Predicción del bajo rendimiento académico en ámbitos rurales de Bolivia mediante modelos de aprendizaje automático

Prediction of poor academic performance in rural areas of Bolivia using machine learning models

Limberg Villca Coraite[1]

Correo de correspondencia: limbervillcacoraite@gmail.com

Resumen

En Bolivia, la desigualdad educativa entre zonas urbanas y rurales continúa siendo un desafío estructural. Este estudio analiza el rendimiento académico en una unidad educativa rural del departamento de Santa Cruz, utilizando registros de calificaciones entre los años 2015 y 2024. El análisis evidenció una concentración significativa de reprobaciones en tres materias clave: Biogeografía/Ciencias Naturales (17.24%), Matemáticas (16.55%) y Comunicación y Lenguajes (13.79%), que en conjunto agrupan el 47.58% de los casos. Estos resultados reflejan debilidades persistentes en competencias científicas y comunicativas.

Con el objetivo de anticipar casos de reprobación, se entrenaron distintos modelos de aprendizaje automático supervisado. Entre ellos, el algoritmo CatBoost alcanzó el mejor desempeño, con un F1-score ponderado de 0.84 y una precisión del 91%, superando a modelos como XGBoost, Random Forest, SVM y redes neuronales. La implementación del modelo permitió generar listas predictivas de estudiantes en riesgo para los siguientes cinco años, lo cual brinda al equipo docente una herramienta operativa para diseñar intervenciones pedagógicas tempranas y focalizadas. Los resultados muestran el potencial del uso de inteligencia artificial en contextos educativos rurales, aportando soluciones basadas en datos para mejorar la equidad y la calidad del aprendizaje.

***Palabras clave*:** analítica de datos, educación rural, F1-score ponderado, modelos predictivos, generación de predicción.

Abstract

In Bolivia, educational inequality between urban and rural areas continues to be a structural challenge. This study analyzes academic performance in a rural educational unit in Sant Cruz city, using grade records between 2015 and 2024. The analysis showed a significant concentration of failures in three key subjects: Biogeography/Natural Sciences (17.24%), Mathematics (16.55%) and Communication and Languages (13.79%), which together account for 47.58% of the cases. These results reflect persistent weaknesses in scientific and communicative competencies.

In order to anticipate failure cases, different supervised machine learning models were trained. Among them, the CatBoost algorithm achieved the best performance, with a weighted F1-score of 0.84 and an accuracy of 91%, surpassing models such as XGBoost, Random Forest, SVM and neural networks. The implementation of the model allowed generating predictive lists of at-risk students for the next five years, providing teachers with an operational tool for designing early and targeted educational interventions. The results demonstrate the potential of using artificial intelligence in rural educational contexts, contributing data-driven solutions to improve equity and the quality of learning.

Keywords: data analytics, rural education, F1-score weighted, predictive models, prediction generation.

1. Introducción

Bolivia enfrenta desafíos estructurales en el ámbito educativo, siendo uno de los más persistentes la marcada disparidad entre la calidad de la educación en zonas urbanas y rurales (Fernandez, 2024). Esta brecha no solo limita

[1] Carrera de Ingeniería de Sistemas, Facultad de Ciencias y Tecnología, Universidad Mayor de San Simón, Cochabamba, Bolivia. https://orcid.org/0009-0006-1428-8828

las oportunidades de aprendizaje, sino que profundiza desigualdades sociales y económicas en el largo plazo. La UNESCO ha señalado que las áreas rurales de América Latina enfrentan condiciones más adversas en términos de acceso, permanencia y calidad educativa, por lo que destaca la necesidad de incorporar herramientas tecnológicas que contribuyan a reducir la brecha de aprendizaje (UNESCO, 2023).

En los últimos años, múltiples diagnósticos advierten sobre el deterioro de la calidad educativa en Bolivia (Amonzabel, 2025). Esto se refleja en los bajos niveles de preparación con los que muchos estudiantes ingresan al nivel universitario, especialmente en las áreas científicas y lógico-matemáticas (Roca, 2025),.

A nivel nacional, estudios recientes han revelado que solo 2 de cada 100 estudiantes lograron aprobar pruebas diagnósticas estandarizadas en 2023 (Trigo, 2025), y que más del 79% de los bachilleres presenta un dominio insuficiente del lenguaje escrito (Santo Bacallao, 2020).

Pese a la existencia de estudios que abordan el bajo rendimiento académico, la mayoría se concentran en contextos urbanos y analizan el fenómeno desde una perspectiva agregada. Predominan los enfoques que utilizan datos estadísticos oficiales para estimar tasas generales de abandono o reprobación, sin llegar a identificar a los estudiantes individuales que presentan mayor riesgo de repetir curso (Laime, 2024).

Existen diversos estudios que han aplicado técnicas de minería de datos para predecir el fracaso escolar en contextos urbanos, destacándose el trabajo doctoral de Márquez Vera (Marquez Vera, Predicción del fracaso y abandono escolar mediante técnica de minería de datos, 2015) y sus contribuciones publicadas en revistas especializadas como IEEE (Marquez Vera, Romero Morales, & Ventura Soto, Predicción del Fracaso Escolar mediante, 2012), centrando su estudio en modelos de clasificación con fines de alerta temprana.

A diferencia de estos enfoques, el presente estudio adopta una perspectiva personalizada. Se propone no solo estimar la proporción de estudiantes en riesgo, sino también identificar específicamente quiénes son esos estudiantes y en qué materias ya han presentado antecedentes de reprobación. Esta aproximación permite anticipar casos críticos y tomar decisiones pedagógicas tempranas y focalizadas, con el fin de prevenir la repetición del año escolar y mitigar los efectos acumulativos del bajo rendimiento académico.

