Universidad Católica Boliviana “San Pablo”
Facultad de Ingeniería
Departamento de Ingeniería de Sistemas
La Paz - Bolivia

PROYECTO FINAL DE MÓDULO

MACHINE LEARNING

Modelo predictivo del riesgo de abandono escolar en Bolivia basado en minería de datos y aprendizaje supervisado

Realizado por:

Jose Ignacio Ríos Villanueva

Saul Enrique Quiroz Castillo

Marco Antonio Villarroel Peña

AÑO 2025

ÍNDICE GENERAL

Introducción

El proyecto tiene como propósito desarrollar un modelo predictivo que permita identificar a estudiantes en Bolivia con alto riesgo de abandonar sus estudios, utilizando técnicas de minería de datos y algoritmos de aprendizaje supervisado. La investigación emplea datos oficiales del Instituto Nacional de Estadística (INE) y busca aportar evidencia empírica y herramientas analíticas que contribuyan a enfrentar la problemática de la deserción escolar. Este fenómeno, al limitar la formación de capital humano y profundizar la pobreza y desigualdad, requiere respuestas estratégicas desde la política pública educativa, donde el uso de ciencia de datos se plantea como una innovación relevante.

En cuanto a los antecedentes, se observa que en la región existen estudios que han aplicado modelos de machine learning (ML) para analizar las determinantes de la deserción escolar, destacando factores como pobreza, nivel educativo de los padres y calidad educativa. Sin embargo, en Bolivia la aplicación de metodologías predictivas es aún incipiente, predominando enfoques descriptivos y retrospectivos. Ante esta carencia, el proyecto propone un modelo de clasificación basado en variables sociodemográficas, con el fin de anticipar el riesgo de abandono escolar de manera individualizada y ofrecer una herramienta novedosa para la prevención en el sistema educativo nacional.

1 Capítulo 1. – MARCO REFERENCIAL

1.1 Referencias técnicas de otros trabajos

En lo referente a aplicación de modelos de ML para la predicción del abandono escolar, no se ha encontrado trabajos específicos en Bolivia. Dos estudios se aproximan a la investigación propuesta: (i) “Determinantes de la deserción escolar y el trabajo adolescente en Bolivia” de Joaquín Morales y Yelussa Vargas, un estudio realizado el año 2018 que utiliza un modelo biprobit para identificar los principales factores que inciden en la decisión de los adolescentes de trabajar, o estudiar, o trabajar y estudiar (Morales & Vargas, 2018); y (ii) una monografía desarrollada por Maya Wara López Laime el año 2024, que busca, no predecir el abandono escolar, pero sí aplica técnicas de ML para identificar patrones de bajo rendimiento escolar (López Laime, 2024).

En la región, se ha encontrado varias investigaciones, en el siguiente cuadro se resume tres referencias técnicas de trabajos que han buscado predecir la deserción estudiantil empleando modelos de ML:

Tabla 1. Referencias técnicas de investigaciones que emplean modelos de ML para predecir la deserción estudiantil

Título	Autor	País	Objetivo general	Resumen	Referencia
Una aplicación de aprendizaje automático (machine learning) en políticas públicas. Predicción de alerta temprana de deserción escolar en el sistema de educación pública de Chile.	Uldall, Jerome Smith; Rojas, Cristián Gutiérrez	Chile	Demostrar que el aprendizaje automático permite dar una alerta temprana más precisa sobre la deserción escolar.	Concluye que las tasas de verdaderos positivos de los algoritmos de ML son significativamente mayores que la de una regresión logit; los modelos que se evaluaron en esta investigación fueron: árbol de decisiones, random forest y redes neuronales.	(Uldall & Rojas, 2021)

Título	Autor	País	Objetivo general	Resumen	Referencia
El machine learning para abordar el abandono escolar: Una revisión de los modelos más innovadores.	Satalaya, Jules Mao Flores	Perú	Sistematizar veinte experiencias de aplicación de modelos de machine learning para predecir el abandono escolar.	Concluye que los modelos: (i) redes neuronales artificiales; (ii) vecinos más cercanos; (iii) regresión lineal y (iv) árboles de decisión, han demostrado adecuados niveles de eficacia en la clasificación y predicción de la deserción escolar, identificando además como predictores clave: (i) el rendimiento académico; (ii) dificultades financieras; y (iii) falta de apoyo social.	(Satalaya, 2024)
Application of the performance of machine learning techniques as support in the prediction of school dropout.	Jiménez-Gutiérrez, Auria Lucia; Mota-Hernández, Cinthya Ivonne; Mezura-Montes, Efrén; Alvarado-Corona, Rafael	México	Construir un modelo de ML con una confiabilidad de 90% para predecir la deserción estudiantil en instituciones de educación superior.	Se recurrió a datos de 32 estados (mexicanos), contenidos en la encuesta intercensal 2015 y los censos de población de 2010 y 2020, logrando consolidar una base de datos de más de 1,08 millones de observaciones y 20 variables. Las técnicas que se emplearon fueron: (i) redes neuronales artificiales; (ii) Support Vector Machines (SVM); (iii) regresión lineal Ridge and Lasso; (iv) optimización bayesiana; y (v) random forest, las dos primeras con una confiabilidad superior al 99% y la última con un 91%	(Jiménez-Gutiérrez, Mota-Hernández, Mezura-Montes, & Alvarado-Corona, 2024)

La principal diferencia e innovación de la presente investigación respecto a los trabajos de referencia, es que considera variables sociodemográficas, mientras que los estudios citados emplean, principalmente, variables socioeconómicas y académicas; asimismo, cabe subrayar que este trabajo es uno de los primeros en emplear técnicas de ML aplicadas a problemáticas sociales en Bolivia, específicamente en el sector educativo.

1.2 Identificación del problema

(i) Pregunta central:

¿Cómo desarrollar un modelo predictivo, basado en técnicas de minería de datos y aprendizaje supervisado, que permita identificar tempranamente a estudiantes en riesgo de abandono escolar en Bolivia a partir de información sociodemográfica disponible?

(ii) Preguntas específicas de investigación:

1. ¿Qué variables sociodemográficas se asocian significativamente con el riesgo de abandono escolar en el contexto boliviano?

2. ¿Qué técnicas de aprendizaje supervisado ofrecen el mejor desempeño predictivo para identificar estudiantes en riesgo de deserción escolar?

3. ¿Qué insights y recomendaciones técnicas pueden emitirse como producto de esta investigación, para contribuir al diseño de estrategias de política pública educativa orientadas a prevenir el abandono escolar?

1.3 Formulación del problema

El abandono escolar es una problemática que tiene incidencias negativas en la formación de capital humano y en la productividad del trabajo en todas las sociedades y, por supuesto, en Bolivia. Si bien, en las últimas dos décadas se ha registrado un descenso importante de la deserción escolar en el país, producto de políticas públicas que han incentivado a los estudiantes a concluir con sus estudios de primaria y secundaria, aún existen niveles significativos de deserción, especialmente en zonas rurales, comunidades indígenas y áreas periurbanas.

Asimismo, la pandemia de Covid-19, y la actual crisis económica por la que atraviesa el país, han sido importantes causas para que el abandono escolar se incremente recientemente.

