el problema de investigación. Persisten preguntas relevantes sobre cómo varían los factores asociados a la no continuidad escolar según el contexto geográfico o sociocultural del estudiante, cómo se distribuye el riesgo en distintos tramos de la trayectoria educativa y en qué medida ciertos patrones se mantienen o cambian entre cohortes.

En consecuencia, el problema de investigación de este trabajo se sitúa en la necesidad de avanzar desde una medición agregada de la desvinculación hacia una comprensión más fina del riesgo

de interrupción educativa observado año a año. Si bien la evidencia nacional e internacional ha mostrado que la continuidad escolar se relaciona con múltiples dimensiones de la experiencia educativa, persiste la necesidad de analizar cómo esas dimensiones se articulan en registros longitudinales, qué señales permiten anticipar episodios de no continuidad y si dichas señales se comportan de manera estable entre distintos momentos del recorrido escolar y entre cohortes de seguimiento.

Desde esta perspectiva, la no continuidad escolar anual se entiende como una manifestación observable de fragilidad dentro de la trayectoria educativa. Su análisis no busca reemplazar la noción de abandono o desvinculación, sino aportar una entrada analítica más cercana a la prevención: observar interrupciones anuales, distinguir tramos del recorrido escolar y evaluar si los patrones de riesgo se mantienen o se modifican entre cohortes. La variable principal de interés corresponde, por tanto, a la ocurrencia de no continuidad escolar anual, mientras que las variables explicativas consideran factores individuales, académicos, escolares, territoriales y de trayectoria.

A partir de ello, la investigación busca analizar cómo se articula el riesgo de interrupción educativa, cómo se expresa en distintos tramos del recorrido escolar y en qué medida sus patrones permanecen o se modifican entre cohortes, con el propósito de aportar evidencia que permita comprender mejor las trayectorias en riesgo y fortalecer una identificación temprana más precisa.

A partir de este marco, las preguntas de investigación que orientan el trabajo son:

1. ¿Qué factores individuales, académicos, escolares y territoriales se asocian con la no continuidad escolar en estudiantes del sistema escolar chileno, y cómo varían según el contexto geográfico o sociocultural del estudiante y del establecimiento educacional?
2. ¿En qué etapas o tramos de la trayectoria educativa se concentra con mayor fuerza el riesgo

de no continuidad escolar, y en qué medida esos patrones se mantienen o se modifican entre distintas cohortes de seguimiento?

3. ¿En qué medida un enfoque predictivo basado en registros administrativos puede aportar evidencia útil para la identificación temprana de trayectorias educativas en riesgo y para una focalización preventiva más oportuna?

2. Objetivos

2.1. Objetivo General

Analizar la no continuidad escolar anual en estudiantes del sistema escolar chileno mediante un enfoque predictivo basado en registros administrativos, identificando los factores individuales, académicos, escolares y territoriales asociados al riesgo de interrupción de la trayectoria educativa, así como sus variaciones según contexto y cohorte, con el fin de aportar evidencia útil para la identificación temprana de trayectorias en riesgo y para el fortalecimiento de decisiones preventivas en el sistema educativo.

2.2. Objetivos específicos

1. Construir un panel longitudinal de estudiantes que integre registros administrativos de matrícula y rendimiento con información de SIMCE (puntajes y cuestionarios), reconstruyendo sus trayectorias educativas desde la educación básica hasta la educación media e identificando episodios de continuidad, interrupción, reingreso y no continuidad observada.
2. Definir y operacionalizar variables individuales, académicas, escolares y territoriales relevantes para el estudio de la no continuidad escolar, incluyendo antecedentes de trayectoria, desempeño, contexto del establecimiento y variables derivadas de SIMCE, resguardando su consistencia temporal y comparabilidad entre cohortes y etapas del ciclo escolar.
3. Estimar y comparar modelos predictivos del riesgo anual de no continuidad escolar, mediante regresión logística y un modelo de aprendizaje automático, utilizando distintas ventanas de observación a lo largo de la trayectoria educativa y evaluando su desempeño fuera de muestra con métricas de discriminación, priorización y calibración.

4. Analizar la relevancia de los factores asociados al riesgo de no continuidad escolar en los modelos predictivos, examinando la contribución de variables académicas, de trayectoria, socioemocionales, escolares y territoriales, y evaluando cómo su importancia varía según el momento del recorrido educativo.
5. Comparar el comportamiento del riesgo y el desempeño predictivo entre cohortes y tramos de la trayectoria escolar, con el fin de identificar patrones de heterogeneidad temporal y posibles diferencias en la configuración de los factores asociados a la no continuidad.
6. Derivar implicancias para el diseño y uso de sistemas de alerta temprana orientados a la continuidad de trayectorias educativas, proponiendo criterios para la focalización preventiva, la priorización de indicadores de seguimiento y el uso de puntajes de riesgo en la toma de decisiones a nivel escolar y de sistema.

3. Marco teórico

El presente marco teórico tiene por finalidad construir el andamiaje conceptual que permite interpretar la no continuidad escolar anual como un fenómeno de trayectoria y no únicamente como un resultado terminal. A diferencia del capítulo anterior, donde se justifica la relevancia del problema y se formula la pregunta de investigación, esta sección organiza la evidencia y los conceptos que sustentan las decisiones analíticas posteriores.

La revisión se estructura a partir de cinco ejes. El primero corresponde al enfoque de trayectorias educativas, que permite comprender la experiencia escolar como un proceso longitudinal compuesto por progresión, rezago, transiciones, movilidad, interrupciones, reingresos y egresos. Desde esta perspectiva, la continuidad escolar no se reduce a la permanencia observada en un momento específico, sino que refiere a la capacidad del sistema educativo y de los entornos sociales para sostener recorridos educativos significativos en el tiempo.

El segundo eje aborda las distinciones entre no continuidad escolar, abandono, desvinculación, interrupción y revinculación. Esta precisión conceptual es necesaria porque cada categoría supone una forma distinta de observar el fenómeno. Mientras algunas aproximaciones privilegian la salida consolidada del sistema, otras permiten atender interrupciones anuales, retornos posteriores y situaciones de fragilidad que pueden anteceder a trayectorias más persistentes de exclusión educativa. En esta investigación, la no continuidad anual ocupa precisamente ese lugar conceptual: no se asume como equivalente a deserción definitiva, sino como una señal observable de interrupción del vínculo escolar en una escala temporal útil para la prevención.

El tercer eje corresponde a los factores asociados al riesgo de no continuidad. La literatura revisada muestra que este fenómeno es progresivo y multicausal, ya que se configura en la

interacción entre antecedentes individuales, trayectoria académica, entorno familiar, experiencia escolar, características del establecimiento y condiciones territoriales. Esta lectura justifica el uso de variables administrativas, académicas, escolares y contextuales, así como la incorporación exploratoria de dimensiones derivadas de SIMCE y sus cuestionarios.

El cuarto eje se refiere a la temporalidad del riesgo. La evidencia sugiere que la probabilidad de interrupción no se distribuye de manera homogénea a lo largo del ciclo escolar, sino que puede intensificarse en ciertos momentos y transiciones. En particular, el paso desde educación básica hacia enseñanza media ha sido descrito como una etapa sensible, en la que se reconfiguran exigencias académicas, expectativas familiares y vínculos institucionales. Por ello, la segmentación por tramos no se entiende como un simple desagregado descriptivo, sino como una decisión analítica coherente con la hipótesis de que el riesgo, sus predictores y el desempeño de los modelos pueden variar según el momento del recorrido educativo.

El quinto eje corresponde a los enfoques predictivos y a los sistemas de alerta temprana. Estos enfoques se fundamentan en la posibilidad de identificar señales observables antes de que una trayectoria se deteriore con mayor intensidad. En Chile, esta discusión dialoga tanto con desarrollos institucionales recientes, como el Sistema de Alerta Temprana y el Sistema de Protección de Trayectorias Educativas, como con estudios académicos que han utilizado registros administrativos, regresión logística y técnicas de aprendizaje automático para anticipar riesgo escolar. En este marco, la presente investigación se sitúa en la intersección entre análisis de trayectorias y modelamiento predictivo, con el propósito de examinar no solo el desempeño de los modelos, sino también la estabilidad y variación de los factores de riesgo entre tramos y cohortes.

En síntesis, este capítulo entrega la base conceptual para comprender por qué la no continuidad anual puede ser tratada como un evento relevante para la prevención, por qué resulta necesario

analizarla desde una perspectiva multicausal y longitudinal, y por qué la comparación entre tramos, cohortes y ventanas de información constituye una parte central del diseño empírico de la memoria.

3.1. Trayectorias educativas, continuidad escolar y derecho a la educación

El concepto de trayectoria educativa permite comprender la experiencia escolar como un proceso extendido en el tiempo, compuesto por ingresos, progresión, transiciones, repitencias, interrupciones y egresos, en lugar de reducirla a un resultado puntual u observado en un solo momento. Desde esta perspectiva, la experiencia escolar no debe entenderse únicamente como una secuencia ideal de cursos aprobados en los tiempos esperados, sino como un recorrido efectivamente vivido por los estudiantes dentro de condiciones sociales, institucionales y familiares específicas. Este enfoque resulta especialmente pertinente para estudiar problemas de abandono o discontinuidad, porque desplaza la atención desde el “evento final” hacia el proceso mediante el cual una trayectoria puede debilitarse, interrumpirse o reorientarse. En Chile, esta aproximación ha sido reforzada tanto por investigaciones sobre trayectorias interrumpidas como por el desarrollo reciente de dispositivos institucionales de seguimiento longitudinal (Romo & Cumsille, 2020).

Entendida de ese modo, la continuidad escolar no debe interpretarse como la simple ausencia de abandono, sino como la capacidad del sistema educativo y de los entornos sociales para sostener trayectorias significativas en el tiempo. Ello incluye no solo acceso inicial a la escuela, sino también asistencia, progresión, aprendizaje, pertenencia institucional y condiciones para que los estudiantes puedan mantenerse vinculados al proceso educativo. Esta lectura es consistente con la Convención sobre los Derechos del Niño (United Nations General Assembly, 1989) y con la formulación más reciente del sistema chileno de protección de trayectorias, que explícitamente vincula asistencia,

permanencia, vinculación y progresión. En otras palabras, la continuidad escolar constituye una expresión concreta del derecho a la educación y, por tanto, su interrupción no puede analizarse solo como un problema individual, sino también como una falla potencial en las capacidades institucionales y sociales para sostener trayectorias educativas.

La utilidad analítica del enfoque de trayectorias radica, además, en que permite reconocer la heterogeneidad del fenómeno. Romo y Cumsille (2020), al estudiar estudiantes de escuelas de reingreso en Chile, muestran que las trayectorias educativas interrumpidas no constituyen un patrón uniforme, sino un conjunto de recorridos diferenciados, marcados por distintas combinaciones de rezago, vulnerabilidad y experiencias escolares previas. Esta evidencia es especialmente relevante para una memoria como la presente, ya que justifica no reducir el análisis a una dicotomía simple entre permanencia y abandono, sino considerar la posibilidad de interrupciones, retornos y distintos momentos de fragilidad a lo largo del recorrido escolar.

En consecuencia, para esta investigación el concepto de trayectoria educativa ofrece el marco general dentro del cual se sitúa la no continuidad escolar. No se trata solo de identificar estudiantes que finalmente salen del sistema, sino de comprender cómo se configuran recorridos más vulnerables, en qué etapas se concentran las señales de riesgo y qué dimensiones del contexto pueden contribuir a sostener o interrumpir la continuidad educativa (Doll et al., 2013). Este enfoque, además, permite que el análisis posterior sobre predicción no quede reducido a una lógica técnica, sino que se comprenda como una herramienta aplicada al estudio de trayectorias educativas y al fortalecimiento de decisiones preventivas.

Si la trayectoria educativa constituye el marco general desde el cual se entiende la permanencia escolar, el paso siguiente consiste en precisar cómo se conceptualizan las interrupciones de dicho proceso, ya que la literatura y la política educativa utilizan términos relacionados, pero no

equivalentes, para referirse a ellas.

3.2. No continuidad escolar, abandono, interrupción y revinculación

Uno de los principales desafíos conceptuales del campo es que los términos utilizados para describir la salida o discontinuidad del sistema escolar no son completamente equivalentes. En la literatura, nociones como deserción, abandono escolar, exclusión educativa, interrupción de trayectoria y revinculación suelen aparecer cercanas entre sí, aunque aluden a matices analíticos distintos. Mendoza Lira et al. (2023) muestran precisamente que, en el caso chileno, la distinción entre deserción y abandono escolar no siempre se trabaja con claridad en los discursos escolares e institucionales, pese a que una diferencia central es que la primera suele implicar mayor permanencia temporal de la salida, mientras que el abandono puede emplearse de manera más amplia o menos precisa. Esta ambigüedad conceptual no es menor, porque afecta directamente el modo en que se construyen las variables de estudio y se interpretan los hallazgos.

La opción conceptual adoptada en esta memoria privilegia la idea de no continuidad escolar, porque permite observar interrupciones del vínculo anual con el sistema sin asumir desde el inicio que toda ausencia constituye una salida definitiva o irreversible. Esta formulación es más coherente con el objetivo preventivo de la investigación, orientado a anticipar quiebres de continuidad antes de que estos se consoliden en trayectorias más prolongadas de desvinculación. Además, se ajusta mejor a un contexto en que la revinculación escolar ha cobrado mayor centralidad institucional y en que la presencia de reingresos observables obliga a tratar con más cuidado la frontera entre interrupción, persistencia fuera del sistema y retorno posterior. Chile Presente, en su formulación oficial, avanza precisamente en esa dirección al situar el problema en términos de asistencia, permanencia y vinculación con el sistema, más que exclusivamente como abandono consumado

(Ministerio de Educación, 2025a).

Esto no significa negar la importancia del abandono escolar como problema sustantivo. Por el contrario, la evidencia muestra que la salida prolongada del sistema sigue teniendo consecuencias significativas para las oportunidades educativas y sociales posteriores. Lo que cambia es el punto de entrada analítico. Mientras un enfoque centrado únicamente en la «deserción definitiva» tiende a observar el resultado final, un enfoque de no continuidad escolar permite capturar interrupciones anuales, trayectorias inestables y retornos posteriores, todos ellos relevantes desde una lógica de alerta temprana. En este marco, la persistencia observada de una interrupción sigue siendo una dimensión importante, pero ya no agota el fenómeno (Lyche, 2010).

Desde esta perspectiva, la revinculación deja de ser una excepción marginal y pasa a entenderse como parte del espectro de trayectorias posibles. Los estudios de O. Espinoza et al. (2020) sobre escuelas alternativas y compromiso de estudiantes que abandonaron el sistema muestran que el reingreso constituye un proceso real y analíticamente relevante, especialmente en contextos de desigualdad educativa. En consecuencia, un trabajo interesado en anticipar la no continuidad escolar debe ser capaz de reconocer que no toda interrupción observada se traduce en una exclusión sostenida, pero sí puede representar una señal crítica que amerita prevención temprana. Ese es precisamente el lugar conceptual que ocupa la no continuidad anual en este estudio.

3.3. Factores asociados a la no continuidad escolar

La literatura especializada coincide en que la no continuidad escolar es un fenómeno multicausal. No responde a una sola variable ni puede entenderse adecuadamente desde una explicación exclusivamente individual. Más bien, se configura en la intersección entre condiciones personales, trayectoria académica previa, entorno familiar, experiencia escolar y contexto territorial o

sociocultural. Esta comprensión multinivel ha sido reforzada tanto por estudios cualitativos y cuantitativos en Chile como por trabajos recientes de modelamiento predictivo (Rodríguez et al., 2023).

En el plano individual y académico, la evidencia chilena ha mostrado de manera consistente la relevancia de variables como rendimiento previo, asistencia, sobreedad, repitencia y trayectoria escolar irregular. El trabajo metodológico de Rodríguez y colaboradores sobre predicción de abandono en sistemas escolares muestra que el uso de trayectorias individuales y acumulación de eventos mejora sustantivamente la capacidad analítica de los modelos, precisamente porque el riesgo no se explica solo por atributos estáticos del estudiante, sino por la historia observada de su recorrido escolar. Esta lógica se alinea directamente con el uso de variables rezagadas y antecedentes de trayectoria en la presente investigación.

La dimensión familiar también ocupa un lugar importante en la literatura. Ó. Espinoza et al. (2014), en un estudio sobre factores familiares asociados a la deserción escolar en niños y niñas mapuche en Chile, muestran que el abandono se vincula con estructura familiar, apoyo a la escolaridad, situación socioeconómica y formas de acompañamiento del hogar. Más allá de las especificidades del caso analizado, este trabajo refuerza una idea central: la continuidad educativa no depende solamente del estudiante o de la escuela, sino también de los recursos materiales y simbólicos disponibles en el entorno familiar. En la misma dirección, estudios posteriores sobre escuelas alternativas en Chile continúan destacando el peso de la situación del hogar, la vulnerabilidad socioeconómica y el compromiso familiar con la escolaridad (O. Espinoza et al., 2020).

En cuanto a la dimensión escolar y relacional, la evidencia para Chile muestra que el clima de relaciones dentro de la escuela importa de manera sustantiva. Contreras et al. (2022), en un estudio

longitudinal de gran escala, encontraron que las percepciones negativas del clima relacional por parte de estudiantes y docentes incrementan la probabilidad de abandono escolar. Este hallazgo es importante porque desplaza la atención desde explicaciones centradas únicamente en déficit individuales hacia la calidad de la experiencia escolar cotidiana, incluyendo relaciones con pares, docentes y entorno institucional. De manera complementaria, Rabuco Hidalgo (2022) destaca que la prevención del abandono escolar también se fortalece a partir de la relación familia-escuela y del despliegue de factores protectores en la comunidad educativa.

Finalmente, el contexto territorial y sociocultural no debe ser tratado como una simple localización geográfica. La literatura chilena muestra que factores como vulnerabilidad social, ruralidad, pertenencia étnica, oferta escolar disponible y expectativas sobre las transiciones educativas se articulan de manera compleja con la continuidad o interrupción del recorrido escolar. Gómez Vera et al. (2021), al estudiar la transición desde educación básica a media en contextos de alta vulnerabilidad, muestran que las expectativas de estudiantes, familias y actores escolares se configuran en relación con el contexto social en que están insertos, y que dichas expectativas pueden sostener o debilitar la continuidad de la trayectoria. En consecuencia, el riesgo de no continuidad debe entenderse como una configuración situada, sensible al territorio y a las condiciones concretas de escolarización.

Como síntesis, la literatura sugiere que la no continuidad escolar es el resultado de una articulación compleja entre factores académicos, familiares, escolares y contextuales. Esta conclusión es central, ya que justifica el uso de un conjunto amplio de variables explicativas y respalda la hipótesis de que la configuración del riesgo puede variar no solo entre estudiantes, sino también entre cohortes y etapas de la trayectoria.

3.4. Transiciones críticas, heterogeneidad temporal y cohortes

Además de ser multicausal, la no continuidad escolar es un fenómeno temporalmente diferenciado. La probabilidad de interrupción no permanece constante a lo largo de la trayectoria educativa, sino que tiende a concentrarse en ciertos tramos y transiciones. Esta idea aparece de manera recurrente en la literatura internacional y chilena, especialmente en relación con el paso entre niveles educativos y con momentos en que confluyen mayores exigencias académicas, cambios institucionales y reconfiguraciones de expectativas.

En el caso chileno, la transición desde educación básica a educación media constituye un punto especialmente relevante. Gómez Vera et al. (2021) muestran que este momento es acompañado por expectativas diferenciadas de estudiantes, familias y actores escolares, y que dichas expectativas pueden cumplir un rol clave en la continuidad del recorrido. Esta evidencia resulta particularmente importante para la presente investigación, porque respalda la decisión de no tratar la no continuidad como un evento indiferenciado en el tiempo, sino de examinar su comportamiento en tramos tempranos, intermedios y tardíos de la trayectoria escolar.

Esta dimensión temporal adquiere aún mayor relevancia al considerar shocks recientes que pudieron alterar las trayectorias educativas de manera diferenciada entre cohortes. En particular, la suspensión prolongada de clases presenciales durante la pandemia de COVID-19 y las dificultades para sostener procesos de enseñanza remota se han asociado con un aumento de la inasistencia grave entre 2019 y 2022, con incrementos más marcados en determinadas regiones y en contextos de mayor vulnerabilidad (Ministerio de Educación, Centro de Estudios, 2023). Este tipo de variaciones históricas refuerza la pertinencia de comparar cohortes expuestas a contextos escolares distintos, ya que el riesgo de no continuidad y la relevancia de sus predictores podrían no permanecer estables cuando el sistema educativo enfrenta perturbaciones de gran escala.