Por motivos de privacidad y viabilidad operativa, el análisis se enfocó en una sola unidad educativa: la Unidad Educativa San José Obrero, ubicada en una zona rural del departamento de Santa Cruz, Bolivia. La base de datos utilizada incluye registros académicos anonimizados de estudiantes de nivel primario y secundario, correspondientes al periodo comprendido entre 2015 y 2024, incluyendo información personal de los estudiantes, así como también datos sobre sus calificaciones por materia, historial de permanencia y evolución del desempeño académico.

Objetivo

Analizar el rendimiento académico de los estudiantes de las zonas rurales de Bolivia, identificando materias que tienen mayor cantidad de reprobados y desarrollando un modelo de predicción basado en técnicas de aprendizaje supervisado para anticipar casos de estudiantes con riesgos de reprobación.

2. Material y métodos

Este estudio adopta un enfoque cuantitativo, de tipo descriptivo y predictivo, cuyo propósito es analizar el rendimiento académico de estudiantes en un contexto rural de Bolivia y construir un modelo que permita anticipar posibles casos de reprobación. Esta investigación se desarrolló utilizando datos provenientes de la Unidad Educativa San José Obrero. Esta unidad fue seleccionada por tratarse de un establecimiento educativo rural con registros completos y continuos de varias gestiones.

Conjunto de datos

Se trabajó con datos académicos correspondientes a estudiantes de los niveles primario y secundario de educación regular, comprendidos entre las gestiones 2015 y 2024. La base de datos inicial estaba compuesta por veinte archivos independientes (10 del nivel primario y 10 del nivel secundario), los cuales fueron consolidados en un único conjunto estructurado en formato Excel. Este conjunto de datos incluyó variables personales y calificaciones por gestión y materia, promedio anual general y datos de permanencia institucional del/la estudiante. La información recopilada permitió construir una matriz de datos apta para su posterior análisis exploratorio y modelado predictivo.

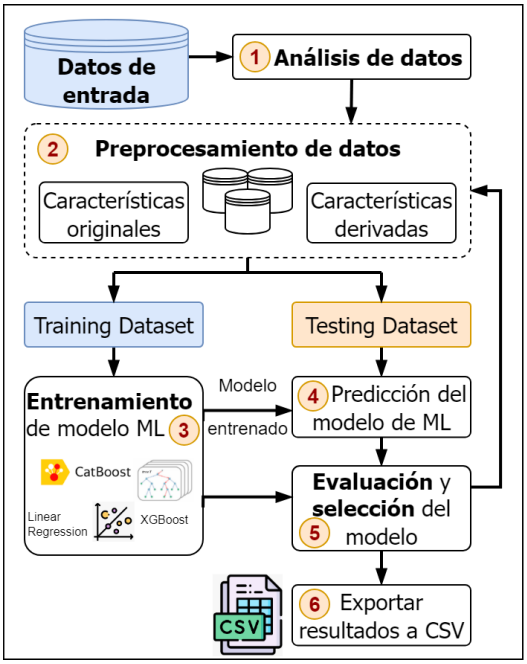
3. Metodología

Este estudio sigue la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), ampliamente reconocida por ofrecer una estructura clara para el desarrollo de proyectos analíticos basados en datos. Sin embargo, considerando las particularidades de este proyecto, se optó por desarrollar un flujograma metodológico adaptado a las necesidades específicas del presente estudio a partir de la metodología ya mencionada.

El diagrama de la Figura 1 resume el flujo lógico de trabajo seguido, facilitando su comprensión tanto para especialistas en el área como para responsables de la toma de decisiones en el ámbito educativo.

Figura 1

Flujograma metodológico propuesto, inspirado en la metodología CRISP-DM



Fuente: Elaboración propia (2025)

Análisis de datos

Una vez consolidado el conjunto de datos, se realizó un análisis exploratorio con el propósito de identificar patrones relevantes en el rendimiento académico de los estudiantes. Este análisis permitió observar tendencias generales y específicas que orientaron el posterior diseño del modelo predictivo.
Entre los principales hallazgos se destacan los siguientes:

Tendencia descendente en el promedio general: Se evidenció una disminución sostenida del promedio anual en la unidad educativa analizada, pasando de 74.24 puntos de promedio general en el año 2015 a 69.91 puntos en 2024 como unidad educativa. Este descenso sugiere un deterioro progresivo del rendimiento académico a lo largo del tiempo (ver Figura 2)

Figura 2

Evolución del promedio por año

Fuente: Elaboración propia (2025)

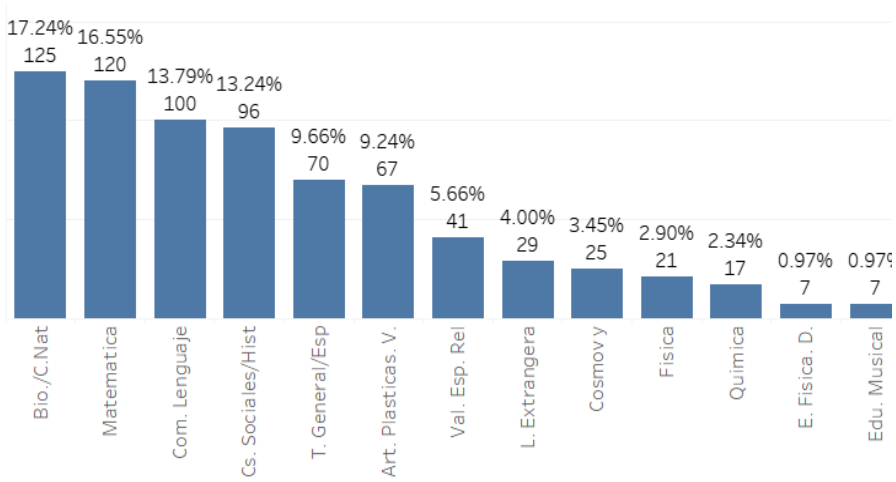
Desempeño por género: Al comparar los promedios por género, se observó que las estudiantes mujeres presentan, en promedio, mejores calificaciones que sus pares varones. El promedio general femenino fue de 73.73 puntos, superando al promedio masculino, lo que refleja una diferencia consistente en el rendimiento académico por género.

