UNIVERSIDAD NACIONAL DEL SANTA

FACULTAD DE INGENIERÍA

Escuela Profesional de Ingeniería de Sistemas e Informática

Modelo predictivo del rendimiento académico en estudiantes de primer año de secundaria a través del aprendizaje automático

Tesis para optar el título profesional de Ingeniero de Sistemas e Informática

Autores:

Bach. Lozano Torres, Jeffry Jeanpool
Bach. Pasache Pasapera, Giancarlo Andree

Asesor:

Ms. Macedo Alcántara, Dayán Fernando
DNI. N° 32941877
Código ORCID: 0000-0003-1190-4032

Nuevo Chimbote - PERÚ
2026

UNIVERSIDAD NACIONAL DEL SANTA
FACULTAD DE INGENIERÍA
Escuela Profesional de Ingeniería de Sistemas e Informática

Modelo predictivo del rendimiento académico
en estudiantes de primer año de secundaria
a través del aprendizaje automático

Tesis para Optar el Título Profesional de Ingeniero de Sistemas e Informática

Revisado y Aprobado por el Asesor:

Ms. Dayán Fernando Macedo Alcántara
DNI: 32978627
Cód. ORCID: 0000-0003-1190-4032
Asesor

UNIVERSIDAD NACIONAL DEL SANTA
FACULTAD DE INGENIERÍA
Escuela Profesional de Ingeniería de Sistemas e Informática

Modelo predictivo del rendimiento académico en estudiantes de primer año de secundaria a través del aprendizaje automático

Tesis para Optar el Título Profesional de Ingeniero de Sistemas e Informática

Revisado y Aprobado por el Jurado Evaluador:

Dr. Juan Pablo Sánchez Chávez
DNI: 17808722
Cód. ORCID: 0000-0002-3521-7037
Presidente

Ms. Mirko Martin Manrique Ronceros
DNI: 32965599
Cód. ORCID: 0000-0002-0364-4237
Secretario

Ms. Dayán Fernando Macedo Alcántara
DNI: 32978627
Cód. ORCID: 0000-0003-1190-4032
Integrante

FACULTAD DE INGENIERÍA
ESCUELA PROFESIONAL INGENIERÍA DE SISTEMAS E INFORMÁTICA

ACTA DE SUSTENTACIÓN INFORME FINAL DE TESIS

A los dieciocho días del mes de diciembre del año dos mil veinticinco, siendo las 11:30 am. En el aula S-2 del Pabellón de la Escuela Profesional de Ingeniería Sistema e Informática-FI-UNS, se instaló el Jurado Evaluador designado mediante Resolución 719-2025-UNS-CFI, y de expedito según Resolución Decanal N° 935-2025-UNS-FI integrado por los docentes: Dr. Juan Pablo Sánchez Chávez **(presidente),** Ms. Mirko Martin Manrique Ronceros **(secretario) y el** Ms. Dayan Fernando Macedo Alcántara **(Integrante)**, para dar inicio a la sustentación de la Tesis intitulada **"MODELO PREDICTIVO DEL RENDIMIENTO ACADÉMICO EN ESTUDIANTES DE PRIMER AÑO DE SECUNDARIA A TRAVÉS DEL APRENDIZAJE AUTOMÁTICO"**, perteneciente a los Bachilleres LOZANO TORRES JEFRY JEANPOOL, con código de matrícula 0201714040 Y PASACHE PASAPERA GIANCARLO ANDREE con código de matrícula 0201714003, quienes fueron asesorado por el MS. DAYAN MACEDO ALCANTARA según Resolución N°835-2023-UNS-FI

El Jurado Evaluador, después de deliberar sobre aspectos relacionados con el trabajo, contenido y sustentación del mismo, y con las sugerencias pertinentes en concordancia con el Reglamento General de Grados y Títulos, vigente, declaran aprobar:

BACHILLER	PROMEDIO VIGESIMAL	PONDERACIÓN
PASACHE PASAPERA GIANCARLO ANDREE	17	BUENO

Siendo las 12.30pm del mismo día, se dio por terminado el acto de sustentación, firmando la presente acta en señal de conformidad.

Nuevo Chimbote, 18 de diciembre de 2025

Dr. Juan Pablo Sánchez Chávez
PRESIDENTE

Ms. Mirko Martin Manrique Ronceros
SECRETARIO

Ms. Dayan Fernando Macedo Alcántara
INTEGRANTE

FACULTAD DE INGENIERÍA
ESCUELA PROFESIONAL INGENIERÍA DE SISTEMAS E INFORMÁTICA

ACTA DE SUSTENTACIÓN INFORME FINAL DE TESIS

A los dieciocho días del mes de diciembre del año dos mil veinticinco, siendo las 11:30 am. En el aula S-2 del Pabellón de la Escuela Profesional de Ingeniería Sistema e Informática-FI-UNS, se instaló el Jurado Evaluador designado mediante Resolución 719-2025-UNS-CFI, y de expedito según Resolución Decanal N° 935-2025-UNS-FI integrado por los docentes: Dr. Juan Pablo Sánchez Chávez **(presidente),** Ms. Mirko Martin Manrique Ronceros **(secretario) y el** Ms. Dayan Fernando Macedo Alcántara **(Integrante)**, para dar inicio a la sustentación de la Tesis intitulada **"MODELO PREDICTIVO DEL RENDIMIENTO ACADÉMICO EN ESTUDIANTES DE PRIMER AÑO DE SECUNDARIA A TRAVÉS DEL APRENDIZAJE AUTOMÁTICO"**, perteneciente a los Bachilleres LOZANO TORRES JEFRY JEANPOOL, con código de matrícula 0201714040 Y PASACHE PASAPERA GIANCARLO ANDREE con código de matrícula 0201714003, quienes fueron asesorado por el MS. DAYAN MACEDO ALCANTARA según Resolución N°835-2023-UNS-FI

El Jurado Evaluador, después de deliberar sobre aspectos relacionados con el trabajo, contenido y sustentación del mismo, y con las sugerencias pertinentes en concordancia con el Reglamento General de Grados y Títulos, vigente, declaran aprobar:

BACHILLER	PROMEDIO VIGESIMAL	PONDERACIÓN
LOZANO TORRES JEFRY JEANPOOL	17	BUENO

Siendo las 12.30pm del mismo día, se dio por terminado el acto de sustentación, firmando la presente acta en señal de conformidad.

Nuevo Chimbote, 18 de diciembre de 2025

Dr. Juan Pablo Sánchez Chávez
PRESIDENTE

Ms. Mirko Martin Manrique Ronceros
SECRETARIO

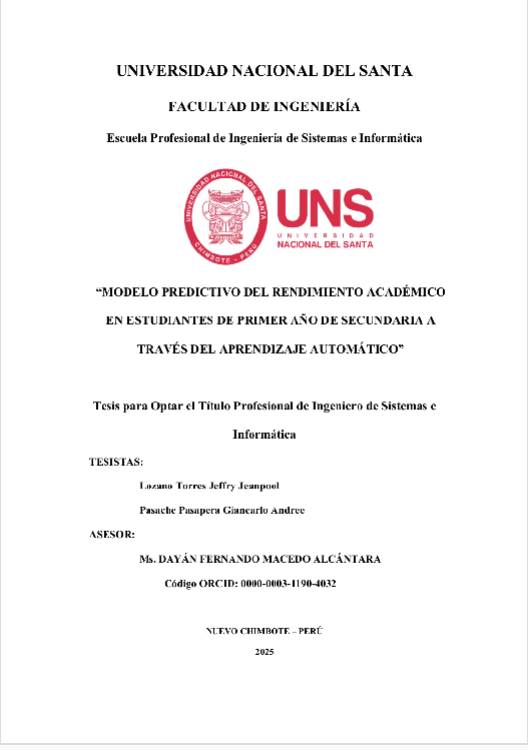
Ms. Dayan Fernando Macedo Alcántara
INTEGRANTE

Recibo digital

Este recibo confirma quesu trabajo ha sido recibido por Turnitin. A continuación podrá ver la información del recibo con respecto a su entrega.

La primera página de tus entregas se muestra abajo.