La segmentación por tramos no solo permite captar variaciones sustantivas del riesgo, sino también mejorar la lectura preventiva del fenómeno. Una interrupción anual en etapas tempranas no tiene el mismo significado ni responde necesariamente a la misma combinación de factores que una interrupción en cursos superiores de enseñanza media. Al mismo tiempo, la comparación entre cohortes permite examinar si esos patrones se mantienen relativamente estables o si cambian en función de condiciones históricas, institucionales o contextuales distintas (Kattan & Székely, 2017). En este sentido, la cohorte no es solo una forma de ordenar la base de datos, sino una dimensión analítica que permite contrastar la estabilidad temporal del riesgo y de sus factores asociados.

Desde esta perspectiva, se asume que la no continuidad escolar anual puede presentar diferencias relevantes según tramo de trayectoria y cohorte de seguimiento. Esa hipótesis no se formula únicamente como un problema descriptivo, sino como una cuestión central para la predicción: si el riesgo cambia según momento del recorrido, entonces también es plausible que cambien la utilidad de los predictores, la prevalencia del evento y el desempeño de los modelos en cada ventana analítica. Por ello, la comparación entre tramos y cohortes constituye una pieza central tanto del marco teórico como del diseño metodológico.

3.5. Enfoques predictivos, sistemas de alerta temprana y posicionamiento de la investigación

La incorporación de enfoques predictivos al estudio de la continuidad escolar se justifica por la existencia de señales observables antes de que ocurra una interrupción del vínculo anual con el sistema. La literatura sobre alerta temprana parte precisamente de esa premisa: ciertos antecedentes académicos, conductuales, relacionales y contextuales permiten identificar estudiantes

con mayor probabilidad de experimentar trayectorias de riesgo antes de que la discontinuidad se consolide. Este desplazamiento desde una lógica correctiva hacia una lógica preventiva resulta especialmente pertinente en contextos donde la revinculación ha adquirido mayor relevancia, pues permite orientar la intervención hacia etapas anteriores del proceso de desvinculación.

Los sistemas de alerta temprana (*Early Warning Systems*, EWS) se han consolidado en la última década como una herramienta central para la prevención del abandono escolar. Su objetivo es identificar con anticipación a estudiantes con mayor probabilidad de interrumpir su trayectoria educativa, utilizando información regularmente disponible en registros administrativos y evaluaciones del sistema para estimar niveles de riesgo y orientar la acción de equipos directivos, docentes y profesionales de apoyo (UNESCO Office Santiago and Regional Bureau for Education in Latin America and the Caribbean, 2021). En sus primeras aplicaciones, estos sistemas se apoyaron en reglas relativamente simples y en modelos de regresión logística basados en un conjunto acotado de indicadores. Su ventaja principal radicó en la facilidad de interpretación y en su aplicabilidad a gran escala (Balfanz & Byrnes, 2019).

Con el desarrollo de infraestructuras de datos más amplias y longitudinales, así como con el avance de las técnicas de aprendizaje automático, la literatura reciente ha incorporado modelos más flexibles, capaces de procesar un mayor número de variables, capturar relaciones no lineales e identificar interacciones entre factores de riesgo (Rodríguez et al., 2023). En términos generales, estos enfoques —como bosques aleatorios, *gradient boosting* o redes neuronales— tienden a mostrar mejoras en capacidad predictiva respecto de modelos lineales tradicionales cuando se dispone de bases extensas y de buena calidad. Sin embargo, la discusión metodológica no se reduce al desempeño estadístico: también incluye problemas de interpretabilidad, equidad, gobernanza de datos y factibilidad de implementación en contextos escolares reales. Por ello, la literatura

especializada coincide en que la utilidad de un modelo predictivo depende no solo de su precisión, sino también de su capacidad para producir alertas comprensibles, accionables y legítimas para los actores del sistema educativo.

A nivel internacional, además, existen experiencias en que los EWS han sido integrados a dispositivos de política pública de alcance sistémico (Haimovich et al., 2021). Estas experiencias muestran que los modelos predictivos pueden contribuir efectivamente a la focalización preventiva, siempre que se acompañen de diseños institucionales claros, herramientas usables y capacidad de respuesta en los establecimientos. En otras palabras, la predicción no reemplaza la intervención educativa, sino que funciona como un insumo para organizarla de manera más oportuna.

En Chile, esta orientación preventiva ha adquirido creciente relevancia tanto en la política pública como en la investigación académica. Aunque la cobertura neta en educación básica es alta, la desvinculación del sistema escolar continúa siendo un problema relevante y heterogéneo. La evidencia reciente muestra que el riesgo no se distribuye de forma uniforme, sino que se concentra con mayor intensidad en determinados niveles educativos y grupos de estudiantes. En este contexto se inserta el desarrollo institucional de sistemas de seguimiento y protección de trayectorias en Chile. El Sistema de Alerta Temprana del Ministerio de Educación, y más recientemente los dispositivos articulados bajo Chile Presente, reflejan la consolidación de una lógica de monitoreo preventivo basada en registros longitudinales. Estas herramientas buscan identificar estudiantes en riesgo y orientar acciones de acompañamiento antes de que la desvinculación se consolide (Ministerio de Educación, 2020). No obstante, la existencia de un sistema institucional no vuelve redundante la investigación académica. Al contrario, abre preguntas adicionales sobre qué variables resultan más útiles para anticipar riesgo, cómo cambia el desempeño predictivo entre cohortes y etapas de la trayectoria, y en qué medida distintas definiciones del evento permiten mejorar la

focalización preventiva.

En el plano académico, el trabajo de Rodríguez et al. (2023) valida la viabilidad técnica de construir modelos predictivos robustos en el contexto nacional y muestra que la predicción de riesgo escolar en Chile ya cuenta con un desarrollo metodológico relevante. De manera complementaria, otros estudios han comparado enfoques de aprendizaje automático con regresión logística tradicional, encontrando mejoras en capacidad de discriminación bajo ciertas configuraciones, aunque insistiendo en la necesidad de equilibrar precisión predictiva e interpretabilidad (Smith-Uldall & Gutiérrez Rojas, 2022).

Una dimensión particularmente relevante en la evidencia chilena es el uso de información SIMCE para enriquecer la predicción de trayectorias educativas. Tanto los puntajes de logro como, en menor medida, los cuestionarios de contexto han sido utilizados como insumos para complementar la información administrativa y construir perfiles de riesgo más completos (Ibáñez Iribarra, 2018). En particular, la incorporación de indicadores académicos estandarizados permite aproximar dimensiones del desempeño escolar comparables entre cohortes y momentos del ciclo educativo, mientras que los cuestionarios entregan información adicional sobre aspectos socioemocionales, escolares y familiares que no suelen estar disponibles en los registros administrativos tradicionales (Agencia de Calidad de la Educación, 2020).

Pese a estos avances, la literatura disponible para Chile presenta todavía vacíos importantes desde una perspectiva de trayectorias. En primer lugar, buena parte de los estudios define la deserción como un desenlace binario en un horizonte fijo, lo que tiende a simplificar recorridos que en la práctica pueden incluir interrupciones, reingresos o distintos momentos de fragilidad. En segundo lugar, la mayor parte de los esfuerzos empíricos se ha concentrado en maximizar la capacidad predictiva en una ventana de observación específica, con menor atención a cómo

cambia el desempeño del modelo según el tramo educativo o el momento en que se observa al estudiante. En tercer lugar, persiste espacio para examinar con mayor detalle si la importancia relativa de los factores de riesgo se mantiene estable entre cohortes o si se modifica en función de cambios contextuales, institucionales o temporales.

En ese marco, el aporte de esta investigación no consiste en inaugurar la predicción de riesgo en el sistema escolar chileno, sino en profundizarla desde un enfoque específico. La memoria se sitúa en la intersección entre análisis de trayectorias y modelamiento predictivo, centrando su atención en la no continuidad escolar anual y en el uso de un enfoque predictivo orientado a la identificación temprana de trayectorias en riesgo.

3.6. Síntesis del marco teórico

La revisión desarrollada permite sostener cuatro ideas principales. En primer lugar, la continuidad escolar debe entenderse como parte del derecho a la educación y como una dimensión central de la inclusión educativa. En segundo lugar, la no continuidad escolar se comprende mejor desde una lógica de trayectorias que desde categorías puramente terminales, ya que ello permite integrar interrupciones, reingresos y persistencia observada. En tercer lugar, el riesgo de no continuidad es multicausal y depende de factores académicos, familiares, escolares y contextuales, cuya relevancia puede variar entre tramos de la trayectoria y entre cohortes. Finalmente, el uso de enfoques predictivos y de alerta temprana se justifica en la medida en que permite desplazar la intervención hacia etapas previas del proceso de exclusión escolar, complementando las estrategias correctivas o de revinculación ya existentes.

En conjunto, este marco teórico ofrece una base conceptual para el análisis empírico posterior, al articular la perspectiva de trayectorias educativas con el estudio predictivo del riesgo y con la

heterogeneidad temporal y contextual del fenómeno.

4. Datos

4.1. Fuentes de información

El estudio se basa en la integración de tres grandes fuentes de datos a nivel individual: registros administrativos de matrícula y rendimiento escolar del Ministerio de Educación de Chile, y resultados de las evaluaciones estandarizadas SIMCE y sus cuestionarios de contexto a estudiantes y apoderados. A partir de estas fuentes se construye un panel longitudinal a nivel estudiante-año, que permite seguir a estudiantes pertenecientes a cohortes definidas según un grado y año de referencia dentro del horizonte temporal observado, con el fin de caracterizar sus trayectorias educativas, su continuidad en el sistema escolar y posibles episodios de interrupción de dicha trayectoria.

4.1.1. Registros administrativos de matrícula y rendimiento escolar

La primera fuente corresponde a los registros administrativos anuales de matrícula por estudiante publicados por el Ministerio de Educación. Estas bases contienen, para cada año escolar, información individual sobre identificación del estudiante (MRUN o identificador provisorio), establecimiento de matrícula (RBD), curso, tipo de establecimiento y de enseñanza, y algunas características sociodemográficas básicas asociadas tanto al estudiante como al establecimiento. Este conjunto de información constituye la base principal del panel longitudinal, ya que permite reconstruir la presencia del estudiante en el sistema educativo en distintos años y seguir su trayectoria mediante un identificador individual anonimizado.

De manera similar se encuentran los archivos de rendimiento, de donde se extrae información sobre la asistencia, la situación final del año (promovido, repitente, retirado, trasladado) y el

promedio general de notas obtenidas en el año escolar. En ambos casos se trabaja con registros administrativos a nivel individual que forman parte de los conjuntos de datos abiertos del Ministerio de Educación, sujetos a procesos estándar de descarga, limpieza y vinculación mediante identificadores comunes.

4.1.2. Evaluaciones SIMCE y cuestionarios de contexto

La segunda fuente corresponde a las bases de datos del Sistema de Medición de la Calidad de la Educación (SIMCE), administrado por la Agencia de Calidad de la Educación. El SIMCE aplica pruebas censales en distintos niveles del sistema escolar —históricamente 4º básico, 6º básico, 8º básico y 2º medio en la mayoría de los casos— que evalúan el logro de los objetivos de aprendizaje del currículum nacional en áreas como Lenguaje, Matemática, Ciencias y otras asignaturas, junto con cuestionarios dirigidos a estudiantes, familias, docentes y directivos que recogen información de contexto escolar y familiar. A diferencia de las bases abiertas de matrícula y rendimiento, estos microdatos a nivel individual (puntajes y respuestas a cuestionarios de contexto) corresponden a información de acceso restringido, que se obtiene mediante una solicitud formal a la autoridad competente y se utiliza bajo condiciones de confidencialidad y resguardo de la identidad de los estudiantes.

Para cada aplicación considerada en este estudio se utilizan dos tipos de información:

- **Puntajes de logro** a nivel de estudiante, estandarizados por prueba y año (en este caso se consideran puntajes en Lenguaje y Matemática en 4º, 6º, 8º básico y 2º medio).
- **Variables de cuestionarios de estudiantes y apoderados**, que aportan información sobre dimensiones socioemocionales y contextuales relevantes para la trayectoria educativa,

tales como percepción de clima y convivencia escolar, vínculos con la escuela, grupo familiar, expectativas educacionales, autoestima y motivación, entre otras. En esta investigación, dichas respuestas se organizan en categorías temáticas que posteriormente se utilizan para construir indicadores comparables entre años y niveles. El procedimiento específico de agrupación, recodificación y estandarización de estas variables se detalla en la sección metodológica correspondiente.

4.1.3. Información de los establecimientos

La base de datos también incorpora información a nivel de establecimiento educacional, la cual permite contextualizar la trayectoria del estudiante dentro del entorno institucional y territorial en que se desarrolla. Entre estos antecedentes se consideran:

- **Ubicación geográfica y ruralidad**, a partir de los códigos de región, provincia y comuna del establecimiento, junto con su clasificación como urbano o rural.
- **Número de estudiantes que rinden SIMCE**, como indicador del volumen de estudiantes evaluados en cada establecimiento.
- **Puntaje promedio del establecimiento en las áreas evaluadas por SIMCE**, como medida agregada de desempeño académico institucional.

En conjunto, estas variables permiten caracterizar atributos del contexto escolar que pueden incidir en la continuidad de las trayectorias educativas. En particular, aportan información relevante para analizar diferencias asociadas a localización territorial, ruralidad y desempeño agregado del establecimiento, dimensiones que la literatura ha relacionado con desigualdades en oportunidades educativas, progresión escolar y riesgo de no continuidad.

4.2. Definición de cohortes de estudio

La investigación considera cohortes de seguimiento entendidas como grupos de estudiantes que cursan un mismo grado en un mismo año calendario y que comparten, por tanto, un mismo año teórico de egreso de la educación media. Esta decisión no responde únicamente a una conveniencia muestral, sino a una motivación analítica: examinar si los factores asociados a la no continuidad escolar presentan patrones relativamente estables o si, por el contrario, varían entre cohortes sometidas a contextos históricos y escolares diferentes.

La comparación entre cohortes permite discutir la robustez temporal de los hallazgos y evita que la interpretación del riesgo quede restringida a una sola generación de estudiantes. En este sentido, la cohorte no se utiliza únicamente como una segmentación descriptiva, sino como una dimensión sustantiva del análisis.

Dado que se busca analizar la no continuidad escolar en distintos momentos de la trayectoria y contrastar la configuración del riesgo entre grupos comparables, la inclusión de cohortes constituye un componente central de la base analítica y no un mero desagregado adicional.

La selección de las cohortes de análisis se realiza siguiendo tres criterios principales:

1. **Disponibilidad de información SIMCE a lo largo de la trayectoria:** se privilegian cohortes para las cuales es posible observar, al menos, una aplicación SIMCE en la educación básica (por ejemplo, en 4º o 6º básico) y, cuando es posible, aplicaciones posteriores en 8º básico y 2º medio. Esto permite comparar la capacidad predictiva de distintas *ventanas de observación* a lo largo del ciclo escolar.
2. **Horizonte suficiente de seguimiento hasta el egreso:** se requieren cohortes que puedan

observarse durante un número de años suficiente como para distinguir entre estudiantes que interrumpen su trayectoria educativa y aquellos que completan la educación media. En la práctica, esto implica trabajar con cohortes cuyo año teórico de egreso ya haya ocurrido dentro del período cubierto por el horizonte de estudio, de modo de minimizar los casos censurados.

3. **Cobertura de contextos pre, durante y post pandemia:** se eligen cohortes cuyos años teóricos de egreso se ubican, respectivamente, antes, durante y después del período de interrupciones y ajustes asociados a la pandemia de COVID-19. De este modo, la comparación entre cohortes no solo captura diferencias temporales en las políticas y condiciones del sistema escolar, sino que también permite examinar en qué medida los factores de riesgo de no continuidad escolar y la capacidad predictiva de los modelos se mantienen estables o se modifican en función de shocks externos al sistema educativo.

De esta manera las cohortes a estudiar son las que cursan 4° básico en los años 2009, 2013 y 2016.

A continuación se detalla la disponibilidad de información SIMCE para cada cohorte teórica:

	4° básico	6° básico	8° básico	2° medio	Año teórico de egreso
Cohorte I	2009	-	2013	2015	2017
Cohorte II	2013	2015	2017	-	2021
Cohorte III	2016	2018	-	2022	2024

Tabla 4.1: Años de aplicación SIMCE observados para cada cohorte de estudio escogida.

Cabe recalcar que las cohortes se definen a partir del año en que los estudiantes cursan 4° básico, seleccionándose como puntos de entrada los años 2009, 2013 y 2016. Estas cohortes deben entenderse como *cohortes teóricas de seguimiento*, en el sentido de que lo que queda fijado es el

momento inicial de observación y no una trayectoria escolar ideal sin repitencia ni interrupciones. En consecuencia, los años teóricos de aplicación SIMCE en 6° básico, 8° básico y 2° medio, así como el año teórico de egreso, funcionan como referencias temporales para organizar el seguimiento y comparar cohortes, pero no suponen que todos los estudiantes progresen regularmente conforme a ese calendario. Si un estudiante repite, presenta rezago, interrumpe su trayectoria o reingresa posteriormente, continúa perteneciendo a la cohorte definida en el año base, ya que el interés analítico del estudio recae precisamente en observar la trayectoria real del estudiante y sus eventuales desviaciones respecto de la progresión esperada. Esta distinción entre cohorte teórica y trayectoria observada resulta consistente con el enfoque de trayectorias educativas adoptado en esta memoria y con la literatura que enfatiza la importancia de reconstruir recorridos individuales a partir de información longitudinal, en lugar de reducir el análisis a estados terminales o a secuencias ideales de avance escolar (Rodríguez et al., 2023; Romo & Cumsille, 2020).

En consecuencia, la selección de cohortes busca equilibrar tres criterios: un punto de entrada comparable para todos los estudiantes, un horizonte suficiente de seguimiento longitudinal y una disponibilidad razonable de información estandarizada. Desde esta perspectiva, el SIMCE no determina por sí solo el diseño del estudio, sino que se incorpora como un insumo adicional dentro de una estrategia más amplia de reconstrucción de trayectorias educativas.

4.3. Construcción del panel longitudinal

A partir de las fuentes de matrícula se organiza la información en un panel en formato *long* mediante el uso de Python, donde cada fila representa a un estudiante en un año determinado. El identificador principal es el *MRUN*, mientras que el eje temporal se define por el año escolar (*AGNO*) y el grado cursado (*grade_num*). El objetivo de esta etapa es consolidar, para cada

estudiante y año, un conjunto de variables de caracterización del estudiante, atributos del establecimiento y medidas de trayectoria escolar, garantizando consistencia temporal y unicidad de observación anual. El procedimiento puede resumirse en los siguientes pasos:

1. **Codificación estandarizada de grado** (*grade_num*). Dado que los registros administrativos reportan el curso mediante los códigos *COD_ENSE* (tipo/modalidad de enseñanza) y *COD_GRADO* (grado dentro de la modalidad), se construye una variable numérica única (*grade_num*) que permite comparar la progresión escolar en una escala común. Este mapeo traduce, de forma consistente en todas las cohortes y años, los cursos de educación básica y media a valores enteros entre 4 y 12 (por ejemplo, 4º básico = 4; 4º medio = 12). En caso de combinaciones *COD_ENSE* – *COD_GRADO* asociadas a modalidades no comparables con la trayectoria regular, se excluyen de este estudio; como los casos asociados a modalidades aceleradas del tipo 2x1 (programas en que el estudiante cursa 3º y 4º medio en un solo año).
2. **Identificación de la cohorte en 4º básico**. Para cada año de referencia (2009, 2013 y 2016) se filtran los registros de matrícula para conservar todas las observaciones con *grado* = 4 y *AGNO* igual al año correspondiente. Los *MRUN* presentes en ese subconjunto definen el universo de la cohorte (por ejemplo, todos los estudiantes que estaban en 4º básico en 2013 para la cohorte 4b2013).
3. **Selección de variables base**: variables estructurales que caracterizan al estudiante y a su establecimiento en cada año escolar. Estas variables sirven como insumos directos para los modelos y como contexto descriptivo de las cohortes analizadas. Entre las principales se encuentran:
 - **Género del estudiante** (*GEN_ALU*), codificado como variable categórica binaria según el registro administrativo.