Materias con mayor índice de reprobación: El análisis identificó tres áreas con mayor concentración de estudiantes reprobados: Biogeografía/Ciencias Naturales (17.24%), Matemáticas (16.55%) y Comunicación y Lenguajes (13.79%). En conjunto, estas materias agrupan el 47.58% del total de reprobaciones, evidenciando debilidades marcadas tanto en competencias científicas como comunicativas. (ver Figura 3).

Aumento de la reprobación anual: Se constató un incremento sostenido en el número de estudiantes reprobados por gestión. Mientras que en 2015 solo se registraron 6 casos, para el año 2024 la cifra ascendió a 31. Cabe destacar que el año 2020 no presenta registros de reprobación debido a las disposiciones excepcionales del Ministerio de Educación durante la pandemia de COVID-19 (ver Figura 4)

Figura 3

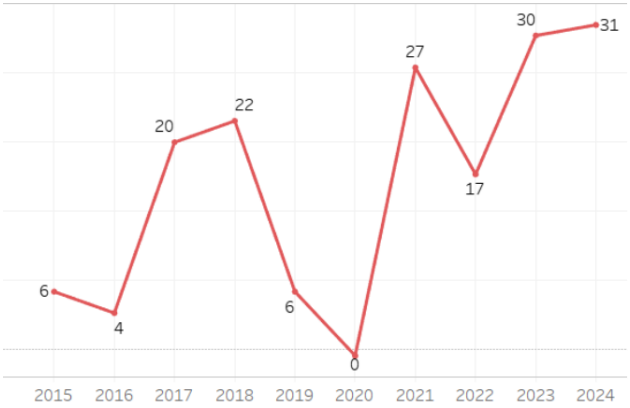
Materias con mayor índice de reprobación



Fuente: Elaboración propia (2025)

Figura 4

Cantidad de reprobados por año



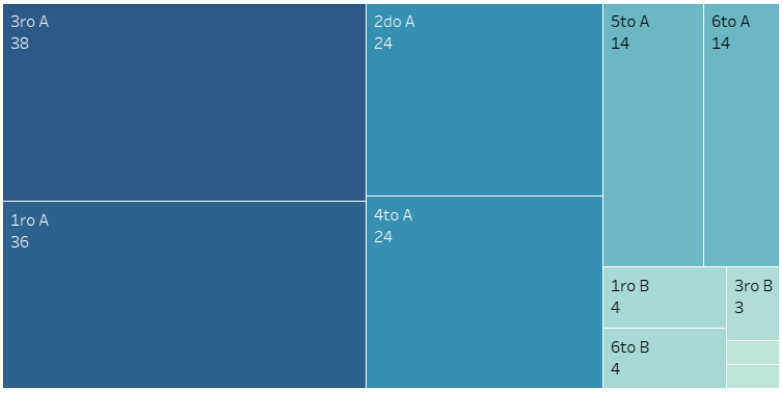
Fuente: Elaboración propia (2025)

Distribución por curso: El curso con mayor número acumulado de reprobaciones fue Tercero de Secundaria “A” (3ro A), con un total de 38 estudiantes reprobados entre 2015 y 2024. Este dato sugiere que existen desafíos estructurales específicos en ese nivel o grupo (ver Figura 5).

Comparación con el promedio nacional: La unidad educativa seleccionada presenta una tasa de reprobación global del 5.51%, superando la media nacional reportada para el año 2023, que fue de 3.61% (Ministerio de educación Bolivia, 2023). Esta diferencia refuerza la necesidad de intervenciones focalizadas en el ámbito rural.

Figura 5

Cantidad de reprobados por curso

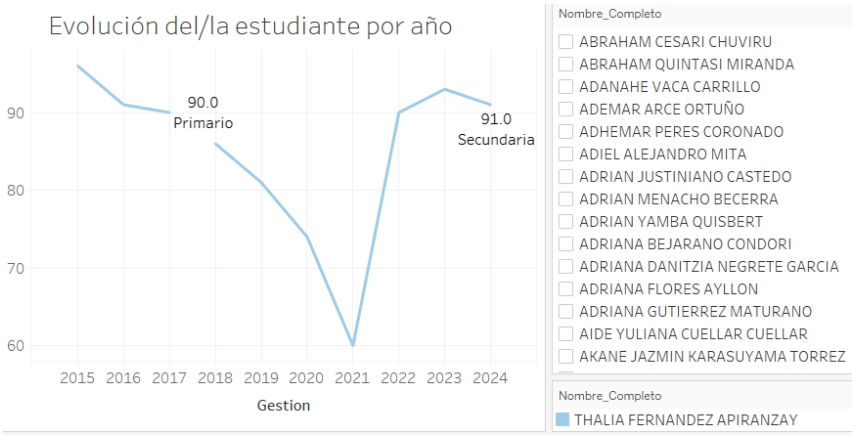


Fuente: Elaboración propia (2025)

Evolución individual del rendimiento estudiantil: Además del análisis agregado, se evaluó la evolución del promedio académico de cada estudiante a lo largo de su trayectoria en la unidad educativa. Este enfoque permitió identificar casos de mejoría, estancamiento o deterioro en el rendimiento, información que luego fue traducida en variables como la tendencia del promedio, el delta de variación interanual y la variabilidad histórica. La Figura 6 muestra algunos ejemplos de estas trayectorias individuales.

Figura 6

Evolución de estudiantes por gestión educativa



Fuente: Elaboración propia (2025)

Estos hallazgos permitieron delimitar con claridad las áreas críticas del rendimiento académico y sentaron las bases para el diseño de un modelo de predicción que no solo cuantifique el riesgo de reprobación, sino que también identifique a los estudiantes más vulnerables con suficiente anticipación para prevenir esta situación con acción temprana.

