Maestría en:

CIENCIA DE DATOS Y MÁQUINAS DE APRENDIZAJE CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

Trabajo previo a la obtención de título de Magister en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

AUTOR/ES:

Manangon Perugachi Pedro Daniel

Martínez Llivicura Luis Felipe

Navarrete López Oscar Fabricio

Salaar Franco Marco Andrés

Villavicencio Proaño María José

TUTOR/ES:

Alejandro Cortés

Karla Mora

TEMA:

Construcción de un modelo predictivo mediante técnicas de machine learning para la proyección de las tasas de deserción escolar de Ecuador en el Periodo (2009–2025)

Quito - Ecuador

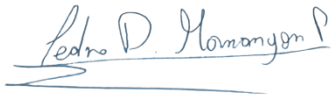
2025

REINVENTEMOS EL FUTURO

Certificación de autoría

Nosotros, Manangon Perugachi Pedro Daniel, Martínez Llivicura Luis Felipe, Navarrete López Oscar Fabricio, Salazar Franco Marco Andrés, Villavicencio Proaño María José, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



Firma

Manangon Perugachi Pedro Daniel

Firma

Martínez Llivicura Luis Felipe

Firma

Navarrete López Oscar Fabricio

Firma

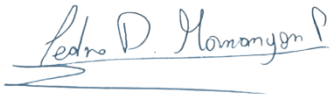
Salazar Franco Marco Andrés

Firma
Villavicencio Proaño María José

Autorización de Derechos de Propiedad Intelectual

Nosotros, Manangon Perugachi Pedro Daniel, Martínez Llivicura Luis Felipe, Navarrete López Oscar Fabricio, Salazar Franco Marco Andrés, Villavicencio Proaño María José, en calidad de autores del trabajo de investigación titulado ***Construcción de un modelo predictivo mediante técnicas de machine learning para la proyección de las tasas de deserción escolar de Ecuador en el Periodo (2009–2025),*** autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, enero de 2026



Firma

Manangon Perugachi Pedro Daniel

Firma

Martínez Llivicura Luis Felipe

Firma

Navarrete López Oscar Fabricio

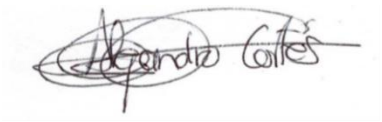
Firma

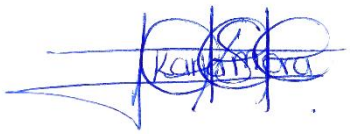
Salazar Franco Marco Andrés

Firma
Villavicencio Proaño María José

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés Director EIG y Karla Mora Coordinadora UIDE ,** declaramos que: **Manangon Perugachi Pedro Daniel, Martínez Llivicura Luis Felipe, Navarrete López Oscar Fabricio, Salazar Franco Marco Andrés, Villavicencio Proaño María José,** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.





---------- ----------

Alejandro Cortés López

Director de la

Maestría en Ciencia de Datos y Maquinas de Aprendizaje con Mención en Inteligencia Artificial

Karla Estefanía Mora Cajas

Coordinadora de la

Maestría en Ciencia de Datos y Maquinas de Aprendizaje con Mención en Inteligencia Artificial

RESUMEN

El Abandono Escolar se considera como uno de los factores con mayor desafío a ser superados por su persistencia debido a varios factores en el sistema educativo del Ecuador, siendo un determinante en el desarrollo socioeconómico del país. Esta investigación se centra en la evolución del abandono escolar ecuatoriano en el periodo 2009-2025 y proyecciones hasta el año 2030. El estudio emplea técnicas de Machine Learning para identificar patrones críticos y determinantes tales como variables geográficas y poblacionales.

La metodología para emplear se enfoca en la recopilación de datos históricos reales, el preprocesamiento de variables requeridas y el entrenamiento de modelos predictivos para estimar las tasas de abandono futuro. Los resultados preliminares obtenidos en el desarrollo nos permiten observar factores como la ubicación geográfica, la temporalidad y otros que se detallan perfectamente, siguen siendo de suma importancia para las predicciones a llevarse a cabo, mientras que las proyecciones a generar al 2030 permiten anticipar escenarios considerados de riesgo para la toma de decisiones preventivas, si asi lo fuera necesario, mediante el uso de una herramienta tecnológica potente y valiosa que conlleva al análisis de cada una de las variables a estudiar y discernir aquellas que no aporten al presente estudio o a sus objetivos planteados.

Los resultados, permitirán llegar a conclusiones que señalen donde hay un posible deterioro fuerte o débil en las tasas de abandono escolar que reflejen la realidad del Ecuador**.**

Palabras Clave: *deserción escolar, aprendizaje automático, proyecciones, variables educativas, modelos predictivos.*

ABSTRACT

School dropout is considered one of the most challenging factors to overcome due to its persistence within the Ecuadorian educational system, acting as a key determinant in the country's socioeconomic development. This research focuses on the evolution of school dropout in Ecuador during the 2009–2025 period, including projections up to the year 2030. The study employs Machine Learning techniques to identify critical patterns and determinants, such as geographical and population variables.

The methodology focuses on the collection of real historical data, the preprocessing of required variables, and the training of predictive models to estimate future dropout rates. Preliminary results obtained during the development phase indicate that factors such as geographical location, temporality, and others described in detail remain of paramount importance for the predictions. Furthermore, the projections generated for 2030 allow for the anticipation of risk scenarios, facilitating preventive decision-making through a powerful and valuable technological tool. This tool enables the analysis of each variable studied, discerning those that do not contribute to the current study or its established objectives. Finally, the results will lead to conclusions that pinpoint areas of strong or weak deterioration in dropout rates, reflecting the educational reality of Ecuador.

Keywords*: school dropout, machine learning, projections, educational variables, predictive models.*

INDICE

FIGURAS

CAPITULO 1

1. INTRODUCCIÓN

1.1. Definición del proyecto

Construcción de un modelo predictivo mediante técnicas de machine learning para la proyección de las tasas de deserción escolar de Ecuador en el Periodo (2009–2025)

Desarrollo de modelo de predicción para determinar el abandono escolar mediante técnicas de ML

1.2. Justificación e importancia del trabajo de investigación

En lo referente a la educación se considera un tema relevante, ya sea, de manera nacional como internacional. Las instituciones y gobiernos hacen su esfuerzo para colocar a la educación como un derecho primordial. La educación es un pilar fundamental para el bienestar de los seres humanos en cualquier ámbito, de forma que puedan adquirir conocimiento y las destrezas necesarias para el desempeño eficiente en sus actividades profesionales y personales.

En los dos últimos años el modelo de enseñanza en todo el mundo se ha transformado como consecuencia de la pandemia por COVID 19, a nivel general todos los países tuvieron que suspender sus actividades educativas de manera presencial y una gran mayoría optaron por incurrir en la educación de forma virtual. (Arteaga-Hernández, 2024)

La deserción escolar se ha consolidado como una problemática central dentro del ámbito educativo debido a su impacto en la continuidad formativa y en las oportunidades de desarrollo social y laboral de los estudiantes. Su persistencia evidencia desigualdades estructurales y brechas de acceso que afectan a gran parte de la población estudiantil, especialmente en contextos vulnerables. Por ello, investigar el fenómeno resulta indispensable para comprender su origen, evolución y las alternativas que permitan disminuirlo, siendo la permanencia escolar un

componente clave para garantizar un proceso educativo efectivo y equitativo. (UNESCO, s.f. (sin fecha))

Desde el plano conceptual, la deserción se define como el abandono anticipado de la educación formal antes de culminar el grado o nivel esperado, lo cual interrumpe el progreso académico del estudiante (UNESCO). Este abandono puede manifestarse como una separación temporal o como una desvinculación definitiva del sistema. En lugar de interpretarse como un hecho repentino, la deserción se reconoce como un proceso gradual y acumulativo, donde el estudiante experimenta señales previas que reflejan su distanciamiento del entorno escolar. Dichas señales suelen relacionarse con el bajo desempeño, la inasistencia continua, el desinterés y la falta de integración en la dinámica institucional (Rumberger, 2011)

En América Latina la pandemia provocó varios efectos, como consecuencia los gobiernos adoptaron medidas para mantener un distanciamiento social y así evitar los contagios masivos, dadas las circunstancias, las instituciones educativas y los métodos de enseñanza tradicionales sufrieron un gran cambio, orientando su manera de impartir clases hacia la educación virtual, que más allá de generar una solución provocó el surgimiento de grandes dificultades, haciendo notar el déficit en infraestructura tecnológica y el desconocimiento de las nuevas tecnologías de la información (TIC). (Arteaga-Hernández, 2024)

Bajo este contexto el problema global que afronta la educación ecuatoriana en tiempos de pandemia es la deserción estudiantil. La educación ha presentado este problema varias décadas atrás, y con la llegada de la pandemia se elevó este índice de una manera descontrolada. (Gómez-Ramírez)La deserción es una problemática que lleva al fracaso en todos los ámbitos del ser humano, sea personal o institucional, debido a que influye en la autoestima y el desempeño en la sociedad impidiendo el desarrollo económico y social.

Varias investigaciones han puesto en evidencia la problemática que se suscita con la educación, durante el cese de las actividades escolares de forma presencial, se calcula que aproximadamente unos 1700 millones de estudiantes dejaron de ir definitivamente a la escuela a nivel global (Prados, 2022)

Diversos estudios coinciden en que el abandono no responde únicamente a características individuales, sino que surge de la interacción entre diversos factores. Según Fortin et al. (2013), tanto la realidad personal del estudiante como su contexto familiar y las condiciones de la institución educativa influyen en la continuidad o el retiro escolar. En la misma línea, Gómez y Belmonte (2020) destacan que las tasas de deserción suelen ser mayores en territorios vulnerables, donde las dificultades económicas y el limitado acceso a recursos educativos aumentan el riesgo de abandono. De esta forma, la deserción se explica mejor desde una mirada multidimensional, en la que confluyen variables sociales, pedagógicas, económicas y emocionales.

Concebir la deserción como un proceso que avanza gradualmente permite entender que muchos casos pueden evitarse si se actúa a tiempo. Tradicionalmente, la intervención se realizaba una vez que el estudiante ya había dejado el sistema, lo cual reduce notablemente la posibilidad de revertir la situación. Bajo un enfoque actual, se prioriza la detección temprana del riesgo, favoreciendo intervenciones oportunas que aumenten la permanencia y el éxito escolar. (Rumberger R. , 2011) (UNESCO)

Comprender los factores importantes asociados con la deserción estudiantil es fundamental para las instituciones y los estudiantes. Las instituciones tienen la responsabilidad de ayudar a los estudiantes a tener éxito. Las altas tasas de deserción indican que muchos estudiantes no reciben el apoyo que necesitan, ya sea académico o financiero. Detectar estos factores puede permitir a las instituciones actuar de forma temprana para controlar o reducir las tasas de deserción, ayudar a los estudiantes a continuar sus estudios y allanar el camino hacia mejores perspectivas personales y profesionales.

Investigaciones para predecir el abandono escolar se han centrado en el uso de técnicas de aprendizaje automático para lograr una mejor precisión en las predicciones y encontrar factores que afecten al riesgo de abandono. Algunos estudios utilizaron modelos de aprendizaje automático para analizar datos de una o varias instituciones específicas, mientras que otros adoptaron un enfoque meta-analítico y examinaron la literatura existente para identificar los factores más comunes (Romero, 2025)

El incremento de datos académicos disponibles —calificaciones, asistencia, participación, registros socioeconómicos y uso de plataformas digitales— facilita la integración de soluciones analíticas más precisas. Con ello, se fortalecen los Sistemas de Alerta Temprana (EWS), los cuales han demostrado ser un medio efectivo para identificar estudiantes en riesgo y prevenir el abandono antes de que ocurra (UNESCO, 2021). Por tanto, la construcción de modelos predictivos resulta relevante no solo a nivel investigativo, sino como una herramienta práctica con alto potencial para mejorar indicadores educativos, promover la equidad y garantizar trayectorias formativas continuas.

Es importante conocer las tasas de deserción estudiantil en las diferentes provincias del Ecuador, tomando en cuenta el tipo de institución educativa a la que pertenecen los alumnos, de igual manera, conocer cómo influye el género y grado en la descensión estudiantil.

1.3. Alcance

Geográfico: El presente proyecto tiene como eje principal el estudio de los alumnos que se han matriculado en las diferentes Entidades Institucionales del Ecuador, por región, que indique un mayor número de alumnos que han desertado en un periodo determinado de tiempo, y de ello, obtener la principal Provincia o provincias que señalen cual ha sido el que mayor porcentaje de abandono escolar generado.

Temporal: El estudio que se llevara a cabo cuenta con datos emitidos por el Ministerio de Educación entre los años 2009 al 2025, lo cual nos permite obtener información histórica y actualizada, con el fin de poder recabar una data valiosa y con un historial amplio de los principales indicadores que representen la obtención de excelentes estimadores para nuestro estudio.

Poblacional: La muestra poblacional la podemos definir como el estudio de alumnos en Ecuador que abarcan el periodo 2009 al 2025 en entidades Fiscales, Particulares, Fiscomisional y Municipal en los primeros años de educación, siendo el enfoque, las mayores tasas de abandono en determinadas provincial del país.

Metodológico: Aplicación de técnicas de Machine Learning con modelos de regresión supervisados para la predicción del abandono escolar. Desarrollo de un prototipo de modelo predictivo basado en algoritmos como Random Forest, Redes Neuronales, etc.

Herramientas: Uso exclusivo de Python y sus respectivas librerías en el empleo de modelos de Machine Learning, para la recolección, preprocesamiento, análisis de datos, desarrollo, evaluación de modelos y predicciones.

Modalidad: Una de las variables que abarca el data set es la modalidad, la cual es una de las variables predictoras – categórica la cual al tener varias clases no se realizara algún tipo de filtro para no alterar los datos existentes.

1.4. Objetivos

1.4.1. Objetivo general

Analizar los factores determinantes del abandono escolar en Ecuador entre 2009 y 2025 a nivel provincial, y desarrollar un modelo predictivo mediante técnicas de machine learning que permita proyectar la probabilidad de deserción escolar.

Desarrollar un modelo predictivo mediante técnicas de machine learning que permita proyectar la probabilidad de deserción escolar en Ecuador a nivel de provincias más importantes como Quito Guayaquil y Cuenca entre 2009 y 2025.

1.4.2. Objetivo especifico

1. Identificar y sistematizar las principales variables relacionadas con el abandono escolar en las diferentes provincias del Ecuador durante el período 2009–2025 emitidos por el Ministerio de Educación.
2. Realizar un análisis exploratorio y estadístico de los datos para determinar patrones, tendencias y correlaciones entre las tasas de abandono escolar y los factores asociados en cada provincia.
3. Diseñar y entrenar modelos predictivos de machine learning para estimar la probabilidad de abandono escolar a nivel provincial.
4. Evaluar y validar los modelos construidos utilizando métricas de desempeño y seleccionar el modelo más adecuado para la proyección de escenarios futuros.4
5. Implementar un modelo predictivo que facilite la visualización por provincia de los resultados obtenidos

CAPITULO 2:

2. REVISIÓN DE LITERATURA

1.1. Estado del Arte

La deserción escolar se ha consolidado como una problemática central dentro del ámbito educativo debido a su impacto en la continuidad formativa y en las oportunidades de desarrollo social y laboral de los estudiantes. Su persistencia evidencia desigualdades estructurales y brechas de acceso que afectan a gran parte de la población estudiantil, especialmente en contextos vulnerables. Por ello, investigar el fenómeno resulta indispensable para comprender su origen, evolución y las alternativas que permitan disminuirlo, siendo la permanencia escolar un componente clave para garantizar un proceso educativo efectivo y equitativo. (UNESCO, 2012)

Desde el plano conceptual, la deserción se define como el abandono anticipado de la educación formal antes de culminar el grado o nivel esperado, lo cual interrumpe el progreso académico del estudiante (UNICEF & UNESCO Institute for Statistics, 2012). Este abandono puede manifestarse como una separación temporal o como una desvinculación definitiva del sistema. En lugar de interpretarse como un hecho repentino, la deserción se reconoce como un proceso gradual y acumulativo, donde el estudiante experimenta señales previas que reflejan su distanciamiento del entorno escolar. Dichas señales suelen relacionarse con el bajo desempeño, la inasistencia continua, el desinterés y la falta de integración en la dinámica institucional. (Rumberger R. , 2011)

Diversos estudios coinciden en que el abandono no responde únicamente a características individuales, sino que surge de la interacción entre diversos factores. Según (Fortin, 2023), tanto la realidad personal del estudiante como su contexto familiar y las condiciones de la institución educativa influyen en la continuidad o el retiro escolar. En la misma línea, (Gómez-Ramírez) destacan que las tasas de deserción suelen ser mayores en territorios vulnerables, donde las dificultades económicas y el limitado acceso a recursos educativos aumentan el riesgo de

abandono. De esta forma, la deserción se explica mejor desde una mirada multidimensional, en la que confluyen variables sociales, pedagógicas, económicas y emocionales.

Concebir la deserción como un proceso que avanza gradualmente permite entender que muchos casos pueden evitarse si se actúa a tiempo. Tradicionalmente, la intervención se realizaba una vez que el estudiante ya había dejado el sistema, lo cual reduce notablemente la posibilidad de revertir la situación. Bajo un enfoque actual, se prioriza la detección temprana del riesgo, favoreciendo intervenciones oportunas que aumenten la permanencia y el éxito escolar. (UNICEF & UNESCO Institute for Statistics, 2012)

En respuesta a esta necesidad, el uso de modelos predictivos se posiciona como una herramienta estratégica para anticipar el abandono escolar. Mediante técnicas estadísticas y algoritmos de aprendizaje automático, es posible evaluar patrones de comportamiento, estimar la probabilidad de deserción y establecer alertas preventivas, de manera que las instituciones puedan tomar decisiones basadas en evidencia (UNESCO, 2021). Estas técnicas permiten priorizar recursos, orientar tutorías, diseñar planes de acompañamiento y, en última instancia, reducir la pérdida estudiantil.

El incremento de datos académicos disponibles —calificaciones, asistencia, participación, registros socioeconómicos y uso de plataformas digitales— facilita la integración de soluciones analíticas más precisas. Con ello, se fortalecen los Sistemas de Alerta Temprana (EWS), los cuales han demostrado ser un medio efectivo para identificar estudiantes en riesgo y prevenir el abandono antes de que ocurra. (UNICEF, 2017). Por tanto, la construcción de modelos predictivos resulta relevante no solo a nivel investigativo, sino como una herramienta práctica con alto potencial para mejorar indicadores educativos, promover la equidad y garantizar trayectorias formativas continuas.