Autor de la entrega:	Giancarlo Pasache Pasapera
Título del ejercicio:	Tesis 2025
Título de la entrega:	TESIS_FINAL.pdf
Nombre del archivo:	TESIS_FINAL.pdf
Tamaño del archivo:	5.13M
Total páginas:	132
Total de palabras:	23,513
Total de caracteres:	137,620
Fecha de entrega:	05-ene-2026 05:25p. m. (UTC-0500)
Identificador de la entrega:	2853024021



UNIVERSIDAD NACIONAL DEL SANTA

FACULTAD DE INGENIERÍA

Escuela Profesional de Ingeniería de Sistemas e Informática

"MODELO PREDICTIVO DEL RENDIMIENTO ACADÉMICO EN ESTUDIANTES DE PRIMER AÑO DE SECUNDARIA A TRAVÉS DEL APRENDIZAJE AUTOMÁTICO"

Tesis para Optar el Título Profesional de Ingeniero de Sistemas e Informática

TESISTAS:

Lozano Torres Jeffry Jeampool

Pasache Pasapera Giancarlo Andree

ASESOR:

Ms. DAYÁN FERNANDO MACEDO ALCÁNTARA

Código ORCID: 0000-0003-1190-4032

NUEVO CHIMBOTE – PERÚ

2025

TESIS_FINAL.pdf

INFORME DE ORIGINALIDAD

15%	15%	5%	%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.uns.edu.pe Fuente de Internet	3%
2	hdl.handle.net Fuente de Internet	1%
3	repositorio.ucv.edu.pe Fuente de Internet	1%
4	www.coursehero.com Fuente de Internet	<1%
5	repositorio.unap.edu.pe Fuente de Internet	<1%
6	repositorio.puce.edu.ec Fuente de Internet	<1%
7	revistasinvestigacion.unmsm.edu.pe Fuente de Internet	<1%
8	repositorio.udh.edu.pe Fuente de Internet	<1%
9	juandomingofarnos.wordpress.com Fuente de Internet	<1%

DEDICATORIA

Dedicamos el presente trabajo de investigación, en primer lugar, a Dios, fuente de nuestra fortaleza, sabiduría y esperanza, por guiarnos en cada paso de este proceso académico y personal, y por permitirnos llegar hasta este logro con determinación y fe.

A nuestras familias, especialmente a nuestros padres, cuyo amor incondicional, apoyo constante y valores inculcados han sido fundamentales para forjar nuestro carácter y disciplina. Su confianza en nosotros ha sido el motor que nos impulsó a no rendirnos ante las dificultades.

A nuestros amigos, compañeros de ruta en esta etapa, por su amistad sincera, por compartir desafíos, alegrías y aprendizajes. Su compañía nos recordó que este camino, aunque exigente, también puede estar lleno de momentos memorables y apoyo mutuo.

Y a todas aquellas personas que, de una u otra forma, nos brindaron palabras de aliento, orientación o una mano solidaria: este logro también les pertenece.
Dedicado a nuestros padres **y** familia, por su inquebrantable apoyo, amor y paciencia a lo largo de todos estos años.

Lozano Torres Jeffry Jeanpool

Pasache Pasapera Giancarlo Andree

AGRADECIMIENTO

Deseamos expresar nuestro más sincero agradecimiento a todas las personas e instituciones que contribuyeron de manera significativa al desarrollo del presente proyecto de investigación.

En primer lugar, extendemos nuestro reconocimiento a nuestras familias, cuyo respaldo constante y compromiso con nuestra formación fueron esenciales para alcanzar los objetivos propuestos.

Agradecemos también a los docentes de la Universidad Nacional del Santa, por su compromiso con la excelencia académica y por brindarnos los conocimientos y herramientas necesarias para nuestra formación profesional.

De manera especial, expresamos nuestra gratitud al **Ing. Ms. Dayán Fernando Macedo Alcántara**, por su valiosa orientación, asesoramiento técnico y académico, así como por su disponibilidad y profesionalismo durante el desarrollo de esta tesis. Su acompañamiento fue determinante para la consolidación y culminación exitosa de este trabajo.

Lozano Torres Jeffry Jeanpool

Pasache Pasapera Giancarlo Andree

INDICE

INDICE DE TABLAS

INDICE DE FIGURAS

RESUMEN

Actualmente, muchas instituciones educativas enfrentan dificultades para monitorear de forma integral el rendimiento académico de sus estudiantes, lo que limita la detección oportuna de patrones de desempeño, especialmente en aquellos con bajo rendimiento, dificultando así la implementación de acciones preventivas o correctivas eficaces.

La carencia de herramientas tecnológicas para el análisis de grandes volúmenes de datos, junto al uso de registros manuales o digitales no especializados, ha generado diagnósticos tardíos o imprecisos. Esto incide negativamente en la calidad del proceso educativo.

Para superar estos desafíos, se propone el proyecto **"Modelo Predictivo del Rendimiento Académico en Estudiantes de Primer Año de Secundaria mediante Aprendizaje Automático"**, cuyo objetivo es anticipar el desempeño académico estudiantil y generar información relevante para la toma de decisiones basadas en datos. El desarrollo del modelo empleó tecnologías como Python, y bibliotecas de ciencia de datos y *machine learning* como Scikit-learn, Matplotlib, Pandas, Numpy, entre otras.

Los resultados obtenidos, diferenciados según asignatura y nivel de logro, evidenciaron que el algoritmo Random Forest fue el más efectivo, alcanzando un 50% de los casos (20 instancias). XGBoost mostró un rendimiento relevante con un 30% de efectividad (12 instancias), mientras que la Regresión Lineal Múltiple se posicionó con un 20% (8 instancias). En función de estos resultados, se decidió implementar el modelo predictivo en instituciones educativas del estudio, con el propósito de optimizar la gestión académica, mitigar riesgos de bajo rendimiento, y mejorar la calidad del sistema educativo, beneficiando tanto a estudiantes como a docentes y administrativos.

Palabras clave: Aprendizaje Automático, Predicción del Rendimiento Académico, Ciencia de Datos, Modelos Predictivos.

ABSTRACT

Currently, many educational institutions face difficulties in comprehensively monitoring the academic performance of their students, which limits the timely detection of performance patterns, especially among those with low achievement, thus hindering the implementation of effective preventive or corrective actions.

The lack of technological tools for analyzing large volumes of data, together with the use of manual or non-specialized digital records, has led to late or inaccurate diagnoses. This negatively affects the quality of the educational process.

To overcome these challenges, the project "Predictive Model of Academic Performance in First-Year Secondary Students through Machine Learning" is proposed, whose objective is to anticipate student academic performance and generate relevant information for data-driven decision making. The model development employed technologies such as Python, and data science and machine learning libraries including Scikit-learn, Matplotlib, Pandas, Numpy, among others.

The results obtained, differentiated by subject and achievement level, showed that the Random Forest algorithm was the most effective, achieving 50% of the cases (20 instances). XGBoost showed significant performance with 30% effectiveness (12 instances), while Multiple Linear Regression accounted for 20% (8 instances). Based on these results, it was decided to implement the predictive model in the studied educational institutions with the purpose of optimizing academic management, mitigating risks of low performance, and improving the quality of the educational system, benefiting students, teachers, and administrative staff alike.

Keywords: Machine Learning, Academic Performance Prediction, Data Science, Predictive Models.

PRESENTACION

Señores miembros del Jurado Evaluador:

En cumplimiento a lo dispuesto en el Reglamento General de Grados y Títulos de la Universidad Nacional del Santa, ponemos a consideración el presente Informe de Tesis intitulado: **"MODELO PREDICTIVO DEL RENDIMIENTO ACADÉMICO EN ESTUDIANTES DE PRIMER AÑO DE SECUNDARIA A TRAVÉS DEL APRENDIZAJE AUTOMÁTICO",** lo que nos permitirá optar por el título profesional de Ingeniero de Sistemas e Informática.

El presente Informe de tesis tiene como objetivo principal desarrollar un modelo basado en técnicas de aprendizaje automático para prever el rendimiento académico de estudiantes de primer año de secundaria. Este modelo permitirá a las instituciones educativas identificar a los estudiantes en riesgo de bajo rendimiento académico y tomar medidas para prevenirlo.

La información recopilada será analizada utilizando diferentes técnicas de aprendizaje automático, proponiendo que el estudio desarrolle un modelo predictivo del rendimiento académico con un alto grado de precisión.

Señores miembros del jurado evaluador, por lo que se les ha expuesto, ponemos este proyecto a su disposición para su revisión, esperando que cumpla con los requisitos mínimos para su aprobación.

Atentamente.

- Bach. Lozano Torres Jeffry Jeanpool
- Bach. Pasache Pasapera Giancarlo Andree

INTRODUCCION

La transformación digital ha modificado sustancialmente la gestión educativa, convirtiendo el manejo eficiente de datos en un componente de mucha importancia para las instituciones escolares. Los modelos predictivos basados en aprendizaje automático permiten anticipar el rendimiento académico con notable precisión, proporcionando a profesores y directivos fundamentos sólidos para decisiones pedagógicas oportunas. Particularmente en educación secundaria, surge la necesidad de modernizar los procesos de evaluación y seguimiento mediante herramientas que faciliten análisis más precisos del desempeño estudiantil. Esta modernización responde a la importancia de obtener diagnósticos tempranos para implementar estrategias de apoyo pertinentes, asegurando así una mejor calidad educativa.

El presente trabajo de investigación está estructurado en cinco capítulos: **CAPÍTULO I: INTRODUCCIÓN**, en este capítulo se contextualiza el entorno educativo del primer año de secundaria, describiendo la problemática actual en la detección temprana del bajo rendimiento académico. Se identifican las limitaciones de los métodos tradicionales, se establecen los objetivos del proyecto, se plantea la hipótesis de mejora mediante aprendizaje automático, y se justifica la importancia del modelo predictivo para optimizar la gestión académica.