Preprocesamiento de datos

El preprocesamiento representa una de las fases más críticas en el desarrollo de modelos predictivos, ya que sienta las bases sobre las cuales se construyen las predicciones. En este estudio, se prestó especial atención a la limpieza, transformación y estructuración temporal de los datos, asegurando su calidad y adecuación al objetivo planteado.

Limpieza y estructuración inicial: Con el conjunto de datos consolidado, se procedió a un riguroso proceso de depuración y transformación para garantizar su calidad analítica. La limpieza incluyó la eliminación de valores nulos y registros duplicados, así como la conversión y normalización de tipos de datos, asegurando la coherencia entre columnas numéricas, categóricas y temporales. Se realizó también una segmentación estructurada entre variables de carácter personal y académico, lo que facilitó su posterior manejo.

Evaluación de la calidad del dato: Una vez depurada la información, se evaluaron aspectos clave como la completitud, consistencia interna y validez semántica de las variables. Esta revisión permitió identificar y excluir columnas con baja densidad informativa o inconsistencias evidentes. Como resultado, el conjunto de datos final del trabajo quedó conformado por 2.958 registros útiles y 31 variables relevantes que se usaron para la parte analítica y predictiva.

División temporal para validación: Para asegurar una validación sólida del modelo predictivo, se aplicó una estrategia de división temporal. Todas las gestiones anteriores a 2024 fueron utilizadas para entrenamiento, mientras que los datos correspondientes a la gestión 2024 se reservaron exclusivamente para validación. Este enfoque simula de forma más realista la aplicación del modelo sobre cohortes futuras, sin riesgo de fuga de información, se contempló también el caso de estudiantes nuevos en la unidad educativa (o con permanencia de una gestión), para este tipo de casos se tomó una característica en donde 1 identifica a estudiante nuevo y 0 en otro caso.

Preparación de variables de entrada: A partir del historial académico de cada estudiante, se construyeron variables derivadas que capturan dimensiones relevantes del desempeño individual. Entre estas se encuentran el promedio global, el promedio de los últimos tres años, la tendencia del rendimiento, la variabilidad interanual, la cantidad acumulada de reprobaciones, los años de permanencia institucional, la condición de ingreso, el delta intergestión del promedio y la ratio personal de reprobación.

Transformaciones para el modelado: Con el fin de adaptar los datos a los algoritmos de clasificación, las variables categóricas fueron transformadas mediante codificación One Hot Encoding. Posteriormente, se aplicaron técnicas de escalado y estandarización a todas las variables numéricas, asegurando la uniformidad de sus rangos. Finalmente, se definió la variable objetivo como binaria, donde el valor 1 indica que el estudiante reprobó en la gestión 2024, y 0 en caso contrario.

Este proceso de preparación no solo permitió estructurar un conjunto de datos consistente y libre de errores, sino que habilitó el desarrollo de modelos de clasificación sobre una base sólida y de calidad. A partir de esta matriz, fue posible avanzar hacia el entrenamiento de distintos algoritmos supervisados, buscando identificar aquellos con mejor capacidad de predicción y generalización para obtener resultados con buena precisión.

Entrenamiento del modelo predictivo

El objetivo central de esta fase fue desarrollar una herramienta capaz de anticipar la probabilidad de que un estudiante repruebe en los siguientes 5 años (configurable a más años), a partir de su historial académico. Para lograrlo, se entrenaron modelos de clasificación binaria, una técnica de aprendizaje automático supervisado.

¿Qué es un modelo supervisado?

Los modelos supervisados aprenden a predecir un resultado específico (llamado variable objetivo) a partir de datos de entrada (variables predictoras), utilizando ejemplos previamente conocidos (IBM, 2024). En este caso, el modelo aprende a clasificar si un estudiante reprobará (1) o no (0), basándose en sus registros históricos a lo largo de 10 gestiones académicas. Esta técnica es ampliamente usada en contextos educativos, médicos y financieros, donde es crucial anticipar eventos adversos para actuar de forma temprana (Marquez Vera, Romero Morales, & Ventura Soto, Predicción del Fracaso Escolar mediante, 2012).

Modelos evaluados

Para asegurar la robustez del análisis, se compararon distintos algoritmos que representan enfoques diversos del aprendizaje automático. A continuación, se describen brevemente los modelos utilizados:

Regresión logística: modelo lineal simple que estima la probabilidad de un evento utilizando una función sigmoide. Es útil como línea base por su interpretabilidad, aunque limitado frente a relaciones complejas (Hosmer, Lemeshow, & Sturdivant, 2013).

Random Forest: conjunto de árboles de decisión entrenados sobre subconjuntos aleatorios del dato. Tiende a ser robusto y menos propenso al sobreajuste que un árbol individual (Breiman, 2001).

Gradient Boosting y sus variantes (XGBoost, LightGBM y CatBoost): algoritmos que construyen secuencialmente árboles, corrigiendo los errores del modelo anterior. XGBoost incluye regularización para evitar sobreajuste (Chen & Guestrin, 2016), LightGBM se enfoca en eficiencia computacional (Ke, y otros, 2017), y CatBoost es especialmente eficaz con variables categóricas (Prokhorenkova, Vorobev, Dorogush, Gulin, & Gusev, 2018).

SVM (Support Vector Machine): modelo que encuentra un límite (hiperplano) óptimo para separar clases, maximizando la distancia entre los grupos. Es efectivo para espacios de alta dimensión, pero sensible al escalado y menos eficiente con grandes volúmenes de datos (Cortes & Vapnik, 1995).

MLP (Perceptrón Multicapa): tipo de red neuronal con capas intermedias que permiten aprender relaciones no lineales. Es flexible, pero puede requerir mucho ajuste de parámetros y más datos (Friedman, 2001).