Ante ello, las políticas educativas han priorizado medidas reactivas, pero carecen de mecanismos sistemáticos de detección temprana que permitan anticipar qué estudiantes presentan mayor probabilidad de abandonar sus estudios. La información disponible suele utilizarse para elaborar diagnósticos generales y no para predecir casos individuales de riesgo.

En otro sentido, el avance tecnológico, permite cada vez un mayor y más sencillo acceso a datos sistematizados, como es el caso de la base de datos íntegra de la EDSA 2023 realizada por el INE, y, además, pone a disposición herramientas de ML que nunca antes habían estado tan cerca. Este contexto representa una importante oportunidad para el desarrollo de modelos predictivos e individualizados de deserción escolar.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un modelo predictivo del riesgo de abandono escolar en Bolivia, utilizando técnicas de minería de datos y aprendizaje supervisado, que permitan identificar tempranamente a los estudiantes en riesgo de deserción para contribuir a la toma de decisiones sobre intervenciones preventivas.

1.4.2 Objetivos específicos

1. Analizar las fuentes de datos disponibles (EDSA 2023) para identificar y seleccionar variables relevantes relacionadas con el abandono escolar.

2. Preparar y limpiar los datos, aplicando técnicas de tratamiento de valores faltantes, valores atípicos, ingeniería de variables y normalización, para garantizar la calidad del conjunto de entrenamiento (EDA).

3. Explorar estadísticamente los datos para detectar patrones y relaciones entre las variables y el abandono escolar.

4. Implementar y comparar diferentes algoritmos de aprendizaje supervisado, para predecir el abandono escolar, optimizar modelos y encontrar el mejor desempeño posible.

5. Evaluar el desempeño de los modelos utilizando métricas de clasificación (matriz de confusión y métricas derivadas, como Accuracy, Precision, Recall, F1-score, métricas de curva, entre otras), priorizando la correcta identificación de estudiantes en riesgo de deserción.

6. Formular insights y recomendaciones técnicas que contribuyan a la formulación de estrategias de política pública educativa orientadas a prevenir el abandono escolar.

1.5 Justificación

Justificación teórica: El abandono escolar ha sido estudiado en la región desde diversas disciplinas —economía, ciencias sociales, pedagogía y sociología— identificando factores como nivel socioeconómico, entorno familiar y desempeño académico. Sin embargo, en Bolivia los estudios existentes son mayormente descriptivos o correlacionales, lo que limita su alcance. Esta investigación busca aportar un marco teórico interdisciplinario que vincule variables sociodemográficas con ciencia de datos, incorporando minería de datos, aprendizaje supervisado y modelos predictivos, ampliando así el conocimiento sobre cómo estas técnicas pueden aplicarse a problemáticas sociales como la deserción escolar.

Justificación metodológica: Metodológicamente, el estudio se sustenta en un enfoque cuantitativo basado en los datos oficiales de la EDSA 2023, aplicando técnicas de minería de datos para procesar y seleccionar variables, posteriormente implementar algoritmos de aprendizaje supervisado, comparar el rendimiento a partir de la técnica crossvalidation, ensamblar dos o más modelos con la técnica Stacking y evaluar el mejor modelo posible. La innovación radica en aplicar ciencia de datos y herramientas de ML a un problema social, y construir un modelo predictivo del abandono escolar con tasas altas de rendimiento.

Justificación práctica: En la práctica, el modelo que se desarrolle tiene un alto potencial de aplicación en el sistema educativo nacional, ya que permitirá identificar tempranamente a estudiantes en riesgo de abandono, priorizar recursos en contextos de mayor vulnerabilidad y brindar herramientas analíticas a las unidades educativas y autoridades. De esta manera, se promueve un cambio de paradigma en la política educativa, pasando de medidas reactivas a un enfoque preventivo y prospectivo basado en evidencia.

Justificación social: El abandono escolar genera consecuencias sociales graves: limita oportunidades de empleo, perpetúa la pobreza y exclusión, y afecta al desarrollo económico y humano del país, especialmente en poblaciones rurales, indígenas y de bajos ingresos. En este contexto, la investigación se justifica socialmente porque busca ofrecer una herramienta que permita actuar de manera preventiva, reduciendo la deserción y promoviendo mayor equidad en el acceso y permanencia escolar de niñas, niños y adolescentes en situación de vulnerabilidad.

1.6 Límites y alcances

1.6.1 Límites

En cuanto a la disponibilidad y características de los datos: (i) Se empleará la base de datos de la última EDSA, limitando el análisis a las variables disponibles en esta base; y (ii) no se recogerán datos primarios mediante encuestas o entrevistas, lo cual restringe la inclusión de factores cualitativos como motivación personal o clima escolar, entre otros.

En cuanto a la cobertura temporal y espacial de los datos: (i) El estudio se centrará en datos disponibles correspondiente a la última EDSA que data del año 2023; asimismo, (ii) las generalizaciones del estudio se harán a nivel nacional, sin entrar al detalle analítico con desagregación a nivel departamental o municipal, aunque el modelo podría aplicarse a esos niveles en investigaciones futuras.

En cuanto al ámbito educativo: (i) El análisis se enfocará en los niveles de primaria y secundaria de educación regular (no se considerará educación superior o profesional, ni modalidades de educación alternativa o especial[1]).

1.6.2 Alcances

Identificación de variables predictoras: se determinarán los principales factores sociodemográficos asociados al abandono escolar en Bolivia.

Comparación de algoritmos: se implementarán distintos modelos de aprendizaje supervisado, evaluando su desempeño con métricas estándar.

Búsqueda del mejor modelo: se buscará el mejor modelo posible, a través de la aplicación de crossvalidatión, optimización de hiperparámetros y técnicas de ensamble de dos o más algoritmos.

Modelo predictivo validado: se generará un modelo con capacidad de anticipar el riesgo de abandono a nivel individual, a partir de información disponible.

Aporte metodológico: se demostrará la aplicabilidad de la ciencia de datos al campo educativo en Bolivia.

Aporte práctico: se formulará insights y recomendaciones para contribuir a la prevención del abandono escolar a partir del uso de este tipo de modelos como herramientas de alerta temprana.

1.7 Metodología de la investigación

1.7.1 Tipo de estudio

El presente trabajo corresponde a un estudio de tipo **cuantitativo, explicativo y aplicado**. Emplea datos provenientes de la Encuesta de Demografía y Salud (EDSA 2023), con el objetivo de identificar patrones y relaciones significativas entre variables sociodemográficas y el riesgo de abandono escolar en Bolivia. El enfoque explicativo permite no solo describir el fenómeno, sino también establecer inferencias causales y construir modelos predictivos que anticipen el riesgo de deserción escolar individual dentro del sistema educativo.

Asimismo, se trata de un estudio aplicado, orientado a generar insights y recomendaciones para contribuir al diseño de políticas públicas de prevención del abandono escolar. La investigación se apoya en técnicas de minería de datos y machine learning, lo que permite abordar el problema desde una perspectiva innovadora y basada en evidencia. El uso de algoritmos de clasificación binaria, validación cruzada, optimización de hiperparámetros y ensable de modelos refuerza el carácter empírico y replicable del estudio, contribuyendo a la generación de aplicaciones técnicas.