2.2. Marco Teórico

Fundamentos de Machine learning

Los términos inteligencia artificial, machine learning y deep learning son frecuentemente usados como sinónimos sin embargo existe una clara distinción y jerarquía entre ellos (Lange, 2024) es importante resaltar que los métodos de machine learning corresponden a una parte de las múltiples estrategias que combinadas permiten extraer información, entender y sacar el máximo provecho a los datos (Amat, 2020)

En los últimos años, la aplicación de machine learning es de gran interés, ha experimentado tal expansión, convirtiéndose en una disciplina que aplicada prácticamente en todos los ámbitos de investigación académica e industrial. Cada vez más personas se dedican a esta disciplina dando como resultado un amplio repertorio de herramientas con las que, perfiles con diferentes niveles de especialización, pueden acceder a métodos predictivos potentes (Amat, 2020)

Machine Learning

Es importante analizar lo que es y no es, referentes a los métodos de Machine Learning, principalmente clasificarse como un subcampo de la inteligencia artificial, aunque esta categorización puede resultar engañoso a primera vista. A pesar de que el machine learning surgió, sin duda de la investigación en este contexto, las aplicaciones de los métodos de machine learning en tratamiento de datos resulta más útil pensar como un medio para construir modelos de datos.

Fundamentalmente machine learning implica la construcción de modelos matemáticos a través de potentes algoritmos para ayudar a comprender los datos, por medio de este aprendizaje los modelos se pueden ajustar automáticamente a través de parámetros que se adaptan a los datos observado, con los modelos ajustados a los datos previamente observados, se puede utilizar para predecir y comprender aspectos de nuevas observaciones (VanderPlas, 2017)

El campo de machine learning surgió en un entorno donde la disponibilidad de datos, los métodos estadísticos, junto con la capacidad de procesamiento evolucionaron rápida y simultáneamente, el creciente volumen de datos requirió mayor capacidad de procesamiento impulsando a su vez el desarrollo de métodos estadísticos para analizar grandes conjuntos de datos, creando un ciclo de avance que permitió la recopilación de cada vez más grandes e interesantes (Lantz, 2013)

Machine learning y su proceso de trabajo

En términos generales un problema típico de machine learning sigue los siguientes pasos:

Definición del problema: necesitamos determinar exactamente cuál es el problema antes de iniciar a resolverlo, sí es que es factible de resolverlo usando machine learning o no.

Recolección de datos: la recolección de datos debe estar basada en la definición del problema, los datos son un aspecto extremadamente importante por tanto deben ser recopilados con cuidado, asegurándonos de tener datos que correspondan con los factores necesarios para nuestro análisis.

Preprocesamiento de datos: para que los datos sean más útiles deben ser depurados, incluyendo tratamiento de valores atípicos, gestión de valores faltantes entre otros, lo cual permitirá reducir posibles errores de nuestro análisis.

Desarrollo del modelo: en esta etapa podemos crear nuestro modelo de machine learning, que se usará para resolver nuestro problema, el modelo usa los datos como entrada, realiza los cálculos respectivos y genera un resultado o salida a partir de ellos.

Evaluación del modelo: Es fundamental evaluar nuestro modelo, verificando su precisión, se debe garantizar que el modelo funcione correctamente con cualquier conjunto nuevo de datos proporcionado. (Silaparasetty, 2020)

Machine learning y tipo de datos

Según (Silaparasetty, 2020), considerando el tipo de datos utilizados, tenemos dos tipos principales de métodos de machine learning:

- Aprendizaje supervisado: método que utiliza datos de etiquetados.
- Aprendizaje no supervisado: método que utiliza datos sin etiquetar.

La tabla 2.1 muestra algunas diferencias entre estos dos métodos de aprendizaje.

Aprendizaje supervisado	**Aprendizaje no supervisado**
Utiliza datos etiquetados.	Utiliza datos sin etiquetar.
No requiere un exceso de datos para lograr precisión.	Requiere un exceso de datos para lograr precisión.
La complejidad computacional es menor, es decir, es más simple.	La complejidad computacional es mayor, es decir, es menos simple.
No encuentra patrones por sí solo en un conjunto de datos.	Encuentra patrones por sí solo en un conjunto de datos determinado

Tabla 2.1: Diferencias entre aprendizaje supervisado y no supervisado

Cada método de aprendizaje cuenta con varios tipos de algoritmos, los cuales pueden usarse para resolver un problema de machine learning. Describiremos las características principales de estos métodos de aprendizaje destacando el aprendizaje supervisado, es el cual es utilizado en este trabajo.

Aprendizaje no supervisado

El aprendizaje no supervisado implica modelar las características de un conjunto de datos sin tener como referencia previa ninguna etiqueta, suele describirse esta característica como: “dejar que el conjunto de datos hable por sí mismo". Tales modelos incluyen tareas como la agrupación en clústeres hoy la reducción de dimensionalidad (VanderPlas, 2017)

Análisis clúster: estos algoritmos identifican distintos grupos de datos, en función de sus similitudes

Reducción de dimensionalidad: los algoritmos buscan representaciones más concisas de los datos manteniendo su relevancia.

Algunos de los algoritmos más importantes usados en aprendizaje no supervisado son (Géron, 2017):

- Clustering
- k-Medias
- Análisis Jerárquico de Agrupamientos (HCA)
- Maximización de Expectativas
- Visualización y reducción de dimensionalidad
- Análisis de Componentes Principales (PCA)
- Kernel PCA
- Incrustación Localmente Lineal (LLE)
- Incrustación Estocástica de Vecinos con Distribución T (t-SNE)
- Aprendizaje de reglas de asociación
- A priori
- Eclat

Aprendizaje Supervisado

El aprendizaje supervisado es utilizado cuando se tienen variables que necesitamos predecir usando otras variables. Situaciones como la regresión lineal, donde tenemos algunas variables de entrada, por ejemplo, x, y se desea un modelo que prediga variables de salida (o respuesta), y = f(x) (Mailund, 2017)

Una vez obtenido el modelo, puede utilizarse para aplicar etiquetas a datos nuevos y desconocidos. Suele subdividirse a su vez en tareas de clasificación y tareas de regresión, en la

clasificación, las etiquetas son categorías discretas, mientras que, en la regresión, son cantidades continuas (VanderPlas, 2017)

Regresión y clasificación

Tenemos dos tipos de aprendizaje supervisado: Regresión y clasificación

Regresión

Principalmente se utiliza cuando la variable de salida buscada es un número, mientras que la clasificación se utiliza cuando buscamos una variable categórica (Mailund, 2017)

Consideramos la regresión lineal (ecuación 2.1), es regresión porque la variable que estamos tratando de buscar es un número. La clase parametrizada de funciones, fθ, son todas lineales, matemáticamente se describe:

$y = \alpha x + ß$, o también $t = \alpha x + ß + \varepsilon$ (2.1)

Haciendo θ = θ1, θ0 y α = θ1, ß = θ0, entonces:

$y(\theta) = f(x; \theta) = \theta 1 (x) + \theta 0$ (2.2)

Ajustar un modelo lineal consiste en encontrar el mejor θ, donde mejor es considerado al θ que acerca y(θ) a t. En regresión lineal la medida de distancia utilizada es la euclidiana al cuadrado.

$\left|y^{(\theta)} - t\right|^2 = \sum_{i=1}^{n} (y_i(\theta) - t_i)^2$ (2.3)

Se utiliza la distancia al cuadrado en lugar de solamente la distancia principalmente por una conveniencia matemática, de esta manera es más fácil maximizar θ, también relacionar la interpretación del término ε distribuido normalmente se utiliza la distancia al cuadrado en lugar de solamente a la distancia.

Cuando ajustamos datos en una regresión lineal, esta distancia es la que se minimiza luego se encuentran los parámetros θ que mejor ajustan los datos en el sentido de:

$$\hat{\theta} = arg\ min_{\theta_1\theta_0} \quad \sum_{i=1}^{n} \quad (\theta_i(x_i) + \theta_0 - t_i)^2 \quad (2.4)$$

En el caso de la clasificación se asume que los objetivos ti son binarios, pudiéndose codificar como 0 y 1, las variables de entrada xi se mantienen como números reales. Una forma común de definir la función f (-; θ) es mapear x al intervalo unitario [0, 1] e interpretar el resultado de y(θ) como la probabilidad de que t sea 1, se predice 0 si f (-; θ)<0.5 y 1 si f (-; θ)>1, En clasificación lineal la función fθ puede tener la siguiente forma:

$$f\ (x;\ \theta)\ =\ \sigma\ (\theta 1(x)\ +\ \theta 0) \qquad (2.5)$$

En (2.5) σ es una función sigmoidea, una elección común de σ es la función logística:

$$\sigma: z \rightarrow \frac{1}{1+e^{-z\prime}} \qquad (2.6)$$

En el caso de (2.6) el ajuste f (-; θ) se denomina regresión logística.

Clasificación

Clasificación involucra la obtención de modelos, mediante los cuales se ordena la información en categorías, clases o valores discretos de acuerdo con las características de los datos, se asigna etiquetas predefinidas o los datos en función de las características.

El objetivo de la clasificación es predecir resultados específicos conocidos como etiquetas de clase, analizando diferentes factores o características discretas, también se requiere definir una separación clara entre los datos lo que se conoce como límite de decisión (Chaturvedi, 2024)

Clasificadores lineales

En machine learning los clasificadores son los que identifican a que clase pertenece una nueva observación, la pertenencia a una clase se determina comparando una combinación lineal de

características con una puntuación (score). En dos dimensiones un clasificador lineal es una línea con la forma funcional:

$$w_1x_1 + w_2x_2 = c \quad (2.7)$$

La regla de clasificación de un clasificador lineal consiste en asignar hago una clase específica sí: $w_1x_1 + w_2x_2 < c$ o $w_1x_1 + w_2x_2 \geq c$.

El algoritmo puede ser escrito de la siguiente manera:

$$puntuación(Score) = \sum_{i=1}^{n} w_1x_1$$
$$Si, Score < c \quad (2.8)$$
$$return + 1$$
$$else - 1$$

Consideremos una característica cuya ocurrencia sea 'bueno' o 'malo', analizando una muestra sí del caso de que del que x_1= 'bueno', x_2 = 'malo', las correspondientes ponderaciones son $w_1 = +1$ y $w_2 = -1$, la puntuación para este caso sería $1 * 1 + (-1) * 1 = 0$. Sin embargo, para otras muestras su puntuación podría ser menor o mayor a cero. Podemos representar esta característica de bueno o malo junto con la puntuación a través de la figura 2.4, lo que se conoce como el límite de decisión.

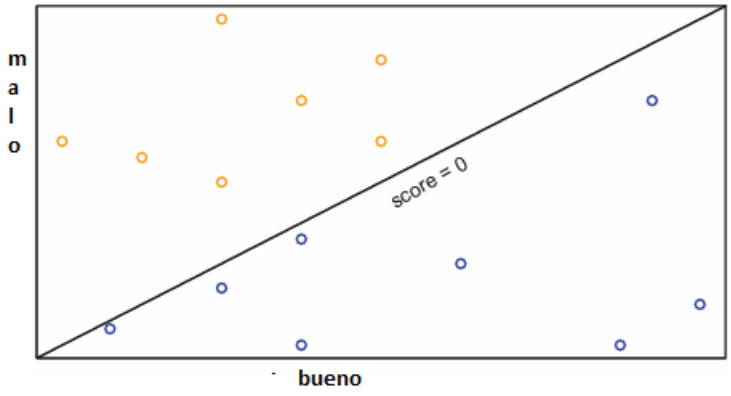


Figura 2.6: Límite de decisión lineal. Fuente: (Ghatak, 2017)

El límite de decisión lineal para estas dos entradas es la línea donde la puntuación (score) es cero, para el caso de dos entradas el límite de decisión es un plano y para más de 3 entradas es un hiperplano. También existen los límites de decisión no lineales.

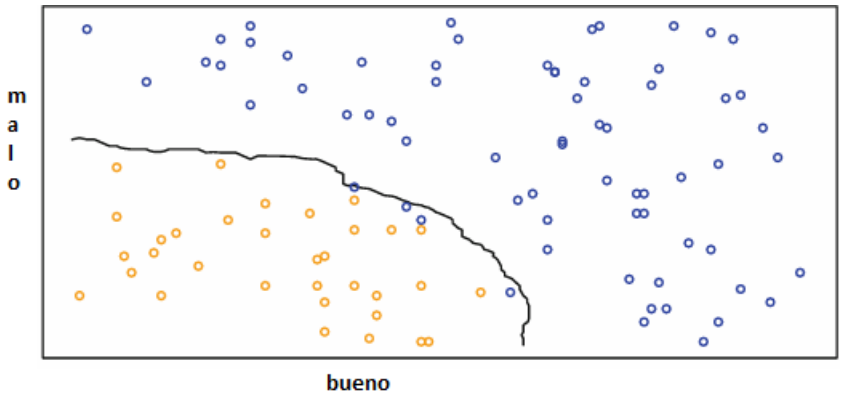


Figura 2.7: Límite de decisión no lineal. Fuente: (Ghatak, 2017)

Modelos de clasificación lineal

Un modelo de clasificación lineal es aquel que se puede entrenar para predecir el signo de la puntuación dadas las entradas, usualmente el signo positivo se representa con factor de "1" signo negativo con factor de "0", se representa de la forma:

$$\hat{y} = sign\big(Score(x_1)\big),$$

$$Score(x_1) = w_0 + w_1 x_i[1] + \cdots + w_n x_i[n] = w^T x_i \qquad (2.9)$$

En el caso de un hiperplano n-dimensional (figura 2.5) puede ser representado de la siguiente manera:

$$Score(x_1) = w_0 + w_1 x_i + \cdots + w_n x_i$$

$$= \sum_{i=0}^{n} w_j h_j x_i \qquad (2.10)$$

$$= w^T h(x_i)$$

Interpretación del Score

La puntuación $w^T h(x_i)$ puede ser interpretada como la clase predicha, además se relacionada con la probabilidad $p(y = +1 \mid x, \hat{w})$, la clase predicha se puede vincular a una escala probabilística, por tanto, se puede predecir las probabilidades que van de 0 a 1, a partir de los valores de las puntuaciones que van de $-\infty$ a $+\infty$, por ejemplo, a través de una función de enlace sigmoidea 'g':

$$\hat{P}(y = +1 \mid x_i, w) = g(w^T h(x_i))$$

Algoritmos de aprendizaje supervisado

Existen una gran variedad de algoritmos de machine learning que son utilizados en diferentes áreas del conocimiento, en esta sección se realizará una descripción de los algoritmos de aprendizaje supervisado, qué son más usuales o relevantes y algunos de los cuales se usan en este trabajo.

Modelos lineales generalizados

Muchos modelos en la vida son modelos lineales generalizados (GLM) tales como: regresión lineal múltiple, análisis de covarianza (ANOVA), modelos basados en familias exponenciales entre otros. Independientemente del tipo de datos que se desee analizar, para una modelización correcta siempre es fundamental hacerse ciertas cuestiones en función de la variable respuesta, homogeneidad de las condiciones de observación, las variables que pueden explicar algo sobre la respuesta, las respuestas a estas interrogantes son esenciales para saber por dónde abordar el análisis de los datos (Martínez, 2001)

Adicionalmente cuando se corrobora la respuesta es necesario verificar la normalidad, homogeneidad de varianza, linealidad de los efectos sistemáticos, se propone como modelización el modelo lineal normal.

$$Y = X\beta + \epsilon \qquad (2.15)$$

Donde $\epsilon \sim N(0, \sigma^2 I)$, de forma que:

$$E(Y) - \mu = X\beta$$

$$Var(Y_i) = \sigma^2$$

Sin embargo, el mundo de los datos no es perfectamente normal en muchas ocasiones los datos provienen de otras distribuciones que no pueden satisfacer los requerimientos del moderado normal.

Los modelos lineales generalizados (GLM) son una alternativa justificada por la falta de linealidad y homogeneidad de la varianza, Aquí las propiedades básicas de los estimadores como puede ser la varianza son insensibles a las distribuciones asumidas dependiendo principalmente:

- La relación asumida entre media-varianza
- Grado de independencia/incorrelación entre observaciones

Las hipótesis básicas de un modelo lineal generalizado son:

- Independencia entre las respuestas
- La respuesta media cambia con las condiciones, pero no la forma funcional de la distribución.
- La respuesta media, o alguna transformación de ella cambian de modo lineal cuando las condiciones cambien

Es como se trabaja con aquella distribución de probabilidad y aquel funcional para el cambio de la respuesta media de la variable observada, que mejora como de los datos.

Los modelos lineales generalizados (GLM) permiten especificar separadamente:

- la distribución de los datos, o lo que es básico, qué es la relación media-varianza (Función de varianza)
- las relaciones de linealidad entre la respuesta media y los predictores (función link)

Algunos elementos de un modelo lineal generalizado (GLM) son:

Las variables respuesta $Y_i, i = 1 \ldots\ldots, n$, comparten la misma distribución en la familia exponencial.

Un conjunto de variables explicativas X y de parámetros β.

Una función monótona llamada link, $g()$ proporciona el predictor lineal η.

$$g(\mu_i) = \eta_i = X_i'\beta \quad (2.16)$$
$$\text{con } E(y_i) = \mu_i$$

Modelos	**Links**
Normal	Identidad μ
Binomial	Inverso $1/\mu$
Poisson	Inverso cuadrático $1/\mu^2$
Gamma	Raíz cuadrada $\sqrt{\mu}$
Gaussiano inverso	Exponencial $(\mu + c_1)^{c_2}$ Log $log\ (\mu)$ Logit $log\ (\frac{\mu}{1-\mu})$ closlog $log\ (-log\ (\mu))$ probit $\Phi^{-1}(\mu)$

Tabla 2.2: Algunos modelos y links usuales en ajustes de GLM's. Fuente: Martínez, 2001

Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) es un potente algoritmo de aprendizaje supervisado que combina múltiples aprendices débiles en un conjunto con un excelente rendimiento predictivo (Friedman). Consideramos un problema de aprendizaje supervisado, con n ejemplos de entrenamiento $(x_i, y_i), i = 1 \ldots\ldots, n$, tal que x_i es el vector de características del i-ésimo ejemplo en R^p y $y_i \epsilon\ R$ es una etiqueta (en un problema de clasificación) o una respuesta continua (en un problema de regresión). En la versión clásica de GBM (Friedman, 2001), la predicción correspondiente a un vector de características x se da mediante un modelo aditivo de la forma:

$$f(x) \coloneqq \sum_{m=1}^{M} \beta_{j_m} b(x; \tau_{j_m}) \quad (2.16)$$

Donde cada función base $b(x;\tau) \epsilon\, R$ (también denominada función de aprendizaje débil) es una función simple del vector de características indexado por un parámetro τ, y β_j es el coeficiente de la j-ésima función de aprendizaje débil. En este caso, β_{j_m} y τ_{j_m} se eligen de forma adaptativa para mejorar la fidelidad de los datos (según una regla específica). Entre los ejemplos de funciones de aprendizaje débil comúnmente utilizados en la práctica (Hastie, Tibshirani, & Friedman, 2008) se incluyen las funciones wavelet, las máquinas de vectores de soporte, los árboles de decisión de profundidad uno y los árboles de clasificación y regresión (CART) (Breiman, 2017), etc. Suponemos que el conjunto de funciones de aprendizaje débil es finito con cardinalidad K; en muchos de los ejemplos mencionados, K puede ser exponencialmente grande, lo que genera dificultades computacionales.