Estas alternativas permitieron evaluar modelos tanto simples como complejos, lineales y no lineales, deterministas y estocásticos, brindando una visión integral del problema.

Validación y métrica de evaluación

El conjunto de datos se dividió de forma temporal: las gestiones académicas hasta 2023 se utilizaron para entrenamiento, mientras que 2024 se reservó para validación. Esta estrategia simula el uso real del modelo, evitando que aprenda sobre casos que debería predecir.

Además, se aplicó validación cruzada, una técnica que consiste en entrenar y probar el modelo múltiples veces sobre diferentes particiones del mismo conjunto de datos. Esto permite obtener un estimado más confiable de su rendimiento (Kohavi, 1995).

Para comparar los modelos, se utilizó el F1-score ponderado como métrica principal. A diferencia del accuracy (que solo mide cuántas veces acierta), el F1-score equilibra precisión y exhaustividad, lo que lo hace más adecuado

cuando las clases están desbalanceadas (Chicco & Giuseppe, 2020). En este caso, había menos casos de reprobación que de aprobación, lo cual hacía necesario un criterio más riguroso.

Contenido generado por IA

Herramienta: ChatGpt: Generó el contenido del Abstract y keywords a partir de original en español (en inglés).
Se usó también Claude para la revisión y mejora de la redacción de todo el documento en conjunto con Chatgpt a partir de la redacción original y gráficos de la investigación.

4. Resultados y Discusión

El análisis del rendimiento académico permitió identificar patrones persistentes de reprobación en la unidad educativa estudiada, así como también validar la utilidad del modelo predictivo desarrollado para anticipar estos casos. Además, los resultados obtenidos fueron contrastados con estudios previos, lo que permitió dimensionar el aporte del presente trabajo frente a enfoques existentes. Los resultados se presentan a continuación:

Identificación de áreas críticas de reprobación

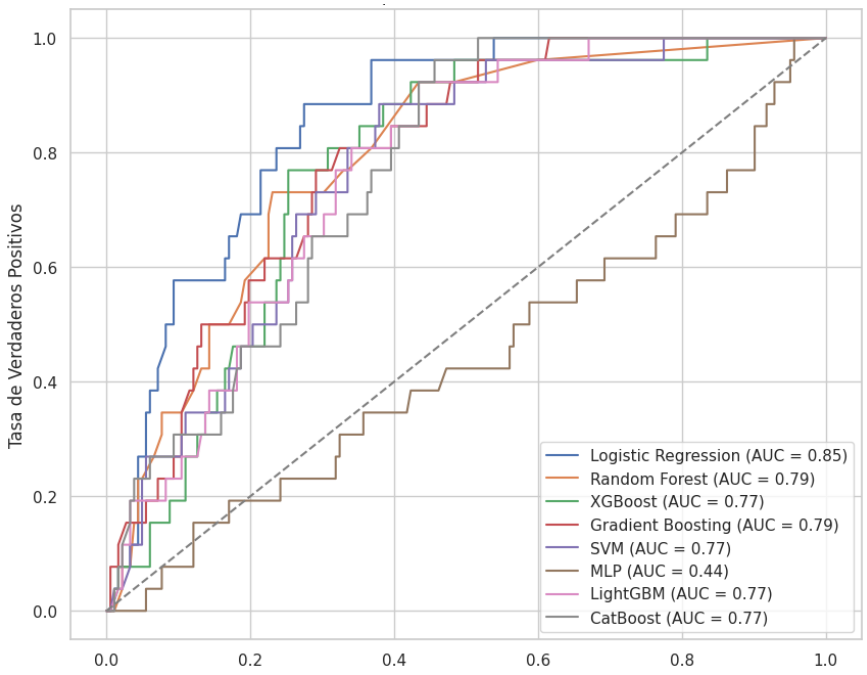
Los datos mostraron una disminución sostenida en el promedio general de la unidad educativa entre los años 2015 y 2024, pasando de 74.24 a 69.91 puntos. Esta tendencia descendente se observó tanto a nivel institucional como individual, donde se evidenció que un número significativo de estudiantes presentaba una caída progresiva en su desempeño a lo largo del tiempo. Este patrón fue especialmente visible en cursos intermedios del nivel secundario, como 3ro de secundaria "A", que concentró la mayor cantidad de estudiantes reprobados en la unidad educativa.
En términos de asignaturas, el análisis permitió identificar tres materias con tasas particularmente altas de reprobación: Biogeografía/Ciencias Naturales (17.24%), Matemáticas (16.55%) y Comunicación y Lenguajes (13.79%). En conjunto, estas tres áreas agruparon el 47.58% del total de reprobaciones, lo cual sugiere la existencia de brechas estructurales tanto en competencias científicas como en habilidades comunicativas.
La concentración de reprobaciones en asignaturas como Matemáticas y Ciencias Naturales adquiere especial relevancia si se considera su relación directa con la continuidad de estudios superiores en áreas científicas y tecnológicas. En muchos casos, las bases que los estudiantes desarrollan durante la educación secundaria en estas materias influyen significativamente en su capacidad para acceder y adaptarse a carreras universitarias vinculadas a las áreas STEM (Science, Technology, Engineering and Mathematics). En contextos rurales, donde las oportunidades educativas suelen ser más limitadas, las debilidades acumuladas en estas áreas pueden reducir las posibilidades de que los estudiantes opten por trayectorias académicas en campos científicos o tecnológicos. En este sentido, la identificación temprana de dificultades en asignaturas clave podría permitir implementar estrategias de apoyo académico antes de que los estudiantes concluyan la educación secundaria, contribuyendo a fortalecer sus competencias y ampliar sus oportunidades de acceso a la educación superior. Estos hallazgos coinciden con estudios previos sobre el bajo dominio del lenguaje y razonamiento lógico en estudiantes egresados de secundaria en Bolivia (Santo Bacallao, 2020).
Asimismo, se evidenció un crecimiento progresivo en el número de reprobaciones por año, pasando de solo seis casos en 2015 a más de treinta en 2024, lo que refuerza la hipótesis de un deterioro acumulativo en el rendimiento académico. Este panorama es especialmente preocupante si se considera que la tasa de reprobación de esta unidad educativa (5.51%) supera la media nacional reportada para el año 2023 (3.61%) por el Ministerio de Educación.