1.7.2 Métodos y técnicas

El abandono escolar es un problema de clasificación (abandona vs. no abandona), asimismo, la mayoría de las variables que tiene la base de datos de la EDSA son variables categóricas. Por lo tanto, el planteamiento, metodología y algoritmos que se emplean en esta investigación son propios de los modelos de clasificación.

La metodología que se implementa es la de Aprendizaje Supervisado, los modelos de aprendizaje supervisado se definen en el libro de Hastie, Tibshirani y Friedman, en los siguientes términos: "*Para cada uno, existe un conjunto de variables que podrían denominarse*

[1] El Ministerio de Educación, se estructura a su vez en tres Viceministerios: (i) Educación Regular; (ii) Educación Alternativa y Especial; y (iii) Educación Superior de Formación Profesional, definiendo las modalidades de educación vigentes en el país.

entradas, las cuales se miden o se establecen previamente. Estas tienen cierta influencia sobre uno o más resultados. Para cada ejemplo, el objetivo es usar las entradas para predecir los valores de los resultados. Este ejercicio se llama aprendizaje supervisado". (Hastie, Tibshirani, & Friedman, 2008)

En este sentido, la metodología básicamente consiste en extraer y cargar el dataset en entorno Python Colab, analizar y adecuar los datos, de tal manera que la variable dependiente, en este caso el abandono escolar, se diferencie de las variables independientes que la explican; posteriormente, se divide el dataset en conjunto de entrenamiento, validación y test, con la finalidad de entrenar cada modelo en el primero y ajustarlo y evaluarlo en los siguientes, para finalmente comparar el rendimiento y elegir el que tenga un mejor desempeño, para emitir insights con base en los resultados del mejor modelo obtenido.

Asimismo, se aplica técnicas frecuentes en la construcción de modelos de aprendizaje supervisado, como el balanceo de clases del target, en este caso, como se cuenta con un dataset de más de 20 mil observaciones (una buena cantidad de datos), se emplea la técnica undersampling, que consiste en extraer una muestra de la clase más grande para equilibrarla con la clase más pequeña; y también se aplica la técnica de escalado de datos, para evitar que las unidades de medida que pueden ser diferentes de una variable a otra, ocasionen distorsiones en los modelos (con excepción del algoritmo árbol de decisiones, donde los datos escalados pueden dificultar la interpretación del árbol).

Adicionalmente, se hace una optimización de hiperparámetros en cada uno de los modelos con el fin de mejorar el desempeño individual de cada uno, se implementa una técnica de validación cruzada o crossvalidation para determinar, comparativamente, qué modelos son los que tienen mejor desempeño, y también se aplica una técnica de ensamble de los modelos más relvantes, para conseguir un modelo final óptimo.

En el siguiente cuadro se resume las técnicas utilizadas:

Tabla 2. Técnicas empleadas en cada fase del proyecto

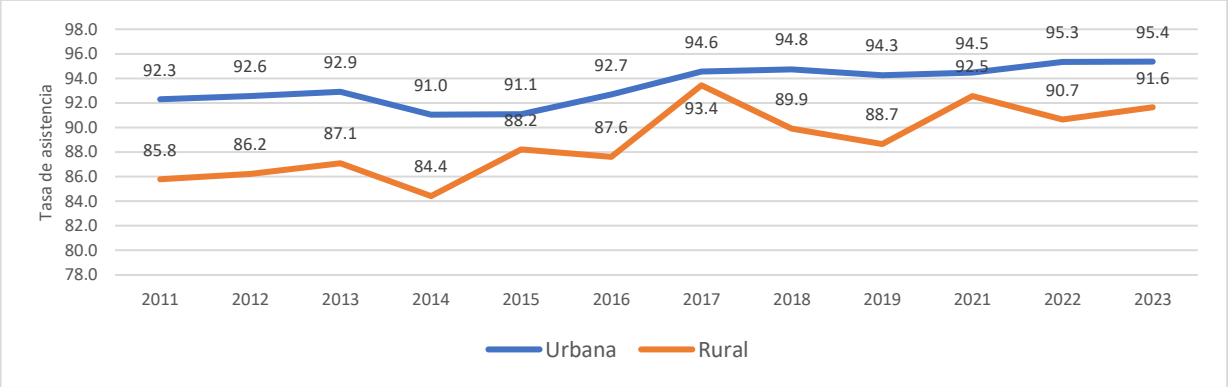
Fase	Técnicas
EDA y preprocesamiento	Dicotomización Obtención de estadísticos descriptivos y frecuencias simples Matriz de correlaciones simples División del dataset en conjuntos de entrenamiento, validación y test con train_test_split Equilibrio de clases con undersampling con RandomUnderSampler Escalado de variables con StandardScaler
Entrenamiento de los modelos	Regresión logística con LogisticRegression Árbol de decisiones con DecisionTreeClassifier Naive Bayes con GaussianNB Redes neuronales artificiales con Sequential, Dense y Dropout
Validación	Crossvalidation con StratifiedKFold y cross_val_predict
Optimización	Búsqueda de los mejores hiperparámetros con GridSearchCV
Ensamble	Stacking con StackingClassifier
Evaluación final	Evaluación comparativa de las métricas de desempeño de cada modelo con accuracy_score y confusion_matrix

1.8 Análisis preliminar

La tasa de asistencia escolar (TAE) en Bolivia muestra una tendencia positiva en los últimos 14 años, pasando de 89,97% el año 2011 a 94,23% el año 2023 (último dato disponible), aunque la brecha entre área urbana y área rural no se ha podido acortar significativa ni sostenidamente, puesto que para la gestión 2023 la TAE en área rural fue de 91,6% frente al 95,4% en área urbana.

El año 2019 se observa una caída en la TAE, y no se cuenta con datos para la gestión 2020, ambos casos atribuibles a la pandemia de Covid-19.

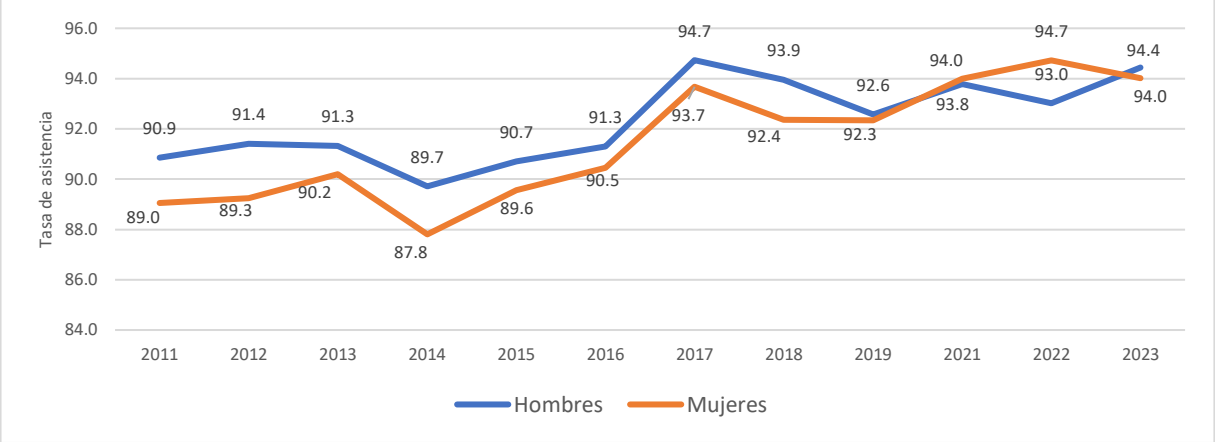
Ilustración 1. Bolivia. Evolución de la tasa de asistencia escolar, por área. Año 2011 – 2023



Fuente: (INE, 2023)

Históricamente, la TAE de las mujeres ha sido inferior a la de los hombres, sin embargo, en los últimos años, la brecha de género se ha acortado significativamente, el año 2023 la diferencia es mínima, siendo la TAE de 94,4% en varones y 94,0% en mujeres.