CAPÍTULO II: MARCO TEÓRICO REFERENCIAL, se desarrolla la base teórica del proyecto, incluyendo investigaciones previas sobre modelos predictivos en educación, conceptos fundamentales del aprendizaje automático, tratamiento de datos educativos y metodologías como la propuesta por Geron. Se presentan estudios de caso similares y se establece el marco conceptual necesario para el desarrollo del modelo. **CAPÍTULO III: METODOLOGÍA**, describe el enfoque metodológico adoptado, detallando el diseño de la investigación, la población y muestra de estudiantes, la

operacionalización de variables relacionadas con el rendimiento académico, y las técnicas e instrumentos para la recolección y análisis de datos escolares.

CAPÍTULO IV: RESULTADOS Y DISCUSIÓN, El modelo predictivo siguió la metodología de Géron, implementando tres algoritmos específicos (Random Forest, XGBoost y Regresión Lineal Múltiple) según la naturaleza de cada asignatura y nivel de logro, donde Random Forest predominó en el 50%, XGBoost fue seleccionado en el 30 %, mientras que la Regresión Lineal Múltiple mostró efectividad del 20%.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES, se exponen las conclusiones de la investigación. Se incluyen recomendaciones para el uso continuo del modelo, su mantenimiento y futuras mejoras, con el fin de asegurar su sostenibilidad y adaptabilidad a nuevas cohortes estudiantiles.

Finalmente, se incluyen las referencias bibliográficas y los anexos correspondientes. Este proyecto busca contribuir significativamente a la modernización de los procesos de seguimiento académico, estableciendo un precedente para la integración de soluciones basadas en inteligencia artificial en el ámbito educativo.

DATOS GENERALES DEL ESTUDIO

- **TITULO DEL PROYECTO**

  "MODELO PREDICTIVO DEL RENDIMIENTO ACADÉMICO EN ESTUDIANTES DE PRIMER AÑO DE SECUNDARIA A TRAVÉS DEL APRENDIZAJE AUTOMÁTICO"

- **TESISTAS**

  - Bach. Lozano Torres Jeffry Jeanpool
  - Bach. Pasache Pasapera Giancarlo Andree

- **ASESOR**

  Ms. Dayán Fernando Macedo Alcántara

- **TIPO DE INVESTIGACION**

  a) Según su fin o propósito

  Aplicada Tecnológica, porque se desarrollaron y aplicaron modelos de aprendizaje automático para predecir el rendimiento académico de estudiantes de primer año de secundaria, proporcionando una solución práctica basada en algoritmos de aprendizaje automático para la identificación de tendencias en el desempeño estudiantil.

  b) Por el nivel de comprensión que se obtiene:

  Descriptiva-Predictiva, porque además de describir y analizar el comportamiento de las variables académicas a través del tiempo (2017-2024), se implementó un modelo predictivo que permite identificar tendencias y realizar pronósticos del rendimiento académico. El estudio no solo describe la situación actual, sino que también desarrolla capacidades predictivas a través de tres algoritmos principales (Random Forest, XGBoost y Regresión Lineal Múltiple) que en conjunto analizaron 40 casos diferentes de cursos y niveles de logro.

- **METODO DE INVESTIGACION**

  El estudio empleó un método cuantitativo con enfoque hipotético-deductivo, basado en el análisis de datos históricos del rendimiento académico de estudiantes de primer año de secundaria. La metodología se centró específicamente en el procesamiento y análisis de información académica previamente registrada. Este proceso sistemático permitió establecer las bases para el desarrollo del modelo predictivo mediante técnicas de aprendizaje automático, siguiendo los principios de objetividad y verificabilidad propios del método científico.

  El análisis se fundamentó en la identificación de tendencias en el rendimiento académico a través del tiempo. Este enfoque metodológico facilitó la construcción de conocimiento mediante un proceso de recopilación y análisis de información existente, ya que se trabajó exclusivamente con registros académicos previos

  Validación de Resultados: Se validaron los resultados mediante métricas de rendimiento (MAE, MSE, RMSE y Precisión), lo que permitió confirmar la capacidad predictiva de los modelos y, por ende, la hipótesis planteada.

CAPITULO I: INTRODUCCION

1.1. PROBLEMA

1.1.1. Realidad Problemática

El rendimiento académico es un concepto con diversos significados, que se refiere a la medida y evaluación del desempeño de los estudiantes en diferentes contextos educativos. Este concepto alude principalmente al éxito o fracaso del estudiante, siendo el objetivo del sistema educativo formar personas que logren culminar con éxito su vida académica. Por ello, es importante identificar a aquellos estudiantes que podrían enfrentar dificultades durante el año académico, para intervenir tempranamente con alternativas de mejora que les permitan superar estos obstáculos.

Touron (1985) afirma que el rendimiento académico "es el resultado del aprendizaje, provocado por la actividad educativa del profesor y con reflejo en el alumno, aunque es evidente que no todo aprendizaje es el resultado de la actividad docente". Además, señala que el desempeño no se deriva de una sola habilidad, sino de una combinación sintética de factores que influyen en y desde la persona que aprende. Esto implica que el rendimiento académico está determinado por una interacción compleja de elementos internos y externos al estudiante.

En la actualidad, el rendimiento académico de los estudiantes de primer año de secundaria es una preocupación constante para profesores, padres y autoridades educativas. Este período representa una transición crítica en la vida académica de los estudiantes, ya que pasan de la educación primaria a la secundaria, lo que conlleva nuevos desafíos y adaptaciones. Diversos factores, como las condiciones socioeconómicas, el entorno familiar, las características

psicológicas y las metodologías pedagógicas, influyen significativamente en el desempeño académico. A pesar de los esfuerzos de las instituciones educativas por mejorar el rendimiento estudiantil, muchos estudiantes continúan enfrentando dificultades que resultan en bajas calificaciones, deserción escolar y problemas de autoestima.

Uno de los principales problemas radica en que muchas instituciones no cuentan con las herramientas necesarias para detectar con anticipación posibles casos de bajo rendimiento académico. Esto deriva en una asignación deficiente o nula de recursos y apoyo para identificar las causas y brindar soluciones efectivas. Las intervenciones suelen aplicarse solo cuando los problemas ya se han manifestado, lo que reduce su efectividad. Esta falta de detección temprana puede llevar a una mayor tasa de deserción escolar, menores tasas de culminación de estudios y un rendimiento general más bajo en comparación con instituciones que implementan sistemas de monitoreo proactivo. Además, las instituciones educativas pueden enfrentar mayores costos a largo plazo debido a la necesidad de programas de recuperación intensiva y apoyo adicional para estudiantes que ya están atrasados. Para los alumnos, esto significa no recibir el apoyo necesario en el momento crítico, lo que puede generar una acumulación de problemas académicos y emocionales, pérdida de motivación, aumento del estrés y la ansiedad, y, en última instancia, abandono escolar.

Entre las problemáticas identificadas se encuentran la falta de identificación temprana de estudiantes en riesgo, las limitaciones para predecir tendencias de calificaciones futuras y la falta de personalización del aprendizaje, lo que impide detectar fortalezas y debilidades en diferentes áreas temáticas. Estas limitaciones

dificultan la implementación de estrategias efectivas para mejorar el rendimiento académico.

Por otro lado, en esta era digital, las instituciones educativas generan y recopilan grandes cantidades de datos sobre los estudiantes, sus entornos y sus interacciones con los sistemas educativos. Sin embargo, gran parte de esta información no se utiliza de manera efectiva para comprender y mejorar el rendimiento académico. En este contexto, el uso de tecnologías avanzadas, como el aprendizaje automático, presenta una oportunidad para mejorar la precisión y eficacia en la predicción del rendimiento académico. Los modelos predictivos basados en aprendizaje automático pueden analizar grandes volúmenes de datos, identificar patrones y tendencias ocultas, y proporcionar predicciones más precisas sobre el desempeño futuro de los estudiantes. Esto permitiría a las instituciones intervenir de manera proactiva, ofreciendo apoyo personalizado y recursos adicionales a aquellos estudiantes que más lo necesitan.

En resumen, la implementación de soluciones tecnológicas avanzadas, como los sistemas predictivos basados en aprendizaje automático, podría transformar la manera en que las instituciones educativas abordan el rendimiento académico. Al identificar tempranamente a los estudiantes en riesgo y personalizar las estrategias de aprendizaje, se podría mejorar no solo el desempeño académico, sino también la retención escolar y el bienestar emocional de los estudiantes. Esto representa un paso importante hacia un sistema educativo más equitativo y efectivo.

Por ello, la presente investigación orientado al desarrollo tiene como objetivo proponer un “Modelo Predictivo del Rendimiento Académico en Estudiantes de Primer Año de Secundaria a través del Aprendizaje Automático”.

1.1.2. Análisis del Problema

El presente estudio aborda la compleja problemática del rendimiento académico en estudiantes de educación secundaria, empleando herramientas de aprendizaje automático para desarrollar modelos predictivos precisos. La literatura científica ha identificado múltiples factores determinantes del desempeño académico, incluyendo variables personales, socioeconómicas, culturales y psicológicas, lo que convierte la predicción del rendimiento en un desafío multidimensional.