Evaluación del modelo predictivo

Con base en los hallazgos anteriores, se desarrollaron modelos predictivos para anticipar casos de reprobación. Se entrenaron y compararon ocho algoritmos de clasificación, representativos de diferentes enfoques del aprendizaje automático supervisado: modelos lineales, de ensamblado, redes neuronales y SVMs como se ve en la Figura 7. Todo el conjunto de datos, proceso y resultados se encuentran disponibles en el repositorio de GitHub agregado en la bibliografía de este artículo (Villca Coraite, 2025).

Figura 7

Cuadro comparativo de entrenamiento de modelos



Los resultados comparativos de los modelos se presentan en la Tabla 1 conjuntamente con sus respectivas métricas.

Tabla 1

Cuadro comparativo de resultado de los modelos

Modelo	Accuracy	F1-Score ponderado	Precision	Recall
Regresión logistica	0,7019	0,7521	0,8892	0,7019
Random forest	0,8702	0,8143	0,7651	0,8702
XGBoost	0,8077	0,8077	0,8077	0,8077
Gradient boosting	0,8702	0,8298	0,8216	0,8702
SVM	0,8077	0,816	0,8256	0,8077
MLP	0,8654	0,8119	0,7646	0,8654
LightGBM	0,8317	0,8239	0,8171	0,8317
CatBoost	0,851	0,8408	0,8331	0,851

Fuente: Elaboración propia (2025)

El algoritmo CatBoost fue el que alcanzó el mejor rendimiento general, destacando especialmente en la métrica F1-score ponderado (0.8408) y en recall (0.8510), lo que indica una alta capacidad para detectar correctamente a los estudiantes que efectivamente reprobaron. Su precisión (0.8331) y su exactitud global (85.1%) también fueron superiores a la mayoría de los otros modelos evaluados.

Con el propósito de comprender mejor cómo el modelo realiza sus predicciones, se analizó la importancia relativa de las variables utilizadas por el algoritmo CatBoost. Este análisis permite identificar cuáles de las características del historial académico influyen con mayor peso en la estimación del riesgo de reprobación. Los resultados muestran que variables relacionadas con la trayectoria académica del estudiante como la tendencia del promedio general, el número acumulado de reprobaciones y el promedio de los últimos años tienen una influencia particularmente alta en la predicción del modelo. Asimismo, factores como los años de permanencia en la unidad educativa y la variación interanual del rendimiento también aportan información relevante para estimar el riesgo académico.

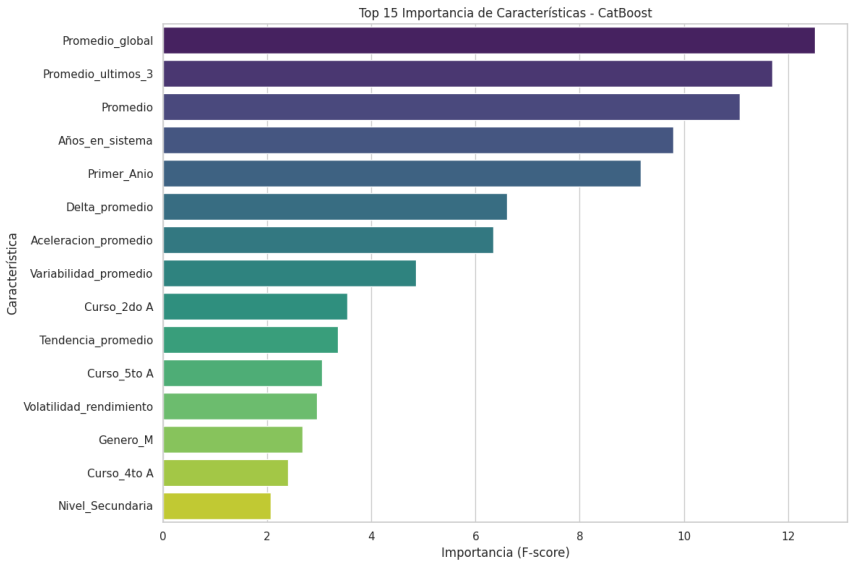
La Figura 8 presenta el ranking de importancia de variables obtenido a partir del modelo entrenado. Este tipo de análisis resulta útil no solo para validar el comportamiento del algoritmo, sino también para ofrecer una interpretación más clara de los factores que están asociados al bajo rendimiento escolar en el contexto analizado.

Estos resultados validan el uso del algoritmo CatBoost, en contextos educativos con datos tabulares y categóricos. Además, la selección de F1-score ponderado como métrica principal resultó adecuada dada la naturaleza desbalanceada del conjunto de datos, donde los casos de reprobación son minoritarios pero críticos (Chicco & Giuseppe, 2020).

El modelo entrenado permitió proyectar el número estimado de estudiantes reprobados para los próximos cinco años, tal como se ilustra en la Figura 9. Además, posibilitó la identificación individual de los estudiantes con mayor riesgo de reprobación (Figura 10). Estas predicciones fueron organizadas en listas anuales y exportadas en un formato accesible para docentes y personal administrativo, con el objetivo de facilitar su uso como herramienta de alerta temprana y apoyar la toma de decisiones pedagógicas orientadas a la prevención.