El objetivo del GBM es obtener una buena estimación de la función f que minimice aproximadamente la pérdida empírica:

$$\sum_{i=1}^{n} l(y_i, f(x_i)) \quad (2.17)$$

Donde, $l(y_i, f(x_i))$ es una medida de la fidelidad de los datos para la i-ésima muestra de la función de pérdida l, que se supone diferenciable en la segunda coordenada. La versión original del GBM (presentada en el Algoritmo 1) puede considerarse como la aplicación de un algoritmo de descenso más pronunciado para minimizar la función de pérdida (2.17). El GBM parte de un modelo nulo $f \equiv 0$ y, en cada iteración, calculamos un aprendiz débil que mejor se ajusta al pseudo-residual actual (es decir, el gradiente negativo de la función de pérdida sobre la predicción) r^m, en términos de la pérdida por mínimos cuadrados, como se indica a continuación:

$$j_m = \; arg\, min_{j\in[K]} \; min_{\sigma} \; \sum_{i=1}^{n} l\left({r_i}^2 - \sigma b\left(x_i; \tau_j\right)\right)^2 \quad (2.18)$$

Donde $[K]$ es una abreviatura del conjunto $\{1, \ldots., K$. (En caso de empates en la operación "argmin", elegimos el que tenga el índice más pequeño. Luego, añadimos el $j_m{}^{th}$ modelo débil al modelo mediante una búsqueda lineal. A medida que avanzan las iteraciones, el GBM genera una secuencia de modelos $\{f^m\}_{m\in[M]}$ (indexada por el número de iteraciones del GBM), donde cada modelo corresponde a una determinada fidelidad de datos y complejidad/reducción [10, 7]; en conjunto, controlan el rendimiento del modelo fuera de muestra. La intención habitual del GBM es detenerse pronto, antes de acercarse al mínimo del Problema (2.18), con la esperanza de que dicho modelo genere un buen rendimiento predictivo.

Deep Learning

Una vez que el aprendizaje automático despegó en el mundo de la tecnología, no hubo forma de detenerlo. Cada día, cada minuto, la gente empezaba a hacer nuevos descubrimientos y a desarrollar modelos más nuevos que funcionaban mejor que los anteriores. Sin embargo, estos modelos de aprendizaje automático seguían sin ser lo suficientemente buenos. También eran bastante eficaces. Pero simplemente no eran lo suficientemente eficientes. Eso fue hasta que se logró desarrollar una técnica de aprendizaje automático que ayudaría a una máquina a resolver problemas por sí misma y, por lo tanto, a resolver problemas extremadamente complejos con gran precisión. De hecho, esta técnica se popularizó tanto que ahora se la conoce como un área independiente de la inteligencia artificial (aunque no está separada del aprendizaje automático). Pronto se le dio el nombre de "Deep Learning" (Silaparasetty, 2020)

El aprendizaje profundo es una rama del aprendizaje automático que utiliza redes neuronales artificiales para ayudar a la máquina a analizar y responder a un problema específico, aunque es muy tentador pensar en el aprendizaje profundo como un área independiente de la inteligencia artificial, definitivamente no lo es. Forma parte integral de la inteligencia artificial y es un subconjunto del aprendizaje automático.

Artificial neural network

Una red neuronal artificial (RNA), o artificial neural network (ANN), es un modelo de aprendizaje automático inspirado en el funcionamiento del cerebro humano. Está compuesta por unidades de procesamiento llamadas neuronas artificiales o nodos, que están organizadas en capas interconectadas.

Algunos conceptos clave relacionados con las redes neuronales artificiales son los siguientes (Berzal, 2019):

Neuronas Artificiales: cada neurona artificial es una unidad de procesamiento que toma una serie de entradas, las procesa y produce una salida. Las entradas se multiplican por pesos, se suman y se pasa el resultado a una función de activación.

Función de activación: introduce no linealidad en la red, permitiendo que las redes neuronales aprendan relaciones y patrones complejos en los datos. Existen varias funciones de activación comunes utilizadas en las redes neuronales, entre las que se incluyen: Sigmoide, ReLU, Tangente Hiperbólica, etc. La elección de la función de activación depende del tipo de problema y la arquitectura de la RNA. La función de activación juega un papel fundamental en la capacidad de la RNA para aprender y generalizar a partir de los datos de entrenamiento.

Capas: las neuronas se organizan en capas. Una red neuronal típica consta de una capa de entrada, una o más capas ocultas y una capa de salida. Las capas ocultas permiten a la red aprender representaciones intermedias y abstraer características de los datos.

Conexiones: cada neurona en una capa está conectada a todas las neuronas de la capa anterior y a todas las de la capa siguiente. Estas conexiones tienen pesos que se ajustan (aprenden) durante el proceso de entrenamiento.

Entrenamiento: el entrenamiento de una red neuronal implica el ajuste de los pesos de las conexiones para que la red pueda hacer predicciones precisas. Esto se hace utilizando algoritmos de optimización y un conjunto de datos de entrenamiento que contiene ejemplos de entrada y las salidas deseadas (etiquetas).

Función de Costo: la función de costo mide la discrepancia (error) entre las predicciones de la red y las salidas reales (ground truth) en el conjunto de entrenamiento. El objetivo del entrenamiento es minimizar esta función de costo.

Backpropagation: es un algoritmo utilizado para propagar el error desde la capa de salida hacia atrás a través de la red, ajustando los pesos de las conexiones en función del error cometido.

Arquitectura: las redes neuronales pueden tener diferentes arquitecturas, como redes neuronales feed-forward redes neuronales recurrentes (RNN) y redes neuronales convolucionales (CNN), cada una adaptada a tareas específicas.

Las redes neuronales artificiales se han destacado en una amplia gama de aplicaciones, incluyendo el procesamiento de imágenes, el procesamiento de lenguaje natural, la visión por computadora, la traducción automática, el reconocimiento de voz, la recomendación de contenido y muchas otras.

Su capacidad para aprender representaciones complejas y realizar tareas sofisticadas las ha convertido en un componente esencial del campo de la inteligencia artificial y el aprendizaje automático.

CAPITULO 3

3. DESARROLLO

3.1. Desarrollo del Trabajo

Enfoque general del desarrollo

Para el desarrollo de modelos de predicción de tasas de abandono escolar se trabajó desde un enfoque aplicado y experimental, basado principalmente en la construcción de un pipeline que permite transformar los datos disponibles en una base estructurada, consistente y apta para el entrenamiento de modelos de machine learning. De esta manera se enfoca en poder reproducir el proceso, las transformaciones y la prevención de sesgos que los diferentes modelos puedan tener, evitando de esa manera errores que afectarían a los resultados obtenidos.

El desarrollo se basó en trabajar con datos históricos reales provenientes del Ministerio de Educación del Ecuador, correspondientes al período 2009–2025, los cuales se encontraban distribuidos en múltiples archivos y presentaban variaciones en estructura, nombres de columnas y formatos de los registros. Esta es la razón por la que antes de una etapa de modelado fue necesario aplicar un procedimiento de estandarización, limpieza y reorganización de los datos, con el objetivo tener datos en una estructura fija para garantizar su coherencia y uso a lo largo del tiempo.

Desde el punto de vista metodológico, el problema abordado corresponde a un escenario de aprendizaje supervisado, en el cual la variable objetivo es la tasa de abandono escolar, calculada a partir de registros de estudiantes promovidos y abandonos en base al total matriculado en un periodo específico.

El proceso completo de desarrollo se basó en una secuencia de etapas, donde cada fase genera resultados para la siguiente. En esta sección se muestran diferentes etapas como un procedimiento, comenzando por la transformación inicial del formato de los datos, la definición

del conjunto de entrenamiento y, finalmente, el entrenamiento y evaluación de los modelos predictivos, los cuales servirán para generar un dashboard interactivo de información.

Transformación y estandarización inicial de los datos

La primera etapa del desarrollo consistió en la transformación del formato original de los datos educativos con el fin de construir una base que permitiera su posterior análisis y modelado. Los datos de entrada se encontraban almacenados en archivos de hojas de cálculo, los cuales presentaban diferencias en la ubicación de los encabezados, en la denominación de las columnas y en la organización de la información por año lectivo.

Para resolver esta diferencia de formato, se implementó un procedimiento que permite identificar dinámicamente la fila de encabezados correcta en cada archivo, utilizando un patrón de lectura con tolerancia a fallos. Este procedimiento asegura que los datos puedan ser procesados incluso cuando existen variaciones entre los archivos correspondientes a distintos períodos académicos.

Una vez identificada la estructura válida de cada archivo, se procedió a la normalización de las columnas, tales como año lectivo, provincia, cantón, área, sostenimiento, jornada y modalidad educativa. Dado que estas variables aparecían con diferentes nombres o variantes ortográficas, se definió un conjunto de claves posible y un se definió un nombre estándar para columna, lo que permitió mapear automáticamente las columnas originales hacia un nombre estandarizado. Este proceso garantiza la consistencia de las variables a lo largo de todo el conjunto de datos y reduce errores que surjan a partir de diferencias en los nombres originales para todos los años disponibles, permitiendo manejar todos los datos de una manera más sencilla.

Después, se identificaron las columnas relacionadas con la información de estudiantes, las cuales incluyen separaciones por sexo, condición académica (promovidos, no promovidos, abandono, no actualizados) y etapa de estudios. Estas columnas fueron analizadas mediante

patrones que permitieron extraer datos relevantes de cada columna, como el sexo del estudiante y la etapa de estudio, facilitando su reorganización en un formato más adecuado para el análisis.

Con el objetivo de preparar los datos para el modelado, se realizó una transformación desde un formato ancho (wide format), en el que cada condición de estudiantes se encontraba representada como una columna independiente, hacia un formato largo (long format). Esta transformación permitió consolidar la información y posteriormente reconstruirla mediante operaciones de agregación que producen una representación por año, provincia, características institucionales y grado académico, definiendo de esta manera características que servirán para el entrenamiento de modelos.

Durante este proceso se aplicaron algoritmos de limpieza para manejar valores faltantes, símbolos no numéricos y registros incompletos, asegurando que las variables cuantitativas pudieran ser utilizadas sin introducir inconsistencias en los cálculos posteriores. Asimismo, se verificó la coherencia entre los totales declarados y los totales calculados a partir de las distintas categorías de estudiantes, priorizando los valores consistentes y descartando registros inválidos como pueden llegar a ser valores nulos.

Como resultado de esta etapa, se obtuvo una base de datos agregada que incluye, para cada combinación de variables, el total de estudiantes, el número de casos de abandono y la tasa de abandono escolar correspondiente. Adicionalmente, se aplicó una transformación logarítmica de la tasa de abandono, la cual se utilizará en etapas posteriores del modelado para estabilizar la varianza y mejorar el comportamiento de ciertos algoritmos predictivos.

Esta base es el punto de partida del proceso de desarrollo del proceso posterior de entrenamiento y análisis de modelos, ya que garantiza que los datos de entrada cumplan con los requisitos de calidad, consistencia y estructura necesarios para un entrenamiento confiable de los modelos de machine learning.

Descripción de las variables finales del conjunto de datos

El conjunto final de datos se encuentra organizado a nivel agregado, donde cada observación representa una combinación específica de año, ubicación geográfica, características institucionales y grado educativo, junto con los indicadores cuantitativos relacionados con la matrícula y el abandono escolar.

A continuación, se describen de manera detallada las variables que conforman la base final utilizada en este estudio.

- **Año:** Corresponde al año de inicio del período lectivo, el cual sale a partir del campo original de año lectivo. Esta variable permite analizar la evolución temporal del abandono escolar y es fundamental para la construcción de series históricas y para la definición de los conjuntos de entrenamiento, validación y prueba en el proceso de modelado.

- **Provincia:** Identifica la provincia del Ecuador a la que pertenece la institución educativa. Esta variable permite capturar diferencias geográficas en las tasas de abandono escolar.

- **Cantón:** Representa el cantón donde se ubica la institución educativa. Su inclusión aporta un nivel adicional de granularidad geográfica, permitiendo identificar patrones locales de abandono escolar dentro de una misma provincia.

- **Área:** Indica el tipo de área en la que se localiza la institución educativa, clasificada como urbana o rural. Esta variable es relevante para analizar diferencias educativas relacionadas al acceso a recursos, infraestructura y conectividad.

- **Sostenimiento:** Describe el tipo de sostenimiento de la institución educativa, como fiscal, particular, fiscomisional o municipal. Esta variable permite evaluar diferencias estructurales en el abandono escolar según el tipo de unidad educativa.

- **Jornada:** Representa la jornada académica bajo la cual se trabaja en la institución, por ejemplo, matutina, vespertina o nocturna. La jornada puede influir en el abandono escolar, especialmente en los estudiantes que combinan estudio y trabajo.

- **Modalidad:** Corresponde a la modalidad educativa o tipo de oferta académica. Esta variable fue conservada en su totalidad sin aplicar filtros, debido a su carácter categórico y a su potencial valor predictivo dentro del modelo.

- **Grado:** Identifica el grado o nivel educativo al que corresponde el registro. Esta variable es fundamental para analizar la distribución del abandono escolar a lo largo de la vida educativa y detectar niveles críticos con mayor riesgo de deserción en algunas etapas del estudio.

- **Total:** Indica el número total de estudiantes matriculados. Este valor se obtiene a partir de la suma de las distintas variables tomadas en cuenta.

- **Promovidos:** Representa la cantidad de estudiantes que aprobaron el período académico correspondiente.

- **No Promovidos:** Corresponde al número de estudiantes que no lograron la promoción al siguiente nivel educativo.

- **Abandono:** Indica la cantidad de estudiantes que abandonaron los estudios durante el período analizado. El valor es el que se utiliza para calcular la tasa de abandono.

- **No Actualizados:** Representa los registros de estudiantes cuyo estado académico no fue actualizado en el sistema.

- **Tasa_Abandono:** Es una variable continua que representa el porcentaje de estudiantes que abandonaron el sistema educativo respecto al total de matriculados. Se calcula como la división entre el número de casos de abandono y el total de estudiantes matriculados, tomando valores en el rango [0, 1]. Es la variable que objetivo del problema.

- **Tasa_Log:** Corresponde a la transformación logarítmica de la tasa de abandono, aplicada con el fin de estabilizar la varianza y mejorar el comportamiento estadístico de ciertos modelos predictivos. Esta transformación resulta especialmente útil donde la tasa de abandono tiene valores cercanos a cero.

Esta estructura de datos constituye la base sobre la cual se desarrollan las etapas posteriores del modelado predictivo, incluyendo la selección de variables, la división del conjunto de datos y el entrenamiento de los modelos de machine learning.

Construcción de la base de datos para el modelado predictivo y prevención de la fuga de información

Una vez obtenida la base de datos consolidada y transformada, se procedió a la construcción de un conjunto de datos específico para el modelado predictivo. Esta etapa tuvo como objetivo principal garantizar que los modelos de machine learning fueran entrenados y evaluados bajo las mismas condiciones, evitando la información que no estaría disponible en un escenario de predicción real. Esta es la razón por la que uno de los principales riesgos a mitigar fue la fuga de información (data leakage), esto ocurre cuando el modelo accede, directa o indirectamente, a datos del futuro durante su entrenamiento.

Para abordar este riesgo, el proceso de construcción de la base de modelado se diseñó de manera explícita como una etapa independiente del procesamiento general de los datos, estableciendo una separación clara entre la base completa y la base destinada al entrenamiento y evaluación de los modelos.

Separación entre base consolidada y base de modelado

A partir de la base completa que contiene la totalidad de los registros históricos, se seleccionó un subconjunto de variables que serían utilizadas exclusivamente para el modelado predictivo. Esta selección se realizó buscando incluir únicamente variables en el mismo período lectivo o en períodos anteriores, excluyendo cualquier información que pudiera depender del

resultado futuro del abandono escolar, como puede ser tal cual el número de estudiantes que abandonaron.

La base de modelado conserva las variables temporales, geográficas, institucionales y académicas descritas en la sección anterior, junto con la variable objetivo-derivada como la tasa de abandono. De esta manera se permite que los modelos aprendan patrones históricos sin incorporar datos que cambien la capacidad de generalización del modelo y que este tenga un sobre entrenamiento.

Selección controlada de variables para el modelado

El conjunto de variables seleccionadas para la base de modelado incluye el año lectivo, la provincia, el cantón, el área, el tipo de sostenimiento, la jornada, la modalidad educativa y el grado académico, además del total de estudiantes y la variable objetivo-relacionada con el abandono escolar. Esta selección se basa en dos criterios: relevancia y disponibilidad durante todos los años disponibles y pensando en los próximos años.

Definición de conjuntos de entrenamiento, validación y prueba

Con el fin de evaluar el desempeño de los modelos de manera objetiva, se implementó una división temporal del conjunto de datos en tres subconjuntos: entrenamiento, validación y prueba. A diferencia de una división aleatoria tradicional, esta separación se realizó en función del año lectivo, lo cual resulta más apropiado en estudios con series temporales, de esta manera los modelos toman en cuenta el año como una variable de entrenamiento más.

Es así que los registros correspondientes a los años más antiguos se asignaron al conjunto de entrenamiento, los años intermedios a validación y los años más recientes al conjunto de prueba. Esta estrategia simula un escenario real de predicción, en el cual el modelo se entrena con información histórica y se evalúa con datos de períodos posteriores no observados durante el entrenamiento.

La asignación explícita de cada registro a un subconjunto específico permite un control de la información durante el entrenamiento y la evaluación, reduciendo el riesgo de fuga de datos y proporcionando una estimación más realista de los modelos.

Control de tipos de datos y consistencia estructural

Como parte de la construcción de la base de modelado, se realizó una validación adicional de los tipos de datos de cada variable, asegurando que las variables temporales, categóricas y numéricas se encuentren correctamente definidas. Este control está para evitar inconsistencias durante el entrenamiento de los modelos y para garantizar la compatibilidad con los algoritmos de machine learning utilizados.

Resultado de la etapa de construcción de la base de modelado

Como resultado de esta etapa, se obtuvo una base de datos específica para el modelado, estructurada, consistente y libre de fuga de información. Esta base constituye la base principal para las etapas posteriores de entrenamiento, ajuste y evaluación de los modelos de machine learning, permitiendo que los resultados obtenidos reflejan de manera fiel la capacidad predictiva de los modelos en escenarios reales.

Almacenamiento de la base

Al realizar todo lo necesario para obtener la base estructurada se optó por almacenar la base en SQLite, esto dado que es mucho más fácil de cargar a los modelos y su gran portabilidad lo que permite la ejecución del entrenamiento en diferentes ambientes.