Ilustración 2. Bolivia. Evolución de la tasa de asistencia escolar, por sexo. Año 2011 – 2023



Fuente: (INE, 2023)

Estos datos, dan cuenta de un 5,8% de niñas, niños y adolescentes, de 6 a 19 años de edad, que el año 2023 no acudieron a la escuela ni a ningún establecimiento de educación regular, lo que –considerando que en Bolivia para entonces se estimaba una población de 3,9 millones de NNA de 6 a 19 años de edad–, significa cerca de 226 mil NNA que no acudieron a la escuela.

A nivel regional, Bolivia se encuentra en un punto intermedio, puesto que países como Argentina y Perú presentan tasas de asistencia superiores, pero otros países, como Chile, más bien tienen tasas de asistencia escolar inferiores.

Tabla 3. Tasa de asistencia escolar por país (últimos datos disponibles)

País	Fuente principal	Referencia	Tasa de asistencia escolar – Educación primaria	Tasa de asistencia escolar – Educación secundaria
Perú	NEI – ENAHO 2024	(gob.pe, 2024)	99.1%	91.5%

País	Fuente principal	Referencia	Tasa de asistencia escolar – Educación primaria	Tasa de asistencia escolar – Educación secundaria
Argentina	INDEC – EPH 2022 / Observatorio de Educación	(LV/12, 2023)	97.6% (4–17 años)	97% (15 años)
Chile	MINEDUC – Reporte 2024	(Ministerio de Educación, Centro de Estudios, 2024)	89% (promedio nacional)	85.7% (2024)

Si se hace un análisis más detallado, considerando el modelo basado en las "Cinco dimensiones de la exclusión – 5DE" que utiliza UNICEF para desarrollar el informe "Iniciativa Mundial Niños y Niñas fuera de la Escuela" Bolivia 2011, los datos se tornan más preocupantes: (i) 93 mil niños y niñas de 5 años de edad no asisten a inicial cuando deberían hacerlo; (ii) 89,5 mil niños de 6 a 11 años están fuera de la escuela; (iii) 17 mil niños y niñas entre 12 y 13 años está fuera de la escuela; (iv) 139,3 mil entre 14 y 17 están fuera de la escuela. (UNICEF, 2011)

Entre los elementos que ayudan a explicar la existencia de niños fuera de la escuela mencionados en este informe, destacan: (i) la limitación de las expectativas educativas; (ii) la discriminación de género; (iii) el embarazo adolescente; (iv) la pobreza; y (v) la educación de los padres.

Otro estudio interesante, que explora los determinantes detrás de la decisión de estudiar o de trabajar de los adolescentes en Bolivia, revela que dicha decisión está fuertemente influenciada por el ingreso, por la educación de los padres, por diferencias en la estructura familiar y por diferencias regionales. (Morales & Vargas, 2018)

2 Capítulo 2.- MARCO PRÁCTICO

2.1 Diseño del proyecto

Como se ha comentado, el problema del abandono escolar es un problema de clasificación, en este sentido, para la correcta interpretación del diseño, desarrollo y conclusiones del proyecto, es fundamental identificar cada clase de manera inequívoca:

(i) clase 0 = NO abandona los estudios; y
(ii) clase 1 = abandona los estudios.

La base de datos de la EDSA 2023 está conformada por nueve archivos SPSS, entre los cuales dos son de interés para el desarrollo del modelo predictivo: (i) EDSA2023_Hombre.sav; y (ii) EDSA2023_Mujer.sav, que contienen información personal de hombres y de mujeres respectivamente, encuestados y encuestadas el año 2023. El primer archivo esta conformado por 5.878 observaciones y 997 atributos, mientras el segundo cuenta con 14.545 observaciones y 1.341 atributos.

Los dos archivos se consolidaron en un solo dataframe, que contiene las 20.423 observaciones de hombres y mujeres. Se identificó las variables sociodemográficas explicativas del abandono escolar, equivalentes entre hombres y mujeres, y se conformó una única base de datos en Excel, misma que fue cargada en Python mediante GitHub, para realizar todos los pasos programados en el notebook.

De un análisis preliminar del dataset, se identifica las siguientes variables sociodemográficas que pueden contribuir a explicar el abandono escolar, y que además existen tanto para hombres como para mujeres:

- Ruralidad
- Región (Altiplano, Valles, Llanos)

- Lengua materna
- Sexo (Hombre, Mujer)
- Identificación cultural
- Edad en que inició la actividad sexual
- Edad a la que tuvo su primer hijo
- Si durante la niñez observó violencia entre sus padres
- Tipos y cantidad de violencias que sufrió durante la niñez

Se ha hecho un trabajo de ingeniería de variables, con la finalidad de adecuar los atributos a las necesidades de un modelo de clasificación. En el siguiente cuadro se resume el diseño de las variables explicativas del abandono escolar:

Tabla 4. Diseño del dataset – Variables explicativas

Criterio (variable original)	Nombre de la variable	Valores
Ruralidad	rural	1 si vive en área rual 0 si NO vive en área rural
Región (Altiplano, Valles, Llanos)	altiplano	1 si vive en la región Altiplano 0 si NO vive en la región Altiplano
	valles	1 si vive en la región Valles 0 si NO vive en la región Valles
	llanos	1 si vive en la región Llanos 0 si NO vive en la región Llanos
Lengua materna	lengua_originaria	1 si la lengua en que aprendió a hablar es originaria (aymara, quechua, guaraní, etc.) 0 si la lengua en que aprendió a hablar NO es originaria
Sexo (Hombre, Mujer)	hombre	1 si nació hombre 0 si NO nació hombre
	mujer	1 si nació mujer 0 si NO nació mujer
Identificación cultural	npioc	1 si se identifica con alguna nación o pueblo indígena orignario campesino 0 si NO se identifica con alguna nación o pueblo indígena originario campesino
Edad en que inició la actividad sexual	ias_<17	1 si declara haber iniciado su actividad sexual antes de los 17 años 0 si NO declara haber iniciado su actividad sexual antes de los 17 años
Edad a la que tuvo su primer hijo	1erhijo_<17	1 si tuvo a su primer hijo/a antes de los 17 años 0 si NO tuvo a su primer hijo/a antes de los 17 años
Si durante la niñez observó violencia entre sus padres	agre_padres	1 si declara haber presenciado violencia entre sus padres durante la niñez 0 si NO declara haber presenciado violencia entre sus padres durante la niñez
Tipos y cantidad de violencias que sufrió durante la niñez	viol_ninez	1 si declara haber sufrido al menos un tipo de violencia durante la niñez 0 si NO declara haber sufrido al menos un tipo de violencia durante la niñez