Figura 8

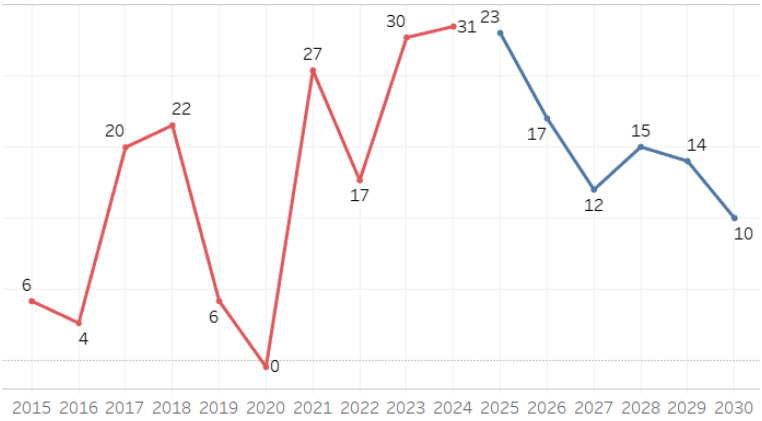
Importancia de variables del modelo CatBoost



Fuente: Elaboración propia (2026)

Figura 9

Cantidad de estudiantes reprobados por año (Rojo real y azul predicción)



Fuente: Elaboración propia (2025)

Figura 10

Estudiantes con mayor probabilidad de reprobar en los siguientes años

Estudiantes con peligro de reprobar

Codigo Rude (pedicc..	Nombre_Completo	
519500012013922	YOLVER CRESPO RAMOS	2,025
519500012016007	MOISES CHOQUE ROCA	2,025
519500012016009	FAVIO ANDRES SURUBI VACA	2,025
519500012017008	RUBEN SOLIZ PEÑA	2,025
519500012017009	GUILLERMO VERA TELMO	2,025
519500012018010	YANETH YAMBAMINI SOLIZ	2,025
519500012020017	DILAN PERES ROCA	2,025
519500012020024	JAZMIN ALBA CORTEZ	2,025
519500012021961A	EZEQUIEL CLAURE CUELLAR	2,025
519500012024471A	VALENTINA SIANCAS CAMACHO	2,025
719500632013951	YESSICA CHAMO ESTRADA	2,025
719500632014275	ADRIANA BEJARANO CONDORI	2,025
719500632014457	BEATRIZ CAMPOS CUELLAR	2,025
719500632014990	JOSE EDUARDO PEDRAZA PEDR..	2,025
719500632016011	HIROSHI ISHIZAKI ORELLANO	2,025
719500632017028	DAYANA FIORELLA ARAUZ APAR	2,025

Fuente: Elaboración propia (2025)

En conjunto, los resultados muestran que es posible anticipar con alta precisión los casos de bajo rendimiento académico en contextos rurales, utilizando técnicas de aprendizaje supervisado sobre datos históricos de calificaciones. Esta capacidad predictiva, cuando es aprovechada de forma adecuada, puede convertirse en un insumo valioso para diseñar estrategias focalizadas de acompañamiento escolar y mejorar la calidad educativa en regiones con limitaciones estructurales.

Comparación con estudios previos

Los resultados obtenidos en este estudio pueden ser mejor comprendidos al contrastarlos con investigaciones previas que abordaron problemáticas similares desde un enfoque predictivo y analítico.

Uno de los referentes más citados en la literatura es el trabajo de Carlos Márquez Vera (2015), quien aplicó técnicas de minería de datos para predecir el abandono y fracaso escolar en contextos urbanos de España. Su modelo logró una precisión del 91% utilizando el algoritmo ICRM (Incremental Cost-sensitive Rule Mining) con un conjunto de datos acotado a una gestión anual. Aunque su enfoque fue eficaz para establecer patrones generales de fracaso escolar, se centró principalmente en el nivel agregado de tasas de abandono, sin identificar individualmente a los estudiantes en riesgo (Marquez Vera, Predicción del fracaso y abandono escolar mediante técnica de minería de datos, 2015).

En el entorno local, el trabajo de Maya Wara López Laime (2024) analizó el rendimiento académico en municipios de Bolivia a través de agrupamiento y aprendizaje automático. Su modelo basado en Random Forest obtuvo una precisión de 58%, centrando su análisis en indicadores globales sin profundizar en predicciones personalizadas por estudiante. Esta investigación utilizó menos variables contextuales para la parte predictiva (Laime, 2024).

En comparación, el presente estudio no solo alcanzó una mayor precisión global (91%), sino que también incorporó una perspectiva individualizada del riesgo, generando listas de estudiantes y por gestión futura en la que muy probablemente este repruebe según indica la predicción. La Tabla 2 resume los resultados comparativos con otras investigaciones similares en el tema educativo.

Tabla 2

Comparativa con otras investigaciones

Comparación	Esta investigación	Carlos Márquez	Maya Wara López
Mejor modelo	CatBoost	ICRM	Random forest
Accuracy	0,9182	0,91	0,58
Tasa de reprobación	0,0551	0,09	0,0702
Gestiones de datos usadas	2015-2024	2009	2006-2019
Aumento de datos el último año	11	0	-576
Cantidad de filas	2978	670	4746

Fuente: Elaboración propia (2025)

Estas diferencias metodológicas refuerzan el valor añadido del presente trabajo: mientras los estudios anteriores se enfocan en identificar patrones generales, este estudio ofrece una herramienta operativa que permite anticipar de forma concreta qué estudiantes podrían reprobar y en qué momento actuar. Esta capacidad resulta especialmente relevante en contextos rurales, donde las intervenciones deben ser altamente focalizadas debido a las limitaciones estructurales y de recursos.

En conjunto, los hallazgos aquí presentados no solo coinciden con la literatura en cuanto a la relevancia del uso de técnicas de aprendizaje supervisado, sino que también avanzan un paso más al integrar predicción personalizada, validación temporal, y una propuesta concreta de aplicación pedagógica a nivel institucional.