- **Región y localización del establecimiento** (*COD_REG_RBD*, *RURAL_RBD*), que permiten distinguir el territorio en el que se ubica el establecimiento y si se trata de una escuela rural o urbana.
 - **Dependencia y tipo de establecimiento**, identificados a partir de los códigos de enseñanza y de dependencia del RBD, lo que permite diferenciar entre establecimientos municipales, subvencionados particulares y particulares pagados, así como modalidades de enseñanza (científico-humanista, técnico-profesional, entre otras).
 - **Promedio general de notas** (*PROM_GRAL*), variable continua que recoge el promedio anual de calificaciones obtenidas por el estudiante en el año escolar correspondiente.
 - **Atención en grupo diferencial** (*GD_ALU*), variable binaria que indica si el estudiante recibe apoyo adicional en el establecimiento por pertenecer a un grupo con necesidades educativas especiales.
4. **Reconstrucción de la trayectoria hacia adelante.** Una vez definido el conjunto de estudiantes (MRUN) pertenecientes a cada cohorte, se recuperan todas sus observaciones disponibles en años posteriores dentro del panel (hasta 2025). Este procedimiento permite reconstruir la trayectoria de matrícula y caracterizar dinámicas relevantes —promoción, repitencia, cambios de establecimiento, interrupciones y eventuales reingresos—, así como identificar el último año y grado en que cada estudiante es observado.
5. **Depuración y criterios de exclusión.** Dado que la unidad de análisis de esta investigación corresponde a la observación estudiante-año, se requiere una única fila por cada par (MRUN, AGNO). En aquellos casos en que la base registra múltiples observaciones para un mismo estudiante en un mismo año, se aplica un criterio de consolidación que privilegia la última observación válida del año, ya que las duplicidades se asocian principalmente a

cambios de establecimiento dentro del mismo período. De este modo, la observación retenida representa la situación escolar final del estudiante en ese año y permite mantener coherencia entre la estructura longitudinal del panel, la construcción de la variable de resultado y los procedimientos de validación del modelo. Cabe mencionar que esta información de movilidad intraanual no se pierde, si no que también se consideran en la variable `n_establecimientos_hasta_anio` dado su relevancia.

6. **Asignación de etiquetas de resultado.** Para cada estudiante de la cohorte, se utiliza la trayectoria completa observada en el panel con el fin de asignar una etiqueta de resultado referida a la continuidad o no continuidad de su trayectoria educativa. En particular, se distingue entre continuidad escolar, no continuidad temprana, no continuidad intermedia y no continuidad tardía, de acuerdo con la definición operacional detallada a continuación. Esta etiqueta se incorpora a la fila correspondiente a la ventana de observación definida para cada estudiante. La utilización de la trayectoria completa observada se limita exclusivamente a la construcción de la variable de resultado y no a la definición de las variables explicativas, con el fin de resguardar la coherencia temporal del análisis predictivo.

4.4. Etiqueta de no continuidad escolar

La variable objetivo de los modelos corresponde a la situación de continuidad o no continuidad escolar observada a lo largo de la trayectoria de cada estudiante. A partir del panel longitudinal se construye una etiqueta que distingue entre estudiantes que mantienen continuidad dentro del horizonte de observación y estudiantes que presentan no continuidad escolar. A su vez, dentro de estos últimos, se clasifica el momento aproximado en que ocurre la interrupción de la trayectoria, distinguiendo entre no continuidad temprana, intermedia y tardía.

4.4.1. Continuidad escolar anual, observaciones elegibles y casos censurados

La variable objetivo principal de esta memoria corresponde a la no continuidad escolar anual. Para cada observación estudiante-año en t , se evalúa si existe o no matrícula observada en el año inmediatamente siguiente ($t + 1$). De este modo, la etiqueta anual toma valor 1 cuando no se observa continuidad en $t + 1$, y valor 0 cuando la matrícula sí se mantiene al año siguiente.

Esta definición no se aplica indistintamente a todas las filas del panel, sino únicamente a aquellas observaciones para las cuales la continuidad en $t + 1$ es analíticamente interpretable. En consecuencia, se consideran *observaciones elegibles* aquellas en que el año siguiente es observable y en que la ausencia posterior de matrícula puede interpretarse como una interrupción de la trayectoria escolar dentro del sistema regular.

Bajo esta lógica, se distinguen dos fuentes principales de censura. En primer lugar, se censuran las observaciones correspondientes al último año disponible del panel, dado que en esos casos no es posible verificar si el estudiante continúa o no en $t + 1$. En segundo lugar, se censuran las observaciones de 4° medio con egreso regular, ya que allí la ausencia de matrícula posterior constituye el desenlace esperado de la trayectoria escolar y no un episodio de no continuidad.

Este criterio difiere de una exclusión uniforme de 4° medio. En efecto, las observaciones de 4° medio que no corresponden a egreso regular sí pueden formar parte de la muestra modelable, siempre que el año $t + 1$ sea observable. En estos casos, la continuidad posterior sigue siendo analíticamente relevante, pues puede expresarse en repetición, revinculación o salida del sistema. De este modo, la variable objetivo no modela la continuidad de todos los estudiantes indistintamente, sino la probabilidad de no continuidad anual dentro del conjunto de observaciones efectivamente expuestas al riesgo de interrumpir su trayectoria escolar en el año siguiente.

Tabla 4.2: Ejemplos de elegibilidad y tratamiento de la etiqueta anual

Caso observado en t	Situación en $t + 1$	Tratamiento	Etiqueta
3° medio en 2024	No aparece en 2025	Elegible	$Y = 1$
3° medio en 2024	Aparece en 2025	Elegible	$Y = 0$
4° medio en 2024, egresa regularmente	No aparece en 2025	Censurado	No definida
4° medio en 2024, no egresa regularmente	No aparece en 2025	Elegible	$Y = 1$
Cualquier grado en 2025	2026 no observable	Censurado	No definida

4.4.2. Clasificación en no continuidad temprana, intermedia y tardía

Con el propósito de caracterizar el momento de ocurrencia de la no continuidad escolar, se utiliza el último grado observado en la trayectoria de los estudiantes clasificados en dicha condición. Sobre la base de la estructura del sistema escolar chileno, se distinguen tres categorías analíticas:

- **No continuidad temprana:** corresponde a estudiantes cuyo último grado observado es inferior a 8° básico. Esta categoría agrupa interrupciones ocurridas en etapas tempranas del recorrido escolar y se asocia a trayectorias que presentan señales de fragilidad desde niveles iniciales de la educación obligatoria.
- **No continuidad intermedia:** corresponde a estudiantes cuyo último registro observado se ubica en 8° básico o 1° medio. Esta categoría busca capturar interrupciones en la transición entre educación básica y educación media, así como en el inicio de este último nivel, etapa que la literatura identifica como especialmente sensible en términos de riesgo escolar.
- **No continuidad tardía:** corresponde a estudiantes cuya trayectoria se interrumpe en cursos superiores de la educación media, tales como 2° o 3° medio, sin alcanzar el grado terminal. Esta categoría agrupa trayectorias que logran avanzar una parte importante del ciclo, pero no culminan la certificación de enseñanza media dentro del horizonte de observación.

A partir de esta clasificación se construyen dos tipos de variables de resultado. En primer lugar, una etiqueta binaria que distingue entre continuidad escolar y no continuidad escolar. En segundo lugar, una variable categórica que diferencia entre continuidad, no continuidad temprana, no continuidad intermedia y no continuidad tardía. Estas etiquetas se utilizan tanto para el análisis descriptivo de trayectorias como para el desarrollo de modelos orientados a examinar si los factores académicos, individuales, escolares y territoriales permiten anticipar distintas formas de interrupción de la trayectoria educativa.

4.5. Variables de trayectoria

A partir de los registros administrativos de matrícula y rendimiento, se busca que las siguientes variables capturen la acumulación de riesgos a lo largo del tiempo en dimensiones como la edad relativa, la movilidad entre establecimientos y la asistencia al establecimiento. Dada la orientación de la investigación hacia sistemas de alerta temprana, estas variables se definen procurando consistencia temporal y evitando el uso de información posterior al año de predicción.

4.5.1. Edad y sobreedad

Para cada observación del panel se verifica la edad del estudiante en el año escolar utilizando la fecha de nacimiento (*FEC_NAC_ALU*) y el año *AGNO*. Cuando la base original incluye una variable de edad declarada, se contrasta con la edad calculada y se depuran los casos con discrepancias extremas.

A partir de la edad se construye la variable de *sobreedad*, definida como la diferencia entre la edad observada y la edad teórica correspondiente al grado cursado, suponiendo ingreso oportuno

y progresión sin repitencias. La edad teórica se aproxima como $5 + grade_num$ años para la educación básica y se ajusta de manera coherente para la educación media. Así, una sobreedad igual a 0 indica trayectoria regular, valores positivos reflejan atraso escolar (por ejemplo, +2 años de sobreedad) y valores negativos se interpretan como adelanto escolar. Los casos que presentan sobreedad igual a 1 positivo o negativo, no se consideran pues puede ser alcance de la exactitud de la fecha de nacimiento y la fecha de cálculo. En particular, se eliminan las trayectorias cuya edad implícita resulta incompatible con el grado cursado (por ejemplo, sobreedades negativas superiores a tres años), de manera de asegurar la consistencia entre edad y nivel educativo.

Se considera que este aspecto acoge a la repitencia y a otros factores que pueden generar desajustes entre la edad y el grado cursado, tales como ingresos tardíos al sistema escolar o interrupciones prolongadas en la trayectoria. Por lo tanto, estos no son agregados al análisis de manera explícita, sino que se asumen como parte del fenómeno de sobreedad.

4.5.2. Movilidad entre establecimientos

La movilidad escolar se operacionaliza principalmente a partir del identificador del establecimiento (*RBD*), comparando el establecimiento observado para un mismo estudiante entre años consecutivos de la trayectoria. De este modo, un cambio de *RBD* entre t y $t + 1$ se interpreta como cambio de establecimiento entre años. De manera complementaria, la variable de situación final (*SIT_FIN_R*) permite identificar casos de «traslado» reportados dentro del mismo año escolar, lo que aporta información adicional sobre movilidad intraanual.

Sobre esta base, se construye una variable acumulada para cada estudiante que contabiliza el número de establecimientos observados hasta el año de referencia ($n_establecimientos_hasta_anio$). Esta variable busca captar trayectorias caracterizadas por mayor inestabilidad institucional, que

la literatura asocia a una mayor probabilidad de interrupción o desvinculación, especialmente cuando los cambios se concentran en momentos críticos de transición educativa.

4.5.3. Interrupciones previas y reingreso

La trayectoria escolar puede presentar interrupciones, entendidas como años en los que el estudiante no aparece registrado en el panel, seguidos de una reaparición posterior. Estas discontinuidades son relevantes porque, desde la perspectiva de un sistema de alerta temprana, la no observación en el año calendario siguiente se interpreta como un evento de no continuidad anual, aun cuando el estudiante pueda reingresar en años posteriores.

No obstante, la ausencia de registros en un año determinado puede deberse tanto a una salida efectiva del sistema como a limitaciones de cobertura o subregistro en las fuentes. Para abordar esta ambigüedad y reducir falsos positivos de “reingreso”, se incorpora un criterio de consistencia entre el gap temporal y el avance en grado. En particular, para cada estudiante se compara: (i) el número de años ausentes entre dos observaciones consecutivas (gap en *AGNO*) y (ii) el número de grados efectivamente avanzados entre dichas observaciones (gap en *grade_num*). Cuando el desfase entre ambos gaps se encuentra dentro de una tolerancia acotada —permitiendo, por ejemplo, hasta un margen compatible con una repitencia— la discontinuidad se interpreta como potencialmente atribuible a registros faltantes más que a una interrupción real de la trayectoria. En cambio, desfases mayores sugieren una interrupción efectiva y posterior reaparición.

Con base en esta regla, se construye la variable ($N_REINGRESOS_HASTA_T$) que utiliza exclusivamente información observada hasta el año t , evitando filtración de información futura. Esta variable corresponde al contador acumulado del número de interrupciones efectivas previas hasta el año t .

4.5.4. Variables académicas rezagadas (enfoque mid-year)

Algunas variables académicas se consolidan típicamente al final del año escolar, por lo que no estarían disponibles en el momento en que se ejecuta una intervención preventiva durante el año t . Para aproximar este escenario, se utilizan rezagos de un año en indicadores de desempeño y participación escolar. En particular, se incorporan *ASISTENCIA_lag1* y *PROM_GRAL_lag1* como aproximaciones al rendimiento y asistencia observados hasta el año $t - 1$. Esta decisión metodológica permite alinear el set de predictores con un uso operativo del modelo en un sistema de alerta temprana.

4.6. Construcción de variables desde SIMCE

4.6.1. Puntajes académicos

Para cada cohorte y nivel evaluado (4º, 6º, 8º básico y 2º medio, cuando corresponde) se utilizan los puntajes SIMCE individuales en Lenguaje y Matemática. Los puntajes se emplean en su escala original estandarizada por la Agencia de Calidad de la Educación, sin transformaciones adicionales, de modo de preservar la comparabilidad entre años y niveles dentro de cada prueba.

4.6.2. Dimensiones psicosociales y de contexto escolar y familiar

Con el fin de incorporar dimensiones psicosociales y de contexto escolar y familiar al análisis, se utilizaron los cuestionarios aplicados en las distintas mediciones disponibles. A partir de dichos instrumentos, se construyeron indicadores sintéticos a nivel de estudiante que resumen información relativa a autoestima académica, apoyo familiar, aspiraciones educativas, clima y convivencia

escolar, apoyo docente, motivación, seguridad, violencia, vinculación con el establecimiento, entre otras dimensiones relevantes para la trayectoria educativa.

La construcción de estos indicadores se realizó mediante un procedimiento de operacionalización temática. En primer lugar, se revisaron los cuestionarios y su documentación asociada para identificar ítems cuyo contenido resultara conceptualmente coherente con las dimensiones de interés. Esta selección no tuvo por objetivo validar escalas psicométricas nuevas, sino construir variables comparables y analíticamente interpretables a partir de la información disponible, en coherencia con la literatura sobre factores psicosociales, escolares y familiares asociados a la continuidad educativa.

En segundo lugar, los ítems seleccionados se agruparon en categorías temáticas según su afinidad conceptual. Dado que las preguntas presentan distintas escalas de respuesta y, en algunos casos, sentidos de valoración opuestos, se aplicó un procedimiento de recodificación para homogenizar la interpretación de los puntajes. Específicamente, para cada ítem se identificó: (i) su escala original de respuesta y (ii) el sentido sustantivo de la medición, de modo que los valores altos representaran de manera sistemática una condición más favorable. En los casos en que un ítem estaba formulado en sentido inverso, la escala se invirtió antes de su incorporación al puntaje compuesto.

En tercer lugar, para cada categoría temática y nivel de aplicación se calculó un puntaje promedio a nivel de estudiante utilizando los ítems válidos disponibles. Cuando una misma dimensión contaba con información proveniente tanto del estudiante como del apoderado, ambos aportes se combinaron mediante el promedio simple de sus respectivos puntajes, con el propósito de obtener una medida sintética que recogiera ambas perspectivas.

Finalmente, los puntajes resultantes se reescalaron linealmente a un rango común de 0 a 100, donde 0 representa la condición menos favorable observada y 100 la más favorable. Esta transformación facilita la interpretación sustantiva de los indicadores y permite compararlos entre categorías y entre distintos años de aplicación, aun cuando los cuestionarios originales no compartan exactamente las mismas escalas de respuesta.

Como resultado de este proceso, se obtuvieron indicadores continuos para las siguientes dimensiones: autoestima académica (AA), apoyo familiar (AF), aspiraciones educativas (ASP), clima de aula y de establecimiento (C), convivencia (CO), conductas de riesgo (CR), apoyo docente (DA), discriminación percibida (DI), higiene del establecimiento (H), ingresos familiares percibidos (IF), involucramiento del estudiante y la familia en la vida escolar (INV), motivación escolar (M), normas y disciplina (N), número de integrantes del hogar (NF), organización del curso (ORG), recursos disponibles (R), seguridad (S), salud (SD), experiencias de violencia (V) y vinculación con el establecimiento (VIN). Estos indicadores se incorporan a las muestras de análisis únicamente cuando la información correspondiente se encuentra disponible en la ventana de observación considerada.

Dado que los cuestionarios experimentan ajustes entre años y niveles, no todas las aplicaciones incluyen exactamente los mismos ítems ni cubren idénticas dimensiones. En consecuencia, la disponibilidad y composición de cada indicador puede variar entre cohortes y ventanas de observación. Por esta razón, el detalle del mapeo entre ítems, categorías temáticas, escalas originales y reglas de recodificación se presenta en el anexo [8.1](#).

Cabe señalar que estos indicadores deben interpretarse como medidas operacionales construidas con fines analíticos y predictivos a partir de la información disponible en los cuestionarios, y no como escalas psicométricas formalmente validadas en el marco de esta investigación.

4.6.3. Cobertura desigual y marcador de disponibilidad

La disponibilidad de SIMCE no es homogénea para todos los estudiantes ni para todas las cohortes, ya sea por diferencias en calendarios de aplicación, no participación, ausencia en el día de la prueba o cambios de instrumento a lo largo del tiempo. Para registrar esta heterogeneidad, se incluye la variable *HAS_SIMCE* como indicador de si la observación estudiante-año cuenta con información SIMCE asociada en la ventana de observación utilizada. Esto permite (i) comparar desempeño de modelos con y sin SIMCE, y (ii) reportar métricas estratificadas según disponibilidad de este conjunto de variables.

Para mantener coherencia con el objetivo de predicción anual y el momento de intervención, la información SIMCE se incorpora bajo un enfoque temporalmente consistente: se utiliza únicamente información disponible hasta el año t para predecir el evento en $t+1$, evitando incorporar mediciones del mismo año cuya disponibilidad efectiva podría ocurrir con posterioridad al periodo de interés. En la práctica, esto se implementa mediante un rezago de un año en la utilización operativa de estos indicadores en el set de predictores, análogo al tratamiento de variables académicas consolidadas al cierre del año.

4.7. Descripción de la muestra y cobertura de datos

La construcción de la base analítica parte de un panel longitudinal consolidado a nivel estudiante-año, obtenido tras la integración y depuración de los registros administrativos y de la información proveniente de SIMCE. Luego de resolver duplicidades dentro de un mismo par estudiante-año, el panel total queda compuesto por 6.231.570 observaciones, correspondientes a 712.969 MRUN únicos, y cubre el período 2009–2025. Sobre esta base se define posteriormente la muestra

modelable para la predicción de no continuidad escolar anual.

La unidad de análisis del estudio corresponde a observaciones estudiante-año. La variable objetivo principal se define como no continuidad escolar anual, entendida como la ausencia de matrícula observada en el año $t + 1$ para un estudiante registrado en el año t . En coherencia con esta definición, se excluyen de la muestra modelable aquellas observaciones ubicadas en el último año observable del panel, dado que en ellas no resulta posible verificar continuidad hacia el año siguiente. Asimismo, se excluyen las filas que representan egreso regular en 4° medio, puesto que en tales casos la salida del sistema corresponde al cierre esperado de la trayectoria escolar y no al evento de interés. En cambio, las observaciones de 4° medio que no corresponden a egreso regular se conservan en la muestra global, ya que su continuidad al año siguiente sigue siendo informativa desde la lógica del estudio.

4.7.1. Tamaño de la muestra modelable

Bajo estos criterios, la muestra analítica principal queda conformada por 5.648.210 observaciones estudiante-año, correspondientes a 712.944 estudiantes únicos. La tasa global de no continuidad anual en esta muestra alcanza 2,84 %. Con el fin de examinar la heterogeneidad del fenómeno a lo largo de la trayectoria escolar, la muestra se segmenta en tres tramos analíticos: temprana, intermedia y tardía. El tramo temprano reúne 2.924.895 observaciones y 712.944 estudiantes, con una tasa de no continuidad anual de 1,34 %; el tramo intermedio comprende 1.424.251 observaciones y 689.919 estudiantes, con una tasa de 3,82 %; y el tramo tardío considera 1.289.340 observaciones y 645.655 estudiantes, con una tasa de 4,68 %.

Tabla 4.3: Cobertura temporal y tamaño del panel total versus muestra modelable

Dataset	Observaciones	Estudiantes únicos	Período
Panel total consolidado	6.231.570	712.969	2009–2025
Muestra modelable anual	5.648.210	712.944	2009–2024

Tabla 4.4: Muestra modelable y tasa de no continuidad anual por tramo educativo

Tramo	Observaciones	Estudiantes únicos	Tasa no continuidad anual
Temprano	2.924.895	712.944	1,34 %
Intermedio	1.424.251	689.919	3,82 %
Tardío	1.289.340	645.655	4,68 %
Total modelable	5.648.210	712.944	2,84 %

La reducción del número de estudiantes únicos entre tramos no debe interpretarse como una pérdida arbitraria de casos, sino como una consecuencia de la estructura longitudinal del panel. El tramo temprano se define sobre la totalidad de la población elegible observada en los primeros grados, mientras que los tramos intermedio y tardío solo incluyen a quienes alcanzan registros en etapas más avanzadas de la trayectoria escolar. En consecuencia, la disminución del universo observable en los tramos superiores refleja la progresión efectiva de los estudiantes dentro del sistema y no un proceso anómalo de depuración muestral.