La variable target ha sido denominada “Abandono” y se ha construido con las siguientes variables:

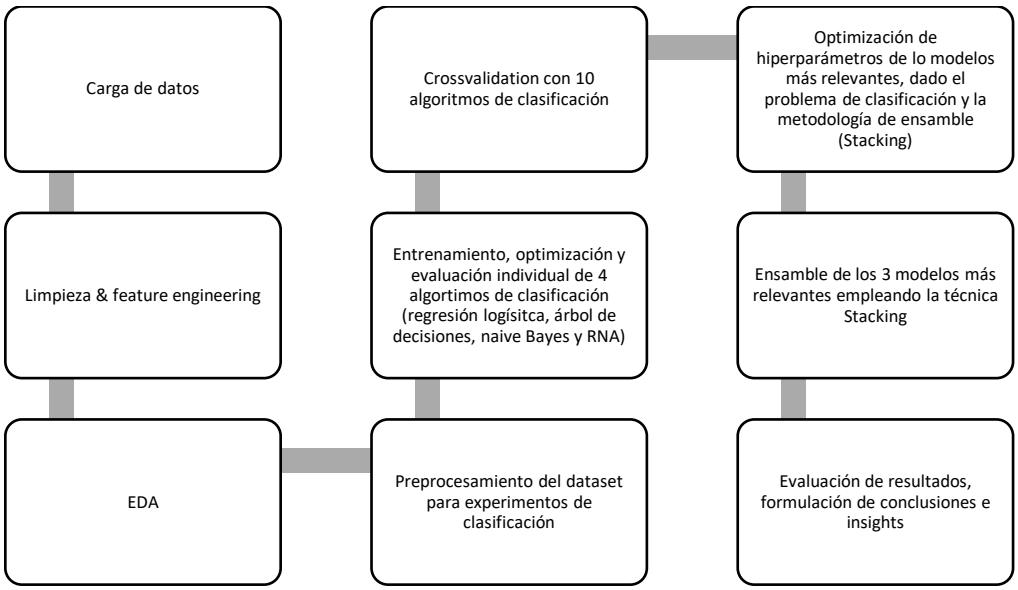
- Edad al momento de responder la encuesta
- Años de estudio
- Último nivel educativo aprobado
- Último curso aprobado

- Si al momento de aplicar la encuesta asistía a la escuela o a alguna institución educativa

Se ha cruzado la variable “Edad al momento de responder la encuesta” con “Años de estudio”; hombres y mujeres mayores de 17 años con menos de 12 años de estudio se han etiquetado como “Posible Abandono”; posteriormente, se ha cruzado este resultado con las variables: “Último nivel educativo aprobado” y “Último curso aprobado”, si éstas no coinciden con 6to. de secundaria, o 4to. se secundaria, o 4to. medio (según el sistema en que hayan estudiado, ya que la nomenclatura de niveles y cursos de primaria y secundaria ha cambiado en Bolivia en varias ocasiones), han sido etiquetados como “Posible Abandono 2”; y finalmente, este resultado se ha cruzado con la variables “Estudia Actualmente”, y si la respuesta era negativa se etiquetó con “Abandono”.

El diseño del proyecto se resume en el siguiente gráfico:

Ilustración 3. Diseño del proyecto



2.2 Desarrollo del Proyecto

(i) **Carga de datos:** La fase de carga de datos consistió en la ingestión del conjunto de datos 'Base_abandono_escolar.xlsx' directamente desde una URL de GitHub utilizando la función pd.read_excel() de la librería pandas. Se implementó un bloque try-except para manejar posibles errores durante la descarga y lectura del archivo, asegurando la robustez del proceso. La carga exitosa se verificó imprimiendo las primeras filas del DataFrame resultante (df.head()) y su información estructural (df.info()).

(ii) **Limpieza e ingeniería de variables:** En esta etapa, se transformaron y crearon nuevas características a partir de las variables originales. Se generaron variables binarias como viol_ninez (sumando y binarizando múltiples columnas de violencia), lengua_originaria (agrupando idiomas), migracion (basada en el lugar de nacimiento), rural (basada en el área), altiplano, valles, llanos (basadas en la región), hombre y mujer (basadas en el sexo), y 1erhijo_<17 e ias_<17 (dicotomizando edades en las que las personas tuvieron a su primer hijo o hija). Se

manejaron valores faltantes en agre_padres reemplazándolos por 0. Finalmente, se eliminaron las columnas originales y auxiliares que no se utilizarían en el modelado.

(iii) **Preprocesamiento:** El preprocesamiento incluyó la separación del conjunto de datos en características (X) y la variable objetivo (y, "Abandono"). Se dividieron estos conjuntos en entrenamiento, validación y prueba utilizando train_test_split con estratificación (stratify=y) para mantener la proporción de clases. Para abordar el desbalance de clases en el conjunto de entrenamiento, se aplicó Random Under-Sampling mediante RandomUnderSampler(random_state=42) para igualar el número de instancias en las clases minoritaria y mayoritaria. Posteriormente, se escalaron las variables explicativas utilizando StandardScaler, ajustándolo únicamente a los datos de entrenamiento undersampled y aplicándolo a los conjuntos de entrenamiento, validación y prueba para estandarizar su rango de valores.

(iv) **Regresión logística:** Para la Regresión Logística, se inicializó un modelo LogisticRegression y se entrenó utilizando los datos de entrenamiento balanceados y escalados (X_train_undersampled_scaled, y_train_undersampled). La sintonización de hiperparámetros (C, penalty, solver) se realizó mediante GridSearchCV con validación cruzada (cv=5) en el conjunto de entrenamiento balanceado. El mejor modelo encontrado se evaluó en los conjuntos de validación y prueba escalados para obtener métricas de rendimiento como Accuracy, Classification Report y Confusion Matrix. Adicionalmente, se analizaron los coeficientes del mejor modelo para interpretar la influencia de cada característica en la predicción del abandono.

(v) **Árbol de decisiones:** El modelo de Árbol de Decisiones se implementó utilizando DecisionTreeClassifier. La optimización de hiperparámetros (max_depth, min_samples_split, min_samples_leaf, criterion) se llevó a cabo con GridSearchCV y validación cruzada (cv=5) en el conjunto de entrenamiento balanceado (X_train_undersampled, y_train_undersampled), ya que los árboles no son sensibles al escalado. El mejor modelo sintonizado se evaluó en los conjuntos de validación y prueba originales (X_val, X_test). Se visualizó el árbol resultante para entender las reglas de decisión aprendidas por el modelo.

(vi) **Naive Bayes:** Se empleó el clasificador GaussianNB, adecuado para datos con distribución aproximadamente gaussiana, utilizando los datos de entrenamiento balanceados y escalados (X_train_undersampled_scaled, y_train_undersampled). Se realizó una búsqueda de hiperparámetros simple para var_smoothing con GridSearchCV (cv=5). El mejor modelo Naive Bayes se evaluó en los conjuntos de validación y prueba escalados. Se analizaron los parámetros aprendidos por el modelo (medias y varianzas por clase, y probabilidades a priori) para interpretar las características de cada grupo.