La evidencia empírica actual demuestra consistentemente que el historial académico constituye uno de los predictores más potentes del desempeño futuro. Schneider y Preckel (2017), en su exhaustiva revisión de meta-análisis sobre factores que influyen en el rendimiento académico en educación superior, confirmaron que el rendimiento académico previo mantiene una correlación significativa con el éxito académico posterior. El estudio destaca específicamente que los promedios generales de calificaciones en educación secundaria presentan un tamaño del efecto de Cohen d = 0.90, indicando una relación excepcionalmente robusta con el rendimiento académico subsecuente. Este hallazgo sugiere que los conocimientos y logros previos no solo reflejan el desempeño pasado, sino que funcionan como predictores confiables del aprendizaje futuro. Consecuentemente, los estudiantes con trayectorias académicas exitosas tienen mayor probabilidad de mantener un alto rendimiento en niveles educativos superiores. Para el presente estudio se utilizarán las calificaciones históricas como variable predictora principal para el rendimiento académico.

Una predicción precisa ayudará a tomar decisiones acertadas como brindar apoyo adicional en forma de tutorías o asesorías al detectar a los estudiantes que tendrían dificultades para superar los cursos.

- ✓ La falta de identificación temprana de estudiantes en riesgo puede llevar a problemas académicos que no se abordan a tiempo, resultando en bajo rendimiento y deserción escolar. Las causas incluyen métodos tradicionales de evaluación, falta de datos integrados y recursos limitados. Las soluciones potenciales abarcan la implementación de sistemas de alerta temprana y la capacitación de docentes en técnicas de identificación de riesgos.
- ✓ Las limitaciones en la identificación de patrones de calificaciones a futuro dificultan la previsión de problemas y la planificación de intervenciones efectivas. Las causas son el análisis insuficiente de datos, la fragmentación de la información y la falta de herramientas avanzadas. Las soluciones incluyen el desarrollo de modelos predictivos utilizando aprendizaje automático y la integración de datos de diversas fuentes.
- ✓ La falta de personalización en el aprendizaje puede resultar en una enseñanza menos efectiva y desmotivación estudiantil. Las causas son un enfoque uniforme en la enseñanza, recursos limitados y falta de datos sobre el estudiante. Las soluciones potenciales son el uso de tecnología educativa personalizada y el desarrollo de estrategias de enseñanza individualizadas para adaptarse a las necesidades de cada estudiante.

1.1.3. Formulación del Problema

¿En qué medida un modelo predictivo basado en aprendizaje automático permitirá predecir el rendimiento académico de los estudiantes de primer año de secundaria?

1.2. OBJETIVOS

1.2.1. Objetivo General

Predecir el rendimiento académico de los estudiantes de primer año de secundaria mediante un modelo de aprendizaje automático.

1.2.2. Objetivo Específicos

- ✓ Determinar la métrica de robustez frente a errores individuales en la predicción de las calificaciones numéricas de los estudiantes de primer año de secundaria mediante el cálculo del Error Absoluto Medio (MAE).
- ✓ Evaluar la precisión predictiva del modelo en la estimación del promedio académico de los estudiantes de primer año de secundaria, utilizando el error cuadrático medio (MSE) como métrica de desempeño.
- ✓ Evaluar la precisión predictiva del modelo en la estimación del promedio académico de los estudiantes de primer año de secundaria, utilizando la Raíz del Error Cuadrático Medio (RMSE) como métrica de desempeño.
- ✓ Comparar los resultados de las métricas (MSE, RMSE y MAE) para seleccionar el algoritmo de regresión más eficiente y preciso.

1.3. HIPÓTESIS

Un modelo predictivo basado en técnicas de aprendizaje automático permite predecir con alta precisión el rendimiento académico de los estudiantes de primer año de secundaria.

1.4. JUSTIFICACIÓN DE LA INVESTIGACIÓN

1.4.1. Justificación operativa

Este proyecto se justifica operativamente por:

- Un modelo predictivo permite asignar de manera óptima recursos materiales, humanos y tecnológicos de manera más eficiente,

centrándose en los estudiantes que más los necesitan.

- Detectar oportunamente distintos niveles de rendimiento académico posibilita intervenciones más precisas, promoviendo el desarrollo integral del estudiantado y favoreciendo el acompañamiento personalizado según sus necesidades y potencialidades.
- El aprendizaje automático automatiza el proceso de predicción, reduciendo carga administrativa y permitiendo a los profesores concentrarse en el diseño y la mejora continua de estrategias pedagógicas.

1.4.2. Justificación social

Este proyecto socialmente se justifica por:

- El éxito educativo de los estudiantes está estrechamente vinculado a su rendimiento académico. Un modelo predictivo permite identificar de forma temprana y precisa los distintos niveles de desempeño, lo que facilita la planificación de intervenciones personalizadas para potenciar las fortalezas y abordar debilidades en cada grupo.
- Al predecir el rendimiento académico de manera integral, se pueden diseñar estrategias pedagógicas diferenciadas que promuevan el desarrollo equitativo de todos los estudiantes. De manera puntual, aquellos con bajo rendimiento pueden ser priorizados con medidas de apoyo adicional, lo que contribuye a reducir la deserción escolar y a mejorar los resultados globales.
- La anticipación de situaciones de riesgo académico permite implementar acciones preventivas y correctivas que fomenten la igualdad de oportunidades, garantizando un acompañamiento oportuno y efectivo para cada estudiante según su nivel de desempeño.

1.4.3. Justificación económica

Este proyecto se justifica económicamente por:

- La identificación temprana de problemas permite reducir costos a largo plazo, reduciendo la inversión en apoyo intensivo y repetición de cursos.
- Se dirigen los recursos financieros y humanos a donde se necesitan más, incrementando la eficiencia de la inversión, maximizando el retorno de la inversión en educación y optimizando los gastos relacionados con la gestión académica.

1.5. IMPORTANCIA DE LA INVESTIGACIÓN

Este proyecto de investigación se presenta como una herramienta clave para la detección temprana del rendimiento académico de los estudiantes, con especial atención a aquellos que muestran bajo desempeño, pero sin dejar de lado una visión integral de todos los niveles de rendimiento. Esta capacidad de anticipación permite actuar a tiempo con medidas adecuadas, evitando que los problemas se agraven y mejorando así las oportunidades de aprendizaje.

El uso de modelos predictivos brinda a las instituciones educativas una base sólida para diseñar e implementar estrategias de intervención más efectivas y focalizadas, lo que contribuye a reducir las brechas de rendimiento entre distintos grupos estudiantiles.

Desde una perspectiva académica, esta investigación aporta valor al validar con datos reales la relación entre el historial académico y el rendimiento futuro. Para ello, se aplican técnicas avanzadas de aprendizaje automático, que superan las limitaciones de los enfoques estadísticos tradicionales al identificar patrones complejos y no lineales en los datos educativos.

Además, la metodología adoptada fomenta la integración entre la educación, la estadística y la informática, generando un enfoque interdisciplinario que enriquece el análisis de los problemas educativos.

1.6. ALCANCES Y LIMITACIONES

- El presente estudio se enfoca en el desarrollo de un modelo predictivo basado en técnicas de aprendizaje automático, utilizando datos históricos provenientes de dos instituciones educativas emblematicas durante el periodo 2017–2024. El objetivo principal es anticipar el rendimiento académico de estudiantes de primer año de secundaria a partir de información cuantitativa previamente registrada (calificaciones).
- El estudio contempla el diseño del modelo, el análisis y limpieza de datos, la selección de algoritmos adecuados, la evaluación de su precisión, y la construcción de un prototipo funcional que pueda ser la base para futuras aplicaciones en entornos reales. Esta investigación proporciona un enfoque preventivo para la identificación temprana de asignaturas con riesgo académico para los estudiantes, facilitando la toma de decisiones pedagógicas más oportunas y eficaces.
- El modelo predictivo desarrollado se basa exclusivamente en datos históricos, lo que garantiza un enfoque riguroso y controlado, aunque no incluye validación en tiempo real en centros educativos. Sin embargo, esta característica permite evaluar su rendimiento con alta precisión antes de su aplicación práctica.
- Asimismo, si bien el estudio utiliza información de dos instituciones educativas, este enfoque permite una mayor profundidad en el análisis y facilita futuras adaptaciones a nuevos contextos. El modelo se concentra en variables objetivas disponibles en los registros escolares, lo cual asegura consistencia y replicabilidad en su desarrollo.

- Al estar orientado específicamente a estudiantes de primer año de secundaria, el modelo ofrece resultados altamente especializados, siendo una base sólida para extender su aplicación a otros niveles educativos en futuras investigaciones.

CAPITULO II: MARCO TEORICO REFERENCIAL

2.1. ANTECEDENTES

A continuación, se enumeran algunos proyectos de investigación recientes; cada uno con aportes valiosos para el tema de estudio de esta presente investigación.