Adicionalmente, un subconjunto acotado de observaciones no se clasifica en los tres tramos principales. Estas corresponden principalmente a registros terminales de 4° medio que no constituyen egreso regular. Dado que su continuidad al año siguiente sigue siendo analíticamente relevante, dichas observaciones se mantienen en la muestra global, aunque se excluyen de la comparación principal por tramo con el fin de preservar la homogeneidad interpretativa entre etapas educativas.

4.7.2. Censura de la etiqueta

La construcción de la etiqueta anual implica además un tratamiento explícito de la censura. Como fue mencionado en la sección anterior, se consideran censuradas las observaciones correspondientes al último año disponible del panel, ya que en esos casos no es posible verificar la continuidad del estudiante en $t + 1$. En segundo lugar, se excluyen de la definición del evento las observaciones que corresponden a egreso regular en 4° medio, dado que allí la salida del sistema representa el cierre esperado de la trayectoria escolar y no un episodio de no continuidad. La combinación de ambos criterios determina el subconjunto de observaciones para las cuales la etiqueta anual no resulta definible. En consecuencia, la muestra modelable utilizada en la estimación de los modelos corresponde a las observaciones estudiante-año no censuradas bajo estas reglas.

Tabla 4.5: Censura de la etiqueta anual: motivos y solapamientos

Categoría	Observaciones
Fin de ventana del panel	20.268
Egreso regular en 4° medio	579.876
Censura total (unión)	583.324
Solo fin de ventana	3.448
Solo egreso regular	563.056
Ambos	16.820

4.7.3. Prevalencia del evento por cohorte teórica

La distribución por cohorte teórica de egreso muestra una composición relativamente equilibrada entre los tres grupos analizados. La cohorte con egreso teórico en 2017 aporta 1.962.351 observaciones correspondientes a 248.270 estudiantes, con una tasa de no continuidad anual de 3,78 %; la cohorte 2021 aporta 1.845.099 observaciones y 231.734 estudiantes, con una tasa de 2,48 %; y la cohorte 2024 aporta 1.840.760 observaciones y 232.940 estudiantes, con una tasa de 2,19 %. En términos agregados, estas cifras sugieren que la incidencia anual del evento tiende a disminuir en

las cohortes más recientes, aunque las diferencias son moderadas cuando se observan a este nivel de agregación.

Tabla 4.6: Muestra modelable y tasa de no continuidad anual por cohorte teórica de egreso

Cohorte teórica	Observaciones	Estudiantes únicos	Tasa no continuidad anual
2017 (4b2009)	1.962.351	248.270	3,78 %
2021 (4b2013)	1.845.099	231.734	2,48 %
2024 (4b2016)	1.840.760	232.940	2,19 %

Este resultado es consistente con la idea de que la prevalencia del evento depende no solo de la cohorte considerada, sino también del momento de la trayectoria escolar en que se observa al estudiante. Por ello, la comparación agregada entre cohortes se complementa con una desagregación por tramo educativo, la cual permite examinar con mayor precisión cómo cambia el riesgo según etapa del recorrido escolar.

4.7.4. Prevalencia del evento por cohorte y tramo educativo

La desagregación por cohorte y tramo educativo muestra un patrón sustantivamente más informativo. En las tres cohortes analizadas, la no continuidad anual es considerablemente menor en el tramo temprano y aumenta de manera importante en los tramos intermedio y tardío. En la cohorte más antigua, la tasa de no continuidad anual alcanza 1,65 % en el tramo temprano, 5,44 % en el intermedio y 6,36 % en el tardío. En la cohorte 2021, las tasas correspondientes son 1,29 %, 3,38 % y 3,62 %. En la cohorte 2024, por su parte, se observan tasas de 1,08 % en el tramo temprano, 2,51 % en el intermedio y 4,04 % en el tardío.

Tabla 4.7: Tasa de no continuidad anual por cohorte teórica y tramo educativo

Cohorte	Tramo	Observaciones	Estudiantes únicos	Tasa no continuidad anual
2017	Temprano	1.025.912	248.270	1,65 %
2017	Intermedio	501.091	237.834	5,44 %
2017	Tardío	431.389	214.399	6,36 %
2017	Fuera de tramo	3.959	3.739	67,16 %
2021	Temprano	953.457	231.734	1,29 %
2021	Intermedio	463.567	224.462	3,38 %
2021	Tardío	423.988	212.026	3,62 %
2021	Fuera de tramo	4.087	3.767	60,46 %
2024	Temprano	945.526	232.940	1,08 %
2024	Intermedio	459.593	227.623	2,51 %
2024	Tardío	433.963	219.230	4,04 %
2024	Fuera de tramo	1.678	1.677	65,67 %

En conjunto, estos resultados refuerzan la idea de que el riesgo de no continuidad anual no se distribuye homogéneamente a lo largo de la trayectoria educativa. Por el contrario, se intensifica a medida que el estudiante avanza hacia etapas más tardías del recorrido escolar. Al mismo tiempo, la comparación entre cohortes sugiere que las diferencias agregadas de prevalencia esconden estructuras internas distintas, particularmente en los tramos intermedio y tardío. Este hallazgo resulta consistente con la hipótesis de heterogeneidad temporal que orienta la investigación y justifica la evaluación posterior del desempeño predictivo por tramo educativo.

4.7.5. Disponibilidad de SIMCE y cobertura de variables seleccionadas

La disponibilidad de información SIMCE no se reporta aquí como mera ausencia de valores perdidos, ya que el indicador *HAS_SIMCE* corresponde a una variable binaria que identifica si la observación cuenta o no con información SIMCE utilizable. En consecuencia, lo relevante no es su completitud formal, sino la proporción efectiva de observaciones con información disponible. A nivel de tramo, dicha proporción alcanza 68,60 % en el tramo temprano, 95,85 % en el tramo intermedio y 98,31 % en el tramo tardío. Por cohorte, la disponibilidad efectiva de SIMCE alcanza

79,21 % en la cohorte 4b2009, 84,39 % en la cohorte 4b2013 y 83,51 % en la cohorte 4b2016. Estos resultados indican que la integración de SIMCE es alta, pero no homogénea, y que su disponibilidad varía tanto según la posición en la trayectoria como según la cohorte de análisis.

Tabla 4.8: Proporción de observaciones con al menos un antecedente SIMCE disponible

Tramo / cohorte	Proporción con SIMCE
Temprano	68,60 %
Intermedio	95,85 %
Tardío	98,31 %
Fuera de tramo	97,82 %
2017	79,21 %
2021	84,39 %
2024	83,51 %

En cuanto a las variables administrativas rezagadas, el promedio general del año anterior (*PROM_GRAL_lag1*) y la asistencia del año anterior (*ASISTENCIA_lag1*) presentan una disponibilidad de 87,33 % en la muestra total. No obstante, esta cobertura es heterogénea entre tramos: en el tramo temprano ambas variables alcanzan 75,56 %, mientras que en los tramos intermedio y tardío su disponibilidad es prácticamente completa, con valores cercanos o superiores a 99,95 %. Esta diferencia responde al hecho de que, en las primeras observaciones de la trayectoria, no siempre existe un año previo utilizable para construir los rezagos académicos (en 4° básico).

La variable de sobreedad presenta cobertura prácticamente total en la muestra analítica, tanto en el total como en cada tramo, lo que confirma su robustez como indicador de trayectoria. Por su parte, las dimensiones psicosociales y de contexto escolar derivadas de SIMCE exhiben una cobertura intermedia, aunque en general elevada. En el total de la muestra, autoestima académica (*AA*), apoyo familiar (*AF*), clima (*C*), motivación (*M*) y vinculación con el establecimiento (*VIN*) presentan coberturas cercanas a 80 %, mientras que apoyo docente (*DA*) y experiencias de

violencia (V) muestran coberturas algo menores, de 64,73 % y 64,36 %, respectivamente.

Tabla 4.9: Cobertura de variables seleccionadas por tramo

Variable	Total	Temprano	Intermedio	Tardío
<i>PROM_GRAL_lag1</i>	87,33 %	75,56 %	99,95 %	99,99 %
<i>ASISTENCIA_lag1</i>	87,33 %	75,56 %	99,95 %	99,99 %
<i>SOBREEDAD</i>	100,00 %	100,00 %	100,00 %	100,00 %
<i>AA</i>	80,86 %	66,78 %	94,47 %	97,66 %
<i>AF</i>	80,62 %	66,70 %	93,96 %	97,33 %
<i>C</i>	80,90 %	66,77 %	94,43 %	97,86 %
<i>DA</i>	64,73 %	45,08 %	77,92 %	94,54 %
<i>M</i>	80,79 %	66,67 %	94,43 %	97,66 %
<i>V</i>	64,36 %	44,76 %	77,52 %	94,10 %
<i>VIN</i>	80,11 %	65,72 %	93,66 %	97,65 %

La heterogeneidad entre tramos vuelve a observarse con claridad en estas dimensiones. En el tramo temprano, la cobertura de *AA*, *AF*, *C*, *M* y *VIN* se sitúa en torno a 66 %–67 %, mientras que *DA* y *V* descienden a aproximadamente 45 %. En cambio, en el tramo intermedio estas mismas variables alcanzan coberturas cercanas o superiores a 94 % para varias dimensiones, y en el tramo tardío superan en la mayoría de los casos 94 %, llegando incluso a valores próximos a 98 %. Este patrón no debe interpretarse como una falla del panel, sino como una consecuencia esperable de las diferencias en aplicación de cuestionarios, niveles evaluados y años de medición disponibles.

En síntesis, la muestra analítica construida ofrece una base amplia, longitudinalmente consistente y metodológicamente adecuada para el estudio de la no continuidad escolar anual. Su tamaño, distribución entre cohortes y segmentación por tramo permiten examinar el fenómeno desde una perspectiva preventiva y diferenciada por etapa educativa. Al mismo tiempo, la revisión de cobertura muestra que las variables administrativas presentan una disponibilidad muy alta,

mientras que la información SIMCE y las dimensiones psicosociales exhiben variación relevante entre cohortes y tramos, lo que justifica su incorporación comparada en ventanas de observación progresivamente más ricas.

5. Metodología

5.1. Diseño predictivo y objetivo del estudio

El objetivo metodológico de esta investigación consiste en estimar el riesgo anual de no continuidad escolar a partir de un panel longitudinal de observaciones estudiante-año. La predicción se formula sobre la transición anual $t \rightarrow t + 1$, de modo que, para cada estudiante observado en el año t , se modela la probabilidad de no registrar matrícula en el año $t + 1$. Este enfoque permite desplazar el análisis hacia una lógica preventiva orientada a detectar quiebres anuales del vínculo escolar antes de que estos se consoliden en trayectorias más persistentes de desvinculación.

La variable objetivo principal corresponde, por tanto, a la no continuidad escolar anual. Se define como un evento binario igual a 1 cuando un estudiante observado en el año t no aparece matriculado en el año $t + 1$, y 0 en caso contrario. Esta definición se construye respetando la secuencia temporal del panel y excluyendo aquellas observaciones para las cuales el evento no resulta definible.

La estrategia metodológica combina tres elementos. En primer lugar, la construcción de niveles de observación con distinta riqueza de información. En segundo lugar, la comparación entre un modelo interpretable de base y un modelo flexible de aprendizaje automático. En tercer lugar, la evaluación fuera de muestra bajo esquemas que reproducen tanto un escenario prospectivo por cohorte como una partición por estudiante. De este modo, la metodología busca no solo maximizar capacidad predictiva, sino también resguardar coherencia temporal, comparabilidad entre cohortes y utilidad operativa para fines de alerta temprana.

5.2. Construcción de la etiqueta anual y tratamiento de censura

Para cada observación estudiante-año en t , se modela la ocurrencia de no continuidad escolar en $t + 1$, entendida como la ausencia de matrícula observada en el año inmediatamente siguiente. Esta definición se aplica únicamente a observaciones elegibles, es decir, aquellas para las cuales el estado en $t + 1$ puede determinarse y corresponde a un episodio analíticamente relevante. Bajo esta regla, se censuran las observaciones del último año del panel y los casos de egreso regular en 4° medio. La estimación y evaluación de los modelos se realiza, por tanto, sobre el subconjunto de observaciones no censuradas.

5.3. Escenario operativo y prevención de filtración de información

Dado que la investigación se orienta a una lógica de alerta temprana, los predictores utilizados en el año t deben corresponder únicamente a información disponible hasta ese momento. En consecuencia, se evita incorporar variables construidas con información futura respecto del momento de predicción, con el fin de prevenir filtración de información y no sobreestimar artificialmente el desempeño de los modelos. En particular, las variables académicas administrativas se utilizan en forma rezagada, mientras que la información SIMCE se incorpora como último antecedente disponible en la trayectoria observada hasta el año t . De este modo, la predicción de no continuidad en $t + 1$ se realiza con antecedentes efectivamente utilizables en un escenario operativo realista.

5.4. Niveles de información

Con el objetivo de comparar sistemáticamente el aporte incremental de distintas fuentes de información, se definen tres niveles de información.

El primer nivel, $W1$, incorpora únicamente variables de trayectoria y contexto administrativo observables en el año t . Entre ellas se incluyen el grado cursado, la cohorte teórica de egreso, el género, la región del establecimiento, la condición de ruralidad, la dependencia administrativa, la sobreedad, el número de establecimientos por los que ha pasado el estudiante hasta el año observado y la disponibilidad de antecedentes SIMCE.

El segundo nivel, $W2$, añade a $W1$ las variables académicas rezagadas del año previo, específicamente promedio general y asistencia. Esta especificación representa un escenario de alerta temprana particularmente plausible desde el punto de vista operativo, ya que incorpora información reciente y sustantivamente relevante sin requerir variables del cierre del mismo año t .

El tercer nivel, $W3$, agrega a $W2$ un conjunto de dimensiones psicosociales y de contexto escolar derivadas de SIMCE. En la especificación principal se consideran autoestima académica, apoyo familiar, clima, apoyo docente, motivación, experiencias de violencia y vinculación con el establecimiento. Estas dimensiones fueron seleccionadas por su relevancia teórica, su interpretabilidad sustantiva y su cobertura suficientemente alta en la muestra analítica. La comparación entre $W1$, $W2$ y $W3$ permite evaluar tanto el valor agregado de los antecedentes administrativos recientes como el aporte incremental de información estandarizada y psicosocial.

5.5. Modelos comparados

La estimación se realiza mediante dos familias de modelos.

En primer lugar, se utiliza una regresión logística penalizada (*Logit*) como línea base interpretable.

Este modelo permite estimar la probabilidad de no continuidad anual a partir de una combinación lineal de predictores, de modo que, para cada observación i , se tiene:

$$P(Y_i = 1 | X_i) = p_i = \frac{1}{1 + \exp(-\eta_i)}, \quad \eta_i = \beta_0 + \sum_{j=1}^k \beta_j X_{ij}$$

De forma equivalente, la especificación puede expresarse en términos del logit de la probabilidad:

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^k \beta_j X_{ij}$$

donde $Y_i = 1$ indica no continuidad anual y $Y_i = 0$ continuidad observada. En esta memoria, la regresión logística se estima con regularización, lo que permite estabilizar los coeficientes y reducir el riesgo de sobreajuste en presencia de múltiples predictores. Su inclusión resulta especialmente valiosa porque ofrece una referencia clara para evaluar la capacidad predictiva de una aproximación lineal ampliamente utilizada en problemas de clasificación binaria, además de facilitar la comparación con estudios previos y con esquemas de alerta temprana basados en modelos simples y fácilmente comunicables.

En segundo lugar, se emplea un modelo de *HistGradientBoosting*, correspondiente a una familia de métodos de aprendizaje automático basados en árboles de decisión potenciados. A diferencia de la regresión logística, este enfoque no impone una relación lineal entre los predictores y el logaritmo

de las razones de probabilidad del evento, sino que permite capturar de manera flexible relaciones no lineales, interacciones entre variables y umbrales complejos de decisión. Estas características resultan especialmente pertinentes en trayectorias escolares marcadas por heterogeneidad de contextos, acumulación de riesgos y posibles efectos diferenciados según momento de la trayectoria educativa. Además, el uso de una implementación basada en histogramas mejora la eficiencia computacional en bases de gran tamaño, lo que la hace adecuada para el volumen de observaciones considerado en este estudio.

La comparación entre ambos enfoques no persigue únicamente identificar el modelo con mejor desempeño, sino también contrastar la ganancia predictiva asociada a una mayor flexibilidad funcional respecto de una especificación más parsimoniosa e interpretable.

5.6. Estrategias de partición y evaluación fuera de muestra

La evaluación se realiza mediante dos esquemas complementarios.

El primero corresponde a una **partición temporal por cohorte**, definida como escenario principal de evaluación. En este caso, el entrenamiento se realiza utilizando las cohortes teóricas de egreso 2017 y 2021, mientras que la cohorte 2024 se reserva para prueba. Este diseño reproduce un escenario prospectivo realista, en el que un sistema de alerta temprana se entrena con datos históricos y luego se aplica a una cohorte futura. Debido a ello, esta estrategia constituye el criterio principal para seleccionar la especificación de referencia del estudio.

El segundo esquema corresponde a un **holdout por estudiante**, implementado a nivel de identificador individual (*MRUN*). Bajo esta estrategia, se realiza una partición 80/20 de estudiantes únicos, asegurando que ningún estudiante aparezca simultáneamente en entrenamiento y prueba.

Este benchmark entrega una referencia complementaria del desempeño predictivo cuando se controla explícitamente la dependencia intraestudiante en el panel longitudinal, aunque sin imponer el desafío adicional de generalización temporal entre cohortes.

En ambos casos, la comparación se realiza de manera consistente sobre los niveles $W1$, $W2$ y $W3$, y sobre los dos modelos definidos previamente.

5.7. Métricas de desempeño y priorización operativa

Dada la naturaleza desbalanceada del evento (baja prevalencia de no continuidad anual), el desempeño de los modelos se evalúa combinando métricas de discriminación global con métricas orientadas a la focalización de intervenciones.

En primer lugar, se reportan métricas de discriminación:

- **AUC-ROC** (área bajo la curva ROC), que resume la capacidad del modelo para ordenar correctamente observaciones positivas y negativas a lo largo de todos los posibles umbrales de decisión.
- **PR-AUC** (Average Precision), que resume la curva Precisión–Recall y resulta especialmente informativa en contextos desbalanceados, ya que enfatiza el comportamiento del modelo sobre la clase positiva.

En este estudio, la evaluación no se limita a la discriminación global del modelo, sino que también considera su capacidad para ordenar observaciones según riesgo. Esta distinción es importante porque, en problemas de ranking, el interés práctico suele concentrarse en la calidad del subconjunto superior del ordenamiento más que en el comportamiento promedio sobre todos

los umbrales posibles. En la literatura de evaluación de sistemas rankeados, métricas como *Precision@k* y *Recall@k* se utilizan precisamente para examinar el desempeño del modelo en el top- k de observaciones priorizadas, es decir, en el segmento del ranking donde la decisión operativa se vuelve más relevante (Manning et al., 2008). De manera complementaria, *Lift@k* resume cuánto mejora la tasa de identificación de casos respecto de una selección aleatoria, por lo que resulta especialmente útil cuando el objetivo es focalizar intervenciones sobre una fracción reducida de estudiantes. En conjunto, estas métricas permiten evaluar no solo la capacidad del modelo para discriminar entre observaciones positivas y negativas, sino también su utilidad operativa como herramienta de priorización (Devriendt et al., 2021).

En segundo lugar, dado que los sistemas de alerta temprana operan típicamente bajo restricciones de recursos (por ejemplo, capacidad limitada de intervención), se incorporan métricas de *priorización* basadas en la selección del subconjunto de estudiantes con mayor riesgo predicho. Sea $p_{i,t}$ la probabilidad predicha de evento para la observación (i, t) . Para un porcentaje fijo $k \in (0, 1)$, se define el conjunto $\mathcal{S}_k$ como el subconjunto correspondiente al porcentaje superior k de observaciones ordenadas de mayor a menor según $p_{i,t}$ (top- k).