(vii) **RNA:** Se construyó una Red Neuronal Artificial secuencial utilizando TensorFlow/Keras. La arquitectura consistió en capas densas (Dense) con activación ReLU y capas de abandono (Dropout) para regularización, finalizando con una capa de salida densa con función de activación sigmoide para la clasificación binaria. El modelo se compiló con el optimizador Adam y pérdida binary_crossentropy. Se entrenó la RNA en los datos de entrenamiento

balanceados y escalados (X_train_undersampled_scaled, y_train_undersampled) utilizando el conjunto de validación escalado para monitorear el rendimiento durante el entrenamiento. Se evaluó el modelo entrenado en el conjunto de prueba escalado y se calculó la importancia de permutación de las características que explican el abandono escolar.

(viii) **Crossvalidation:** La validación cruzada estratificada (Stratified K-Fold) con K=5 folds se utilizó para obtener estimaciones más robustas del rendimiento de varios modelos. Se implementó un bucle manual para K-Fold que permitió aplicar el escalado (StandardScaler) y el sobremuestreo (SMOTE) dentro de cada fold de entrenamiento, evitando la fuga de información. Para cada fold y modelo, se calcularon métricas como AUC, Accuracy, Precision, Recall y F1-Score. Los resultados promedio y la desviación estándar de estas métricas a través de los folds se resumieron en un DataFrame para comparar el rendimiento general de los modelos.

(ix) **Optimización de hiperparámetros:** La optimización de hiperparámetros se realizó para la Regresión Logística y Random Forest utilizando GridSearchCV. Se definió un espacio de búsqueda de hiperparámetros (param_grid) para cada modelo. GridSearchCV exploró sistemáticamente combinaciones de estos hiperparámetros, entrenando y evaluando el modelo con validación cruzada (cv=5) en el conjunto de entrenamiento balanceado. El objetivo era encontrar la combinación de hiperparámetros que maximizara la puntuación de validación cruzada. El mejor estimador encontrado por GridSearchCV se utilizó para la evaluación final.

(x) **Stacking:** Se implementó un modelo de ensamblaje utilizando StackingClassifier. Este enfoque combina las predicciones de varios modelos base (en este caso: LogisticRegression, SVC y RandomForest) utilizando un modelo meta (LogisticRegression) para hacer la predicción final. Los modelos base se entrenaron en los datos de entrenamiento balanceados y escalados, y sus predicciones (probabilidades) se utilizaron como características de entrada para entrenar el meta-modelo. El modelo Stacking completo se evaluó en los conjuntos de validación y prueba escalados para determinar si la combinación de modelos mejoraba el rendimiento individual. Se visualizó la matriz de confusión del modelo Stacking final.

2.3 Validación y pruebas

Las métricas de desempeño de los modelos implementados son las siguientes:

- **Accuracy:** Es la proporción de predicciones correctas sobre el total de predicciones, es decir, la cantidad de niños y niñas que el modelo predijo que iban a abandonar los estudios que relamente abandonaron, más la cantidad de niñas y niños que el modelo predijo que NO iban a abandonar los estudios que realmente no lo hicieron, respecto al total de casos predichos.
- **Precision:** Responde a la pregunta: De todos los casos que el modelo predijo que iban a abandonar los estudios, ¿cuántos realmente abandonaron la escuela?
- **Recall:** Responde a la pregunta: De todos los casos que realmente abandonaron la escuela, ¿cuántos predijo el modelo que iban a abandonar los estudios?
- **F1-score:** Es la media armónica de Precision y Recall, proporciona un balance entre ambas métricas.

En este caso, el costo social y económico del abandono escolar se materializa cuando un o una estudiante abandona realmente los estudios, se expresa en años de gasto en educación que no llega a consolidarse como inversión, por lo tanto, el Recall es la métrica que se ajusta mejor a la necesidad de predecir los casos que realmente abandonan la escuela. Sin embargo, en todo modelo de ML, es deseable que un buen desempeño se exprese en una combinación de varias o todas las métricas con valores cercanos a 1.

Se puede observar que, en este caso, ninguna de las métricas supera el umbral de 0,75, es decir que, en general, el desempeño de los modelos de clasificación es bueno pero no excelente ni óptimo, todos los modelos consiguen predecir, como máximo, tres cuartos de los casos.

Por otro lado, la validación en los conjuntos de entrenamiento, validación y test, muestran métricas muy similares, lo que quiere decir que no existe sobreajuste en los modelos, es decir que han aprendido a clasificar correctamente aproximadamente entre 71% y 75% de datos vistos por primera vez, en este caso, niñas y niños inscritos en el sistema educativo boliviano.

Tabla 5. Métricas de desempeño de los modelos evaluados

Modelo	Accuracy mean	Precision mean	Recall mean	F1 mean
Stacking	**0.74**	**0.69**	**0.73**	**0.70**
RNA	**0.73**	**0.69**	**0.74**	**0.70**
GradientBoosting	0.72	0.49	**0.74**	**0.59**
SVC	0.72	0.48	**0.74**	**0.58**
LightGBM	0.72	0.48	0.73	**0.58**
LogisticRegression	0.73	**0.50**	0.69	0.58
MLP	0.72	0.48	0.71	0.57
XGBoost	0.72	0.48	0.71	0.57
RandomForest	0.71	0.48	0.71	0.57
DecisionTree	0.72	0.48	0.70	0.57
GaussianNB	**0.75**	**0.54**	0.57	0.55
KNN	**0.74**	**0.52**	0.46	0.49

De todos los modelos entrenados, dos son los que presentan una mejor combinación de métricas de Accuracy, Precision, Recall y F1-score, estos son: (i) el meta modelo que surge del ensamble de regresión logística simple, SVC y Random Forest usando la técnica Stacking; y (ii) el modelo de Redes Neuronales Artificiales (RNA), en ambos casos se observan métricas muy similares, lo que hace pensar que los modelos han alcanzado el mejor nivel de desempeño posible con los datos disponibles.

2.4 Análisis del proyecto

Este proyecto ha permitido hacer un análisis profundo de las determinantes sociodemográficas que inciden en el abandono escolar a nivel nacional. Los modelos paramétricos desarrollados individualmente: regresión logística y naive Bayes, así como el modelo de redes neuronales artificiales, permiten calcular los coeficientes o pesos específicos, así como la importancia relativa –en el caso de RNA– de cada uno de los atributos, lo que a su vez facilita la interpretación de las relaciones de causalidad del abandono escolar.