3.2. Procedimiento experimental

Análisis exploratorio de datos (EDA) y control de calidad

Una vez construida la base consolidada y la base específica de modelado, se creó una etapa de análisis exploratorio de datos (EDA) con el objetivo de comprender la estructura del conjunto

de datos, identificar patrones relevantes, y verificar la calidad de la información antes de entrenar los modelos predictivos.

Inspección del dataset

En primer lugar, se realizó una inspección general de la base almacenada en SQLite, verificando dimensiones del dataset, tipos de datos por columna y estadísticas descriptivas globales. Adicionalmente, se construyó un perfil básico por variable que incluye número de valores nulos y número de valores únicos, lo cual permitió detectar variables categóricas de alta cardinalidad y cuantificar el nivel de completitud de cada campo.

Figura 3.1

Descripción de las columnas de la base.

	Provincia	Cantón	Área	Sostenimiento	Jornada	Modalidad	Grado	Anio	Total	Promovidos	NoPromovidos	Abandono	NoActualizados	Tasa_Abandono
count	313429	313429	313429	313429	313429	313429	313429	313429.000000	313429.000000	313429.000000	313429.000000	313429.000000	313429.000000	313429.000000
unique	25	222	2	4	7	45	30	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	GUAYAS	QUITO	Urbana	Fiscal	Matutina	Presencial	DecimoAñoEGB	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	41414	10816	213787	184061	180728	275874	23179	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2015.777659	212.308325	195.225241	3.300263	6.511302	5.474197	0.044745
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.313709	820.002950	765.264474	24.102178	26.798854	58.568489	0.101446
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2009.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2012.000000	21.000000	17.000000	0.000000	0.000000	0.000000	0.000000
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2016.000000	61.000000	54.000000	0.000000	1.000000	0.000000	0.005974
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2020.000000	166.000000	152.000000	1.000000	5.000000	0.000000	0.041667
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2023.000000	25025.000000	24514.000000	2327.000000	1737.000000	5278.000000	1.000000

Nota. Elaboración propia con datos procesados en Python

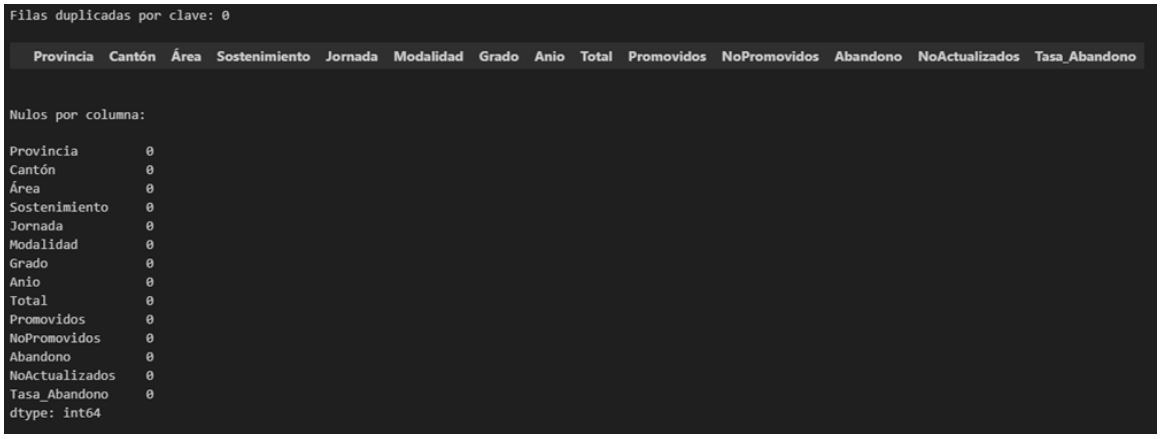
Como parte de esta inspección, se validó el rango temporal del conjunto de datos mediante la identificación del año mínimo y máximo disponibles, garantizando la cobertura histórica necesaria para el entrenamiento y evaluación posterior.

Verificación de unicidad por clave y consistencia lógica

Posteriormente, se definió una clave compuesta basada en las variables contextuales y temporales (provincia, cantón, área, sostenimiento, jornada, modalidad, grado y año), con el fin de comprobar la existencia de duplicados a nivel de registro agregado. Esta verificación fue utilizada para determinar si existían combinaciones repetidas de la misma unidad de análisis, lo que podría distorsionar métricas agregadas y el aprendizaje del modelo.

Figura 3.2

Descripción de las columnas de la base.



```
Filas duplicadas por clave: 0

   Provincia  Cantón  Área  Sostenimiento  Jornada  Modalidad  Grado  Anio  Total  Promovidos  NoPromovidos  Abandono  NoActualizados  Tasa_Abandono


Nulos por columna:

Provincia         0
Cantón            0
Área              0
Sostenimiento     0
Jornada           0
Modalidad         0
Grado             0
Anio              0
Total             0
Promovidos        0
NoPromovidos      0
Abandono          0
NoActualizados    0
Tasa_Abandono     0
dtype: int64
```

Nota. Elaboración propia con datos procesados en Python

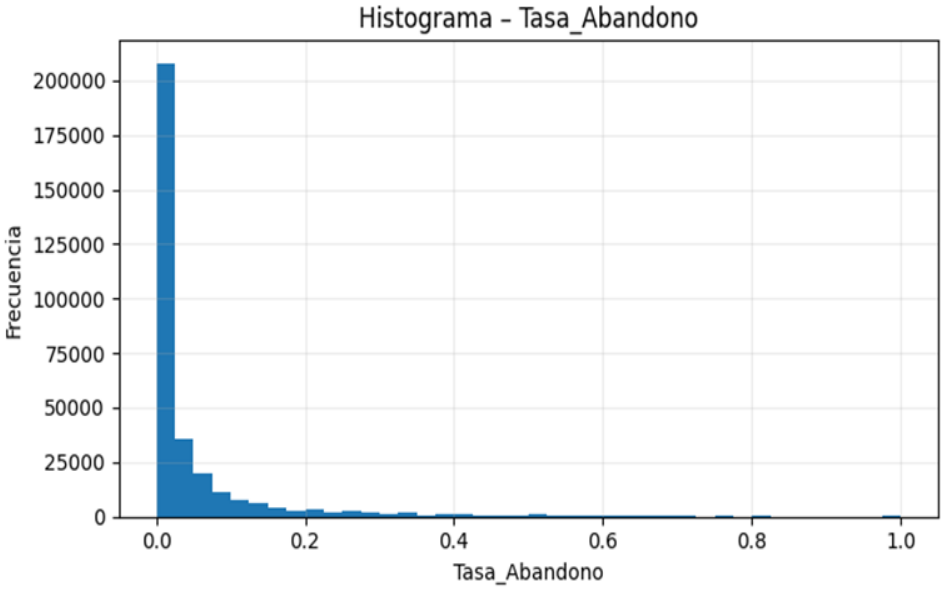
Asimismo, se aplicaron reglas de consistencia sobre las variables numéricas, se realizó la verificación de que el total de estudiantes tenga sentido con la suma de las otras variables que lo componen (promovidos, no promovidos, abandono y no actualizados). De esta manera se garantiza que la variable objetivo (tasa de abandono) esté calculada sobre bases numéricamente consistentes. Adicionalmente, se comprobó que la tasa de abandono se encuentre dentro del rango teórico válido [0, 1].

Análisis exploratorio de distribuciones y patrones temporales

Con el fin de comprender el comportamiento estadístico de las variables numéricas, se generaron histogramas para variables de total de estudiantes matriculados y sus componentes, así como para la tasa de abandono. Este análisis permitió observar asimetrías, concentraciones en valores bajos y dispersión por magnitud del total de estudiantes.

Figura 3.3

Histograma de la tasa de abandono de todos los años.



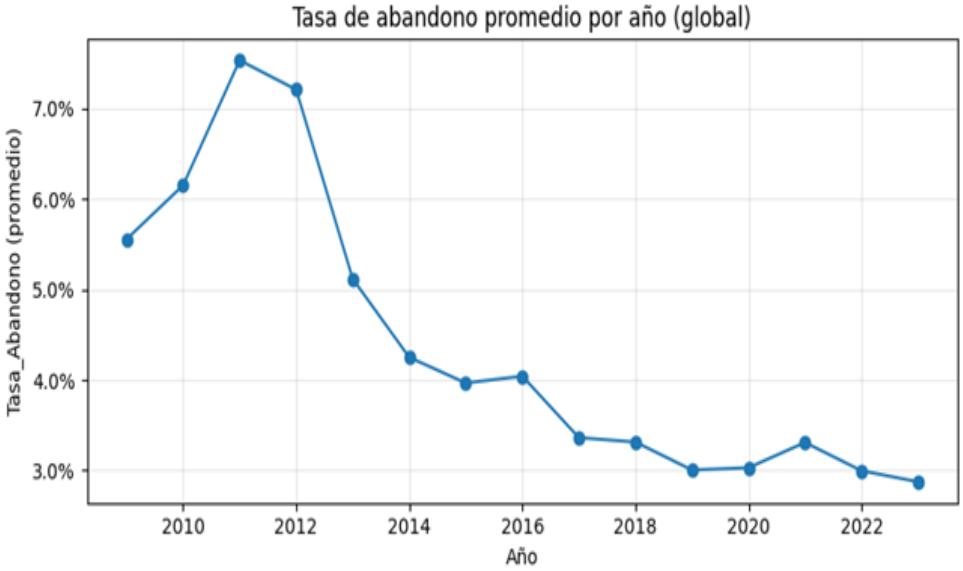
Nota. Elaboración propia con datos procesados en Python

Al analizar el resultado de la Figura c se puede observar que hay una gran concentración de valores de tasa de abandono cercanas al 0, esto se debe principalmente a que los valores de tasa de abandono no son altos, algo a tomar en cuenta para evitar que los modelos predigan siempre valores bajos.

A nivel temporal, se construyó una serie global de la tasa de abandono promedio por año, permitiendo identificar tendencias históricas y cambios relevantes en el comportamiento del indicador.

Figura 3.4

Gráfica de línea sobre tasa de abandono promedio por año.



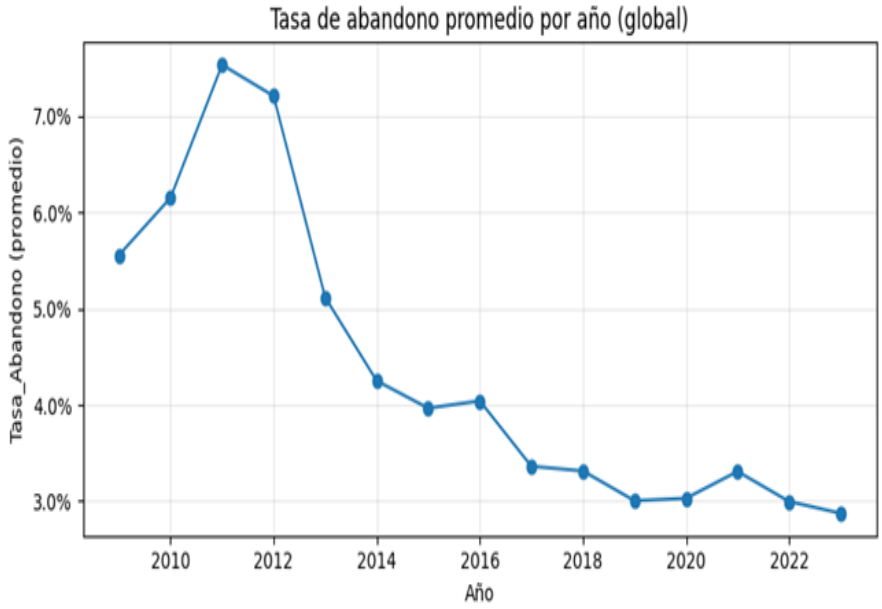
Nota. Elaboración propia con datos procesados en Python

Mediante esta gráfica se puede observar la tendencia a la baja de la tasa de abandono que se ha tenido al paso de los años.

Complementariamente, se realizó una comparación temporal por provincias con mayor número de estudiantes matriculados, con el propósito de identificar divergencias regionales y potenciales patrones territoriales.

Figura 3.5

Gráfico de líneas de la tasa de abandono promedio por provincia con mayores matriculados



Nota. Elaboración propia con datos procesados en Python

También se evaluó el comportamiento de la tasa de abandono en función de categorías institucionales (por ejemplo, sostenimiento, área y jornada) mediante gráficos comparativos (boxplots), lo cual facilitó la identificación de grupos con mayor dispersión o mayores medianas de abandono.

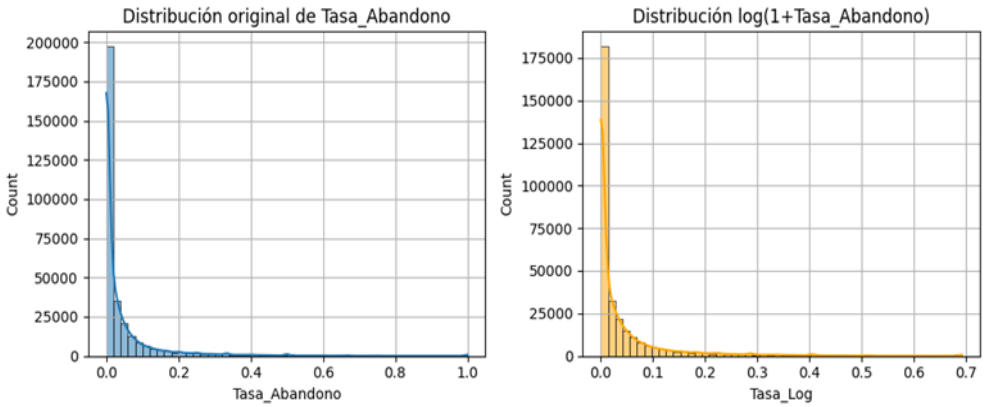
Tratamiento de ceros y transformación logarítmica de la variable objetivo

Durante el EDA se constató la presencia de una proporción significativa de registros con tasa de abandono igual a cero. Este comportamiento es relevante porque puede reflejar contextos reales de bajo abandono (especialmente en grupos pequeños) y no necesariamente errores de registro, lo que sabe ser común en datasets grandes. Sin embargo, desde la perspectiva del modelado, una alta concentración de ceros puede introducir sesgos o dificultar el ajuste de modelos que asumen distribuciones aproximadamente normales.

Por esta razón, se consideró la transformación logarítmica de la variable objetivo mediante log(1+Tasa_Abandono) , generando la variable Tasa_Log como objetivo alternativo. Esta transformación tiende a reducir la asimetría y estabilizar la varianza, lo cual resulta útil para ciertos enfoques predictivos.

Figura 3.6

Distribución de tasa de abandono y tasa log



Nota. Elaboración propia con datos procesados en Python

Detección de valores atípicos (outliers)

Con el objetivo de identificar valores extremos que pudieran influir desproporcionadamente en el entrenamiento, se aplicaron múltiples criterios de detección de outliers sobre la tasa de abandono:

1. Método IQR global, para detectar valores extremos respecto a la distribución general.
2. Método IQR por provincia, para capturar valores atípicos considerando variabilidad territorial.
3. Z-score por año, para detectar extremos dentro de cada período, evitando comparar años con distribuciones diferentes.

4. Isolation Forest usando variables numéricas seleccionadas, con el propósito de identificar observaciones anómalas multivariadas (por ejemplo, combinaciones inusuales entre matrícula total y tasa de abandono).

En lugar de eliminar automáticamente estos registros, se optó por generar “banderas” (variables binarias) y mantenerlos dentro del conjunto de datos, debido a que pueden representar escenarios reales de alto abandono escolar. Esta decisión permite entrenar modelos más robustos, a la vez que habilita análisis posteriores de sensibilidad o entrenamiento con y sin outliers.

Finalmente, para mantener trazabilidad y facilitar revisiones, los resultados del análisis de outliers se almacenaron en tablas específicas dentro de SQLite: una tabla a nivel de fila con banderas de outliers y puntajes de anomalía, y un resumen agregado por año y variables contextuales. Este registro constituye una evidencia reproducible del control de calidad aplicado antes del modelado.

Síntesis de hallazgos relevantes para el modelado

En conjunto, el EDA permitió: confirmar la coherencia estructural del dataset, comprender la evolución temporal del abandono escolar, identificar diferencias relevantes por variables institucionales y territoriales, justificar la incorporación de una variable objetivo transformada (Tasa_Log). Con base en estos hallazgos, el proceso avanza a la etapa de preparación de variables para entrenamiento y selección/entrenamiento de modelos predictivos.

CAPITULO 4:

4. ANÁLISIS DE RESULTADOS

4.1. Pruebas de Concepto

Modelo 1: Modelo lineal regularizado (Ridge, Lasso y ElasticNet)

Objetivo del modelo

El primer modelo desarrollado corresponde a un modelo baseline lineal regularizado, cuyo objetivo es establecer un punto de referencia inicial para la predicción de la tasa de abandono escolar. Este modelo permite evaluar hasta qué punto una relación lineal entre las variables es capaz de explicar el comportamiento del abandono escolar, antes de introducir modelos de mayor complejidad.

El modelo se plantea como un problema de regresión supervisada, utilizando como variable objetivo la tasa de abandono escolar. Con el fin de mejorar la estabilidad numérica y el ajuste del modelo, el entrenamiento se realizó sobre la variable transformada

$$Tasa\_Log = log(1 + Tasa\_Abandono),$$

mientras que la evaluación de resultados se efectuó en la escala original de la tasa de abandono, revirtiendo la transformación mediante la función exponencial. Esto ayuda a que el entrenamiento del modelo sea más preciso en valores bajos.

Este enfoque permite comparar los resultados con métricas directamente interpretables en términos del indicador educativo original.

Preparación de los datos para el modelo

El modelo se entrenó utilizando la tabla base modelado, construida en las etapas previas del desarrollo. Esta base ya incorpora una división temporal explícita de los datos en conjuntos de entrenamiento, validación y prueba, identificados mediante la variable split_set, lo que garantiza la ausencia de fuga de información.

Las variables predictoras utilizadas se agrupan en dos tipos:

- **Variables categóricas**: provincia, cantón, área, sostenimiento, jornada, modalidad y grado.
- **Variables numéricas**: total de estudiantes matriculados y año lectivo.

Las variables categóricas fueron transformadas mediante codificación One-Hot, mientras que las variables numéricas fueron escaladas utilizando estandarización, dentro de un mismo pipeline de preprocesamiento. Este pipeline se ajusta únicamente con los datos de entrenamiento y luego se aplica de forma consistente a los conjuntos de validación y prueba.

Configuración del modelo y estrategia de regularización

Se evaluaron tres variantes de modelos lineales regularizados:

- **Ridge Regression**, que penaliza la magnitud de los coeficientes mediante regularización L2.
- **Lasso Regressio**n, que introduce regularización L1 y permite realizar selección implícita de variables.
- **ElasticNet**, que combina regularización L1 y L2, ofreciendo un compromiso entre estabilidad y selección de variables.