Con ello, se calculan:

- **Recall@k**: proporción de eventos correctamente identificados dentro del top- k ,

$$\text{Recall@}k = \frac{\sum_{(i,t) \in \mathcal{S}_k} y_{i,t+1}}{\sum_{(i,t)} y_{i,t+1}}. \quad (1)$$

- **Precision@k**: proporción de observaciones verdaderamente positivas dentro del top- k ,

$$\text{Precision@}k = \frac{\sum_{(i,t) \in \mathcal{S}_k} y_{i,t+1}}{|\mathcal{S}_k|}. \quad (2)$$

- **Lift@k**: ganancia relativa respecto de una selección aleatoria, definida como el cociente entre $\text{Precision@}k$ y la tasa base del evento (*event rate*) en el conjunto evaluado,

$$\text{Lift@}k = \frac{\text{Precision@}k}{\bar{y}}, \quad \text{donde} \quad \bar{y} = \frac{1}{N} \sum_{(i,t)} y_{i,t+1}. \quad (3)$$

En la práctica, se evalúan estos indicadores para $k = 5\%$ y $k = 10\%$, lo que permite interpretar el rendimiento del modelo en escenarios de focalización de distinta intensidad. La elección de $k = 5\%$ y $k = 10\%$ busca representar escenarios de focalización relativamente intensiva y moderada, respectivamente, permitiendo evaluar la sensibilidad del modelo a distintos niveles de capacidad de intervención.

En consecuencia, estas métricas complementan a AUC-ROC y PR-AUC porque responden a una pregunta distinta. Mientras las métricas globales indican qué tan bien discrimina el modelo en términos generales, $\text{Recall@}k$, $\text{Precision@}k$ y $\text{Lift@}k$ permiten responder qué ocurre cuando solo se puede actuar sobre una fracción acotada de estudiantes. Desde la perspectiva de un sistema de alerta temprana, esta es una condición especialmente relevante, ya que la utilidad del modelo no depende únicamente de ordenar correctamente el riesgo, sino de concentrar casos positivos en el subconjunto efectivamente intervenible.

5.8. Calibración del riesgo

Además de la capacidad de ranking, se evalúa la calibración de las probabilidades predichas. Para ello se utiliza el *Brier score*, que resume el error cuadrático medio entre probabilidades estimadas y resultados observados, y se construyen tablas por deciles de riesgo predicho, comparando la tasa observada del evento con la probabilidad media predicha en cada grupo.

La calibración se considera un complemento relevante de la discriminación, ya que un modelo puede ordenar adecuadamente el riesgo sin necesariamente producir probabilidades absolutas bien ajustadas. Esta dimensión adquiere especial importancia cuando se pretende interpretar el score como probabilidad o utilizar umbrales de decisión explícitos en la gestión de alertas.

5.9. Análisis por tramo educativo

Con el fin de examinar la heterogeneidad del desempeño a lo largo de la trayectoria escolar, los resultados se desagregan por tramo educativo. La segmentación principal distingue entre tramo temprano, intermedio y tardío, definidos sobre grados 4 a 11. Las observaciones terminales de 4° medio no clasificadas en estos tres bloques se mantienen en la muestra global, pero no forman parte de la comparación principal por tramo, dado que su interpretación es sustantivamente distinta.

Esta desagregación permite evaluar en qué etapas el modelo discrimina mejor el riesgo de no continuidad y en cuáles la priorización resulta más efectiva. Dado que la prevalencia del evento varía de manera importante entre tramos, la lectura de las métricas se realiza considerando explícitamente esta heterogeneidad de tasas base.

5.10. Síntesis del flujo metodológico

En síntesis, el flujo metodológico del estudio se compone de: (i) la construcción de una etiqueta anual de no continuidad escolar a partir del panel longitudinal, con tratamiento explícito de censura; (ii) la definición de ventanas de observación de complejidad creciente; (iii) la comparación entre una línea base interpretable y un modelo flexible de aprendizaje automático; (iv) la evaluación

fuera de muestra bajo particiones temporal y por estudiante; y (v) el análisis del desempeño en términos de discriminación, priorización operativa, calibración y heterogeneidad por tramo educativo.

Esta estrategia permite articular coherencia temporal, consistencia con el objetivo preventivo del estudio y capacidad de comparación entre especificaciones, ofreciendo una base metodológica robusta para el análisis de resultados que se presenta en el capítulo siguiente.

6. Resultados

6.1. Comparación de ventanas y modelos

Esta sección presenta los resultados de desempeño predictivo obtenidos al comparar las tres ventanas de observación definidas ($W1$, $W2$ y $W3$) y las dos familias de modelos consideradas: regresión logística penalizada y *HistGradientBoosting* (HGB). La evaluación se reporta bajo dos esquemas complementarios. El primero corresponde al escenario principal $E1$, en el cual el modelo se entrena con las cohortes teóricas de egreso 2017 y 2021 y se evalúa fuera de muestra en la cohorte 2024. El segundo corresponde a un benchmark de partición 80/20 por estudiante (*holdout* por MRUN), que entrega una referencia adicional cuando se controla la dependencia intraestudiante en el panel longitudinal.

La comparación bajo $E1$ constituye el ejercicio más cercano a un escenario operativo realista, ya que reproduce la situación en la cual un sistema de alerta temprana se entrena con datos históricos y luego se aplica a una cohorte futura. En este sentido, la evaluación no solo exige discriminar entre observaciones positivas y negativas, sino también generalizar hacia una generación de estudiantes observada en un contexto temporal posterior.

Los resultados del escenario principal se presentan en la Tabla 6.1. En términos generales, HGB supera sistemáticamente a la regresión logística en las tres ventanas, tanto en métricas de discriminación global como en métricas de priorización. El principal salto de desempeño se observa al pasar desde $W1$ a $W2$. Para HGB, el AUC aumenta desde 0,8190 a 0,8366, mientras que el $PR-AUC$ aumenta desde 0,1998 a 0,2377. De manera consistente, el $recall@5$ aumenta desde 0,4260 a 0,4721 y la $precision@5$ desde 0,1867 a 0,2069. En contraste, el paso desde $W2$ a $W3$ entrega mejoras marginales: el AUC aumenta levemente a 0,8375, pero el $PR-AUC$ disminuye

levemente y las métricas de priorización al 5 % superior permanecen prácticamente estables.

Tabla 6.1: Desempeño fuera de muestra por ventana y modelo en el escenario principal E1 (entrenamiento en cohortes 2017 y 2021; prueba en cohorte 2024).

Ventana	Modelo	AUC	PR-AUC	event_rate	recall@5	precision@5	lift@5	recall@10	precision@10
W1	Logit	0.7983	0.1848	0.0219	0.4173	0.1829	8.3451	0.5464	0.1198
W1	HGB	0.8190	0.1998	0.0219	0.4260	0.1867	8.5191	0.5565	0.1220
W2	Logit	0.8086	0.2083	0.0219	0.4486	0.1967	8.9722	0.5764	0.1263
W2	HGB	0.8366	0.2377	0.0219	0.4721	0.2069	9.4411	0.5937	0.1301
W3	Logit	0.8087	0.2070	0.0219	0.4466	0.1958	8.9325	0.5748	0.1260
W3	HGB	0.8375	0.2372	0.0219	0.4711	0.2065	9.4213	0.5940	0.1302

La comparación interna entre ventanas sugiere, por tanto, que la principal ganancia predictiva proviene de incorporar antecedentes académicos recientes del estudiante, en particular promedio general y asistencia del año previo. La información adicional proveniente de SIMCE y de dimensiones psicosociales construidas a partir de sus cuestionarios agrega señal, pero no modifica sustantivamente el ranking de riesgo una vez consideradas las variables de trayectoria y desempeño rezagado.

Desde una perspectiva más amplia, este patrón es consistente con la literatura reciente sobre predicción de abandono escolar en Chile y otros contextos, donde los enfoques de aprendizaje automático suelen superar a la regresión logística y las variables académicas recientes tienden a concentrar una parte importante del poder predictivo del modelo (Rodríguez et al., 2023; Smith-Uldall & Gutiérrez Rojas, 2022). En ese sentido, los resultados del presente estudio no contradicen la evidencia previa, sino que la extienden a una definición alternativa del evento centrada en la no continuidad anual.

6.2. Generalización temporal entre cohortes y benchmark por estudiante

La comparación entre el escenario principal $E1$ y el benchmark por MRUN permite evaluar con mayor precisión el significado de la generalización temporal entre cohortes. Mientras que en el *holdout* por MRUN el entrenamiento y la prueba contienen estudiantes distintos pero provenientes del mismo universo temporal, en $E1$ el entrenamiento se realiza con cohortes anteriores y la evaluación se aplica exclusivamente a una cohorte posterior. En otras palabras, $E1$ exige que el modelo mantenga su utilidad predictiva al pasar desde cohortes históricas hacia una nueva generación de estudiantes.

Los resultados del benchmark por MRUN se reportan en la Tabla 6.2. Como era esperable, el desempeño bajo este esquema es superior al observado en $E1$. Para HGB, el AUC aumenta hasta aproximadamente 0,879 en $W2$ y $W3$, y el $PR-AUC$ alcanza valores cercanos a 0,327. Del mismo modo, el $lift@5$ supera 10, lo que indica una concentración del evento en el grupo priorizado aún mayor que en la evaluación temporal por cohorte.

Tabla 6.2: Desempeño bajo benchmark de partición 80/20 por estudiante (holdout por MRUN).

Ventana	Modelo	AUC	PR-AUC	event_rate	recall@5	precision@5	lift@5	recall@10	precision@10
W1	Logit	0.8456	0.2587	0.0287	0.4566	0.2622	9.1314	0.6181	0.1775
W1	HGB	0.8631	0.2834	0.0287	0.4655	0.2673	9.3097	0.6345	0.1821
W2	Logit	0.8541	0.2804	0.0287	0.4830	0.2773	9.6594	0.6477	0.1860
W2	HGB	0.8791	0.3273	0.0287	0.5146	0.2955	10.2911	0.6748	0.1937
W3	Logit	0.8548	0.2790	0.0287	0.4829	0.2773	9.6582	0.6476	0.1859
W3	HGB	0.8791	0.3271	0.0287	0.5149	0.2956	10.2973	0.6752	0.1938

La comparación entre ambos esquemas permite extraer una conclusión metodológicamente importante. El hecho de que el desempeño sea sistemáticamente menor en $E1$ confirma que la transferencia temporal entre cohortes constituye un escenario más exigente que la simple

separación por estudiante. Por lo tanto, los resultados obtenidos en $E1$ deben interpretarse como una estimación más conservadora y más cercana al uso prospectivo real del modelo.

Aun así, el patrón cualitativo es estable entre ambos ejercicios: HGB supera a la regresión logística, $W2$ mejora claramente respecto de $W1$, y $W3$ no entrega una ganancia sustantiva sobre $W2$. Esta consistencia fortalece la elección de $W2$ con HGB como especificación principal del estudio.

6.3. Desempeño global del modelo principal

Dado el patrón descrito anteriormente, se selecciona como modelo principal la especificación HGB con ventana $W2$, evaluada en el escenario temporal $E1$. Esta elección privilegia el mejor equilibrio entre desempeño predictivo, parsimonia y factibilidad operativa. La Tabla 6.3 resume las métricas principales de esta especificación.

Tabla 6.3: Desempeño global del modelo principal seleccionado ($E1$, $W2$, HGB).

Métrica	Valor
AUC	0.8366
PR-AUC	0.2377
event_rate	0.0219
recall@5	0.4721
precision@5	0.2069
lift@5	9.4411
recall@10	0.5937
precision@10	0.1301
lift@10	5.9367
Brier score	0.0190

En la cohorte 2024, el modelo alcanza un AUC de 0,8366 y un $PR-AUC$ de 0,2377, con una prevalencia del evento de 2,19 %. En términos de priorización operativa, al focalizar el 5 % superior del ranking de riesgo se obtiene un $recall@5$ de 0,4721, una $precision@5$ de 0,2069 y un $lift@5$ de 9,44. Esto implica que, aun cuando la no continuidad anual es un evento relativamente infrecuente,

el modelo logra concentrar una fracción relevante de los casos en un subconjunto acotado del ranking predicho.

Desde el punto de vista aplicado, este resultado es especialmente importante. En un contexto de capacidad limitada de intervención, el valor del modelo no depende únicamente de su capacidad para discriminar globalmente entre casos positivos y negativos, sino de su utilidad para ordenar el riesgo y priorizar estudiantes. En este sentido, las métricas de ranking muestran que el modelo principal cumple adecuadamente con una función de focalización preventiva.

6.4. Curvas ROC, PR y distribución de puntajes

Las Figuras 6.1, 6.2 y 6.3 muestran, respectivamente, la curva ROC, la curva precisión–recall y la distribución de puntajes del modelo principal. La curva ROC evidencia una separación consistente entre clases, mientras que la curva precisión–recall confirma que el modelo mantiene utilidad predictiva aun en un contexto de baja prevalencia del evento.

La distribución de puntajes muestra una fuerte asimetría hacia valores bajos, con una cola derecha donde se concentran los casos de mayor riesgo. Este patrón es coherente con la naturaleza poco frecuente de la no continuidad anual y ayuda a interpretar las métricas de priorización reportadas anteriormente: la utilidad del modelo radica precisamente en identificar esa fracción superior del ranking donde el riesgo se acumula de manera desproporcionada.

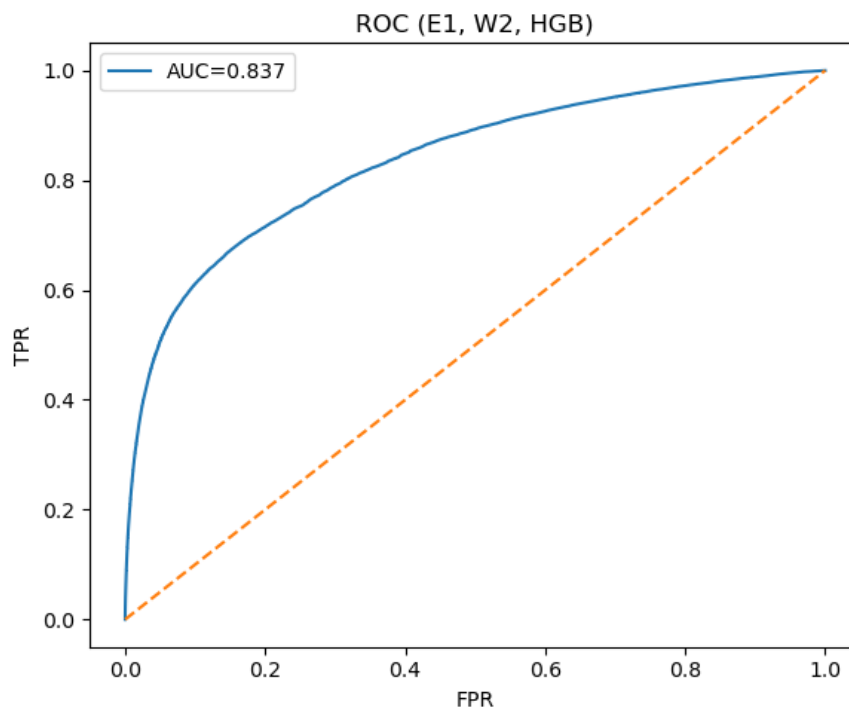


Figura 6.1: Curva ROC del modelo principal (E1, W2, HGB). La figura resume la capacidad de discriminación global del modelo sobre la cohorte de prueba 2024.

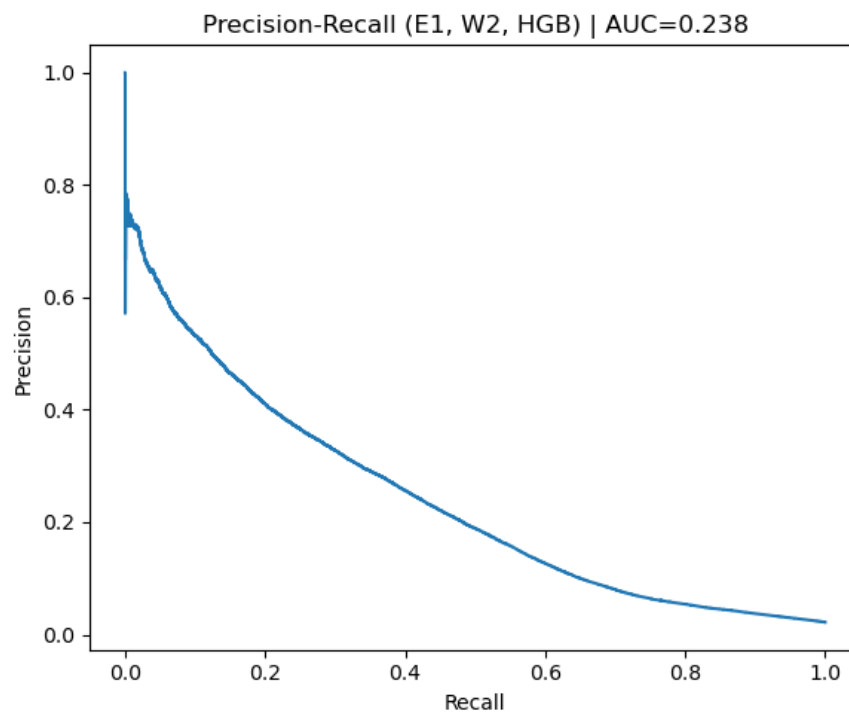


Figura 6.2: Curva precisión–recall del modelo principal (E1, W2, HGB). Esta representación resulta especialmente informativa dada la baja prevalencia del evento en el conjunto de prueba.

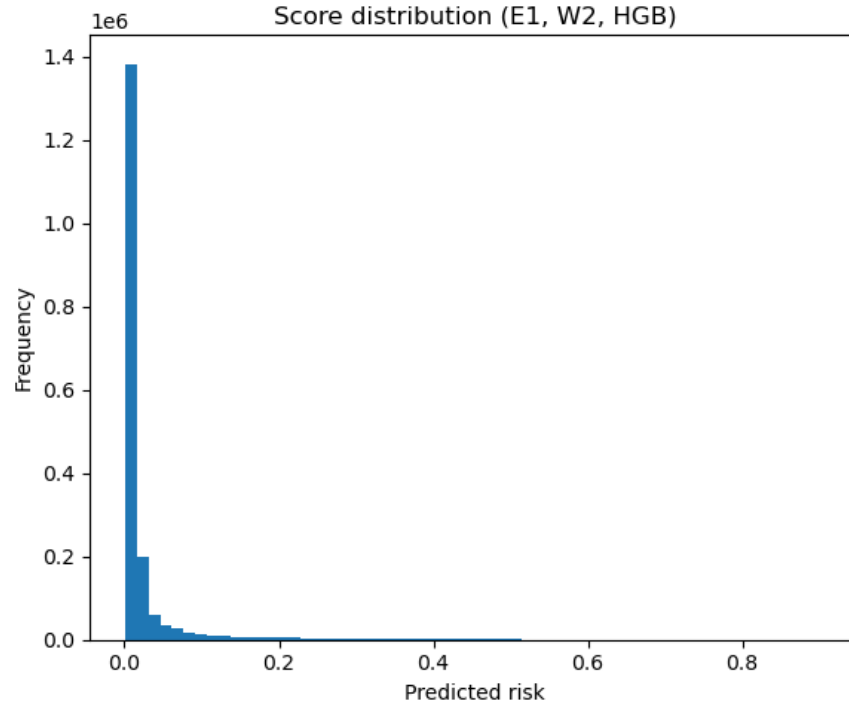


Figura 6.3: Distribución de puntajes predichos del modelo principal (E1, W2, HGB). La asimetría observada refleja la baja prevalencia del evento y la concentración del riesgo en la cola superior del ranking.

6.5. Heterogeneidad del desempeño por tramo educativo

Los resultados desagregados por tramo educativo se presentan en la Tabla 6.4 y la Tabla 6.5. Esta desagregación permite evaluar hasta qué punto la utilidad del modelo varía a lo largo de la trayectoria escolar.

Tabla 6.4: Desempeño por tramo educativo del modelo principal (E1, W2, HGB).

Tramo	n obs.	n estud.	AUC	PR-AUC	event_rate	recall@5	precision@5	lift@5
Temprana	945 526	232 940	0.7623	0.0791	0.0108	0.3004	0.0647	6.0080
Intermedia	459 593	227 623	0.8297	0.2517	0.0251	0.4658	0.2341	9.3150
Tardía	433 963	219 230	0.8443	0.2772	0.0404	0.3872	0.3125	7.7437

En el tramo temprano, el modelo alcanza un AUC de 0,7623 y un $PR-AUC$ de 0,0791, con una prevalencia del evento de 1,08 %. Bajo priorización al 5 %, el $recall@5$ es 0,3004, la $precision@5$ 0,0647 y el $lift@5$ 6,01. Estas métricas son inferiores a las observadas en los otros tramos, pero

Tabla 6.5: Métricas complementarias al 10 % por tramo educativo del modelo principal (E1, W2, HGB).