En la siguiente tabla se resume los pesos específicos que cada variable tiene en la explicación del target en cada uno de los modelos:

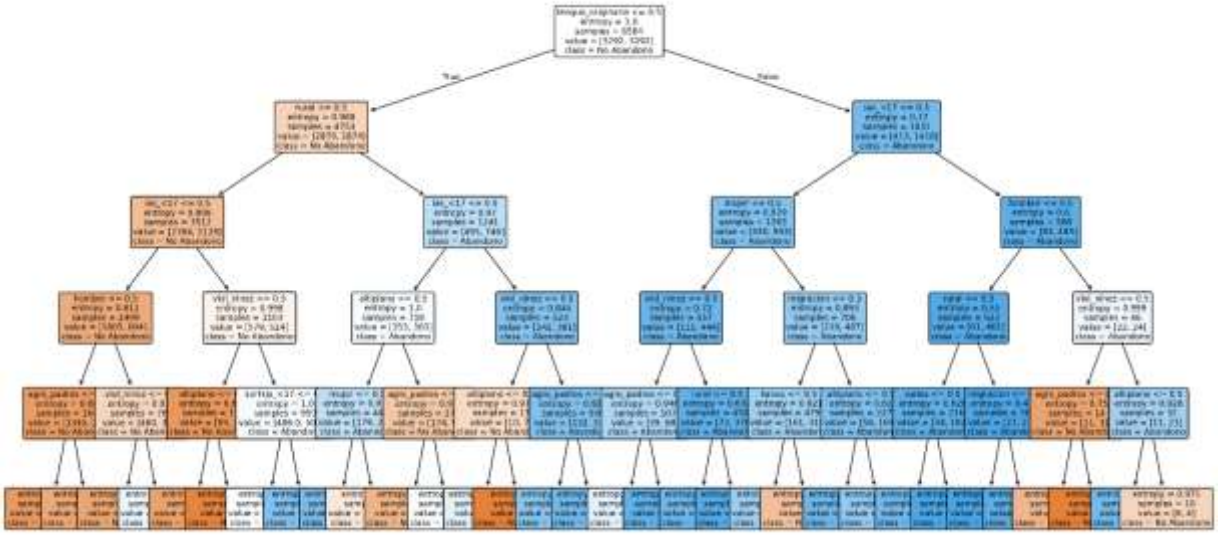
Tabla 6. Pesos específicos de las variables, en la explicación del target, por modelo

Atributo	Coeficientes - Regresion Logistica	Media para la clase 0 - Naive Bayes	Media para la clase 1 - Naive Bayes	Importancia de permutación - RNA
lengua_originaria	**0.6240**	**-0.3407**	**0.3407**	**0.0790**
rural	**0.3930**	**-0.2773**	**0.2773**	**0.0190**
ias_<17	**0.4130**	**-0.1888**	**0.1888**	**0.0232**
viol_ninez	**0.3797**	**-0.1346**	**0.1346**	**0.0209**

Atributo	Coeficientes - Regresion Logistica	Media para la clase 0 - Naive Bayes	Media para la clase 1 - Naive Bayes	Importancia de permutación - RNA
mujer	-0.1928	0.0651	-0.0651	-0.0058
hombre	0.1928	-0.0651	0.0651	-0.0003
agre_padres	0.1358	-0.1132	0.1132	0.0008
npioc	0.0600	-0.1786	0.1786	0.0063
migracion	0.1173	-0.1002	0.1002	0.0080
1erhijo_<17	0.1138	-0.0878	0.0878	0.0095
altiplano	-0.1054	0.0558	-0.0558	-0.0028
llanos	0.0850	0.0095	-0.0095	-0.0005
valles	0.0250	-0.0679	0.0679	-0.0003

Por otro lado, el algoritmo no paramétrico árbol de decisiones revela que las variables más importantes son lengua_originaria, rural y ias_<17, tal como se puede apreciar en el gráfico:

Ilustración 4. Árbol de decisiones



En todos los modelos entrenados y evaluados independientemente, las variables más importantes son:

(i) **lengua_originaria:** quiere decir que las niñas y niños que aprenden a hablar en lengua originaria (aymara, quechua, guaraní, entre otras) tienen mayor probabilidadde abandonar la escuela;

(ii) **rural:** es decir que las niñas y niños que viven en área rural tienen mayor probabilidad de abandonar los estudios;

(iii) **ias_<17:** las y los adolescentes que inician su actividad sexual antes de los 17 años tienen una mayor probabilidad de abandonar la escuela;

(iv) **viol_ninez:** las niñas y niños que sufren algún tipo de violencia, tienen mayor probabilidad de abandonar los estudios.

Para el algoritmo naive Bayes también es importante la variable “npioc”, es decir que, según este algoritmo, las niñas y niños que pertenecen a alguna nación o pueblo indígena originario campesino, tienen mayor probabilidad de abandonar la escuela.

Considerando los modelos que tienen el mejor performance: (i) redes neuronales artificiales (RNA) y (ii) el metamodelo Stacking (ensamble de regresión logística + random forest + VSC) se hizo un ejercicio de sensibilidad del abandono escolar respecto a las variables más relevantes. En el siguiente cuadro se resume los resultados:

Tabla 7. Análisis de sensibilidad

	Probabilidad de abandono		Variación de la probabilidad de abandono respecto al perfil inicial	
	RNA	Stacking	RNA	Stacking
Perfil inicial	0,0653	0,1254		
lengua_originaria	0,3342	0,3427	0,2689	0,2173
rural	0,1425	0,1551	0,0772	0,0297
ias_<17	0,2125	0,1826	0,1472	0,0572
viol_ninez	0,1493	0,1701	0,0840	0,0447
lengua_originaria + rural	0,4674	0,4538	0,4021	0,3284
Las cuatro	0,8588	0,8540	0,7935	0,7286

Partiendo de un perfil inicial, que es: (i) una niña en edad escolar; (ii) que vive en la región Altiplano; (iii) que su lengua materna no es originaria; (iv) que no vive en área rural; (v) que no ha iniciado su actividad sexual; (vi) que no sufre ningún tipo de violencia; y (vii) manteniendo las demás variables constantes (no pertenece a una nación o pueblo indígena originario campesino, no presencia violencia entre sus padres, no ha migrado, no tiene hijos), cuya probabilidad de abandonar la escuela es baja: 6,53% según el modelo RNA y 12,54% según el modelo Stacking, se ha modificado cada una de las cuatro variables que inciden con mayor fuerza en el abandono escolar, se ha observado e interpretado los resultados.

Según el modelo RNA, la probabilidad de que una niña del Altiplano cuya lengua materna sea aymara o quechua, solamente por esta característica, tiene una probabilidad de 33,42% de abandonar la escuela, es decir, su probabilidad de deserción escolar se incrementa en 26,89% respecto a una niña en la misma región cuya lengua materna no es originaria. Según el modelo Stacking la probabilidad de abandono de esta niña sería 34,27%, es decir, 21,73% más que su perfil inicial.

El escenario más probable en la realidad, es que las variables “lengua_originaria” y “rural” estén relacionadas. Se observa que, según el modelo RNA, la probabilidad de una niña del occidente boliviano, que tiene lengua materna originaria y que vive en área rural, tiene una probabilidad 40,21% más alta de abandonar la escuela, que una niña que habla castellano y que vive en área urbana; según el modelo Stacking, esta probabilidad es 32,84% más alta que la de su par en área urbana.

Finalmente, la misma niña del ejemplo anterior, si ya ha iniciado su actividad sexual y si sufre algún tipo de violencia, según el modelo RNA, tiene una probabilidad 79,35% mayor que una niña en área urbana, que habla castellano, que no ha iniciado su vida sexual y que no sufre violencia; de acuerdo al modelo Stacking, esta probabilidad es 72,86% más alta que la de su par en la ciudad.