5. Conclusión

Este estudio logró cumplir su objetivo de analizar el rendimiento académico en un contexto rural boliviano, identificando materias con alta tasa de reprobación e implementando un modelo de predicción que permite anticipar casos de estudiantes en riesgo. A partir de un conjunto de datos consolidada y preprocesada con criterios de calidad, se identificó una tendencia sostenida de deterioro en los promedios académicos entre 2015 y 2024, con especial concentración de reprobaciones en materias como Biogeografía/Ciencias Naturales, Matemáticas y Comunicación y Lenguajes.

La aplicación de técnicas de aprendizaje automático supervisado, específicamente el modelo CatBoost, permitió generar predicciones individualizadas con alta precisión (F1-score ponderado de 0.84 y accuracy del 91%), superando enfoques previos centrados únicamente en tasas globales de abandono o reprobación. Este modelo permitió identificar variables clave como la tendencia del promedio, el historial de reprobaciones y los años que el estudiante prevaleció en la unidad educativa, lo que facilita una comprensión más completa del desempeño académico individual.

Al comparar los resultados con investigaciones anteriores, se destaca que este trabajo avanza en la capacidad de predicción personalizada, aportando una herramienta concreta para la toma de decisiones pedagógicas en contextos de alta vulnerabilidad educativa. La posibilidad de generar listas anticipadas de estudiantes en riesgo representa una innovación útil para la gestión educativa rural, donde los recursos son limitados y las intervenciones deben ser focalizadas cuanto antes.

Como recomendación, se sugiere replicar este enfoque en otras unidades educativas del país rurales y urbanas, incorporando además variables socioeconómicas, emocionales y contextuales que puedan enriquecer la capacidad predictiva del modelo. Asimismo, la participación de equipos multidisciplinarios, como psicopedagogos o especialistas en gestión educativa, podría potenciar la aplicación práctica de los resultados, garantizando no solo la identificación de riesgos, sino también la implementación de estrategias de acompañamiento efectivas.

Aunque el análisis se desarrolló a partir de los datos de una sola unidad educativa rural, el enfoque metodológico empleado presenta el potencial de ser aplicado en otros contextos educativos del país. El modelo se basa principalmente en registros académicos históricos, información que generalmente está disponible en las instituciones educativas, lo que facilita su posible implementación en otras regiones.

No obstante, Bolivia presenta una gran diversidad de realidades socioeducativas entre departamentos, municipios y comunidades rurales. Por esta razón, la replicación del modelo en otros contextos requeriría procesos de validación y ajuste que consideren particularidades locales, como diferencias en estructura curricular, condiciones socioeconómicas o dinámicas institucionales. Evaluar el desempeño del modelo en distintos entornos permitiría determinar su capacidad de generalización y fortalecer su utilidad como herramienta de apoyo para la toma de decisiones educativas a mayor escala.

Agradecimientos

Agradezco a Ing. Evelyn Cusi e Ing. Ariel Mamani por guiarme y asesorarme para que este artículo se lleve a cabo, a mi querida tía Delia Villca por siempre apoyarme y animarme en los aspectos académicos, a mi madre y mis hermanas Zunilda y Luz Darley por darme ese apoyo en los momentos difíciles.

Dedico este artículo al P. Hermann Nigris SDB (+), gracias a sus consejos y ejemplo de vida como misionero salesiano que trabajó incansablemente por ayudar a los demás.

6. Referencias bibliográficas

Amonzabel, M. A. (5 de marzo de 2025). YouTube. Obtenido de youtube.com: https://www.youtube.com/watch?v=gImhqTufEMQ

Breiman, L. (2001). Random forest. Machine Learning. Springer nature link, 5-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD, 785-794.

Chicco, D., & Giuseppe, J. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 6.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning. Springer nature link, 273-297.

Fernandez, M. (20 de febrero de 2024). La educación rural en Bolivia: desafíos y oportunidades. El país.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. (2013). Applied Logistic Regression. Hoboken, New Jersey: Wiley.

IBM. (28 de diciembre de 2024). ibm.com. Obtenido de www.ibm.com: https://www.ibm.com/es-es/think/topics/supervised-learning

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NeurIPS, 3146-3154.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai , 1137-1145.

Laime, M. W. (2024). ddigital. Obtenido de ddigital.umss.edu: http://ddigital.umss.edu/bitstream/123456789/43657/1/MONOGRAFIA_LOPEZ%20LAIME%20MAYA%20WARA.pdf

Marquez Vera, C. (2015). Predicción del fracaso y abandono escolar mediante técnica de minería de datos. Córdoba: Servicio de Publicaciones de la Universidad de Córdoba.

Marquez Vera, C., Romero Morales, C., & Ventura Soto, S. (2012). Predicción del Fracaso Escolar mediante. IEEE-RITA.

Ministerio de educación Bolivia. (2023). Tasa de reprobación anual. La paz: minedu.gob.bo.

Prokhorenkova, L., Vorobev, A., Dorogush, A. V., Gulin, A., & Gusev, G. (2018). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31.

Roca, E. (1 de marzo de 2025). La calidad de la educación en Bolivia. Opinión.

Santo Bacallao, A. (2020). Rendimiento escolar en la asignatura de lenguaje y comunicación en el nivel secundario de Bolivia. Scielo.

Trigo, M. S. (2 de marzo de 2025). Infobae. Obtenido de infobae.com: https://www.infobae.com/america/america-latina/2025/03/02/preocupacion-por-el-nivel-educativo-en-bolivia-solo-tres-de-cada-100-estudiantes-aprobaron-matematicas-y-quimica-en-un-examen-de-diagnostico/

UNESCO. (2023). Informe de seguimiento de la educación en el mundo – América Latina. París: UNESCO.

Villca Coraite, L. (agosto de 2025). GitHub. Obtenido de GitHub.com: https://github.com/LimbergVillcaCoraite/Proyecto-Dip.-Ciencia-de-datos.git