Para cada uno de estos modelos se definió una grilla de hiperparámetros, centrada principalmente en el parámetro de regularización α, y en el caso de ElasticNet, en el parámetro l1_ratio. El ajuste de hiperparámetros se realizó mediante Grid Search.

Validación temporal y proceso de entrenamiento

El entrenamiento del modelo se llevó a cabo utilizando validación cruzada, construido a partir del año lectivo. En este esquema, los datos se organizan en ventanas deslizantes de varios

años, donde los primeros años de cada ventana se utilizan para entrenamiento y el último año para validación.

Esta estrategia nos sirve dado que se tiene un problema de predicción temporal, ya que respeta el orden cronológico de los datos. La métrica principal utilizada durante el proceso de ajuste fue el error cuadrático medio (RMSE) sobre la variable transformada (Tasa_Log), utilizando su versión negativa para maximizar el desempeño durante la búsqueda de hiperparámetros.

Como resultado del proceso de ajuste, se seleccionó automáticamente el modelo y la configuración de hiperparámetros que ofrecieron el mejor desempeño promedio en los folds temporales.

Evaluación del modelo en validación y prueba

Una vez entrenado el mejor modelo lineal, se generaron predicciones para los conjuntos de validación y prueba. Las predicciones obtenidas en escala logarítmica fueron transformadas nuevamente a la escala original de la tasa de abandono mediante la función exponencial inversa, asegurando que los valores finales se encuentren dentro del rango válido [0, 1].

El desempeño del modelo se evaluó utilizando las siguientes métricas en escala original:

- Error cuadrático medio (RMSE)
- Error absoluto medio (MAE)
- Error porcentual absoluto medio (MAPE)
- Coeficiente de determinación (R2)

Estas métricas permiten analizar tanto la magnitud del error como la capacidad explicativa del modelo. Los resultados muestran que el modelo lineal logra capturar tendencias generales del abandono escolar, aunque presenta limitaciones para ajustar valores extremos o cambios abruptos en la tasa de abandono.

A continuación se muestran los resultados obtenidos por el mejor modelo conseguido mediante el Grid Search:

Métricas VALIDACIÓN (escala original):

val_rmse: 0.077458

val_mae: 0.033899

val_mape: 122.985117

val_r2: 0.237884

clipped_to_[0,1]: 14219

Métricas TEST (escala original):

test_rmse: 0.073286

test_mae: 0.031124

test_mape: 101.584746

test_r2: 0.220690

clipped_to_[0,1]: 14006

Modelo 2: Modelos tabulares basados en árboles (XGBoost)

Objetivo del modelo

El segundo modelado se orientó a evaluar algoritmos no lineales para la predicción de la tasa de abandono escolar, con el fin de superar las limitaciones del modelo baseline lineal. Para ello, se implementó un conjunto de modelos tabulares basados en árboles, utilizando variantes de XGBoost con aceleración por GPU. El objetivo es capturar relaciones complejas e interacciones entre variables territoriales, institucionales y temporales que no pueden representarse adecuadamente con un modelo lineal.

Al igual que en el baseline, se trabajó con la base base_modelado. El entrenamiento se realizó sobre la variable transformada Tasa_Log y la evaluación se enfocó en métricas sobre la variable en escala original Tasa_Abandono.

Preparación de datos y partición temporal

La tabla base_modelado fue cargada desde SQLite y se aplicaron conversiones de tipo y filtros mínimos para garantizar consistencia:

- Las variables categóricas se estandarizaron a tipo string.
- Las variables numéricas (Total, Anio) y objetivos (Tasa_Abandono, Tasa_Log) se convirtieron a formato numérico, eliminando registros con valores faltantes.
- Se restringió Tasa_Abandono al rango válido [0, 1].

A diferencia del Modelo 1 (donde split_set ya venía definido), en este notebook se construyó nuevamente un esquema de partición temporal explícito por año, con cortes definidos para:

- **Entrenamiento:** hasta el año 2018
- **Validación:** 2019 a 2021
- **Prueba:** desde 2022 en adelante

Esta estrategia asegura una evaluación realista para proyección temporal, ya que el modelo se entrena con años históricos y se valida/prueba con años futuros no observados durante el entrenamiento.

Preprocesamiento de variables

Se utilizó un esquema de preprocesamiento común para todos los modelos candidatos, implementado mediante ColumnTransformer:

- Categóricas: codificación One-Hot con tolerancia a categorías no vistas, y salida dispersa para reducir consumo de memoria.
- Numéricas: escalado con StandardScaler para mantener compatibilidad con matrices dispersas.

Adicionalmente, se configuró la codificación One-Hot en float32 para optimizar rendimiento en ambientes con grandes volúmenes de categorías (por ejemplo, cantones o modalidades con alta cardinalidad).

Modelos candidatos evaluados

Se definieron tres candidatos principales, todos basados en XGBoost y ejecutados con GPU, representando estrategias distintas de aprendizaje:

1. RandomForest_GPU (XGB con num_parallel_tree): Random Forest usando múltiples árboles paralelos dentro de XGBoost.

2. GradientBoosting_GPU (XGB estándar): árboles secuenciales, con ajuste de profundidad, tasa de aprendizaje y número de estimadores.

3. XGBoost_GPU (DART): variante con dropout de árboles, orientada a mejorar generalización en presencia de potencial sobreajuste.

Cada modelo se acompañó de una cantidad de hiperparámetros centrada en los principales controladores de complejidad (por ejemplo, max_depth, learning_rate, rate_drop, n_estimators).

Estrategia de tuning: HalvingGridSearchCV con validación temporal

Para optimizar hiperparámetros con un costo computacional razonable, se utilizó HalvingGridSearchCV, una estrategia de búsqueda sucesiva que asigna recursos progresivamente a los mejores candidatos. En este caso, el recurso controlado fue el número de árboles (n_estimators), explorando valores desde un mínimo de 200 hasta un máximo de 1200.

La validación cruzada se realizó con TimeSeriesSplit, garantizando que en cada split se respete el orden temporal. La métrica utilizada para seleccionar el mejor modelo durante tuning fue el RMSE (en el objetivo de entrenamiento), utilizando su versión negativa para maximizar el score.

Este diseño busca equilibrar eficiencia (menos combinaciones completas entrenadas), consistencia temporal y robustez del modelo seleccionado.

Evaluación del desempeño en validación y prueba

Una vez entrenado cada candidato con sus mejores hiperparámetros, se generaron predicciones para:

- el conjunto de validación (2019–2021),
- y el conjunto de prueba (desde 2022).

Para cada uno se calcularon métricas clave:

- RMSE
- MAE
- R^2

y se construyó un leaderboard ordenado por val_rmse, seleccionando como "mejor modelo" aquel con menor error en validación. Esta elección se alinea con el criterio de generalización temporal: el modelo debe desempeñarse correctamente en años inmediatamente posteriores al entrenamiento.

Tabla 4.1. Mejores resultados del entrenamiento de los modelos

Modelo	**Parámetros óptimos**	**RMSE (Val)**	**MAE (Val)**	**R² (Val)**	**RMSE (Test)**	**MAE (Test)**	**R² (Test)**	**Tiempo (min)**
RandomForest_GPU (XGB)	colsample_bynode=0.8, max_depth=6, num_parallel_tree=8, subsample=0.632, n_estimators=200	0.07616	0.03276	0.26315	0.07644	0.03376	0.15209	1.56
GradientBoosting_GPU (XGB)	learning_rate=0.06, max_depth=6, n_estimators=600	0.07491	0.03284	0.28726	0.07350	0.03317	0.21609	0.26

XGBoost_GPU (DART)	learning_rate=0.06, max_depth=6, rate_drop=0.0, n_estimators=600	0.07499	0.03282	0.28566	0.07356	0.03315	0.21490	76.42

Nota: En la tabla se muestran los resultados del mejor modelo de cada tipo en cuanto a sus diferentes métricas en los distintos conjuntos de evaluación.

Como se puede ver en la Tabla X se puede ver que el modelo que obtuvo el mejor resultado es GradientBoosting, esta es la razón por la que es el candidato almacenado y con el que se hace los siguientes pasos.

Persistencia del mejor modelo y trazabilidad de resultados

Con fines de reproducibilidad y despliegue, se implementa una estrategia completa de persistencia:

- Guardado del mejor estimador entrenado mediante joblib.
- Generación de un archivo de metadata en JSON que incluye:
- algoritmo seleccionado,
- hiperparámetros óptimos,
- número de folds temporales,
- rango de años por split,
- métricas finales en validación y prueba,
- y el leaderboard completo de candidatos evaluados.
- Escritura en SQLite de:
- una tabla de predicciones (validación + prueba) con y_true y y_pred,
- y una tabla con el leaderboard de métricas de la corrida.

Modelo 3: Red neuronal tabular (Embeddings + MLP) en PyTorch

Objetivo del modelo

El tercer enfoque de modelado incorpora una arquitectura de red neuronal profunda diseñada específicamente para datos tabulares con alta presencia de variables categóricas. El

objetivo es mejorar la capacidad predictiva respecto a modelos lineales y, potencialmente, respecto a modelos basados en árboles, capturando relaciones no lineales e interacciones complejas entre variables territoriales, institucionales y temporales.

El entrenamiento se realiza sobre la variable transformada:

$$Tasa\_Log = log(1 + Tasa\_Abandono),$$

mientras que la evaluación se reporta en la escala original Tasa_Abandono y restringiendo el resultado final al rango válido [0,1].

Datos de entrada y partición temporal

El modelo utiliza la tabla base_modelado, almacenada en SQLite, la cual ya contiene la variable split_set con particiones temporales definidas (train, val, test). Se aplican chequeos de integridad para asegurar la existencia de las columnas requeridas y se estandarizan tipos:

· Variables categóricas (Provincia, Cantón, Área, Sostenimiento, Jornada, Modalidad, Grado) convertidas a texto y normalizadas (por ejemplo, eliminando espacios).

· Variables numéricas permitidas: Total y Anio.

· Variable objetivo: Tasa_Abandono (saneada a rango [0,1]) y su transformación Tasa_Log.

La selección de variables se mantiene bajo la política “sin fuga” al excluir componentes que podrían estar directamente vinculados al cálculo del abandono (por ejemplo, promovidos o abandonos absolutos), utilizando únicamente contexto y magnitud (Total) junto con la componente temporal (Año).

Arquitectura del modelo neuronal

La arquitectura propuesta corresponde a un modelo tabular híbrido:

1. Embeddings para categóricas: una capa de embedding por cada columna categórica.

2. Normalización de numéricas: se aplica LayerNorm a las variables numéricas (Total y Anio) para estabilizar el entrenamiento.

3. Red MLP: concatenación de embeddings + numéricas normalizadas y posterior paso por una red multicapa con activaciones ReLU y regularización mediante dropout.

La configuración principal del MLP incluye:

· una capa densa inicial de 256 neuronas,

· capas intermedias de 128 y 64 neuronas,

· dropout (p=0.15) para mitigar sobreajuste,

· y una salida escalar que representa la predicción en espacio logarítmico (Tasa_Log).

Estrategia de entrenamiento y control de generalización

El entrenamiento se realiza con las siguientes decisiones técnicas:

· Optimizador AdamW, que combina Adam con regularización por weight decay.

· Función de pérdida: MSE sobre Tasa_Log, favoreciendo estabilidad numérica.

· Scheduler OneCycleLR, para ajustar la tasa de aprendizaje de manera suave y eficiente a lo largo de las épocas.

· Entrenamiento en GPU cuando está disponible, con mixed precision (torch.cuda.amp) para acelerar y reducir consumo de memoria.

Para prevenir sobre ajuste, se implementa early stopping con paciencia de 20 épocas, reteniendo el mejor estado del modelo según la pérdida en validación.

Modelo 4: Modelo basado en embeddings preentrenados y redes neuronales profundas

Objetivo del Modelo

En los modelos previamente evaluados, las variables categóricas de contexto educativo (provincia, cantón, área, sostenimiento, jornada, modalidad y grado) fueron representadas mediante esquemas tradicionales como codificación One-Hot o embeddings aprendidos desde cero. Si bien estos enfoques permiten capturar información categórica, presentan limitaciones cuando el número de categorías es elevado o cuando existen combinaciones poco frecuentes en los datos.

Con el objetivo de explorar una representación más rica y semánticamente informativa del contexto educativo, se propone un modelo que utiliza embeddings preentrenados basados en transformadores. Estos embeddings permiten mapear combinaciones de atributos categóricos a un espacio vectorial denso, donde relaciones de similitud entre contextos pueden ser capturadas de forma implícita.

Este enfoque se inspira en avances recientes del procesamiento de lenguaje natural, donde modelos entrenados sobre grandes volúmenes de texto logran representaciones generales reutilizables en tareas supervisadas posteriores.

Construcción de la representación de entrada

Para cada observación del conjunto de datos, las variables categóricas permitidas se integran en una representación textual estructurada, sin incluir variables que generen fuga de información. Específicamente, se construye una cadena de texto que concatena:

- Provincia
- Cantón
- Área
- Sostenimiento
- Jornada

- Modalidad
- Grado

Cada registro se transforma en una secuencia textual del tipo:

Provincia = … | Cantón = … | Área = … | Sostenimiento = … | Jornada = … | Modalidad = … | Grado = …

Esta representación textual se utiliza como entrada para un modelo SentenceTransformer multilingüe, el cual genera un vector denso de dimensión fija que resume el contexto categórico completo de la observación.

Adicionalmente, se incorporan variables numéricas de contexto permitidas (Total y Anio), las cuales son estandarizadas y concatenadas a los embeddings textuales, conformando así la matriz final de entrada al modelo predictivo.

Arquitectura del modelo predictivo

Sobre la representación combinada (embeddings textuales + variables numéricas), se entrena una red neuronal multicapa (MLP) orientada a regresión. La arquitectura utilizada consta de:

- Una capa de entrada con dimensión igual al tamaño del embedding más las variables numéricas.
- Capas densas intermedias con activación ReLU y regularización mediante dropout.
- Una capa de salida lineal que estima directamente la tasa de abandono escolar en escala original.

El modelo se entrena utilizando el optimizador Adam y la función de pérdida de error cuadrático medio (MSE), incorporando early stopping basado en el desempeño del conjunto de validación para evitar sobreajuste.

Este diseño permite que el modelo aprenda relaciones no lineales complejas entre contextos educativos representados semánticamente y la tasa de abandono observada.

Estrategia de entrenamiento y evaluación

El entrenamiento y la evaluación del modelo respetan la partición temporal definida previamente en el proyecto:

- Conjunto de entrenamiento: años históricos
- Conjunto de validación: años intermedios
- Conjunto de prueba: años más recientes

Esta estrategia garantiza que la evaluación refleje un escenario realista de proyección temporal, evitando que el modelo se beneficie de información futura.

Las métricas utilizadas para evaluar el desempeño incluyen:

- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)
- MAPE (Mean Absolute Percentage Error)
- R^2 (coeficiente de determinación)

Todas las métricas se calculan en la escala original de la tasa de abandono, facilitando su interpretación desde el punto de vista educativo.

Modelo predictivo basado en LightGBM con variables rezagadas

Descripción general del modelo

Una vez incorporadas las variables rezagadas que capturan la dinámica temporal del abandono escolar, se emplea el algoritmo Light Gradient Boosting Machine (LightGBM) como modelo predictivo principal para este enfoque.

LightGBM es un método de ensemble basado en árboles de decisión, optimizado para manejar grandes volúmenes de datos, variables categóricas y relaciones no lineales complejas. A diferencia de los modelos lineales evaluados previamente, este algoritmo permite capturar interacciones no explícitas entre variables territoriales, institucionales y temporales, lo cual resulta especialmente adecuado para fenómenos educativos de naturaleza multicausal.

El modelo se entrena utilizando como variables explicativas:

- Las variables categóricas de contexto educativo y territorial.
- Las variables numéricas originales permitidas (año y total de estudiantes).
- El conjunto de variables rezagadas de la tasa de abandono y del tamaño de la población estudiantil, descritas en la sección anterior.

El objetivo del modelo es predecir la tasa de abandono escolar en escala original, manteniendo el rango válido entre 0 y 1.

Esquema de entrenamiento y validación

El entrenamiento del modelo sigue un esquema temporal estricto, consistente con los enfoques anteriores:

- Conjunto de entrenamiento: años históricos hasta 2018.
- Conjunto de validación: período intermedio 2019–2021, utilizado para ajuste de hiperparámetros y parada temprana.
- Conjunto de prueba: años a partir de 2022, reservado exclusivamente para evaluación final.

Se emplea un mecanismo de early stopping sobre el conjunto de validación, con el fin de evitar sobreajuste y seleccionar automáticamente el número óptimo de iteraciones del modelo.

Las variables categóricas se tratan de forma nativa por LightGBM, lo que permite preservar su estructura sin necesidad de codificaciones expansivas, mientras que las variables numéricas y rezagadas se incorporan directamente al proceso de entrenamiento.

Métricas de desempeño

El desempeño del modelo se evalúa mediante múltiples métricas, tanto en validación como en prueba, con el objetivo de obtener una visión integral de su capacidad predictiva:

- RMSE (Root Mean Squared Error): penaliza errores grandes y mide precisión global.
- MAE (Mean Absolute Error): mide el error promedio en términos absolutos.
- MAPE (Mean Absolute Percentage Error): facilita la interpretación relativa del error.
- R^2 (Coeficiente de determinación): cuantifica la proporción de variabilidad explicada por el modelo.

Modelo predictivo basado en CatBoost con variables rezagadas

Descripción general del modelo

Sobre la base del conjunto de variables enriquecido con información temporal mediante rezagos, se evalúa un segundo modelo basado en CatBoost, un algoritmo de boosting de gradiente diseñado específicamente para manejar variables categóricas de forma nativa.

CatBoost resulta particularmente adecuado en este contexto debido a dos características clave:

1. su capacidad para modelar relaciones no lineales complejas, y
2. su tratamiento interno de variables categóricas sin necesidad de transformaciones explícitas como one-hot encoding.

El modelo utiliza el mismo conjunto de predictores definido en el enfoque con lags, integrando información histórica de la tasa de abandono, variaciones interanuales y tendencias recientes, junto con las variables territoriales e institucionales.

El objetivo del modelo es estimar la tasa de abandono escolar en escala original, manteniendo coherencia con los enfoques previos y permitiendo una comparación directa de resultados.

Esquema de entrenamiento y validación

El entrenamiento del modelo CatBoost se realiza bajo un esquema de partición temporal consistente, idéntico al utilizado en el modelo LightGBM con lags:

- **Entrenamiento**: datos históricos hasta el año 2018.
- **Validación**: período 2019–2021, utilizado para control del sobreajuste mediante parada temprana.
- **Prueba:** datos a partir de 2022, reservados para evaluación final.