2.1.1. Antecedentes a Nivel Internacional

- **Antecedente 01:**

  López-García, L., Lino-Ramírez, C., Zamudio-Rodríguez, V. M., & Del Valle-Hernández, J. (2022). *Predictive model for the analysis of academic performance and preventing student dropout using machine learning techniques***.** Revista de Educación Técnica, 1-5. https://doi.org/10.35429/jote.2022.16.6.1.5

  En el artículo: "Modelo predictivo para el análisis del rendimiento académico y prevenir la deserción estudiantil utilizando técnicas de aprendizaje automático", los autores afirman que uno de los mayores problemas en México es la deserción escolar, la cual es causada por una variedad de factores, por lo que es necesario desarrollar estrategias y medidas para disminuirla. Esta investigación analiza una base de datos con datos demográficos y sociales de los estudiantes de secundaria, que se recopilaron mediante cuestionarios e informes escolares. El objetivo de este análisis es identificar los factores que contribuyen al abandono escolar y identificar a tiempo a los estudiantes que necesitan asesoramiento personalizado para ofrecerles orientación educativa y evitar el abandono escolar. El análisis se llevó a cabo utilizando técnicas de aprendizaje automático para crear un modelo predictivo utilizando un algoritmo de descenso de gradiente. Los resultados demostraron los errores de pronóstico

utilizando la métrica de error cuadrático medio para estimar los posibles errores de predicción del modelo. Se espera que el uso de estas técnicas de aprendizaje automático en la comunidad educativa tenga un impacto social significativo, ya que permitirá que los estudiantes puedan fortalecer su formación integral, además de orientar sus talentos e intereses.

- **<u>Antecedente 02</u> :**

  (Guamán Luna et al., 2023). *Comparación entre Modelos de Regresión Lineal Múltiple Vs Redes Neuronales Artificiales Supervisadas en la Predicción de Calificaciones Ser Bachiller 2018-2019 del Ecuador*. Revista Iberoamericana de la educación, 7(2). https://doi.org/10.31876/ie.v7i2.249.

  En este artículo se compara los modelos de regresión lineal múltiple y las redes neuronales artificiales supervisadas para predecir el rendimiento académico en forma de calificaciones de la evaluación Ser-Bachiller en Ecuador.

  El estudio utiliza datos de las pruebas Ser-Bachiller de Ecuador en el ciclo 2018-2019 y evalúa los modelos que predicen los puntajes en los ámbitos de las matemáticas, la lingüística, las ciencias y las ciencias sociales.

  El artículo concluye que, si bien los modelos de regresión lineal son ligeramente más precisos en las predicciones, no cumplen con los supuestos de linealidad, homocedasticidad e independencia. Por lo tanto, se recomiendan las redes neuronales artificiales supervisadas para predecir las calificaciones de Ser Bachiller en Ecuador.

  El estudio reconoce las limitaciones a la hora de ajustar los modelos de redes neuronales para grupos más pequeños debido a las limitaciones de la

capacidad computacional y sugiere que en el futuro se estudien las agregaciones de nivel superior en las instituciones educativas.

- **Antecedente 03:**

  Henríquez Cabezas, N., & Vargas Escobar, D. (2022). *Modelos predictivos de rendimiento y deserción académica en estudiantes de primer año de una universidad pública chilena*. Revista de Estudios y Experiencias en Educación, 21(45), 299-316. https://doi.org/10.21703/0718-5162.v21.n45.2022.015.

  En este artículo se concluye que la falta de educación universitaria se ha convertido en un tema de gran importancia para el público en general, debido a los recursos que el gobierno y las familias destinan a la educación de los jóvenes en Chile. Por lo tanto, el objetivo de este estudio es modelar un sistema de alerta temprana para prevenir la deserción académica analizando el rendimiento académico. La investigación utiliza una estrategia asociativa y es de tipo cuantitativo, no predictivo experimental. La muestra se extrajo de la población de estudiantes que comenzaron su primer año en la Prueba de Selección Universitaria Chilena en 2014 (N=739), quienes se sometieron a análisis diferenciados en función de las cuatro facultades de una universidad pública en Chile. Los modelos predictivos se obtuvieron utilizando modelamiento logístico y se establecieron puntos de cohorte académicos para cada facultad. Los resultados del análisis indican que las dos facultades utilizaron la curva ROC como método para obtener criterios de detección y discriminación. Sin embargo, en el primer semestre de la Facultad de Filosofía y Educación, no fue posible crear un modelo. Se

descubrió que los modelos exhiben capacidad predictiva en función del porcentaje de riesgo académico de los estudiantes.

2.1.2. Antecedentes a Nivel Nacional

- **Antecedente 04:**

Garcia, J., (2021). Machine learning para predecir el rendimiento académico de los estudiantes universitarios. (Informe final de Trabajo de Grado). Universidad Cesar Vallejo. Lima.

En este estudio se creó un modelo de aprendizaje automático para predecir el rendimiento académico de los estudiantes universitarios. Se utilizó la metodología KDD y herramientas como SPSS y SPSS Modeler para crear el modelo predictivo. El objetivo de esta investigación es determinar en qué porcentaje el aprendizaje automático tiene la capacidad de predecir el rendimiento académico con precisión, sensibilidad y especificidad, para que se pueda determinar la probabilidad de que los estudiantes tengan éxito o fracaso.

Esta investigación utilizó una población de 87 estudiantes como muestra. Por otro lado, el estudio es de tipo aplicada, con un diseño de investigación experimental de tipo preexperimental de un solo grupo, ya que se podrán observar los resultados y realizar la medición después de implementar el aprendizaje automático. Se ha demostrado que el aprendizaje automático puede predecir el rendimiento académico de los estudiantes universitarios debido a su precisión, sensibilidad y especificaciones para los algoritmos de árbol de decisión, vectores y K-NN. Además, se ha demostrado que el algoritmo de vectores (SVM) con un valor de 100% fue el mejor en esta situación.

- **<u>Antecedente 05</u>:**

  Vega García (2019). *Modelo de pronóstico de rendimiento académico de alumnos en los cursos del programa de estudios básicos de la Universidad Ricardo Palma usando algoritmos de Machine Learning*. Universidad Ricardo Palma, Maestría en Ciencia de los Datos.

  Se desarrolló un modelo de pronóstico basado en algoritmos de Machine Learning con el objetivo de predecir el rendimiento académico de los estudiantes, específicamente la cantidad de alumnos aprobados y desaprobados. La población estudiada estuvo conformada por 9118 estudiantes pertenecientes a los períodos académicos 2015-I al 2019-0, cuya información fue recopilada mediante técnicas de recolección de datos. El modelo demostró una diferencia del 4.9% en la predicción de estudiantes aprobados y del 28% en la predicción de alumnos desaprobados en comparación con los datos reales. Como conclusión, se determinó que es viable aplicar un modelo de Machine Learning para la determinación del rendimiento académico. Este trabajo proporciona una base sólida para futuras investigaciones en el ámbito del pronóstico académico utilizando técnicas de aprendizaje automático.

- **<u>Antecedente 06</u>:**

  Candia Oviedo (2019). *Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático*. Universidad Nacional de San Antonio Abad del Cusco, Maestría en Ciencias mención Informática.

Se desarrolló un modelo predictivo basado en algoritmos de aprendizaje automático con el objetivo de predecir el rendimiento académico de los estudiantes utilizando sus datos de ingreso. La población estudiada estuvo conformada por una muestra probabilística de 12,698 estudiantes ingresantes desde el semestre 2014-I hasta 2018-I, cuyos datos fueron obtenidos a través de las bases de datos del Centro de Cómputo de la UNSAAC, aplicando técnicas de extracción, transformación y carga de datos. El algoritmo de árboles de decisión Random Forest demostró el mayor porcentaje de predicción acertada, alcanzando un 69.35% de efectividad. Como conclusión, se determinó que es posible predecir el rendimiento académico a partir de los datos de ingreso o admisión a la UNSAAC, utilizando algoritmos de aprendizaje automático con una efectividad del 69%. Este trabajo proporciona una base metodológica y técnica para futuras investigaciones en el ámbito de la predicción del rendimiento académico.

2.1.3. Antecedentes a Nivel Local

- **Antecedente 07:**

Cahuana, J., (2021). *Factores determinantes asociados al rendimiento académico mediante machine learning en estudiantes de la asignatura de matemática I, UNASAM – 2019*. (Informe final de Trabajo de Maestría). Universidad Nacional Santiago Antúnez de Mayolo, Huaraz.

El estudio examinó los factores determinantes para el rendimiento académico de los estudiantes de la asignatura de Matemática I. Se identificaron tres categorías de factores: personales, sociales e institucionales.

Dentro de los factores personales se incluye lo siguiente: edad menor o igual a 18 años, sexo femenino, procedencia fuera de Huaraz.

En los factores sociales se encontró: pertenecer a una familia extensa/monoparental, trabajar actualmente, padecer enfermedad, tipo de alimentación, tener pareja y que los padres estén separados/viudos, relaciones familiares en casa regulares, número de personas en casa mayor a 4, asistir a una secundaria estatal.