3 Capítulo 3. – RESULTADOS

3.1 Conclusiones

La investigación resulta altamente pertinente en el contexto boliviano, dado que aborda una problemática crítica como el abandono escolar desde una perspectiva predictiva. Al incorporar variables sociodemográficas y utilizar datos oficiales del INE, el estudio aporta evidencia empírica novedosa que permite anticipar riesgos individuales y ofrece un insumo valioso para el diseño de políticas públicas educativas más preventivas y focalizadas.

El uso de técnicas de minería de datos y algoritmos de aprendizaje supervisado —incluyendo regresión logística, árbol de decisiones, naive Bayes, redes neuronales artificiales y ensamble mediante Stacking— demuestra la solidez metodológica del proyecto. La aplicación de procedimientos como el balanceo de clases, la optimización de hiperparámetros y la validación

cruzada garantiza la calidad del modelo final y refuerza la confiabilidad de los resultados obtenidos.

El modelo predictivo desarrollado tiene un potencial significativo de aplicación práctica en el sistema educativo y en instancias de protección social. Al ilustrar cómo interactúan distintos factores de riesgo —lengua originaria, ruralidad y violencia— desde un enfoque interseccional, la investigación ofrece una herramienta que puede orientar intervenciones más integrales y efectivas, contribuyendo a reducir la vulnerabilidad de niñas, niños y adolescentes frente al abandono escolar.

Este trabajo constituye un aporte pionero en Bolivia al vincular el machine learning con una problemática social de alto impacto. La investigación evidencia que las técnicas de aprendizaje automático no solo son útiles en campos tecnológicos o empresariales, sino que también pueden generar conocimiento aplicable a la transformación de políticas públicas y a la mejora de condiciones sociales. De esta manera, se abre un camino para futuras investigaciones que integren ciencia de datos con desafíos estructurales en educación, salud y protección social.

3.2 Insights

La naturaleza del dataset constituye la principal limitación para alcanzar niveles de desempeño superiores a un Accuracy = 0,75. Con base en los hallazgos de esta investigación, se recomienda ampliar la base de datos mediante el uso de registros administrativos del sistema educativo (anonimizados), lo que podría aportar cientos de miles, e incluso millones de observaciones. Con ello sería posible construir modelos más consistentes y alcanzar niveles de rendimiento más elevados.

Los coeficientes de la regresión logística del ensamble Stacking muestran que los modelos base con mayor contribución son la regresión logística simple (2,79) y el Random Forest (1,53), mientras que el SVC aporta únicamente 0,43. Para futuros ejercicios orientados a profundizar este análisis, se sugiere explorar nuevas combinaciones de algoritmos mediante la técnica de Stacking, buscando un mejor desempeño y una contribución más equilibrada de los modelos base.

El análisis de la importancia de los atributos en la predicción del abandono escolar revela que el sistema educativo boliviano, históricamente, ha sido y continúa siendo discriminador hacia las poblaciones de habla aymara, quechua, guaraní y otras lenguas originarias. Resulta prioritario impulsar una reforma educativa que lo haga más inclusivo, de manera que niñas y niños puedan: (i) acceder a conocimiento y formación integral en su lengua nativa; y (ii) contar con mecanismos efectivos de castellanización y aprendizaje de lenguas extranjeras a partir de sus lenguas originarias. El hecho de que la lengua materna de miles de estudiantes sea originaria no debería constituir un factor determinante para el abandono escolar.

La violencia contra niñas, niños y adolescentes también se refleja en los resultados de la investigación, tanto en su forma explícita —representada por la variable viol_ninez— como en la violencia sexual —representada por la variable ias_<17—, ambas con alta incidencia en el abandono escolar según los modelos entrenados. La recomendación en este ámbito es promover programas e intervenciones, tanto desde el sector público como desde la cooperación internacional, que fortalezcan la prevención de toda forma de violencia contra niñas, niños y adolescentes. Estas acciones deben incluir medidas directas y concretas, así como iniciativas de incidencia política orientadas a ajustar el marco normativo y las políticas públicas vigentes, haciéndolas más efectivas.

Finalmente, los modelos de clasificación entrenados y evaluados poseen un importante potencial de aplicación práctica en el ámbito educativo y de protección, ya que permiten ilustrar cómo interactúan los factores de riesgo desde un enfoque interseccional. Es decir, cómo

distintas vulnerabilidades que afectan a niñas, niños y adolescentes se combinan y potencian entre sí: la lengua originaria se suma a la ruralidad, y ambas se entrelazan con expresiones de violencia, generando una vulnerabilidad mayor que desemboca en el abandono escolar. Se recomienda fomentar el uso de modelos de clasificación en instancias del sector educativo y de protección, como herramienta para profundizar en el enfoque de interseccionalidad que actualmente guía el trabajo de muchas agencias de cooperación y entidades públicas.

4 Referencias

gob.pe. (10 de Septiembre de 2024). Se incrementó asistencia escolar de educación primaria y secundaria. *Plataforma Digital Única del Estado Peruano*. Obtenido de https://www.gob.pe/institucion/inei/noticias/1020110-se-incremento-asistencia-escolar-de-educacion-primaria-y-secundaria

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning* (Second ed.). Standford California: Springer.

INE. (2023). Estadísticas sociales - Educación. *Bolivia: Tasa de asistencia de la población entre 6 y 19 años de edad por sexo, según área, 2011 - 2023*. Bolivia.

Jiménez-Gutiérrez, A. L., Mota-Hernández, C. I., Mezura-Montes, E., & Alvarado-Corona, R. (2024). Application of the performance of machine learning techniques as support in the prediction of school dropout. *Scientific reports*.

López Laime, M. W. (2024). *Análisis de indicadores del rendimiento académico en municipios de Bolivia, por agrupamiento y Machine Learning. Monografía presentada para obtener el certificado de diplomado estadística aplicada a la toma de decisiones-Segunda versión.* Cochabamba: Universidad Mayor de San Simón.

LV/12. (7 de Septiembre de 2023). Argentina: casi un 98% de niños y niñas, asisten a la escuela. *Portal web de radio LV/12*. Obtenido de https://www.lv12.com.ar/ninos/argentina-casi-un-98-ninos-y-ninas-asisten-la-escuela-n145073#:~:text=El%2097%2C6%25%20de%20la%20poblaci%C3%B3n%20en%20edad%20escolar%2C,trav%C3%A9s%20de%20la%20Encuesta%20Permanente%20de%20Hogares%20%28EPH%29.

Ministerio de Educación, Centro de Estudios. (16 de Octubre de 2024). Reporte nacional de asistencia agosto de los años 2018, 2023 y 2024. *Apuntes 68*. Obtenido de https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/21242/APUNTES%2068_2024_fd01.pdf?sequence=1

Morales, J., & Vargas, Y. (2018). *Determinantes de la deserción escolar y el trabajo adolescente en Bolivia.* La Paz: UPB.

Satalaya, J. M. (2024). El machine learning para abordar el abandono escolar: Una revisión de los modelos más innovadores. *Ciencia Latina Internacional, 8*(6), 10993 - 11027.

Uldall, J. S., & Rojas, C. G. (2021). Una aplicación de aprendizaje automático (machine learning) en políticas públicas. Predicción de alerta temprana de deserción escolar en el sistema de educación pública de Chile. *Multidisciplinary Business Review, 15*(1), 20-35.

UNICEF. (2011). *Iniciativa Mundial Niños y Niñas Fuera de la Escuela.*