Se emplea un mecanismo de early stopping basado en el desempeño en validación, lo que permite seleccionar automáticamente el número óptimo de iteraciones y reducir el riesgo de sobreajuste, especialmente relevante dado el uso de un número elevado de árboles.

Las variables categóricas son procesadas directamente por el algoritmo, mientras que las variables numéricas —incluyendo los rezagos— se incorporan sin transformaciones adicionales, aprovechando la robustez del modelo frente a escalas heterogéneas.

Métricas de desempeño

El desempeño del modelo se evalúa utilizando las mismas métricas definidas para los modelos anteriores, tanto en validación como en prueba:

- RMSE, para medir la precisión global penalizando errores grandes.

- MAE, para evaluar el error promedio absoluto.
- MAPE, como indicador relativo del error.
- R^2, para cuantificar la proporción de variabilidad explicada.

Los resultados obtenidos reflejan un desempeño competitivo dentro del enfoque basado en lags, confirmando que la combinación de información temporal y modelado no lineal permite capturar mejor la estructura del fenómeno respecto a los modelos sin memoria histórica.

Modelo XGBoost con variables rezagadas

Descripción del modelo

Bajo el enfoque basado en variables rezagadas (lags), se entrena un modelo XGBoost (Extreme Gradient Boosting) con el objetivo de capturar relaciones no lineales e interacciones complejas entre:

- las variables contextuales territoriales e institucionales (Provincia, Cantón, Área, Sostenimiento, Jornada, Modalidad, Grado),
- las variables numéricas permitidas (Total, Año),
- y las características derivadas de rezagos (lags, diferencias y estadísticas de ventana) construidas previamente sobre el panel.

Debido a que XGBoost no consume variables categóricas de forma nativa en su implementación estándar, se aplica codificación one-hot (OneHotEncoder) para las variables categóricas. Esta transformación se encapsula dentro de un preprocesador reproducible que se guarda junto con el booster entrenado para asegurar trazabilidad y capacidad de despliegue.

El entrenamiento se realiza con parada temprana (early stopping) utilizando el conjunto de validación temporal, lo que permite seleccionar automáticamente el número óptimo de iteraciones y controlar el sobreajuste.

Esquema de validación temporal

Se utiliza un esquema consistente con los experimentos previos:

- Entrenamiento: hasta 2018
- Validación: 2019–2021 (para early stopping y selección de iteración óptima)
- Prueba: 2022 en adelante (evaluación final)

Este diseño refleja condiciones realistas de proyección, ya que el modelo solo ve pasado para predecir el futuro.

4.2.Análisis de Resultados

Modelo 1: Modelo lineal regularizado (Ridge, Lasso y ElasticNet)

En el conjunto de validación, el modelo alcanzó un error cuadrático medio (RMSE) de 0.0775 y un error absoluto medio (MAE) de 0.0339. Estos valores indican que, en promedio, la diferencia absoluta entre la tasa de abandono real y la predicha se sitúa alrededor de 3 a 4 puntos porcentuales, lo cual resulta razonable para un modelo baseline lineal aplicado a un caso complejo.

El coeficiente de determinación obtenido ($R^2 = 0.2379$) sugiere que el modelo es capaz de explicar aproximadamente el 24% de la variabilidad en la tasa de abandono dentro del conjunto de validación. Si bien este valor no es elevado, es consistente con la naturaleza exploratoria y de referencia del modelo, y confirma que existe una relación parcialmente lineal entre las variables explicativas utilizadas y el comportamiento del abandono escolar.

Por otro lado, el MAPE elevado (122.99%) explica principalmente por la presencia de un gran número de registros con valores reales de tasa de abandono cercanos a cero, lo que distorsiona la interpretación de esta métrica. Por esta razón, el MAPE no se considera una métrica robusta en este conjunto de datos.

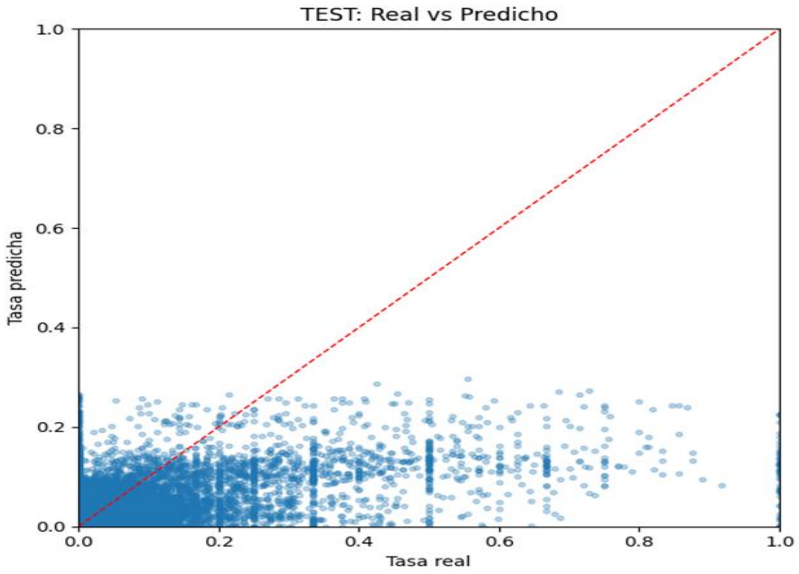
En el conjunto de prueba, el modelo mostró un desempeño ligeramente superior al observado en validación, con un RMSE de 0.0733 y un MAE de 0.0311. Esta reducción del error sugiere una buena estabilidad temporal del modelo y ausencia de sobreajuste significativo durante el entrenamiento y el ajuste de hiperparámetros.

El valor de $R^2 = 0.2207$ indica que el modelo explica aproximadamente el 22% de la variabilidad en datos no observados durante el entrenamiento. La cercanía entre los valores de R^2 en validación y prueba refuerza la consistencia del modelo y confirma que el desempeño observado no es producto de un ajuste excesivo a los datos de entrenamiento.

El MAPE en prueba (101.58%), aunque menor que en validación, continúa siendo elevado por las mismas razones estructurales previamente descritas: una alta proporción de registros con tasa de abandono igual o cercana a cero. En este contexto, métricas basadas en errores absolutos (RMSE y MAE) ofrecen una evaluación más fiable del desempeño del modelo.

Figura 4.1

Comparación entre la tasa predicha y tasa real del modelo



Nota. Elaboración propia con datos procesados en Python

Como se puede ver en la Figura X el modelo no obtiene muy buenos resultados ya que lo común es que ponga los valores de la tasa predicha en valores cercanos a 0, por lo que el modelo no logra predecir de manera correcta los valores reales de tasa de abandono.

En conjunto, los resultados confirman que el modelo lineal regularizado cumple adecuadamente su rol como modelo baseline. Su desempeño evidencia que existe una señal predictiva real en las variables contextuales e institucionales utilizadas, aunque la relación entre estas variables y la tasa de abandono no puede ser capturada completamente mediante un enfoque lineal.

Las métricas obtenidas muestran que el modelo:

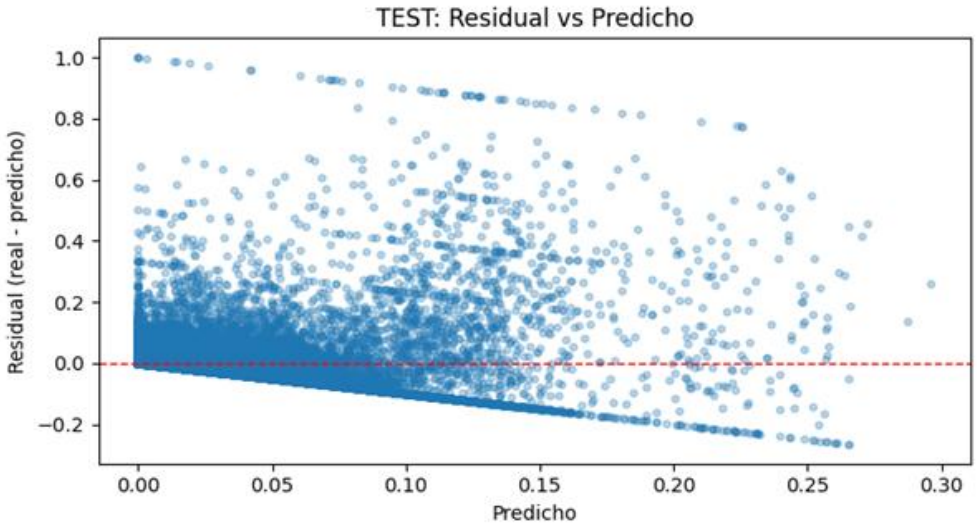
- Presenta errores absolutos moderados.
- Mantiene estabilidad temporal, con resultados comparables entre validación y prueba.
- Tiene capacidad explicativa limitada.

Análisis de residuos y comportamiento del modelo

Para profundizar en el análisis del desempeño, se evaluaron los residuos del modelo en el conjunto de prueba. El análisis incluyó la inspección de la distribución de los residuos, su relación con los valores predichos y la evolución del error absoluto medio por año.

Figura 4.2

Valores residuales vs valor predicho.



Nota. Elaboración propia con datos procesados en Python

Como se ve en la Figura 4.2 los resultados indican que los residuos se concentran mayoritariamente alrededor de cero, lo que sugiere ausencia de sesgos sistemáticos importantes. No obstante, se observa una mayor dispersión en los extremos de la distribución, particularmente en registros con tasas de abandono elevadas o con bajo número total de estudiantes, lo cual es consistente con las limitaciones inherentes a los modelos lineales.

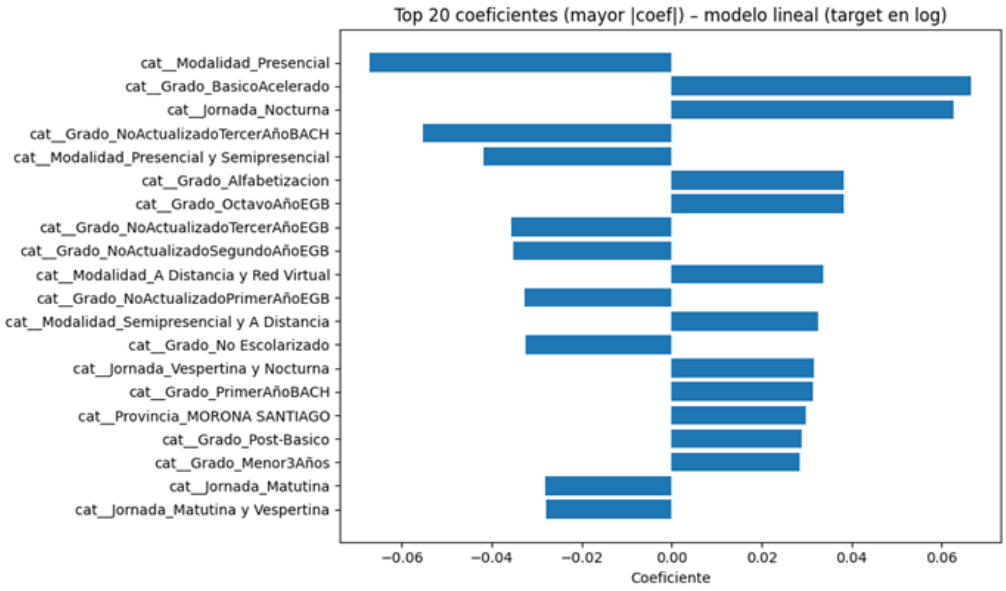
El análisis por año evidencia que el error medio no es constante en el tiempo, lo que refuerza la importancia de considerar modelos más flexibles capaces de capturar relaciones no lineales y cambios estructurales en el comportamiento del abandono escolar.

Interpretación de coeficientes y relevancia de variables

Una ventaja clave de los modelos lineales es su interpretabilidad. En este caso, se analizaron los coeficientes del modelo entrenado para identificar las variables con mayor influencia en la predicción de la tasa de abandono (en escala logarítmica).

Figura 4.3

Comparación de los coeficiente con mayor influencia en el modelo.



Nota. Elaboración propia con datos procesados en Python

Como se observa en la Figura 4.3 los coeficientes de mayor magnitud corresponden principalmente a determinadas categorías de variables territoriales, institucionales y de nivel educativo, lo que sugiere que el contexto geográfico y el tipo de institución desempeñan un papel relevante en el comportamiento del abandono escolar. Esta información resulta valiosa como insumo para el análisis cualitativo y para la formulación de políticas educativas focalizadas.

Análisis del proceso de ajuste de hiperparámetros

Se exploraron los resultados completos del proceso de Grid Search, analizando la relación entre los valores del parámetro de regularización y el desempeño del modelo. Este análisis permitió verificar que el modelo presenta un comportamiento estable frente a diferentes niveles de

penalización, y que el uso de regularización es fundamental para evitar sobreajuste en un conjunto de datos con alta dimensionalidad derivada de la codificación One-Hot.

Conclusiones del modelo baseline

El modelo lineal regularizado cumple adecuadamente su función como modelo baseline, proporcionando un punto de referencia claro para la comparación con modelos posteriores. Sus principales fortalezas radican en su simplicidad, estabilidad e interpretabilidad; sin embargo, su capacidad predictiva es limitada frente a relaciones no lineales y patrones complejos presentes en los datos educativos.

En consecuencia, los siguientes modelos explorarán algoritmos más flexibles, con el objetivo de mejorar el desempeño predictivo manteniendo un equilibrio razonable entre precisión e interpretabilidad.

Modelo 2: Modelos tabulares basados en árboles (XGBoost)

Análisis de resultados del modelo seleccionado

El modelo tabular basado en árboles seleccionado como mejor candidato fue evaluado en los conjuntos de validación (2019–2021) y prueba (desde 2022), utilizando métricas calculadas en la escala original de la tasa de abandono escolar.

En validación, el modelo obtuvo un RMSE = 0.0749 y MAE = 0.0328, lo que indica un error absoluto promedio cercano a 3.3 puntos porcentuales en la tasa de abandono. El coeficiente de determinación fue $R^2 = 0.2873$, lo cual implica que el modelo explica aproximadamente el 29% de la variabilidad observada en la tasa de abandono durante el periodo de validación.

En prueba, el modelo alcanzó un RMSE = 0.0735 y MAE = 0.0332, con $R^2 = 0.2161$. Aunque el valor de R^2 disminuye respecto a validación, las métricas basadas en error (RMSE y MAE) se mantienen relativamente cercanas, lo que sugiere que el modelo conserva una estabilidad temporal razonable al proyectarse hacia años más recientes.

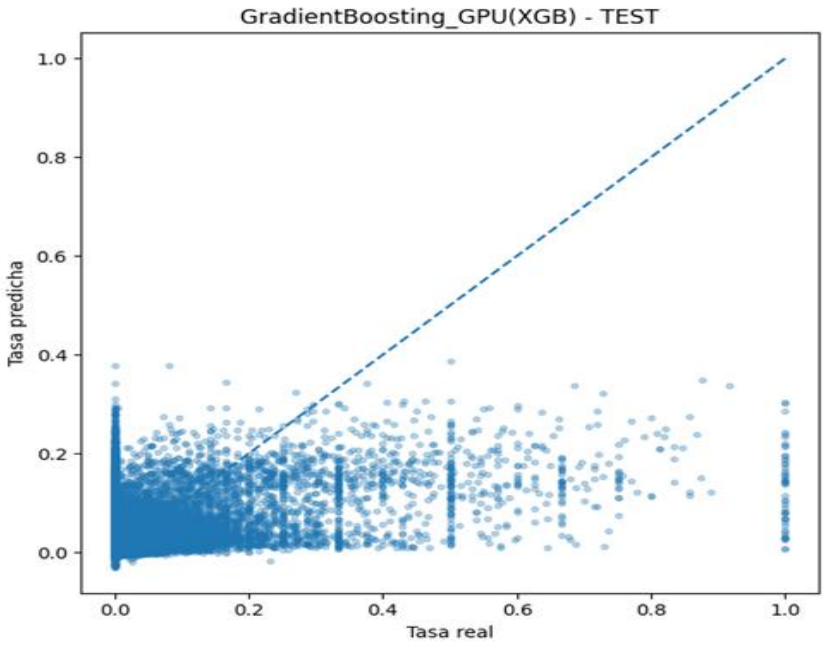
En términos comparativos frente al baseline lineal, el modelo tabular muestra una mejora clara en validación (especialmente en R^2), lo que es consistente con la capacidad de los modelos basados en árboles para capturar no linealidades e interacciones entre variables categóricas territoriales e institucionales.

Análisis gráfico y diagnóstico del modelo seleccionado

Además de las métricas numéricas, el desempeño del modelo se evaluó mediante visualizaciones diagnósticas, las cuales permiten identificar patrones sistemáticos de error, sesgos potenciales y estabilidad de las predicciones.

Figura 4.4

Gráfico de tasa predicha vs tasa real (TEST)



Nota. Elaboración propia con datos procesados en Python

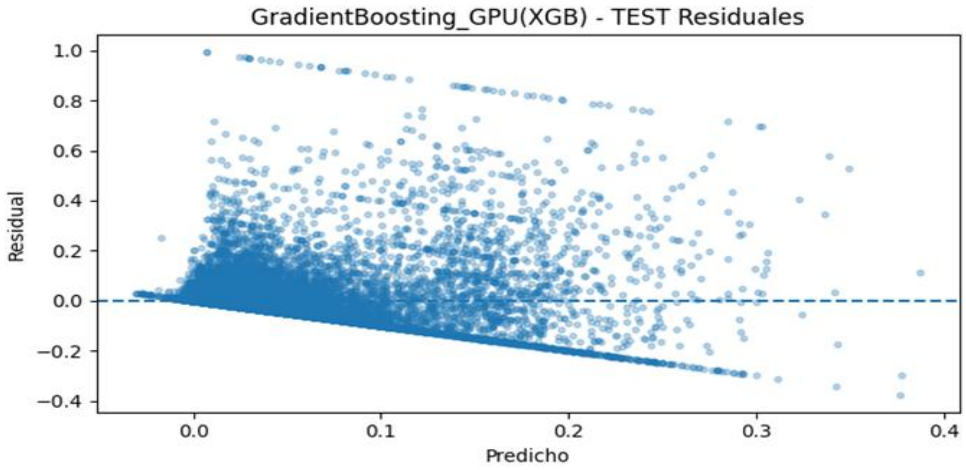
En la Figura 4.4 se puede contrastar el ajuste del modelo respecto a la diagonal ideal, el gráfico de prueba evidencia que el modelo logra capturar el rango general de valores observados, aunque tiende a presentar mayor dispersión en extremos mas cercano al 0. Este comportamiento es consistente con la naturaleza del problema, en el cual existe una alta proporción de valores

cercanos a cero y un conjunto reducido de registros con tasas elevadas, que suelen ser más difíciles de predecir con precisión.