También se detectó factores institucionales como: Infraestructura educativa inadecuada en la UNASAM, falta de libros adecuados de Matemática I en la biblioteca, poca preocupación de las autoridades de la UNASAM por el aprendizaje de los estudiantes, adaptabilidad de las aulas para recibir clases, calificación deficiente del servicio de bienestar social del estudiante.

- **Antecedente 08:**

  Espinoza, G., León, E. (2020). *Modelo de Machine Learning para la clasificación de estudiantes de acuerdo con su rendimiento académico en el Centro de Idiomas de la Universidad Nacional del Santa* (Informe final de Trabajo de Grado). Universidad Nacional del Santa. Chimbote.

  El objetivo principal de esta tesis es mejorar el proceso de clasificación de los estudiantes del centro de idiomas utilizando el Aprendizaje Automático, centrándose en diferentes niveles, para así cumplir con el objetivo principal de la institución, que es brindar enseñanza de un idioma extranjero o nativo. Las predicciones de clasificación que se pueden hacer en el futuro en base a los registros de datos históricos que se pueden obtener gracias a las nuevas tecnologías de información, especialmente a los sistemas inteligentes. Esto permite crear una solución que plasme esta información y sirva como

herramienta de conocimiento en el proceso de clasificación de los estudiantes según su rendimiento académico en el CEIDUNS.

Como resultado, se logró una reducción en el tiempo promedio de clasificación de los estudiantes según su rendimiento académico en un 74.60% (de 218.19 segundos a 55.42 segundos), un aumento en el número de clasificaciones acertadas en un 82.08%), y un aumento en el nivel de satisfacción del personal docente en un 71.35% y el nivel de satisfacción de los estudiantes en un 66.30% utilizando el modelo predictivo. El personal de CEIDUNS pudo identificar a los estudiantes, asignarles aulas y aumentar el número de docentes gracias al modelo de aprendizaje automático. El sistema de predicción de clasificación logró su objetivo principal, que era mejorar el proceso.

- **Antecedente 09:**

Caselli (2021). *Modelo predictivo basado en Machine Learning como soporte para el seguimiento académico del estudiante universitario.* Repositorio Institucional Universidad Nacional del Santa.

Se desarrolló un modelo predictivo basado en Machine Learning con el objetivo de optimizar la gestión del seguimiento académico. La población estudiada estuvo conformada por los estudiantes de la Universidad Nacional del Santa, cuya información fue recopilada a través de las distintas oficinas de información de la universidad, utilizando técnicas de extracción, transformación y carga de datos. El modelo del experimento Nº 13 demostró una menor diferencia de precisión entre el conjunto de entrenamiento (98.97%) y el conjunto de prueba (81.73%), con una variación de 14.24 puntos porcentuales. Como conclusión, se determinó que la aplicación del

modelo predictivo basado en Machine Learning permitió mejorar significativamente el seguimiento académico de los estudiantes. Este trabajo proporciona una base clara para el diseño de futuros modelos predictivos.

2.2. MARCO CONCEPTUAL

2.2.1. Rendimiento Académico

Según (Himmel, 2018), "El rendimiento académico es un componente clave de la integración académica del estudiante en una institución educativa, queinfluye en la reevaluación de su compromiso con la meta de graduarse. Un buen rendimiento académico refuerza este compromiso, mientras que un bajo rendimiento tiene un efecto negativo, incrementando la probabilidad de deserción o abandono de los estudios"

Existe una variedad de categorías de modelos que explican el rendimiento académico y la posterior retención o deserción estudiantil enfatizando factores psicológicos, económicos, sociológicos, organizacionales e interaccionales.

- Los modelos psicológicos se enfocan en las características de personalidad, actitudes, intenciones, objetivos y valores que distinguen a los estudiantes que terminan sus estudios de los que abandonan.
- Los modelos sociológicos enfatizan elementos externos como la integración social y el contexto familiar.
- Los modelos organizacionales se enfocan en características institucionales como la calidad de la enseñanza, los servicios estudiantiles y los recursos.
- Los modelos integrados más recientes incorporan una variedad de factores, incluida experiencia académica y social, finanzas, características institucionales, experiencias académicas y sociales, etc., en un modelo procesal longitudinal.
- Entender cómo estas variables se combinan en diferentes instituciones y grupos de estudiantes puede ayudar a identificar los elementos críticos para

reducir la deserción

2.2.2. Metodología de Desarrollo.

(Géron, 2022, p. 37-85) propone una serie de etapas surgidas de la práctica yla experiencia acumulada en el campo de la ciencia de datos que se describena continuación:

a) Comprender el panorama general:

- Comprender el objetivo del negocio del proyecto y como Machine Learning puede contribuir a ese objetivo.
- Definir el tipo de problema que se va a abordar (por ejemplo, clasificación, regresión, clustering, etc.).
- Seleccionar las métricas de rendimiento adecuadas para evaluar el modelo.

b) Obtener los datos:

- Recopilar los datos necesarios para el proyecto, ya sea de bases de datos existentes, fuentes en línea o generación interna.
- Asegurarse de que los datos están en un formato adecuado para su procesamiento.

c) Descubrir y visualizar los datos para obtener información:

- Explorar los datos para comprender su estructura, distribuciones, correlaciones y posibles patrones descubriendo posibles datos estructurados, no estructurados, continuos, discretos o de series temporales.
- Visualizar los datos utilizando gráficos y estadísticas descriptivas para identificar tendencias y anómalas.

d) Preparar los datos para los algoritmos de Machine Learning:

   - Limpieza de datos, incluida la gestión de valores atípicos, valores faltantes y datos erróneos.
   - Selección y transformación de características relevantes para el modelo.
   - División de los datos en conjuntos de entrenamiento y prueba.

e) Seleccionar un modelo y entrenarlo:

   - Elegir un algoritmo de Machine Learning adecuado según el tipo de problema y los datos disponibles.
   - Entrenar el modelo utilizando el conjunto de entrenamiento y ajustar los hiperparámetros según sea necesario.

f) Ajustar el modelo:

   - Validar el rendimiento del modelo utilizando el conjunto de prueba y técnicas como la validación cruzada.
   - Ajustar el modelo según los resultados de la validación para mejorar su rendimiento.

g) Presentar la solución:

   - Comunicar los hallazgos y resultados del modelo a las partes interesadas de manera clara y comprensible.
   - Demostrar como el modelo aborda el problema del negocio y proporciona valor.

h) Implementar, monitorear y mantener el sistema:

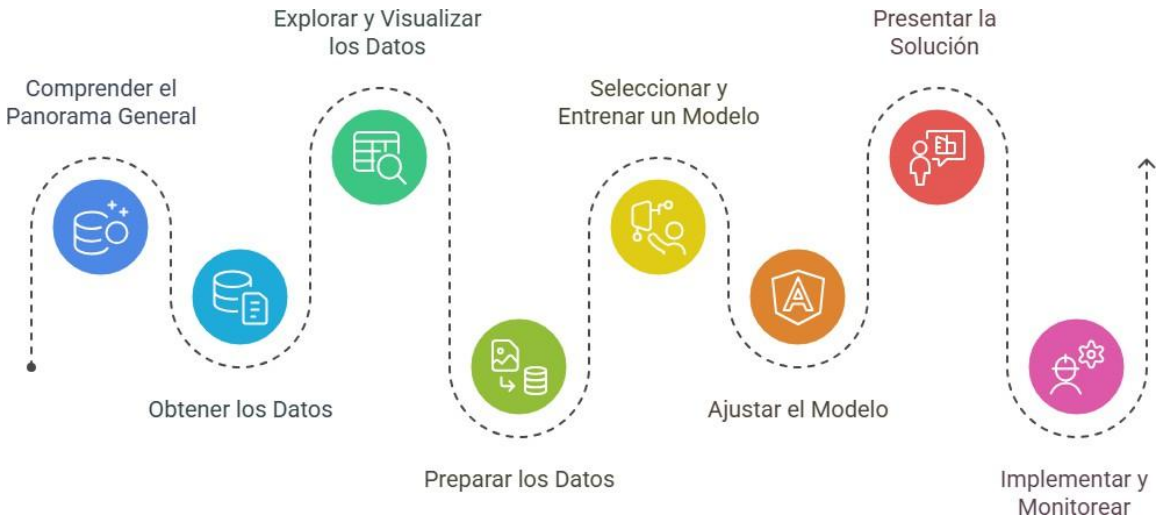
   - Desplegar el modelo en un entorno de producción y establecer un

sistema de monitoreo para evaluar su rendimiento continuamente.

- Realizar ajustes y actualizaciones según sea necesario para mantener la eficacia del modelo a lo largo del tiempo.

Figura 1

Metodología de desarrollo de Géron



Nota. Adaptado de (Géron, 2022, p. 37-85)

2.2.3. Machine Learning

El término "aprendizaje automático" se refiere a las técnicas de análisis de datos que automatizan la creación de modelos analíticos. Es una rama de la inteligencia artificial que se basa en la idea de que podemos permitir que las computadoras accedan a los datos y los ordenen por sí mismas. (Mueller & Massaron, 2016, p.3).