Tramo	recall@10	precision@10	lift@10
Temprana	0.3924	0.0423	3.9242
Intermedia	0.5990	0.1505	5.9900
Tardía	0.5590	0.2256	5.5895

ello debe interpretarse a la luz de la muy baja tasa base del evento en esta etapa.

En el tramo intermedio, el desempeño mejora sustantivamente. El modelo alcanza un AUC de 0,8297 y un *PR-AUC* de 0,2517, con una prevalencia de 2,51 %. En este tramo, el *recall@5* asciende a 0,4658, la *precision@5* a 0,2341 y el *lift@5* a 9,32. Esta combinación de métricas sugiere que el tramo intermedio constituye un punto especialmente relevante para la focalización preventiva, ya que el modelo logra recuperar una fracción importante de los casos manteniendo una precisión alta en el subconjunto priorizado.

En el tramo tardío, el AUC aumenta a 0,8443 y el *PR-AUC* a 0,2772, con una prevalencia de 4,04 %. La *precision@5* alcanza 0,3125, la más alta de los tres tramos, aunque el *recall@5* desciende a 0,3872. Esto indica que en etapas tardías el modelo identifica un subconjunto de riesgo particularmente concentrado, aunque recupera proporcionalmente menos casos que en el tramo intermedio.

En conjunto, estos resultados confirman que la dificultad del problema predictivo depende del momento de la trayectoria. El modelo conserva señal útil en todos los tramos, pero su utilidad operativa es mayor en las etapas intermedia y tardía, donde la prevalencia del evento es mayor y la priorización resulta más precisa. Este hallazgo es coherente con la hipótesis teórica de heterogeneidad temporal y refuerza la pertinencia de desagregar resultados según etapa educativa.

6.6. Calibración del riesgo

La calibración del modelo principal se evaluó mediante el *Brier score* y una tabla por deciles de riesgo predicho. El *Brier score* obtenido fue 0,0190, lo que indica un error cuadrático medio bajo en la estimación de probabilidades y sugiere una calibración razonable en promedio.

La Tabla 6.6 muestra una relación monotonica clara entre la probabilidad media predicha y la tasa observada del evento por deciles de riesgo. En los grupos de menor riesgo ambas magnitudes se mantienen relativamente cercanas. En cambio, en la cola superior del ranking aparece una tendencia a sobreestimar el riesgo: en el decil más alto la tasa observada del evento alcanza 0,1301, mientras que la probabilidad media predicha asciende a 0,1772.

Tabla 6.6: Calibración por deciles de riesgo del modelo principal (E1, W2, HGB).

Decil de riesgo	n	event_rate	p_mean
1	184 085	0.0026	0.0025
2	184 083	0.0035	0.0036
3	184 066	0.0047	0.0045
4	184 072	0.0059	0.0054
5	184 844	0.0074	0.0068
6	183 306	0.0105	0.0084
7	184 078	0.0130	0.0113
8	184 074	0.0166	0.0162
9	184 078	0.0247	0.0308
10	184 074	0.1301	0.1772

Este patrón sugiere que el modelo funciona adecuadamente como instrumento de ranking, pero que la interpretación del score como probabilidad absoluta debe realizarse con cautela, particularmente en el extremo superior de riesgo. En otras palabras, el modelo ordena bien los casos, pero tiende a asignar probabilidades algo mayores que las efectivamente observadas en la cola alta.

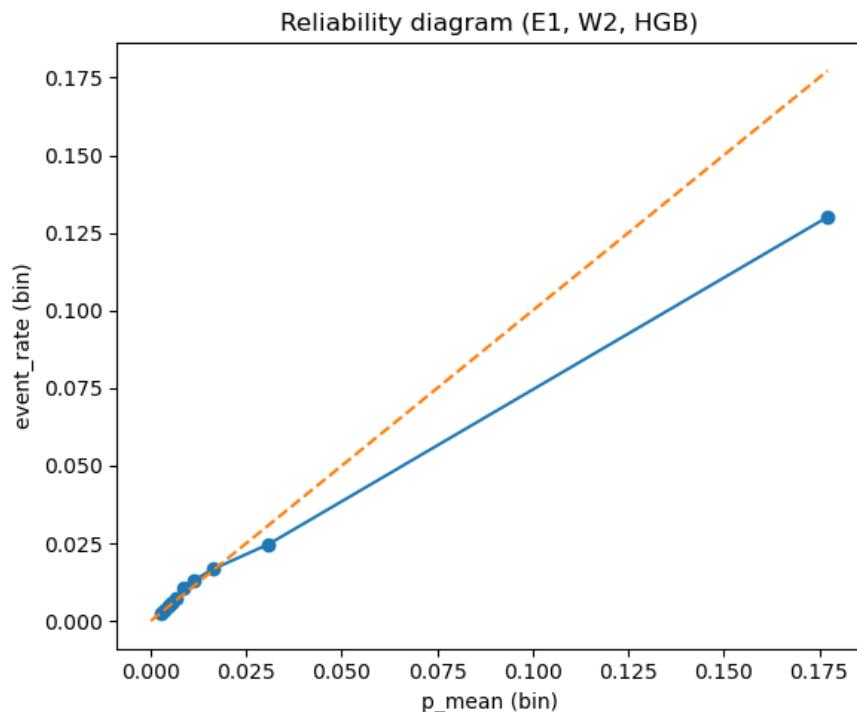


Figura 6.4: Diagrama de confiabilidad del modelo principal (E1, W2, HGB). La comparación entre probabilidad media predicha y tasa observada por grupos de riesgo permite evaluar la calibración del modelo.

El diagrama de confiabilidad correspondiente se presenta en la Figura 6.4. En conjunto con la tabla por deciles, este resultado refuerza la idea de que el modelo está bien alineado para tareas de priorización, aun cuando una eventual implementación operativa basada en umbrales probabilísticos podría beneficiarse de procedimientos adicionales de recalibración.

6.7. Heterogeneidad del desempeño entre cohortes

Con el fin de profundizar en la comparación temporal, se realizaron ejercicios adicionales manteniendo fijo el conjunto de variables y el modelo principal. Los resultados muestran que la capacidad predictiva no es homogénea entre cohortes. En los ejercicios de transferencia temporal, el escenario 2017 → 2021 alcanza mejores métricas que los ejercicios con prueba en 2024, lo que sugiere que esta última cohorte constituye un caso más exigente desde el punto de vista predictivo.

Al mismo tiempo, la comparación entre el escenario principal $2017 + 2021 \rightarrow 2024$ y el ejercicio $2021 \rightarrow 2024$ muestra que entrenar con dos cohortes históricas mejora el desempeño sobre 2024, lo que sugiere que una mayor profundidad temporal en el entrenamiento favorece la generalización hacia cohortes más recientes.

Este patrón también se confirma al evaluar cada cohorte por separado bajo un esquema de *holdout* por estudiante. Manteniendo fijo el modelo y el set de variables, la cohorte 2017 presenta la mayor capacidad de discriminación global, mientras que 2024 aparece como la más difícil de predecir. En consecuencia, la cohorte no opera únicamente como una forma de ordenar temporalmente los datos, sino como una dimensión analítica relevante para examinar la estabilidad del fenómeno y la dificultad relativa del problema predictivo.

La Tabla 6.7 resume los ejercicios adicionales de transferencia temporal entre cohortes.

Tabla 6.7: Desempeño del modelo principal (W2, HGB) en ejercicios de transferencia temporal entre cohortes.

Experimento	Cohorte(s) de entrenamiento	Cohorte de prueba	AUC	PR-AUC	lift@5
E1	2017, 2021	2024	0.8366	0.2377	9.4411
E2	2017	2021	0.8694	0.2834	10.1141
E4	2021	2024	0.8334	0.2189	9.0763
E3	2021	2017	0.8813	0.3533	9.5300

Nota: E3 corresponde a un ejercicio retrospectivo de robustez y no a un escenario prospectivo de aplicación.

De manera complementaria, la Tabla 6.8 muestra la capacidad predictiva obtenida al mantener fijo el modelo y el conjunto de variables dentro de cada cohorte.

Tabla 6.8: Capacidad predictiva por cohorte manteniendo fijo el modelo y el conjunto de variables (W2, HGB).

Cohorte	event_rate	AUC	PR-AUC	recall@5	lift@5
2017	0.0377	0.8925	0.3782	0.4925	9.8506
2021	0.0247	0.8784	0.3045	0.5287	10.5727
2024	0.0223	0.8487	0.2531	0.4821	9.6409

6.8. Estabilidad y variación de asociaciones entre cohortes

Como análisis complementario, se estimaron modelos de regresión logística parsimoniosos por cohorte con el fin de examinar la estabilidad de las asociaciones entre variables clave y no continuidad anual. Los resultados muestran una regularidad importante: las variables estructurales y académicas centrales conservan el mismo signo en las tres cohortes consideradas. En particular, el grado cursado, la sobreedad y el número de establecimientos acumulados presentan asociación positiva con el riesgo de no continuidad, mientras que la asistencia rezagada y el promedio general rezagado presentan asociación negativa en 2017, 2021 y 2024.

Este patrón sugiere que, aun cuando la capacidad predictiva agregada varía entre cohortes, las señales principales del riesgo mantienen una dirección consistente en el tiempo. En contraste, algunas variables contextuales muestran mayor variabilidad relativa entre cohortes, tanto en magnitud como en significancia estadística. En conjunto, estos resultados refuerzan la idea de que la estructura básica del riesgo es estable en sus componentes centrales, pero no completamente invariante en sus dimensiones de contexto.

La Tabla 6.9 presenta los odds ratios de las variables clave en los modelos logísticos parsimoniosos estimados por cohorte.

Tabla 6.9: Odds ratios de variables clave en modelos logísticos parsimoniosos por cohorte.

Variable	2017	2021	2024
grade_num	1.378***	1.306***	1.347***
SOBREEDAD	2.908***	2.797***	2.560***
n_establecimientos_hasta_anio	1.359***	1.322***	1.435***
ASISTENCIA_lag1	0.987***	0.994***	0.990***
PROM_GRAL_lag1	0.836***	0.736***	0.760***

Nota: *** $p < 0,01$. Se reportan odds ratios para facilitar la comparación entre cohortes. Valores mayores que 1 indican asociación positiva con la no continuidad anual; valores menores que 1 indican asociación negativa.

Un aspecto sustantivamente relevante de estos resultados es la diferencia entre variables que describen condiciones estructurales de la trayectoria y variables más próximas a ámbitos de acción escolar. Entre las primeras se cuentan el grado cursado, la sobreedad y la acumulación de cambios de establecimiento, que capturan procesos ya instalados o de más difícil modificación en el corto plazo. En cambio, la asistencia y el desempeño académico reciente, aun cuando también expresan trayectorias acumuladas, se vinculan de manera más directa con ámbitos donde el establecimiento puede ejercer mayor injerencia mediante monitoreo, acompañamiento y apoyo pedagógico. Esta distinción no implica que unas variables sean causalmente modificables y otras no, pero sí permite interpretar el modelo no solo como un instrumento descriptivo del riesgo, sino también como una herramienta útil para orientar focos de intervención preventiva.

6.9. Relevancia de variables y lectura sustantiva del riesgo

Con el fin de complementar la comparación de desempeño entre modelos, se examinó la relevancia relativa de las variables en la especificación principal seleccionada, correspondiente a HistGradientBoosting con la ventana $W2$. Este análisis permite distinguir entre variables que concentran una fracción importante de la señal predictiva y variables cuya contribución es más acotada una vez consideradas las demás dimensiones del modelo.

Los resultados muestran que la señal predictiva se concentra principalmente en variables de trayectoria y desempeño reciente. En particular, la sobreedad, el grado cursado, el promedio general rezagado y la asistencia rezagada aparecen entre los predictores más relevantes del riesgo de no continuidad anual. A ellas se suma el número de establecimientos acumulados hasta el año observado, lo que sugiere que la inestabilidad institucional también constituye una dimensión importante dentro de la configuración del riesgo.

Desde una perspectiva sustantiva, este patrón es consistente con la literatura revisada y con el enfoque de trayectorias adoptado en esta memoria. La no continuidad anual no parece estar determinada por un único rasgo individual aislado, sino por la acumulación de señales de fragilidad escolar observables en la trayectoria. En este sentido, la importancia relativa de la sobreedad, la movilidad entre establecimientos y el desempeño reciente refuerza la idea de que el riesgo se construye progresivamente y se expresa con distinta intensidad según el momento del recorrido educativo.

Al mismo tiempo, estos resultados permiten distinguir entre variables más estructurales y variables más próximas a ámbitos de acción escolar. Entre las primeras se cuentan el grado cursado, la sobreedad y la acumulación de cambios de establecimiento, que describen procesos ya instalados o de más difícil modificación en el corto plazo. En cambio, la asistencia y el desempeño académico reciente, aun cuando también reflejan trayectorias acumuladas, se vinculan más directamente con ámbitos donde el establecimiento puede ejercer mayor injerencia mediante monitoreo, acompañamiento y apoyo pedagógico. Esta distinción no implica interpretar causalmente el modelo, pero sí aporta una lectura más útil para fines de alerta temprana y focalización preventiva.

En conjunto, la evidencia sugiere que el modelo principal no se apoya en una única variable dominante, sino en una combinación de factores de trayectoria, rendimiento y contexto escolar. No obstante, la concentración de importancia en un conjunto relativamente acotado de predictores indica que la capacidad predictiva del modelo depende en gran medida de señales longitudinales observables y operativamente plausibles dentro del sistema escolar.

La Tabla 6.10 resume la importancia por permutación de las variables del modelo principal.

La inclusión de AÑO_EGRESO respondió inicialmente al propósito de capturar diferencias agregadas

Tabla 6.10: Variables más relevantes del modelo principal según importancia por permutación.

Variable	Importancia media	Desv. est.
SOBREEDAD	0.1305	0.0019
grade_num	0.1107	0.0005
PROM_GRAL_lag1	0.0396	0.0029
ASISTENCIA_lag1	0.0241	0.0007
n_establecimientos_hasta_anio	0.0127	0.0009
COD_REG_RBD	0.0025	0.0008
COD_DEPE2	0.0022	0.0005
HAS_SIMCE	0.0021	0.0003
GEN_ALU	0.0013	0.0007
RURAL_RBD	0.0004	0.0002
AÑO_EGRESO	0.0000	0.0000

entre cohortes teóricas de seguimiento. No obstante, su importancia predictiva resultó nula en la especificación principal, lo que sugiere que la mayor parte de la heterogeneidad temporal relevante ya queda recogida por otras variables de trayectoria, desempeño y contexto incorporadas en el modelo.

6.10. Estabilidad y variación de asociaciones entre tramos educativos

Como análisis complementario, se estimaron modelos logísticos parsimoniosos por tramo educativo con el fin de examinar si la dirección y magnitud de las asociaciones entre variables clave y no continuidad anual permanecen estables a lo largo de la trayectoria escolar. La Tabla 6.11 resume los *odds ratios* obtenidos para las variables numéricas principales en los tramos temprano, intermedio y tardío.

Los resultados muestran una regularidad importante: las variables núcleo conservan el mismo signo en los tres tramos considerados. En particular, la sobreedad, el grado cursado y el número de establecimientos acumulados presentan asociación positiva con la no continuidad anual, mientras que la asistencia rezagada y el promedio general rezagado presentan asociación negativa en las tres etapas. Este patrón refuerza la idea de que la estructura básica del riesgo mantiene una

Tabla 6.11: Odds ratios de variables numéricas clave en modelos logísticos parsimoniosos por tramo educativo.

Variable	Temprano	Intermedio	Tardío
SOBREEDAD	2.555***	3.383***	2.886***
n_establecimientos_hasta_anio	1.461***	1.334***	1.389***
ASISTENCIA_lag1	0.982***	0.990***	0.995***
grade_num	1.112***	1.448***	1.080***
PROM_GRAL_lag1	0.917***	0.802***	0.678***

Nota: *** $p < 0,01$. Se reportan *odds ratios*. Valores mayores que 1 indican asociación positiva con la no continuidad anual; valores menores que 1 indican asociación negativa.

dirección consistente a lo largo de la trayectoria escolar.

Sin embargo, la intensidad de estas asociaciones no es homogénea entre tramos. La sobreedad presenta su mayor magnitud relativa en el tramo intermedio, lo que sugiere que el desfase entre edad y grado adquiere especial relevancia en una etapa marcada por transiciones y reorganización de expectativas escolares. De manera similar, el promedio general rezagado exhibe una asociación protectora más fuerte en el tramo tardío, indicando que el deterioro del desempeño reciente adquiere mayor peso como señal de riesgo en etapas más avanzadas del recorrido escolar.

La asistencia rezagada mantiene una asociación protectora en los tres tramos, aunque su magnitud es relativamente mayor en la etapa temprana. Esto sugiere que, en los primeros años de la trayectoria observada, la desvinculación del proceso escolar cotidiano constituye una señal particularmente sensible de fragilidad. Por su parte, la movilidad entre establecimientos presenta asociación positiva en todos los tramos, con una magnitud relativamente más alta en la etapa temprana, lo que es consistente con la idea de que la inestabilidad institucional temprana puede debilitar el vínculo escolar desde fases iniciales del recorrido educativo.

En conjunto, estos resultados sugieren que la no continuidad anual no solo varía en prevalencia y dificultad predictiva entre tramos, sino también en la intensidad relativa de los factores que

configuran el riesgo. Así, aunque el núcleo de señales relevantes permanece estable, su peso analítico cambia según el momento de la trayectoria, lo que refuerza la pertinencia de una lectura temporalmente diferenciada del fenómeno.

6.11. Análisis de sensibilidad: incorporación exploratoria de dimensiones psicosociales

Como complemento a la comparación entre ventanas presentada anteriormente, se realizó un ejercicio exploratorio de sensibilidad agregando individualmente distintas dimensiones psicosociales sobre la especificación W2 del modelo principal en el escenario E1. El objetivo de este análisis no fue redefinir exhaustivamente la construcción de W3, sino examinar si algunas dimensiones contextuales podían aportar señal predictiva adicional una vez consideradas las variables de trayectoria y desempeño reciente.

Los resultados muestran que varias dimensiones contienen información adicional, aunque con ganancias marginales en términos absolutos. Entre las mejoras más favorables destacan seguridad percibida, salud y bienestar, ingresos familiares percibidos y aspiraciones educativas, todas con aumentos pequeños en PR-AUC y en lift@5 respecto de la especificación base. En contraste, otras dimensiones como discriminación percibida, organización y participación estudiantil, higiene e infraestructura básica, recursos disponibles y apoyo familiar no producen mejoras y, en algunos casos, muestran leves deterioros del desempeño.

En conjunto, este ejercicio refuerza la idea de que las dimensiones psicosociales y de contexto no carecen de señal sustantiva, pero sí presentan un aporte incremental acotado una vez consideradas variables administrativas y académicas recientes. Por ello, la especificación W3 debe entenderse como una selección parsimoniosa y plausible de variables contextuales, más que como el resultado

de una búsqueda exhaustiva entre todas las dimensiones construidas. El detalle completo de este screening se presenta en el Anexo 8.2.

6.12. Robustez: particiones alternativas y estabilidad estadística

La robustez del modelo principal se evaluó mediante dos procedimientos complementarios. En primer lugar, se comparó el desempeño entre el escenario temporal $E1$ y el benchmark por MRUN. Como ya se indicó, el segundo presenta métricas más altas, lo que confirma que la generalización temporal entre cohortes constituye un escenario más exigente y, por tanto, más informativo para fines prospectivos.

En segundo lugar, se implementó un procedimiento de remuestreo bootstrap sobre el conjunto de prueba del escenario $E1$. Los intervalos resultantes se presentan en la Tabla 6.12. La variación observada en las métricas principales es acotada: el AUC se ubica entre 0,8346 y 0,8383, el PR -AUC entre 0,2338 y 0,2423, y el $recall@5$ entre 0,4675 y 0,4769. Del mismo modo, la $precision@5$ y el $lift@5$ permanecen dentro de rangos estrechos. En conjunto, estos resultados sugieren que las conclusiones obtenidas no dependen de forma sensible de pequeñas perturbaciones muestrales en el conjunto de evaluación.

Tabla 6.12: Intervalos bootstrap para métricas seleccionadas del modelo principal ($E1$, $W2$, HGB).