Análisis de residuales (TEST)

Figura 4.5

Residual vs valor predicho GradientBoosting



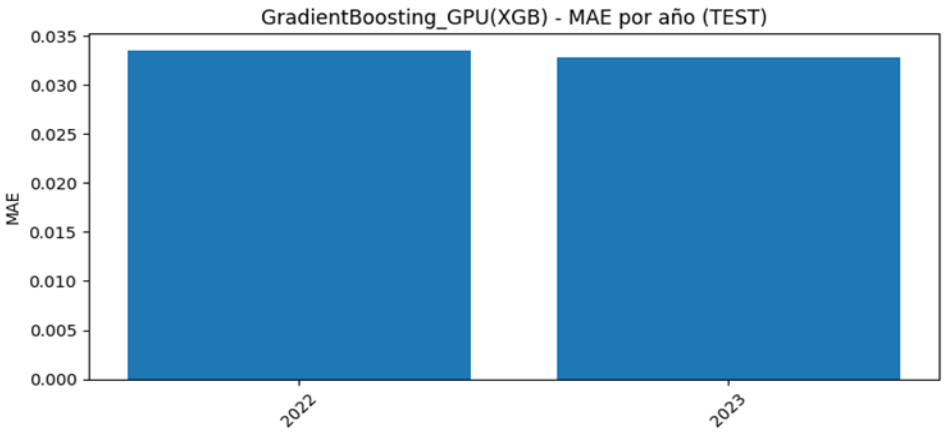
Nota. Elaboración propia con datos procesados en Python

El diagnóstico de residuales se evaluó mediante el gráfico Residual vs valor predicho, el cual es útil para detectar heterocedasticidad o sesgos sistemáticos.

En este caso, la dispersión creciente con el valor predicho sugiere heterocedasticidad, mientras que una tendencia clara arriba o abajo de cero indicaría sesgo, esto es especialmente relevante porque, aun cuando RMSE y MAE reflejan buen desempeño promedio, la estructura de residuales no aleatoria indica oportunidades de mejora mediante variables adicionales o modelos con mayor capacidad de generalización.

Figura 4.6

Gráfico de MAE por Año



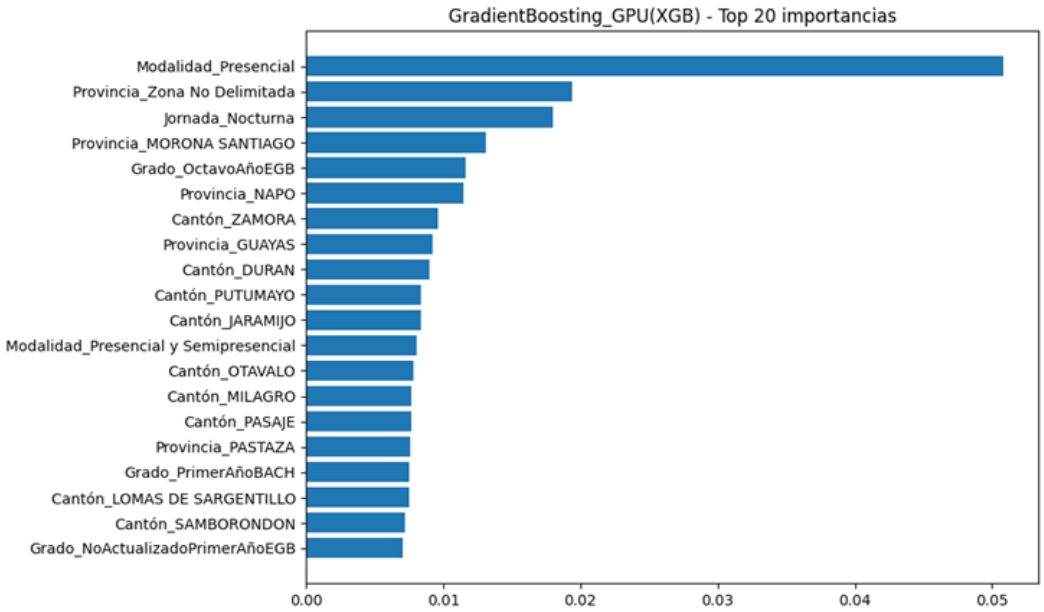
Nota. Elaboración propia con datos procesados en Python

El gráfico de la Figura 4.6 en el conjunto de prueba permite evaluar estabilidad temporal: si el error se mantiene relativamente constante por año, se sugiere que el modelo generaliza de manera consistente; si hay picos, puede indicar cambios estructurales en el fenómeno educativo o diferencias en la composición de datos.

Este análisis es valioso para un problema de proyección, ya que identifica periodos específicos donde el modelo pierde precisión y donde podrían requerirse estrategias adicionales, como ajustes por cambio de régimen, variables externas o re entrenamiento periódico.

Figura 4.7

Importancia de características



Nota. Elaboración propia con datos procesados en Python

La Figura 4.7 permite identificar qué categorías territoriales, qué variables institucionales y qué factores estructurales contribuyen más al poder predictivo.

Conclusiones del enfoque tabular basado en árboles

El modelo tabular basado en árboles representa un avance metodológico respecto al baseline lineal, al permitir capturar relaciones no lineales e interacciones entre variables categóricas y numéricas. Los resultados cuantitativos muestran una mejora en el poder explicativo durante validación ($R^2 = 0.2873$) y un desempeño estable en términos de error absoluto y cuadrático medio al pasar a prueba (RMSE = 0.0735, MAE = 0.0332).

El análisis gráfico complementa las métricas al revelar el comportamiento del modelo en diferentes rangos de la variable objetivo, la estructura de los residuales y la estabilidad temporal del error. Finalmente, la interpretación basada en importancias de características proporciona un componente explicativo adicional, útil para sustentar la relevancia de variables territoriales e institucionales asociadas al abandono escolar.

Con estos resultados, el modelo tabular se consolida como un candidato competitivo para comparación con modelos posteriores, manteniendo la lógica de evaluación temporal y trazabilidad del pipeline.

Modelo 3: Red neuronal tabular (Embeddings + MLP) en PyTorch

Análisis de resultados

El desempeño del modelo neuronal tabular basado en embeddings y una red multicapa (MLP) fue evaluado en los conjuntos de validación y prueba utilizando métricas calculadas en la escala original de la tasa de abandono escolar. Los resultados obtenidos evidencian una mejora progresiva respecto a los modelos previamente evaluados.

En el conjunto de validación, el modelo alcanzó un RMSE de 0.0736 y un MAE de 0.0313, lo que indica un error absoluto promedio cercano a 3.1 puntos porcentuales en la predicción de la tasa de abandono. El coeficiente de determinación obtenido fue $R^2 = 0.3290$, lo que implica que el modelo es capaz de explicar aproximadamente el 33% de la variabilidad observada en la tasa de abandono durante el periodo de validación.

Este incremento en el poder explicativo, respecto al modelo lineal y al modelo tabular basado en árboles, sugiere que la arquitectura neuronal logra capturar de forma más efectiva relaciones no lineales e interacciones complejas entre las variables categóricas territoriales, institucionales y las variables numéricas de contexto.

En el conjunto de prueba, el modelo mantuvo un desempeño consistente, con un RMSE de 0.0709 y un MAE de 0.0303, valores que representan una reducción adicional del error absoluto promedio frente a los modelos anteriores. El coeficiente de determinación en prueba fue $R^2 = 0.2704$, lo cual indica que el modelo conserva una capacidad explicativa cercana al 27% en datos no observados durante el entrenamiento.

La cercanía entre las métricas de validación y prueba sugiere una buena estabilidad temporal y ausencia de sobreajuste significativo, lo cual es especialmente relevante en un problema de proyección educativa donde los patrones pueden variar a lo largo del tiempo.

Interpretación comparativa de los resultados

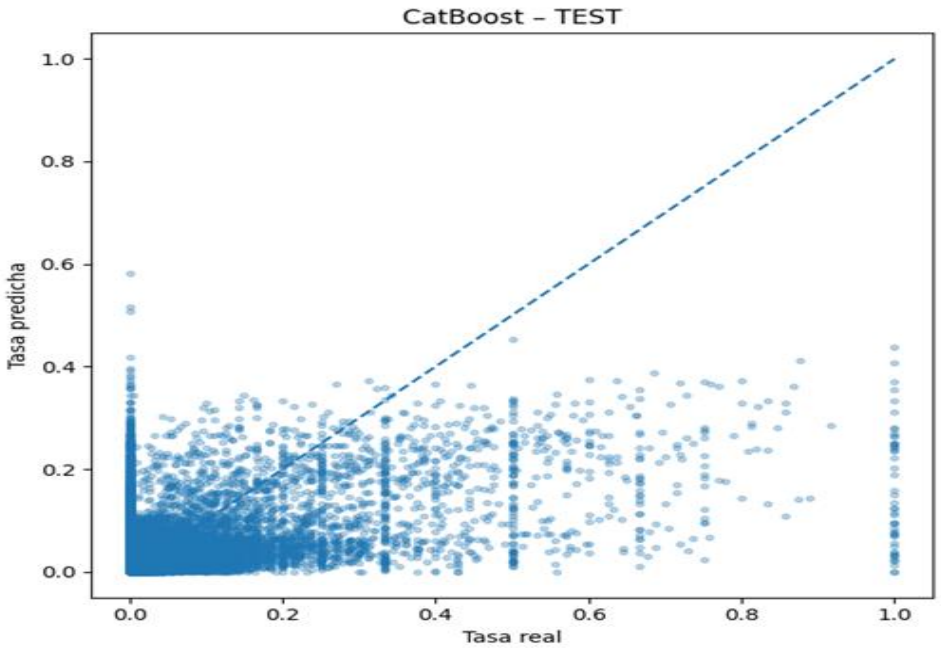
Comparado con el modelo baseline lineal, el modelo neuronal tabular presenta una mejora sustancial tanto en reducción de error (RMSE y MAE) como en poder explicativo (R^2). Frente al modelo tabular basado en árboles, también se observa una ganancia moderada pero consistente, particularmente en el conjunto de prueba, donde el modelo neuronal alcanza el menor RMSE y MAE entre los enfoques evaluados hasta este punto.

Estos resultados sugieren que el uso de embeddings para variables categóricas, en lugar de codificaciones One-Hot, permite una representación más compacta y expresiva del contexto educativo, especialmente en variables de alta cardinalidad como cantón o modalidad.

Para interpretar el desempeño del modelo de forma más allá de las métricas agregadas, se utilizan visualizaciones diagnósticas basadas en las predicciones:

Figura 4.8

Tasa real vs tasa predicha en prueba



Nota. Elaboración propia con datos procesados en Python

La Figura 4.8 representa dispersión donde el alineamiento de puntos alrededor de la diagonal y representa buena capacidad predictiva. La dispersión y los desvíos sistemáticos permiten identificar sesgos, especialmente en rangos bajos (cercanos a cero) y en valores altos de abandono. Mediante esta gráfica se puede corroborar los valores de las métricas, ya que el modelo aplicado no tiene los mejores resultados.

Modelo 4: Modelo basado en embeddings preentrenados y redes neuronales profundas

Análisis de resultados

El desempeño del modelo basado en embeddings preentrenados multilingües y una red neuronal multicapa fue evaluada sobre los conjuntos de validación y prueba, utilizando métricas calculadas en la escala original de la tasa de abandono escolar.

En el conjunto de validación, el modelo alcanzó un RMSE de 0.0758 y un MAE de 0.0352, lo que indica un error absoluto promedio cercano a 3.5 puntos porcentuales en la estimación de la

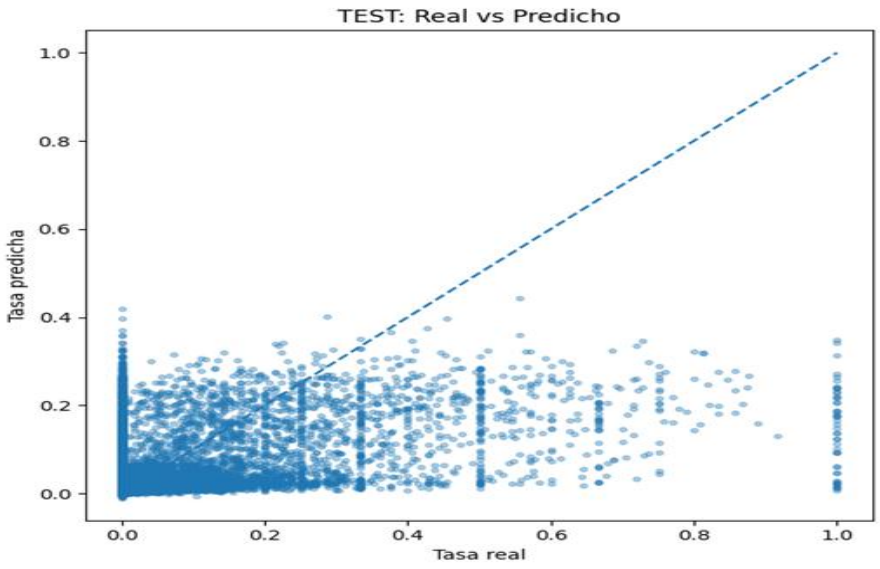
tasa de abandono. El coeficiente de determinación obtenido fue $R^2 = 0.2883$, lo que sugiere que el modelo logra explicar aproximadamente el 29% de la variabilidad observada en los datos de validación.

En el conjunto de prueba, el modelo mostró un comportamiento consistente, con un RMSE de 0.0718 y un MAE de 0.0319, reflejando una ligera mejora respecto al conjunto de validación. El valor de $R^2 = 0.2512$ indica que el modelo mantiene una capacidad explicativa cercana al 25% en datos no observados durante el entrenamiento, lo cual es relevante en un escenario de proyección temporal.

El MAPE, tanto en validación como en prueba, presenta valores elevados, lo cual se explica por la presencia de observaciones con tasas de abandono cercanas a cero, donde pequeñas desviaciones absolutas generan errores porcentuales elevados. Por esta razón, el MAPE se considera una métrica complementaria y se interpreta con cautela en este contexto.

Figura 4.9

Gráfico de tasa real vs tasa predicha



Nota. Elaboración propia con datos procesados en Python

La Figura 4.9 permite evaluar visualmente la calibración del modelo y la magnitud de las desviaciones. En el conjunto de prueba, se observa una concentración importante de puntos en valores bajos de la tasa de abandono, lo cual es coherente con la distribución real del fenómeno, se puede ver que muchos de estos valores no corresponden a los reales, ya que este modelo sigue sin poder diferenciar valores mas altos y los acumula cerca al 0.

Síntesis del desempeño del modelo

En conjunto, los resultados indican que el modelo basado en embeddings preentrenados ofrece un desempeño casi igual o peor que los otros modelos, con errores promedio y una capacidad explicativa no muy buena del conjunto de prueba. La utilización de representaciones semánticas densas permite capturar similitudes entre contextos territoriales e institucionales, favoreciendo la generalización en combinaciones categóricas poco frecuentes.

Si bien el modelo no supera ampliamente a las redes neuronales tabulares entrenadas con embeddings aprendidos desde cero, sí demuestra que el uso de embeddings preentrenados constituye una alternativa válida y conceptualmente sólida para representar variables categóricas complejas en problemas educativos de carácter territorial y temporal.

Cambio de enfoque: incorporación de variables rezagadas (lags)

Motivación del uso de información temporal

Los modelos evaluados previamente —basados en variables estáticas de contexto institucional y territorial— mostraron una capacidad limitada para explicar lo valores de la tasa de abandono escolar. Si bien capturan diferencias estructurales entre provincias, cantones y modalidades educativas, estos enfoques no incorporan de manera explícita la dinámica temporal del fenómeno.

El abandono escolar es un proceso que normalmente dependiente del tiempo, donde los valores observados en un año están fuertemente condicionados por el comportamiento de años

anteriores. Ignorar esta dependencia temporal limita la capacidad predictiva del modelo y reduce su sensibilidad ante cambios graduales o persistentes en el sistema educativo.

Por este motivo, se introduce un nuevo enfoque basado en variables rezagadas (lags), cuyo objetivo es incorporar información histórica directa de la tasa de abandono y del tamaño de la población estudiantil, manteniendo al mismo tiempo un diseño libre de fuga de información.

Definición del panel temporal

Para construir las variables rezagadas, los datos se organizan como un panel temporal definido por la combinación de las siguientes dimensiones categóricas:

- Provincia
- Cantón
- Área (urbana/rural)
- Tipo de sostenimiento
- Jornada
- Modalidad
- Grado

Cada combinación de estas variables define una unidad homogénea de análisis, sobre la cual se observa la evolución anual de la tasa de abandono escolar.

La estructura del panel permite capturar trayectorias temporales específicas para cada contexto educativo, evitando mezclar dinámicas de unidades no comparables.

Construcción de variables rezagadas

Sobre cada panel, se generan variables que resumen el comportamiento histórico reciente de la tasa de abandono y del tamaño de la población estudiantil. Estas variables se calculan

ordenando cronológicamente los registros por año y aplicando desplazamientos temporales que garantizan que la información utilizada corresponde únicamente a años anteriores.

Las variables construidas incluyen:

a) Rezagos de la tasa de abandono

- Tasa de abandono con un año de rezago, que representa el valor observado en el período inmediatamente anterior.

- Tasa de abandono con dos años de rezago, que permite capturar persistencias de más largo plazo.

Estas variables introducen memoria explícita del sistema y permiten al modelo aprender patrones de continuidad o reversión.

b) Cambios interanuales en la tasa de abandono

- Diferencia interanual, calculada como la variación entre el valor actual y el valor del año previo.

Esta variable captura aceleraciones o desaceleraciones en el abandono, permitiendo identificar contextos donde la situación mejora o empeora rápidamente.

c) Promedio móvil histórico

- Promedio móvil de tres años, calculado exclusivamente sobre valores pasados.

Este indicador suaviza fluctuaciones anuales y aporta una señal de tendencia estructural, especialmente útil en contextos con alta variabilidad.

d) Rezagos y variaciones del tamaño de la población estudiantil

- Total de estudiantes del año previo.

- Cambio absoluto interanual del total de estudiantes.

- Cambio porcentual interanual, cuando el denominador lo permite.

Estas variables permiten capturar efectos demográficos que pueden influir en la tasa de abandono, como expansiones o contracciones del sistema educativo.

Control de fuga de información

Algo que se toma en cuenta en este nuevo enfoque es la prevención de fuga de información temporal. Todas las variables rezagadas se calculan utilizando exclusivamente valores anteriores al año objetivo de predicción. En particular:

- No se utilizan valores contemporáneos ni futuros para construir los predictores.
- Los promedios móviles y diferencias se calculan sobre series desplazadas.
- Las observaciones sin historia suficiente (por ejemplo, el primer año de cada unidad del panel) son excluidas del entrenamiento.

Este diseño garantiza que el modelo simula un escenario real de predicción, donde solo se dispone de información histórica al momento de estimar la tasa de abandono de un año determinado.