"Los algoritmos utilizados en el aprendizaje automático tienen la capacidad de inferir datos y aprender de ellos. Estos algoritmos hacen posible que las computadoras descubran patrones ocultos en los datos sin necesidad de programarse explícitamente. Las redes neuronales, los árboles de decisión, el

naive Bayes, los k-means y otros algoritmos son parte del aprendizaje automático. Estos algoritmos hacen predicciones sin estar programados para realizar la tarea, utilizando un modelo matemático basado en muestras de datos conocido como conjunto de entrenamiento." (Raschka, 2022, p.1)

2.2.4. Coeficiente de Variación

Según Gómez Barrantes, M. (2012)., en el análisis de datos y la investigación existe frecuentemente la necesidad de comparar la variabilidad entre diferentes conjuntos de datos. Aunque la desviación estándar es una herramienta útil cuando los datos tienen unidades y promedios similares, pierde efectividad cuando estas condiciones no se cumplen. El autor identifica dos situaciones problemáticas principales: cuando los datos están expresados en diferentes unidades de medida (por ejemplo, kilogramos vs. centímetros) y cuando existe una diferencia significativa entre los promedios de los conjuntos de datos. Para resolver esta problemática, introduce el coeficiente de variación como una medida de dispersión relativa, que se define como el cociente entre la desviación estándar y la media aritmética, multiplicado por 100.

$$Coeficiente\ de\ Variacion = \frac{Desviacion\ estandar}{Media\ Aritmetica}\ x\ 100$$

Esta solución se justifica por dos razones principales: proporciona valores independientes de las unidades de medida y neutraliza el efecto de la magnitud general de los datos. En términos matemáticos, se presentan dos fórmulas según el caso: para población **(C.V. = (σ/μ) × 100)** y para muestra **(C.V. = (s/x̄) × 100).** El autor concluye que esta medida resulta efectiva porque elimina el problema de las unidades al dividir la desviación estándar entre la media, controla el efecto de la magnitud general de los datos, y la multiplicación por 100 facilita la interpretación al convertir el resultado en un número relativo más cómodo de usar. Este coeficiente se presenta como una solución práctica para comparar la variabilidad entre conjuntos de datos con diferentes características, superando las limitaciones de la desviación estándar como medida absoluta de dispersión.

Rangos Generales del CV

Según Posada Hernández, G. J. (2016), considera los siguientes criterios para calificar estadísticamente la calidad de las estimaciones.

Tabla 1:

Rango e interpretación del Coeficiente de Variación

Rango de CV	Interpretación CV	Interpretación	Significado en Contexto Educativo
≤ 0.07 (≤7%)	Variabilidad muy baja	Estimación precisa	Altísima precisión en las calificaciones; resultados muy consistentes
[0.08 - 0.14] (8% - 14%)	Variabilidad baja	Precisión aceptable	Buen nivel de estabilidad; el rendimiento presenta ligeras variaciones
[0.15 - 0.20] (15% - 20%)	Variabilidad moderada	Precisión regular	Fluctuaciones más evidentes; se requiere atención a posibles tendencias
> 0.20 (>20%)	Variabilidad alta	Estimación poco precisa	Alta inestabilidad en el rendimiento; difícil establecer patrones claros

Nota. Adaptado de Posada Hernández, G. J. (2016)

2.2.5. Métodos de Suavización

Según Anderson D., Sweeney D. y Williams T. (2008), los métodos de suavizamiento tienen la finalidad principal "suavizar" las variaciones aleatorias provocadas por el componente irregular presente en las series temporales, razón por la cual se les denomina métodos de suavización. Los métodos de suavización se caracterizan por su facilidad de uso y, en general, ofrecen un elevado nivel de precisión en pronósticos a corto plazo, como aquellos destinados al período inmediato siguiente. Entre los métodos más conocidos tenemos los promedios móviles, promedios móviles ponderados y el suavizamiento exponencial.

Suavizamiento Exponencial

El suavizamiento exponencial es un método para predecir valores futuros en una serie de tiempo. Utiliza un promedio ponderado de los valores pasados, donde los datos más recientes tienen más peso que los más antiguos. Solo necesitas elegir un peso para la observación más reciente, y los pesos para los demás datos se calculan automáticamente.

La fórmula básica del suavizamiento exponencial es:

$F_{t+1} = \alpha Y_t + (1-\alpha)F_t$

donde:

- F_{t+1} es el pronóstico para el próximo periodo.
- Y_t es el valor real en el periodo actual.
- F_t es el pronóstico para el periodo actual.
- α es una constante de suavizamiento entre 0 y 1.

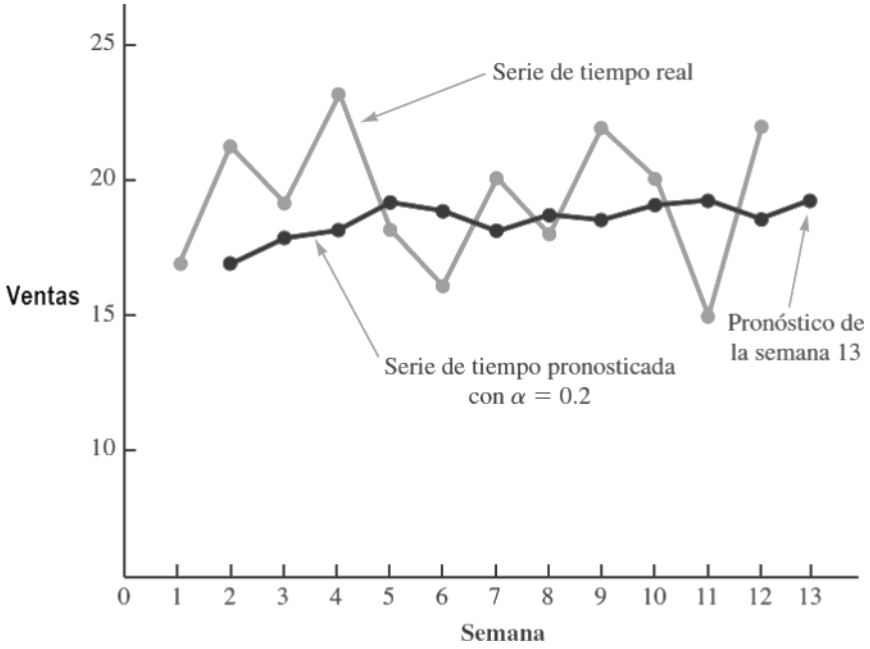
En este método, el pronóstico para el próximo periodo es un promedio ponderado del valor real actual y el pronóstico actual. El peso dado al valor real actual es α, y el peso dado al pronóstico actual es $1-\alpha$

Un ejemplo con tres datos $Y1, Y2, Y3$ muestra que el pronóstico para cualquier periodo es un promedio ponderado de todos los valores reales anteriores. No necesitas guardar todos los datos pasados para hacer el pronóstico; solo necesitas el valor real y el pronóstico del periodo actual, junto con la constante de suavizamiento α.

El suavizamiento exponencial se distingue por requerir una cantidad mínima de datos, lo que la convierte en una opción idónea cuando se necesitan pronósticos para un gran número de artículos.

Figura 2

Métodos de Suavización



Nota. Adaptado de Anderson D., Sweeney D. y Williams T. (2008)

2.2.6. Regresión Lineal Múltiple

Según Roback y Legler (2020), La regresión lineal múltiple es una técnica estadística que modela la relación entre una variable dependiente continua y múltiples variables predictoras (independientes). El modelo se expresa como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

Donde:

- Y es la variable respuesta
- $X_1, X_2, ..., X_p$ son las variables predictoras
- $\beta_0, \beta_1, \beta_2, ..., \beta_p$ son los coeficientes de regresión
- ε representa el error aleatorio

los supuestos clave para la regresión lineal por mínimos cuadrados son:

1. Linealidad: La relación entre predictores y la variable respuesta es lineal
2. Independencia: Las observaciones son independientes entre sí

3. Homocedasticidad: Varianza constante en los errores
4. Normalidad: Los errores siguen una distribución normal

Métodos de Estimación y Validación

1. Métodos robustos de estimación de parámetros
2. Técnicas de validación cruzada para evaluar el rendimiento predictivo
3. Métodos bootstrap para la inferencia estadística
4. Diagnósticos avanzados para verificar supuestos del modelo

Selección de Variables y Construcción del Modelo

El texto destaca estrategias contemporáneas para la construcción de modelos:

- Métodos de selección automática (forward, backward, stepwise)
- Criterios de información (AIC, BIC)
- Validación cruzada para evitar el sobreajuste
- Regularización (ridge, lasso) para manejar la multicolinealidad

Interpretación y Visualización

- Interpretación contextual de los coeficientes
- Visualización de efectos parciales
- Análisis de interacciones mediante gráficos
- Evaluación de la importancia relativa de predictores

2.2.7. Random Forest

Según (Genuer & Poggi, 2020, p. 33), Random Forest es un método de aprendizaje estadístico ampliamente utilizado en diversos campos gracias a su excelente rendimiento predictivo y su flexibilidad, que impone pocas restricciones sobre la naturaleza de los datos utilizados.