Métrica	p2.5	Mediana	p97.5
AUC	0.8346	0.8363	0.8383
PR-AUC	0.2338	0.2374	0.2423
recall@5	0.4675	0.4716	0.4769
precision@5	0.2041	0.2069	0.2094
lift@5	9.3495	9.4325	9.5379

6.13. Alcance conceptual de la etiqueta anual

Como verificación complementaria, se examinó el grado de superposición entre la no continuidad anual y trayectorias que culminan en desvinculación persistente del sistema. Los resultados muestran una superposición alta entre ambas condiciones, mientras que la etiqueta anual incorpora además interrupciones no terminales que posteriormente pueden revertirse. Como se muestra en el Anexo 8.6, el 12,29% de quienes presentan no continuidad anual no queda clasificado como desvinculación persistente, lo que evidencia que la etiqueta anual también captura interrupciones potencialmente reversibles.

Este patrón sugiere que la no continuidad anual conserva buena parte de la señal estructural de riesgo asociada a trayectorias altamente vulnerables, pero la amplía al incorporar quiebres del vínculo escolar que no necesariamente culminan en salida definitiva. Desde el punto de vista sustantivo, ello refuerza la pertinencia de la etiqueta anual como aproximación más alineada con una lógica preventiva y de alerta temprana.

6.14. Síntesis de resultados

En síntesis, los resultados permiten establecer siete hallazgos principales. Primero, el modelo HGB supera de manera consistente a la regresión logística, tanto en el escenario temporal por cohorte como en el benchmark por MRUN. Segundo, la incorporación de variables académicas rezagadas produce una mejora sustantiva respecto de una especificación basada solo en trayectoria y contexto. Tercero, la información adicional de SIMCE y de dimensiones psicosociales aporta mejoras marginales, por lo que $W2$ emerge como la ventana más equilibrada entre desempeño, parsimonia y cobertura.

Cuarto, el desempeño del modelo es heterogéneo entre tramos, con mejores resultados operativos en las etapas intermedia y tardía que en la etapa temprana. Quinto, la calibración del modelo principal es razonable en promedio, aunque con cierta sobreestimación en la cola superior del ranking. Sexto, la capacidad predictiva no es homogénea entre cohortes: 2024 aparece como la cohorte más exigente, mientras que 2017 presenta la mayor capacidad de discriminación global bajo el mismo conjunto de variables. Séptimo, aun cuando el desempeño agregado cambia entre cohortes, las variables centrales del riesgo conservan una dirección estable en el tiempo, especialmente en lo relativo a sobreedad, grado, movilidad entre establecimientos y desempeño académico reciente.

En conjunto, estos resultados refuerzan la idea de que la no continuidad anual puede anticiparse mediante señales longitudinales observables, pero también que su configuración depende del momento de la trayectoria y del contexto de cohorte en que se observe al estudiante.

7. Conclusiones

7.1. Conclusiones generales

La presente investigación tuvo como propósito predecir la no continuidad escolar anual a partir de información longitudinal del sistema escolar chileno, examinando además cómo varía el riesgo según tramo educativo y cohorte teórica de egreso. A diferencia de enfoques centrados exclusivamente en la deserción final o persistente, el estudio adoptó una definición del evento orientada a capturar quiebres anuales del vínculo escolar, en coherencia con una lógica preventiva y de alerta temprana. Esta decisión conceptual permitió situar el análisis en un punto intermedio entre la observación de trayectorias efectivamente interrumpidas y la necesidad operativa de detectar episodios de riesgo antes de que se consoliden en salidas más prolongadas del sistema.

Los resultados obtenidos muestran que la no continuidad escolar anual constituye un fenómeno empíricamente identificable y predictivamente abordable mediante registros administrativos y antecedentes educativos acumulados. En el escenario principal de evaluación temporal por cohorte, el modelo *HistGradientBoosting* con la ventana $W2$ alcanzó un desempeño sólido tanto en métricas globales de discriminación como en métricas de priorización operativa. Este hallazgo confirma que el uso de aprendizaje automático puede mejorar la capacidad predictiva respecto de una línea base de regresión logística, especialmente cuando se incorporan antecedentes académicos recientes del estudiante.

Al mismo tiempo, la investigación muestra que el mayor aumento de desempeño no proviene de incorporar toda la información disponible, sino de seleccionar cuidadosamente aquella que agrega señal de manera consistente y operativamente plausible. En efecto, la transición desde una ventana basada en trayectoria y contexto ($W1$) hacia una que incorpora variables académicas rezagadas

(W2) produce una mejora clara en el rendimiento del modelo, mientras que la incorporación adicional de dimensiones SIMCE y psicosociales (W3) entrega ganancias marginales. Este resultado es metodológicamente importante, ya que sugiere que, para fines de focalización preventiva, una especificación parsimoniosa y bien alineada con la disponibilidad temporal de los datos puede ser tan o más útil que una configuración más compleja.

7.2. Hallazgos principales del estudio

El primer hallazgo central es que la no continuidad escolar anual no se distribuye homogéneamente a lo largo de la trayectoria educativa. La prevalencia del evento aumenta a medida que el estudiante avanza en su recorrido escolar, y ello se refleja tanto en los análisis descriptivos como en el desempeño diferenciado de los modelos por tramo. En términos operativos, el modelo principal muestra una capacidad de priorización más efectiva en los tramos intermedio y tardío que en el tramo temprano, lo que es coherente con la mayor prevalencia observada del evento en esas etapas y con la acumulación progresiva de señales de riesgo.

El segundo hallazgo es que las variables con mayor aporte predictivo corresponden principalmente a dimensiones de trayectoria y desempeño reciente, en particular sobreedad, grado cursado, promedio general rezagado y asistencia rezagada. Este patrón refuerza la idea de que el riesgo de no continuidad anual está estrechamente ligado a procesos acumulativos de fragilidad escolar más que a un único evento puntual. En consecuencia, la predicción del riesgo no debe interpretarse como una clasificación estática de estudiantes, sino como una lectura dinámica del momento de la trayectoria y de las señales que anticipan un posible quiebre del vínculo con el sistema.

El tercer hallazgo es que la no continuidad anual presenta una superposición empírica alta con trayectorias que culminan en desvinculación persistente, pero no se reduce a ellas. La etiqueta

anual conserva buena parte de la señal estructural de riesgo asociada a trayectorias altamente vulnerables, aunque incorpora además interrupciones no terminales que posteriormente pueden revertirse. Esta característica constituye uno de los principales aportes conceptuales del estudio, ya que desplaza el foco desde la detección tardía de salidas consolidadas hacia una lectura más sensible al momento en que ocurre la interrupción del vínculo escolar.

El cuarto hallazgo es que la evaluación temporal por cohorte constituye un criterio más exigente y más informativo que la simple partición por estudiante. El hecho de que el desempeño bajo el escenario principal $E1$ sea menor que en el benchmark por MRUN confirma que la generalización hacia cohortes futuras involucra un desafío adicional, asociado a cambios de composición, prevalencia y contexto entre generaciones de estudiantes. En ese sentido, la decisión de privilegiar la evaluación temporal fortalece la validez aplicada del estudio y evita sobreestimar la utilidad del modelo bajo condiciones más favorables que las de un uso prospectivo real.

Un quinto hallazgo es que la capacidad predictiva no permanece constante entre cohortes. Manteniendo fijo el modelo y el conjunto de variables, la cohorte 2024 aparece como la más exigente desde el punto de vista predictivo, mientras que la cohorte 2017 presenta la mayor capacidad de discriminación global. Este resultado refuerza la idea de que la cohorte no constituye únicamente una forma de ordenar temporalmente los datos, sino una dimensión analítica relevante para examinar la estabilidad del fenómeno y la generalización del modelo en contextos históricos distintos.

Un sexto hallazgo se relaciona con la lectura sustantiva de las variables más relevantes. Los resultados muestran que el modelo principal combina señales estructurales de trayectoria —como la sobreedad, el grado cursado y la movilidad entre establecimientos— con indicadores más próximos a ámbitos de acción escolar, como la asistencia y el desempeño académico reciente. Esta

distinción no transforma el modelo en causal, pero sí aporta criterios útiles para interpretar el riesgo y orientar prioridades de monitoreo y acompañamiento preventivo.

Un séptimo hallazgo es que, además de variar en prevalencia y desempeño predictivo, la no continuidad anual también presenta diferencias en la intensidad relativa de sus factores asociados según tramo educativo. Los análisis logísticos parsimoniosos por tramo muestran que las variables núcleo conservan una dirección estable a lo largo de la trayectoria, pero con magnitudes distintas: la sobreedad presenta una asociación especialmente fuerte en el tramo intermedio, mientras que el desempeño académico reciente adquiere mayor peso relativo en el tramo tardío. Esto sugiere que la estructura básica del riesgo es estable, aunque su configuración interna cambia según el momento del recorrido escolar.

7.3. Aporte de la investigación

El aporte de este trabajo no radica en introducir por primera vez el uso de aprendizaje automático para problemas de riesgo escolar en Chile, ya que la literatura reciente ha avanzado de manera relevante en esa dirección. Su contribución específica consiste, más bien, en tres aspectos complementarios.

En primer lugar, el estudio desplaza el foco desde la deserción final hacia la no continuidad escolar anual como evento objetivo, lo que permite alinear la predicción con una lógica más preventiva y sensible a trayectorias interrumpidas o potencialmente reversibles. En segundo lugar, incorpora una lectura explícita de heterogeneidad temporal mediante el análisis por tramo educativo y por cohorte, mostrando que el riesgo y el desempeño predictivo no permanecen constantes a lo largo del recorrido escolar. En tercer lugar, compara sistemáticamente modelos, ventanas de observación y esquemas de evaluación fuera de muestra, lo que permite seleccionar una especificación principal

no solo por su rendimiento, sino también por su parsimonia, su estabilidad y su factibilidad de implementación. Asimismo, la memoria no solo compara desempeño predictivo entre modelos, ventanas y cohortes, sino que también examina cómo cambia la intensidad relativa de factores clave según tramo educativo, aportando una lectura más fina de la heterogeneidad temporal del riesgo.

Adicionalmente, la memoria examina no solo diferencias de desempeño entre cohortes, sino también la estabilidad de asociaciones entre variables clave y riesgo de no continuidad, mostrando que algunos predictores centrales mantienen una dirección consistente en el tiempo, mientras que otros presentan mayor sensibilidad al contexto de cohorte.

Desde esta perspectiva, la memoria se sitúa en continuidad con la literatura de sistemas de alerta temprana, pero aporta una forma distinta de definir el evento de interés y de estructurar la evaluación. En otras palabras, no se propone reemplazar la discusión existente sobre abandono o deserción, sino complementarla con una aproximación más cercana al problema operativo de anticipar quiebres del vínculo escolar en tiempo oportuno.

7.4. Implicancias sustantivas y operativas

Los resultados obtenidos tienen implicancias tanto analíticas como de política educativa. Desde una perspectiva sustantiva, refuerzan la idea de que la continuidad escolar debe entenderse como una dimensión dinámica de la trayectoria educativa y no únicamente como el resultado terminal de permanecer o salir del sistema. La existencia de interrupciones anuales observables, incluso entre estudiantes que no terminan en desvinculación persistente, sugiere que la fragilidad del vínculo escolar puede manifestarse antes y de forma más sutil que la deserción consolidada.

Desde una perspectiva operativa, el modelo principal muestra que es posible priorizar un subconjunto relativamente pequeño de observaciones en el cual la incidencia del evento es sustantivamente mayor que la prevalencia base. Esto resulta especialmente relevante para contextos de intervención con recursos limitados, en los cuales la utilidad del modelo no depende de clasificar exhaustivamente a todos los casos, sino de ordenar el riesgo de manera que la capacidad institucional pueda focalizarse de forma más eficiente.

No obstante, la utilidad del modelo debe entenderse como complemento y no como sustituto de la intervención educativa. La predicción no identifica causas individuales ni prescribe automáticamente una respuesta, sino que ofrece una herramienta de apoyo para orientar procesos de monitoreo, acompañamiento y focalización preventiva. En ese sentido, el valor del enfoque depende tanto del desempeño estadístico del modelo como de la capacidad institucional para traducir la alerta en acciones pedagógicas y de apoyo pertinentes.

En este marco, la distinción entre variables más estructurales y variables más próximas a ámbitos de acción escolar resulta especialmente útil. Mientras algunas señales describen trayectorias ya acumuladas y de más difícil modificación inmediata, otras —como la asistencia y el desempeño reciente— pueden funcionar como focos de seguimiento más directamente utilizables por establecimientos y equipos de apoyo.

7.5. Limitaciones del estudio

Aun cuando los resultados son consistentes y metodológicamente sólidos, el estudio presenta limitaciones que deben ser reconocidas explícitamente.

En primer lugar, las cohortes analizadas son comparables en términos analíticos, pero no estric-

tamente equivalentes en su contexto histórico. La selección de cohortes se basa en criterios de seguimiento longitudinal, disponibilidad de SIMCE y observabilidad hasta años teóricos de egreso, pero ello no elimina completamente diferencias asociadas a cambios de contexto, shocks históricos o variaciones administrativas. En particular, parte de la heterogeneidad observada entre cohortes podría reflejar no solo diferencias en el comportamiento del fenómeno, sino también cambios en el entorno educativo o en la composición de las trayectorias.

En este sentido, las diferencias de desempeño observadas entre cohortes no deben interpretarse automáticamente como cambios causales en los mecanismos subyacentes, sino como evidencia de que la generalización temporal entre cohortes está sujeta tanto a cambios en la estructura del fenómeno como a variaciones del contexto histórico y administrativo.

En segundo lugar, la variable objetivo se construye a partir de registros administrativos y, por tanto, identifica la ausencia de matrícula observada en el año siguiente, pero no permite distinguir directamente los motivos de dicha interrupción. La no continuidad anual puede estar asociada a trabajo, responsabilidades de cuidado, embarazo, movilidad, problemas económicos u otros factores no observables en la base utilizada. En consecuencia, el modelo predice riesgo de interrupción del vínculo escolar, pero no explica causalmente las razones subyacentes de ese quiebre.

En tercer lugar, aunque la incorporación de información SIMCE y de dimensiones psicosociales enriquece el análisis, su disponibilidad no es perfectamente homogénea entre cohortes y tramos. Además, la construcción del antecedente SIMCE responde a una lógica de último registro disponible en la trayectoria, lo que exige cautela en la interpretación de estas variables, especialmente en etapas avanzadas del recorrido educativo. Adicionalmente, aunque a partir de los cuestionarios SIMCE se construyó un conjunto amplio de dimensiones psicosociales y de contexto escolar, la especificación W3 incorporó solo un subconjunto de ellas, definido con criterios de plausibilidad

teórica, interpretabilidad y cobertura razonablemente alta. En consecuencia, el aporte marginal observado para estas variables debe interpretarse como referido a esa selección principal y no como resultado de una comparación exhaustiva entre todas las dimensiones contextuales potencialmente disponibles.

En cuarto lugar, la evaluación principal se concentra en tres cohortes teóricas específicas. Si bien ello permite una comparación temporal razonable y metodológicamente consistente, la validación externa del modelo sobre otras generaciones, territorios o contextos institucionales queda fuera del alcance del presente estudio. En consecuencia, las conclusiones deben entenderse como válidas para el universo analizado, sin extrapolar automáticamente su desempeño a otros escenarios sin verificación adicional.

Finalmente, aunque la calibración del modelo principal resulta razonable en promedio, se observa cierta sobreestimación del riesgo en la cola superior del ranking. Esto no invalida su utilidad para tareas de priorización, pero sí sugiere que una implementación operativa basada en probabilidades absolutas o en umbrales explícitos podría requerir procedimientos adicionales de recalibración.

7.6. Líneas futuras de investigación

A partir de estas limitaciones, se abren varias líneas de profundización futura.

Una primera línea consiste en extender la validación temporal y externa del modelo, incorporando nuevas cohortes y contrastando su desempeño en otros contextos o ventanas históricas. Esto permitiría evaluar con mayor precisión la estabilidad del enfoque ante cambios de contexto y reforzar la discusión sobre generalización temporal entre generaciones de estudiantes.

Una segunda línea consiste en enriquecer la representación temporal del riesgo. En lugar de

utilizar únicamente antecedentes rezagados simples, futuros estudios podrían explorar secuencias más largas de asistencia, rendimiento, movilidad escolar e interrupciones previas, con el fin de capturar patrones dinámicos más complejos dentro de la trayectoria.

Una tercera línea se relaciona con la interpretación y uso institucional del modelo. Sería especialmente valioso complementar el enfoque predictivo con evidencia sobre intervención, evaluando qué tipo de acciones resultan más efectivas una vez que un estudiante es identificado como de alto riesgo. En otras palabras, la predicción puede mejorarse, pero también debe integrarse a una agenda más amplia sobre respuesta educativa, revinculación y sostenimiento de trayectorias.

Finalmente, una cuarta línea consiste en profundizar el análisis de las diferencias entre tramos educativos. Los resultados de esta memoria sugieren que el problema predictivo cambia según la etapa de la trayectoria, por lo que futuras investigaciones podrían estimar modelos específicos por tramo o incluso por grado, con el fin de identificar configuraciones de riesgo más finas y adaptar las estrategias de focalización a distintos momentos del recorrido escolar.

7.7. Cierre

En conclusión, la memoria muestra que la no continuidad escolar anual constituye un evento metodológicamente definible, sustantivamente relevante y predictivamente abordable en el sistema escolar chileno. La evidencia obtenida indica que el riesgo puede anticiparse con un grado importante de precisión utilizando información longitudinal administrativa y antecedentes educativos acumulados, y que dicha predicción resulta particularmente útil cuando se orienta a la priorización operativa de casos.

El estudio confirma que la incorporación de variables académicas rezagadas mejora de manera

sustantiva el desempeño predictivo y que los modelos de aprendizaje automático superan a la regresión logística en este contexto. Al mismo tiempo, muestra que la utilidad del modelo depende del momento de la trayectoria y que la etiqueta anual permite capturar una parte del fenómeno que permanece invisibilizada cuando solo se observa la desvinculación final.

En ese sentido, el principal valor de la investigación no consiste únicamente en alcanzar un buen desempeño estadístico, sino en proponer una forma de medir y anticipar el riesgo que resulta más coherente con una lógica de prevención temprana y con una comprensión longitudinal de las trayectorias educativas. Desde esa perspectiva, la memoria ofrece una contribución empírica, metodológica y aplicada al estudio de la continuidad escolar en Chile.

Referencias

Agencia de Calidad de la Educación. (2020). *Propuesta de plan de evaluaciones 2021–2025*.