Integración con los modelos predictivos

Las variables rezagadas se combinan con las variables categóricas de contexto y el total de estudiantes, conformando un conjunto de predictores enriquecido que incorpora tanto estructura espacial como dinámica temporal.

Este conjunto de variables se utiliza de forma consistente en los modelos posteriores, lo que permite:

- Comparar distintos algoritmos bajo un mismo esquema de información.
- Evaluar el impacto real de incorporar memoria temporal en el desempeño predictivo.
- Evitar la repetición conceptual de la construcción de lags en cada modelo individual.

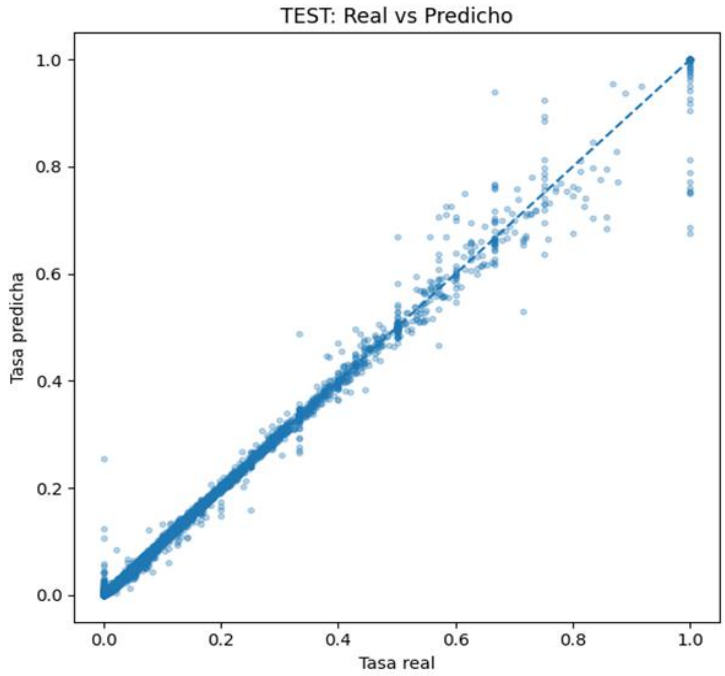
A partir de este punto, los modelos evaluados comparten esta base temporal común y difieren únicamente en el algoritmo de aprendizaje utilizado.

Modelo predictivo basado en LightGBM con variables rezagadas

Los resultados obtenidos muestran una mejora clara respecto a los enfoques previos sin información temporal, evidenciando que la incorporación de lags aporta información relevante para la predicción del abandono escolar.

Figura 4.10

Gráfica valores reales vs predicho del modelo



Nota. Elaboración propia con datos procesados en Python

La Figura 4.10 muestra el resultado de la predicción del modelo en el conjunto de entrenamiento con respecto al valor real de la tasa.

Conclusiones del modelo LightGBM con lags

El modelo LightGBM con variables rezagadas representa un cambio cualitativo respecto a los enfoques anteriores, al integrar explícitamente la dimensión temporal del fenómeno.

La mejora observada en las métricas y en los diagnósticos gráficos sugiere que la dinámica histórica del abandono escolar es un componente esencial para su modelado. Este enfoque establece una base sólida para evaluar modelos adicionales que reutilizan la misma construcción de lags, permitiendo comparaciones consistentes entre distintos algoritmos.

En las siguientes secciones se presentan otros modelos entrenados sobre este mismo conjunto de variables rezagadas, con el objetivo de analizar si arquitecturas alternativas permiten capturar aún mejor la complejidad del proceso.

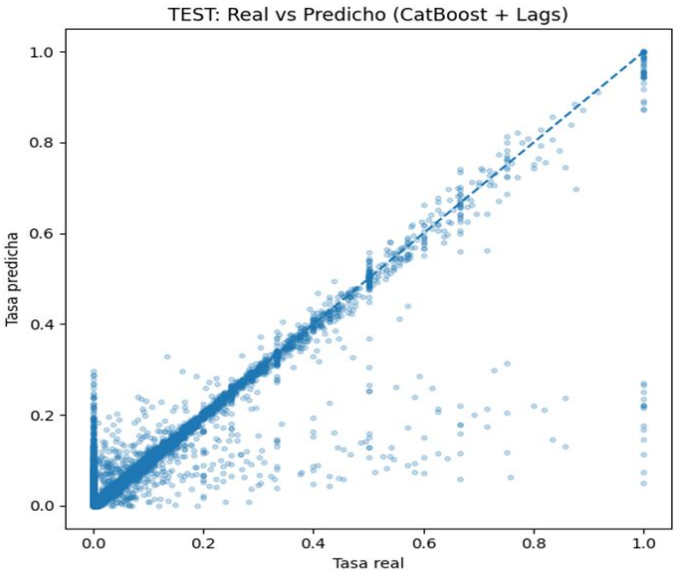
Modelo predictivo basado en CatBoost con variables rezagadas

Análisis gráfico de resultados

El análisis cuantitativo se complementa con visualizaciones que permiten evaluar el comportamiento del modelo en el conjunto de prueba.

Figura 4.11

Valores reales versus valores predichos



Nota. Elaboración propia con datos procesados en Python

La Figura 4.11 de tasa de abandono real frente a tasa predicha muestra una alineación razonable de los puntos alrededor de la diagonal, lo que indica que el modelo logra reproducir adecuadamente la relación general entre valores observados y estimados.

Se observa una mayor dispersión en niveles elevados de abandono, fenómeno consistente con la mayor heterogeneidad estructural de estos casos. No obstante, el modelo mantiene un ajuste estable en la mayor parte del rango de valores.

Consideraciones sobre el modelo CatBoost con lags

El modelo CatBoost con variables rezagadas confirma la relevancia de incorporar información histórica directa en la predicción del abandono escolar. En comparación con enfoques sin lags, se observa una mejora sustancial en la capacidad explicativa, mientras que frente a otros modelos con lags ofrece un equilibrio adecuado entre flexibilidad y estabilidad.

La capacidad del algoritmo para manejar variables categóricas de alta cardinalidad resulta especialmente útil en un contexto educativo con múltiples dimensiones territoriales e institucionales, reduciendo la necesidad de preprocesamientos complejos.

Conclusiones del modelo CatBoost con lags

El enfoque basado en CatBoost con variables rezagadas constituye una alternativa sólida dentro de la familia de modelos temporales evaluados. Su desempeño confirma que la dinámica histórica del abandono escolar es un factor determinante y que los modelos de boosting son especialmente adecuados para explotar esta información.

En conjunto con el modelo LightGBM con lags, este enfoque refuerza la evidencia empírica a favor de incorporar memoria temporal explícita en la modelización del abandono escolar. En la siguiente sección se presenta el último modelo evaluado bajo este mismo esquema de lags, seguido de una comparación global entre los distintos enfoques.

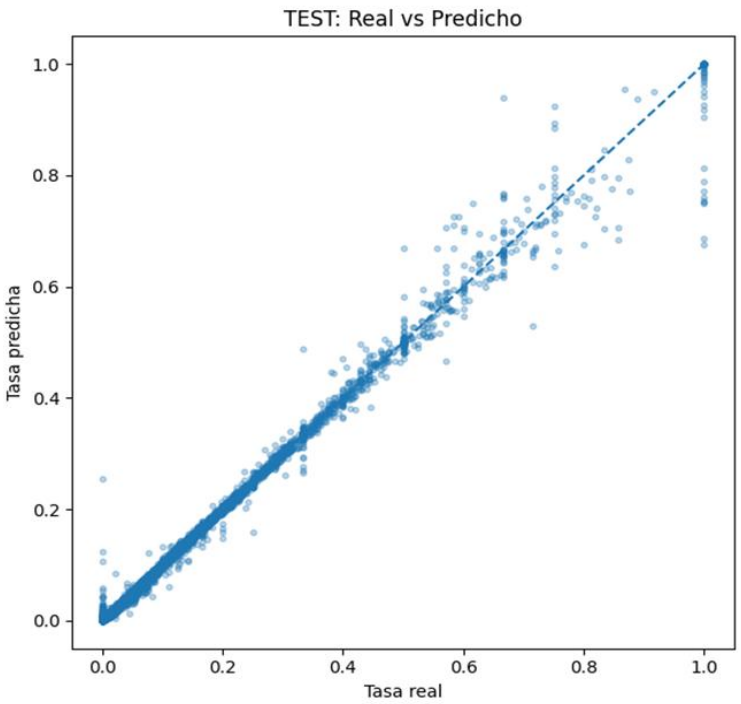
Modelo XGBoost con variables rezagadas

Análisis gráfico y diagnóstico del modelo

Se incorporan dos diagnósticos visuales para comprender el comportamiento del modelo en prueba.

Figura 4.12

Real vs Predicho en prueba



Nota. Elaboración propia con datos procesados en Python

El gráfico Tasa real vs Tasa predicha (TEST) permite inspeccionar la calibración del modelo:

· Una concentración cercana a la diagonal indica buena correspondencia entre observaciones y predicciones.

· Una dispersión creciente en valores altos sugiere mayor dificultad para estimar escenarios de abandono elevado, típicamente más heterogéneos y con menor frecuencia.

· Si se aprecia compresión hacia el centro (predicciones acumuladas en torno a un rango estrecho), puede interpretarse como un sesgo hacia la media.

Este gráfico es clave para identificar desviaciones sistemáticas como sobreestimación de valores bajos o subestimación de valores altos.

Conclusión del modelo XGBoost con lags

El modelo XGBoost con variables rezagadas extiende el enfoque temporal introducido previamente, combinando:

- memoria histórica explícita (lags y tendencias),
- alta capacidad no lineal,
- y validación temporal realista.

En la siguiente etapa se contrastan estos resultados con los otros modelos basados en lags (LightGBM y CatBoost), con el fin de seleccionar el enfoque con mejor compromiso entre precisión, estabilidad y capacidad de generalización.

Procedimiento general para la proyección de la tasa de abandono en datos futuros

Una vez entrenados los modelos predictivos con información histórica y variables rezagadas, se implementó un procedimiento específico para la proyección de la tasa de abandono en años futuros, comprendidos entre 2025 y 2030. Este proceso difiere del entrenamiento y evaluación tradicional, ya que no se dispone de valores observados de la variable objetivo, lo que obliga a reconstruir de manera controlada las variables temporales utilizadas por los modelos.

El punto de partida del procedimiento consiste en estructurar los datos bajo un enfoque de panel temporal, donde cada observación pertenece a ciertas características territoriales e institucionales (provincia, cantón, área, sostenimiento, jornada, modalidad y grado). Cada panel representa una unidad a lo largo del tiempo, para la cual existe una serie anual histórica de matrícula total y tasa de abandono. Esta definición permite tratar el problema como un conjunto de series temporales paralelas, manteniendo la coherencia estructural entre años consecutivos.

Dado que los modelos entrenados incorporan variables con memoria temporal —como rezagos de la tasa de abandono, diferencias interanuales y promedios móviles—, la proyección futura se realiza mediante un esquema recursivo año a año. De esta manera, la predicción de un

año futuro se utiliza posteriormente como insumo para generar las variables del año siguiente, replicando de manera controlada el flujo natural de la información temporal.

Antes de estimar la tasa de abandono futura, se requiere disponer de valores proyectados de la variable Total, ya que esta forma parte del conjunto de variables explicativas y de sus respectivos rezagos. Para ello, se aplica un procedimiento de pronóstico variado por panel, utilizando modelos de series de tiempo robustos y adaptables al tamaño de cada serie.

Con los valores históricos y proyectados disponibles, se inicializa un estado temporal para cada panel, que conserva los últimos valores conocidos de la tasa de abandono y del total, así como una medida resumida de tendencia reciente. Este estado funciona como memoria mínima para construir las variables rezagadas requeridas por los modelos predictivos en el primer año del horizonte de proyección.

A partir de esta inicialización, la predicción se realiza de forma iterativa para cada año futuro. En cada iteración se construye un conjunto de variables explicativas, combinando la información con las variables numéricas dinámicas, tales como rezagos, diferencias y promedios. Estas variables se organizan respetando estrictamente la estructura y el orden utilizados durante el entrenamiento de los modelos, asegurando así la consistencia del proceso de inferencia.

El conjunto de variables resultante es transformado mediante los mismos mecanismos de preprocesamiento empleados en la etapa de entrenamiento, incluyendo la codificación de variables categóricas y el tratamiento de valores faltantes. Posteriormente, el modelo predictivo correspondiente genera la estimación de la tasa de abandono para el año en cuestión. Esta estimación es acotada al rango válido de la variable y se incorpora inmediatamente al estado del panel, actualizando los rezagos y medidas de tendencia que serán utilizados en la siguiente iteración.

Este procedimiento se repite secuencialmente para cada año, de modo que las predicciones se encadenan temporalmente y reflejan la dinámica aprendida por los modelos a partir de los datos históricos. El resultado final es un conjunto coherente de predicciones anuales por panel, que mantiene la estructura territorial e institucional original y permite su posterior análisis agregado o desagregado.

Finalmente, las proyecciones obtenidas se almacenan en una tabla específica dentro de la base de datos, preservando tanto las claves del panel como los valores proyectados de la tasa de abandono y las variables auxiliares. Esta persistencia facilita la trazabilidad del proceso, la comparación entre distintos enfoques de modelado y el uso.

CAPITULO 5

5. CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

Se consiguió sistematizar un conjunto robusto de variables predictoras tales como: la provincia, el área (urbana/rural), el tipo de sostenimiento, la jornada y el grado académico, identificando que la deserción escolar puede ser visto como un proceso acumulativo influenciado por factores estructurales y territoriales.

El análisis exploratorio reveló una alta concentración de tasas de abandono cercanas a cero en ciertos grupos, debido a que los valores de tasa de abandono no son muy altos, así mismo, se puede notar ciertos picos picos significativos vinculados principalmente a la pandemia, lo que permite contemplar el el uso de transformaciones logarítmicas para estabilizar la varianza en los modelos.

Se diseñaron diversas arquitecturas, determinando que los modelos lineales (baseline) capturan tendencias generales del abandono escolar y presentan limitaciones para capturar relaciones no lineales complejas, especialmente en la presencia de valores extremos o cambios abruptos en la tasa de abandono, las cuales son mejor abordadas por redes neuronales y modelos de ensamble como LightGBM.

El modelo neuronal tabular alcanzó un R^2 de aproximadamente 0.33 en validación, lo que implica que el modelo es capaz de explicar el 33% de la varianza observada, demostrando una estabilidad temporal aceptable con un error absoluto promedio (MAE) de 3.1 puntos porcentuales,el incremento del R^2 respecto al modelo lineal, sugiere que la arquitectura neuronal logra capturar de forma más efectiva relaciones no lineales e interacciones complejas entre las variables categóricas territoriales, institucionales y las variables numéricas de contexto.

Se estableció la base para un dashboard interactivo que permita visualizar las proyecciones por provincia, facilitando la identificación de zonas críticas para la intervención institucional

5.2 Recomendaciones.

Debido a que el modelo puede fallar cuando se presentan picos en años con cambios drásticos (como la pandemia), se recomienda analizar lo referente al ajuste de cambios estructurales, que puede ser con reentrenamientos periódicos e incorporar variables externas socioeconómicas para tratar de mejorar la precisión.

Las proyecciones del modelo pueden ser útiles para implementar Sistemas de Alerta Temprana, las cuales pueden ayudar a las instituciones educativas a intervenir proactivamente antes de que el estudiante abandone definitivamente el sistema escolar.

Referencias

Amat, J. M. (2020). *Machine learning para análisis de datos con Python y R...* CreateSpace Independent Publishing Platform.

Arteaga-Hernández, K. M. (2024). Impacto de la enseñanza aprendizaje y rendimiento académico en tiempos de covid y post covid-19 en niños y adolescentes. págs. 4712-4728. Obtenido de https://www.investigarmqr.com/ojs/index.php/mqr/article/view/1137

Breiman, L. (2017). *Classification and regression trees.* Routledge.

Chaturvedi, A. &. (2024). *Introduction and types of Machine Learning.* 181-186.

Chaturvedi, A. &. (2024). *Machine learning classification algorithms: Theory and implementation.* Springer.

Chen, T. &. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, págs. 785-815.

Fortin, L. M. (2023). Multidimensional analysis of student dropout risk: Individual, family, and school factors. *Journal of School Psychology*, págs. 102-118.

Friedman, J. H. (s.f.). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, pág. 1189.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media.

Gómez-Ramírez, I. &.-V. (s.f.). Impacto de la pandemia COVID-19 en el rendimiento académico de estudiantes de educación básica. *Ciencia Latina Revista Científica Multidisciplinar*, pág. Artículo 4640.

Hastie, T. T. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer.

Lange, T. (2024). *Inteligencia artificial, machine learning y deep learning: Jerarquías conceptuales y aplicaciones prácticas en análisis predictivo.* CreateSpace Independent Publishing Platform.

Lantz, B. (2013). *Machine Learning with R.* Packt Publishing.

Mailund, T. (2017). *Introduction to Data Science: A Python Approach to Machine Learning.* Springer.

Prados, M. L. (2022). Acompañamiento, vínculo pedagógico e imaginarios sobre el primer año en tiempos de virtualización forzosa desde la perspectiva de estudiantes de primer año. *Praxis Educativa*, págs. 1-24.

Romero, C. &. (2025). Machine learning applications for dropout prediction: A systematic review of institutional and meta-analytic approaches. *Educational Data Mining Journal*.

Rumberger, R. &. (2000). The distribution of dropout and turnover rates among urban and suburban high schools. *Sociology of Education*, págs. 39-67.

Rumberger, R. (2011). Dropping out: Why students drop out of high school and what can be done about it. Harvard University Press.

Rumberger, R. A.-G. (2017). *Preventing dropout in secondary schools.* Washington, DC (opcional): National Center for Education Evaluation and Regional Assistance.

Silaparasetty, V. (2020). *Hands-on machine learning with Java for beginners.* Packt Publishing.

UNESCO. (2012). *Oportunidades perdidas: El impacto de la deserción escolar y el abandono temprano.* UNESCO.

UNESCO. (2021). *Sistemas de alerta temprana (SAT) basados en los sistemas de información para la gestión educativa (SIGED).* UNESCO.

UNESCO. (s.f. (sin fecha)). *Educación para niños*. Obtenido de https://www.unesco.org/es/gender-equality/education/boys

UNICEF & UNESCO Institute for Statistics. (2012). *Fixing the broken promise of education for all: Find out what works!* UNESCO Institute for Statistics.

UNICEF. (2017). *Early warning systems in education: Tools for dropout prevention.* United Nations Children's Fund.

VanderPlas, J. (2017). *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media.

Vásconez Altamirano, G. E. (2023). Modelo de predicción de deserción escolar en los estudiantes de la unidad educativa Los Andes por impacto de la pandemia. *Ciencia Latina Revista Científica Multidisciplinar*, pág. 3038.

Apéndice

Documentación y código fuente formato GitHub:

https://github.com/xPandy12/TrabajoFinalMaestria.git