Agencia de Calidad de la Educación. Consultado el 14 de abril de 2026, desde https://archivos.agenciaeducacion.cl/Propuesta_de_Plan_de_Evaluaciones_del_Consejo_ACE_FINAL.pdf

Balfanz, R., & Byrnes, V. (2019). Early warning indicators and intervention systems: State of the field. En J. A. Fredricks, A. L. Reschly & S. L. Christenson (Eds.), *Handbook of Student Engagement Interventions: Working with Disengaged Students* (pp. 45-55). Academic Press. <https://doi.org/10.1016/B978-0-12-813413-9.00004-8>

Contreras, D., González, L., Láscar, S., & López, V. (2022). Negative teacher–student and student–student relationships are associated with school dropout: Evidence from a large-scale longitudinal study in Chile. *International Journal of Educational Development*, 91, 102576. <https://doi.org/10.1016/j.ijedudev.2022.102576>

Devriendt, F., Guns, T., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, 548, 497-515. <https://doi.org/10.1016/j.ins.2020.10.084>

Doll, J. J., Eslami, Z., & Walters, L. (2013). Understanding why students drop out of high school. *Children & Schools*, 35(4), 219-230. <https://doi.org/10.1093/cs/cdt016>

Espinoza, Ó., Castillo, D., González, L. E., & Loyola, J. (2014). Factores familiares asociados a la deserción escolar en los niños y niñas mapuche: un estudio de caso. *Estudios Pedagógicos (Valdivia)*, 40(1), 97-112. <https://doi.org/10.4067/S0718-07052014000100006>

- Espinoza, O., González, L. E., McGinn, N., & Castillo, D. (2020). What factors predict the engagement of dropouts in alternative secondary schools in Chile? *Improving Schools*, 23(1), 47-67. <https://doi.org/10.1177/1365480219864835>
- Gómez Vera, G., Rivas Muená, M., & Lobos Guerrero, C. (2021). Expectativas sobre la transición desde la educación básica a la educación media de estudiantes provenientes de contextos de vulnerabilidad social. *Pensamiento Educativo. Revista de Investigación Educativa Latinoamericana*, 58(1), 1-13. <https://doi.org/10.7764/PEL.58.1.2021.1>
- Haimovich, F., Vazquez, E. J., & Adelman, M. A. (2021). *Scalable Early Warning Systems for School Dropout Prevention: Evidence from a 4,000-School Randomized Controlled Trial* (Policy Research Working Paper N.º 9685). World Bank. <https://doi.org/10.1596/1813-9450-9685>
- Ibáñez Iribarra, D. E. (2018). *Predicción y descripción de la exclusión educativa del sistema escolar regular chileno, ciencia de datos para la innovación pública* [Memoria para optar al título de Ingeniero Civil Industrial]. Universidad de Chile. Consultado el 14 de abril de 2026, desde <https://repositorio.uchile.cl/handle/2250/172696>
- Kattan, R. B., & Székely, M. (2017). Analyzing Upper Secondary Education Dropout in Latin America through a Cohort Approach. *Journal of Education and Learning*, 6(4), 12-39. <https://doi.org/10.5539/jel.v6n4p12>
- Lyche, C. S. (2010). *Taking on the completion challenge: A literature review on policies to prevent dropout and early school leaving* (OECD Education Working Papers N.º 53). OECD Publishing. Paris. <https://doi.org/10.1787/5km4m2t59cmr-en>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

- Mendoza Lira, M., Ballesta Acevedo, E., Muñoz Jorquera, S., & Covarrubias Apablaza, C. G. (2023). Conceptualizations of School Dropout and Retention from Chilean Educational Actors' Perspectives. *Participatory Educational Research*, 10(2), 216-235. <https://doi.org/10.17275/per.23.37.10.2>
- Ministerio de Educación. (2020). *Propuestas Mesa Técnica para la prevención de la deserción escolar* (Informe técnico). Ministerio de Educación. Santiago, Chile. https://www.mineduc.cl/wp-content/uploads/sites/19/2020/07/propuestas_desercionescolar.pdf
- Ministerio de Educación. (2025a, septiembre). *Chile Presente: Protegiendo trayectorias educativas. Guía de orientación para establecimientos educativos* (Guía de orientación). Ministerio de Educación, División de Educación General. Santiago, Chile. https://chilepresente.mineduc.cl/wp-content/uploads/sites/140/2025/12/Orientaciones_EE_Chile_Presente.pdf
- Ministerio de Educación. (2025b, octubre). *Desvinculación alcanza su mínimo histórico: 2024 es el año con menos estudiantes que abandonan el sistema escolar desde que existen registros*. Ministerio de Educación. <https://www.mineduc.cl/desvinculacion-alcanza-su-minimo-historico-2024-es-el-ano-con-menos-estudiantes-que-abandonan-el-sistema-escolar-desde-que-existen-registros/>
- Ministerio de Educación, Centro de Estudios. (2020, abril). *Medición de la exclusión escolar en Chile* (Documento de trabajo N.º 20). Ministerio de Educación. Santiago, Chile. https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2020/04/DOCUMENTO-DE-TRABAJO-20_2020_f01.pdf
- Ministerio de Educación, Centro de Estudios. (2023). *Situación de la desvinculación y la asistencia, año 2022* (Apuntes N° 25). Ministerio de Educación, Centro de Estudios. Santiago, Chile. https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2023/04/APUNTES-25_2023_fd01.pdf

- Ministerio de Educación, Centro de Estudios. (2025, noviembre). *Revinculación en el sistema escolar del año 2025* (Apuntes N.º 99). Ministerio de Educación. Santiago, Chile. https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/21919/APUNTES%2099_2025_fd01.pdf?isAllowed=y&sequence=1
- Rabuco Hidalgo, A. (2022). Factores asociados a la prevención del abandono escolar: una mirada desde la implementación del programa tutorías pedagógicas. *Sophia Austral*, 28(8), 1-21. <https://doi.org/10.22352/SAUSTRAL20222808>
- Rodríguez, P., Villanueva, A., Dombrowskaia, L., & Valenzuela, J. P. (2023). A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: The case of Chile. *Education and Information Technologies*, 28(8), 10103-10149. <https://doi.org/10.1007/s10639-022-11515-5>
- Romo, J., & Cumsille, P. (2020). Caracterización de estudiantes con trayectorias educativas interrumpidas en una muestra de escuelas de reingreso a través de análisis de clases latentes. *Pensamiento Educativo. Revista de Investigación Educativa Latinoamericana*, 57(1), 1-21. <https://doi.org/10.7764/pel.57.1.2020.4>
- Smith-Uldall, J., & Gutiérrez Rojas, C. (2022). An Application of Machine Learning in Public Policy: Early Warning Prediction of School Dropout in the Chilean Public Education System. *Multidisciplinary Business Review*, 15(1), 20-35. <https://doi.org/10.35692/07183992.15.1.4>
- UNESCO. (2026). *Access and equity* (Report). United Nations Educational, Scientific y Cultural Organization. <https://www.unesco.org/reports/gem-report/en/2026-access-equity>
- UNESCO Office Santiago and Regional Bureau for Education in Latin America and the Caribbean. (2021). *Early Warning Systems for School Dropout Prevention in Latin America and*

the Caribbean. UNESCO. Santiago, Chile. https://unesdoc.unesco.org/ark:/48223/pf0000380354_eng

United Nations General Assembly. (1989). Convention on the Rights of the Child [Adopted by General Assembly resolution 44/25 of 20 November 1989]. *United Nations Treaty Series*, 180, 159-184. <https://www.ohchr.org/sites/default/files/crc.pdf>

8. Anexos

8.1. Mapeo de dimensiones psicosociales

Tabla 8.1: Resumen del mapeo de dimensiones psicosociales y de contexto construidas a partir de cuestionarios SIMCE.

Cód.	Dimensión	Informante(s)	Disponib.	Nº ítems	Ejemplos abreviados de contenido
AA	Autoestima académica	Est.	9/9 aplic.	138	Capacidad para aprender contenidos difíciles; percepción de logro; satisfacción con las notas.
AF	Apoyo familiar	Est./Apod.	9/9 aplic.	55	Ayuda para estudiar; apoyo con tareas; seguimiento familiar del desempeño escolar.
ASP	Aspiraciones educativas	Est./Apod.	9/9 aplic.	15	Nivel educacional que el estudiante espera completar o que la familia proyecta para él/ella.
C	Clima de aula y establecimiento	Est./Apod.	9/9 aplic.	159	Participación en clases; confianza en el curso; ambiente general del establecimiento.
CO	Convivencia	Est./Apod.	9/9 aplic.	120	Respeto entre estudiantes; apoyo ante problemas; trato cotidiano dentro del curso.
CR	Conductas de riesgo	Est./Apod.	5/9 aplic.	27	Consumo de sustancias; conductas problemáticas; exposición a situaciones de riesgo.
DA	Apoyo docente	Est./Apod.	8/9 aplic.	95	Docentes que explican cuando no se entiende; felicitación; preocupación por el aprendizaje.

Cód.	Dimensión	Informante(s) Disponib.		N°	Ejemplos abreviados de contenido ítems
DI	Discriminación percibida	Est./Apod.	8/9 aplic.	146	Sentirse excluido, discriminado o “mirado en menos” en el establecimiento.
H	Higiene e infraestructura básica	Est./Apod.	8/9 aplic.	25	Estado y limpieza de baños; condiciones básicas de infraestructura escolar.
IF	Ingresos familiares percibidos	Apod.	8/9 aplic.	8	Tramo de ingresos del hogar o situación económica percibida.
INV	Involucramiento escolar	Est./Apod.	8/9 aplic.	102	Participación en talleres; actividades extracurriculares; apoyo a actividades del establecimiento.
M	Motivación escolar	Est./Apod.	9/9 aplic.	111	Esfuerzo en clases; interés por aprender; persistencia frente a tareas difíciles.
N	Normas y disciplina	Est./Apod.	9/9 aplic.	146	Percepción de reglas; gravedad atribuida a conductas como copiar o hacer la cimarra; aplicación de normas.
NF	Número de integrantes del hogar	Apod.	5/9 aplic.	5	Cantidad de personas que viven en el hogar del estudiante.
ORG	Organización y participación estudiantil	Est./Apod.	5/9 aplic.	22	Consejo de curso; directiva; centro de alumnos; organización del curso.
R	Recursos disponibles	Est./Apod.	8/9 aplic.	24	Acceso a internet y dispositivos; facilidad para realizar actividades escolares con recursos digitales.
S	Seguridad percibida	Est./Apod.	6/9 aplic.	22	Seguridad en sala, patio, baños, pasillos y otros espacios del establecimiento.
SD	Salud y bienestar	Est.	8/9 aplic.	198	Hábitos alimentarios; educación sexual; autocuidado y otros aspectos de bienestar.

Cód.	Dimensión	Informante(s) Disponib.		Nº	Ejemplos abreviados de contenido
				ítems	
V	Experiencias de violencia	Est./Apod.	8/9 aplic.	104	Peleas, insultos, amenazas, robos y otras situaciones de violencia escolar.
VIN	Vinculación con el establecimiento	Est./Apod.	9/9 aplic.	158	Satisfacción con el establecimiento; sentido de pertenencia; valoración de la experiencia escolar.

Nota. “Disponib.” indica en cuántas de las 9 aplicaciones revisadas aparece la dimensión. “Nº ítems” corresponde al total de asignaciones ítem–aplicación–informante identificadas en el archivo de mapeo, no al número de preguntas únicas. Los ejemplos abreviados son ilustrativos y no exhaustivos para no extender de sobremanera el archivo.

8.2. Screening exploratorio de dimensiones psicosociales sobre W2

Tabla 8.2: Screening exploratorio de dimensiones psicosociales sobre la ventana W2 en el escenario principal E1.

Dimensión	Cob. train	Cob. test	PR-AUC	Δ PR-AUC	lift@5	Δ lift@5
S	0.5254	0.4986	0.2388	0.0011	9.4887	0.0476
SD	0.5566	0.8187	0.2384	0.0006	9.4619	0.0208
IF	0.5218	0.7019	0.2384	0.0006	9.4847	0.0436
ASP	0.7788	0.7805	0.2383	0.0005	9.4490	0.0079

Continúa en la página siguiente

Dimensión	Cob. train	Cob. test	PR-AUC	Δ PR-AUC	lift@5	Δ lift@5
V	0.5534	0.8301	0.2381	0.0004	9.4327	-0.0084
AA	0.7988	0.8289	0.2381	0.0004	9.4545	0.0134
N	0.7849	0.8308	0.2381	0.0003	9.4669	0.0258
CR	0.2972	0.5545	0.2379	0.0002	9.4580	0.0169
VIN	0.7876	0.8290	0.2379	0.0001	9.4456	0.0045
M	0.7999	0.8244	0.2379	0.0001	9.4426	0.0015
NF	0.2794	0.7787	0.2378	0.0000	9.4475	0.0064
C	0.7984	0.8308	0.2376	-0.0002	9.4366	-0.0045
DA	0.5586	0.8307	0.2375	-0.0003	9.4530	0.0119
CO	0.8002	0.8279	0.2374	-0.0003	9.4609	0.0198
INV	0.5589	0.8279	0.2373	-0.0004	9.4436	0.0025
AF	0.7972	0.8247	0.2368	-0.0010	9.4233	-0.0178
R	0.7921	0.7708	0.2367	-0.0010	9.4242	-0.0169
H	0.5575	0.8184	0.2366	-0.0012	9.4347	-0.0064
ORG	0.3978	0.0873	0.2358	-0.0019	9.4188	-0.0223
DI	0.5617	0.8300	0.2341	-0.0037	9.4084	-0.0327

Nota: El screening agrega individualmente cada dimensión psicosocial sobre la especificación W2 del modelo HGB en el escenario principal E1. Las mejoras observadas deben interpretarse como aportes marginales exploratorios y no como evidencia de una selección exhaustiva de variables para la construcción de W3.

8.3. Resultados complementarios de modelos logísticos por tramo

Este anexo presenta resultados ampliados de los modelos logísticos parsimoniosos estimados por tramo educativo. Se incluye, en primer lugar, un resumen de ajuste por tramo y, en segundo lugar, la tabla completa de coeficientes, *odds ratios* y niveles de significancia para las covariables incluidas. Estas tablas complementan la discusión del cuerpo principal, donde se privilegia una presentación más sintética centrada en las variables núcleo de trayectoria y desempeño reciente.

Tabla 8.3: Resumen de ajuste de modelos logísticos parsimoniosos por tramo educativo.

Tramo	n muestra	N° columnas modelo	Pseudo- R^2	Log-verosimilitud	AIC
Temprano	200 000	11	0.0307	-9 071.6753	18 165.3506
Intermedio	200 000	12	0.0842	-23 481.5260	46 987.0520
Tardío	200 000	12	0.0733	-30 164.6486	60 353.2972

Tabla 8.4: Resultados completos de modelos logísticos parsimoniosos por tramo educativo.

Variable	Temprano			Intermedio			Tardío		
	Coef.	OR	p-valor	Coef.	OR	p-valor	Coef.	OR	p-valor
const	-4.2082	0.0149	$5,42 \times 10^{-117}$	-5.9472	0.0026	$1,65 \times 10^{-120}$	-2.6612	0.0699	$1,74 \times 10^{-26}$
ASISTENCIA_lag1	-0.0178	0.9824	$4,00 \times 10^{-21}$	-0.0104	0.9897	$6,24 \times 10^{-26}$	-0.0055	0.9945	$1,27 \times 10^{-10}$
grade_num	0.1061	1.1119	$6,53 \times 10^{-5}$	0.3700	1.4477	$1,57 \times 10^{-37}$	0.0773	1.0803	$8,46 \times 10^{-4}$
SOBREDAD	0.9379	2.5546	$< 0,001$	1.2189	3.3834	$< 0,001$	1.0600	2.8862	$< 0,001$
PROM_GRAL_lag1	-0.0867	0.9169	$7,76 \times 10^{-3}$	-0.2209	0.8018	$5,66 \times 10^{-38}$	-0.3883	0.6782	$1,25 \times 10^{-148}$
n_establecimientos_hasta_anio	0.3793	1.4612	$1,49 \times 10^{-35}$	0.2882	1.3340	$3,11 \times 10^{-106}$	0.3286	1.3890	$5,62 \times 10^{-212}$
GEN_ALU_2	-0.0939	0.9104	$4,78 \times 10^{-2}$	-0.0083	0.9917	$7,57 \times 10^{-1}$	-0.2029	0.8163	$9,60 \times 10^{-18}$
RURAL_RBD_1	-0.2991	0.7415	$1,56 \times 10^{-4}$	0.0703	1.0728	$1,76 \times 10^{-1}$	-0.1414	0.8682	$2,22 \times 10^{-2}$
COD_DEPE2_2	-0.0749	0.9279	$1,39 \times 10^{-1}$	-0.2299	0.7946	$4,71 \times 10^{-16}$	-0.1541	0.8572	$8,73 \times 10^{-10}$
COD_DEPE2_3	0.9079	2.4792	$3,11 \times 10^{-29}$	-0.0283	0.9721	$6,47 \times 10^{-1}$	-0.0824	0.9209	$1,18 \times 10^{-1}$
COD_DEPE2_4	—	—	—	-0.1262	0.8815	$6,43 \times 10^{-2}$	-0.3175	0.7279	$1,08 \times 10^{-9}$
COD_DEPE2_5	-0.1022	0.9029	$7,26 \times 10^{-1}$	-0.3757	0.6868	$2,34 \times 10^{-4}$	0.0040	1.0040	$9,53 \times 10^{-1}$

Nota: OR = *odds ratio*. Valores mayores que 1 indican asociación positiva con la no continuidad anual; valores menores que 1 indican asociación negativa. Las variables categóricas se interpretan respecto de la categoría de referencia omitida. En el tramo temprano, una categoría agrupada de COD_DEPE2 presentó separación simple y fue excluida automáticamente del ajuste, por lo que COD_DEPE2_4 no aparece estimada en ese modelo.

8.4. Resultados complementarios de modelos logísticos por cohorte

Este anexo presenta resultados ampliados de los modelos logísticos parsimoniosos estimados por cohorte teórica. Se incluye, en primer lugar, un resumen del ajuste de los modelos por cohorte y, en segundo lugar, la tabla completa de coeficientes, *odds ratios* y niveles de significancia para las covariables incluidas. Estas tablas complementan la discusión del cuerpo principal, donde se privilegia una presentación más sintética centrada en las variables núcleo y en la estabilidad de asociaciones entre cohortes.

Tabla 8.5: Resumen de ajuste de modelos logísticos parsimoniosos por cohorte teórica.

Cohorte	n muestra	Pseudo- R^2	Log-verosimilitud	AIC
2017	250 000	0.2668	-27 754.3066	55 534.6132
2021	250 000	0.2368	-20 597.9867	41 221.9733
2024	250 000	0.1979	-19 489.1282	39 004.2563

8.5. Especificación operativa de variables por ventana de información

Este anexo resume la composición de las tres ventanas de información utilizadas en la estimación de modelos predictivos. La tabla distingue entre variables incorporadas en W1, W2 y W3, su fuente de origen, tipo analítico y el criterio general de disponibilidad temporal. El objetivo es transparentar la lógica incremental del diseño metodológico y mostrar qué bloques de información se agregan a medida que aumenta la riqueza de la ventana observacional.

Tabla 8.6: Especificación operativa de variables por ventana de información.

Variable / bloque	Fuente	W1	W2	W3
grade_num	Matrícula / trayectoria	Sí	Sí	Sí
AÑO_EGRESO	Cohorte teórica	Sí	Sí	Sí
GEN_ALU	Registro administrativo	Sí	Sí	Sí
COD_REG_RBD	Establecimiento	Sí	Sí	Sí
RURAL_RBD	Establecimiento	Sí	Sí	Sí
COD_DEPE2	Establecimiento	Sí	Sí	Sí
SOBREEDAD	Trayectoria	Sí	Sí	Sí
n_establecimientos_hasta_anio	Trayectoria	Sí	Sí	Sí
HAS_SIMCE	Disponibilidad SIMCE	Sí	Sí	Sí
PROM_GRAL_lag1	Rendimiento rezagado	No	Sí	Sí
ASISTENCIA_lag1	Asistencia rezagada	No	Sí	Sí
AA	Cuestionarios SIMCE	No	No	Sí
AF	Cuestionarios SIMCE	No	No	Sí
C	Cuestionarios SIMCE	No	No	Sí
DA	Cuestionarios SIMCE	No	No	Sí
M	Cuestionarios SIMCE	No	No	Sí
V	Cuestionarios SIMCE	No	No	Sí
VIN	Cuestionarios SIMCE	No	No	Sí

Nota: W1 incorpora trayectoria y contexto administrativo observables en el año t ; W2 añade antecedentes académicos rezagados del año $t - 1$; y W3 agrega un subconjunto parsimonioso de dimensiones psicosociales y de contexto derivadas de SIMCE. La disponibilidad efectiva de estas variables depende de la posición del estudiante en la trayectoria y de la cobertura temporal de las fuentes, tal como se discute en el Capítulo 4.

8.6. Superposición entre no continuidad anual y desvinculación persistente

Como complemento al análisis del alcance conceptual de la etiqueta anual, se comparó la ocurrencia de al menos un episodio de no continuidad anual con la variable **DESERTA**, entendida como indicador de desvinculación persistente observada dentro del horizonte del panel. La comparación se realizó a nivel estudiante y permite distinguir entre estudiantes sin ninguna condición, estudiantes con solo no continuidad anual y estudiantes que presentan ambas condiciones.

Tabla 8.7: Superposición entre no continuidad anual y desvinculación persistente observada a nivel estudiante.

Grupo	N° estudiantes	% del total
Ninguna condición	561.579	78,77 %
Solo no continuidad anual	18.237	2,97 %
Ambas condiciones	130.171	18,26 %

Nota: la tabla se construye a nivel estudiante. La categoría “solo no continuidad anual” identifica estudiantes con al menos un episodio anual de no continuidad, pero sin clasificación de desvinculación persistente bajo la variable **DESERTA**. La categoría “ambas condiciones” identifica estudiantes que presentan al menos un episodio de no continuidad anual y que además son clasificados como casos de desvinculación persistente dentro del horizonte observado. La comparación no distingue la cantidad de episodios de no continuidad anual, sino únicamente si el estudiante presentó al menos uno.

Los resultados muestran que el 99,7 % de los estudiantes clasificados con desvinculación persistente presenta al menos un episodio de no continuidad anual. Esta alta superposición es esperable, dado que la desvinculación persistente implica dejar de observar matrícula en el sistema. Por ello, el interés de la comparación no está en demostrar que ambas medidas se relacionan, sino en delimitar el alcance de la etiqueta anual.

En particular, se observa que 18.237 estudiantes presentan no continuidad anual sin quedar clasificados como casos de desvinculación persistente. Este grupo representa el 12,29 % de quienes presentan al menos un episodio de no continuidad anual, lo que confirma que la etiqueta anual también captura interrupciones reversibles o trayectorias con retorno posterior al sistema. Desde una perspectiva preventiva, este subconjunto es relevante porque permite observar quiebres del vínculo escolar antes de que necesariamente se consoliden como una desvinculación persistente.