"Los bosques aleatorios son parte de la familia de métodos basados en árboles. Pueden adaptarse tanto a problemas de clasificación supervisada como a problemas de regresión. Además, permiten considerar variables explicativas cualitativas y cuantitativas juntas, sin preprocesamiento”

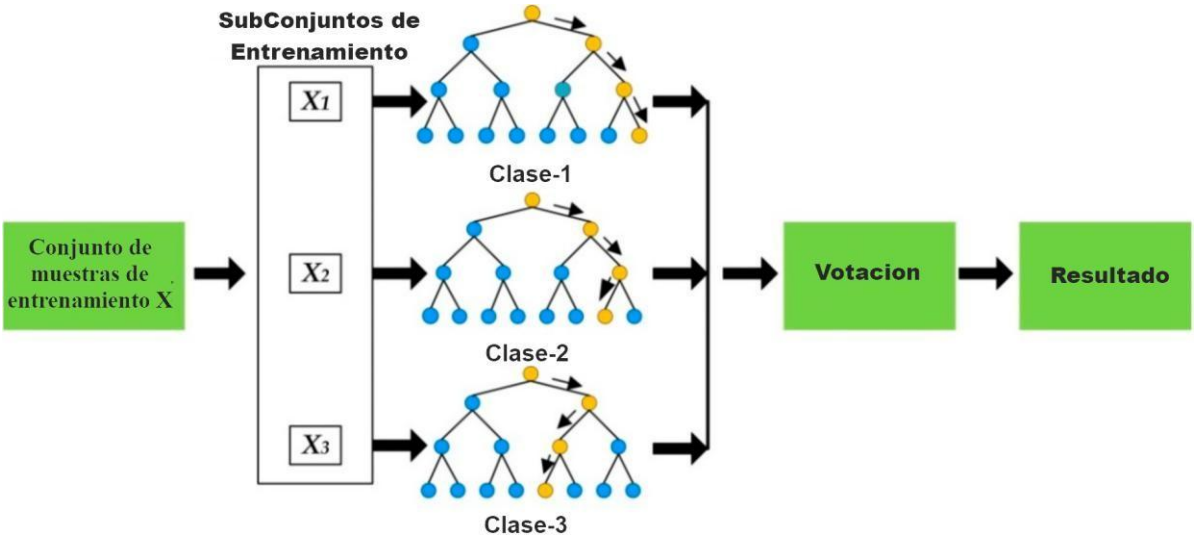
Características Principales

- **Métodos basados en árboles**: Los Random Forests pertenecen a una categoría de algoritmos que utilizan árboles de decisión como componentes básicos. Estos árboles dividen los datos recursivamente en subconjuntos más homogéneos basándose en reglas de decisión simples.
- **Versatilidad en tipos de problemas:**
  - **Clasificación supervisada**: Pueden predecir categorías o clases (como "sí/no", "bueno/malo", o múltiples categorías).
  - **Regresión:** Pueden predecir valores numéricos continuos (como precios, temperaturas o cualquier variable cuantitativa).
- **Manejo flexible de variables:**
  - **Variables cualitativas:** Categorías, etiquetas o variables nominales (como género, color, tipo de producto).
  - **Variables cuantitativas:** Valores numéricos continuos o discretos (como edad, precio, conteos).
  - **Sin preprocesamiento:** A diferencia de otros algoritmos que requieren transformaciones específicas (como normalización, codificación one-hot, etc.), los Random Forests pueden trabajar con variables mixtas directamente.
- **Ventajas adicionales no mencionadas en la cita:**

- **Robustez** ante valores atípicos: Son menos sensibles a valores atípicos que muchos otros métodos.
- **Manejo eficiente de datos faltantes:** Pueden trabajar con conjuntos de datos incompletos.
- **Estimación incorporada de importancia de variables:** Proporcionan medidas de cuán relevante es cada variable para la predicción.
- **Reducción del sobreajuste**: La combinación de múltiples árboles reduce el riesgo de sobreajuste en comparación con un solo árbol de decisión.

Figura 3

El proceso de producción de resultados de Random Forest



Nota. Adaptado de Wu, X., Gao, Y. y Jiao, D. (2019).

2.2.8. Xgboost

Según (Chen & Guestrin, 2016), XGBoost es un método de aprendizaje automático supervisado, que se utiliza tanto para tareas de clasificación como de regresión. Este potente algoritmo fundamenta su funcionamiento en árboles de decisión y presenta tres características fundamentales:

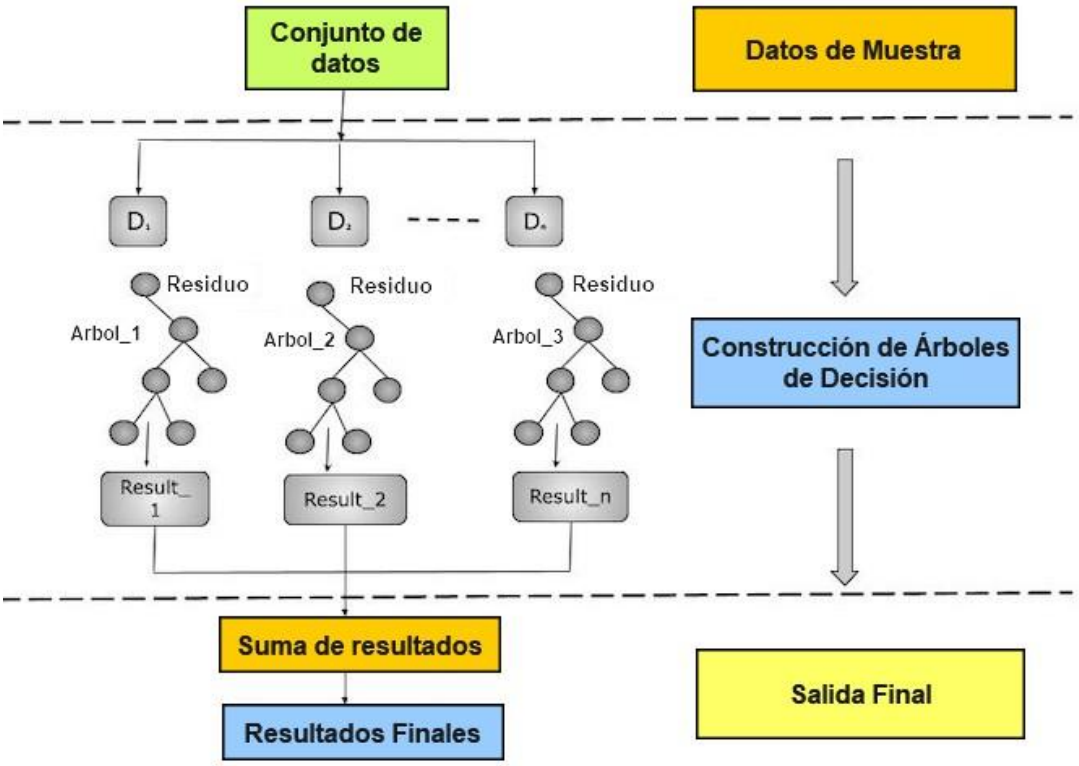
En primer lugar, funciona como un ensamblado secuencial de árboles de decisión. El sistema va incorporando nuevos árboles de manera progresiva, aprendiendo de los resultados anteriores y ajustando los errores detectados hasta alcanzar una precisión óptima.

En segundo lugar, XGBoost se distingue por su sofisticada implementación técnica. Incorpora procesamiento paralelo, realiza poda de árboles de manera eficiente, gestiona valores faltantes y aplica técnicas de regularización. Estas características permiten optimizar los modelos mientras previenen el sobreajuste.

Finalmente, este algoritmo se presenta como un paquete de código abierto mantenido por la comunidad "xgboost developers". Según sus creadores, XGBoost es una biblioteca de gradient boosting distribuida y optimizada, diseñada para ofrecer máxima eficiencia, flexibilidad y portabilidad. Su arquitectura implementa algoritmos de Machine Learning siguiendo el paradigma de Gradient Boosting.

Figura 4

Arquitectura XGBoost



Nota. Adaptado de tutorialspoint.com,

https://www.tutorialspoint.com/xgboost/xgboost-architecture.htm

2.2.9. Métricas de Evaluación

Según Géron, A. (2022), la elección entre estas métricas debe basarse en consideraciones específicas del dominio. El MAE es más robusto frente a valores atípicos y proporciona una interpretación directa. El RMSE es más sensible a errores grandes y tiene propiedades matemáticas deseables para optimización

MAE (Error Absoluto Medio)

El MAE es una métrica de evaluación que mide la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección. Mide el error promedio absoluto entre las predicciones y los valores reales.

Fórmula:

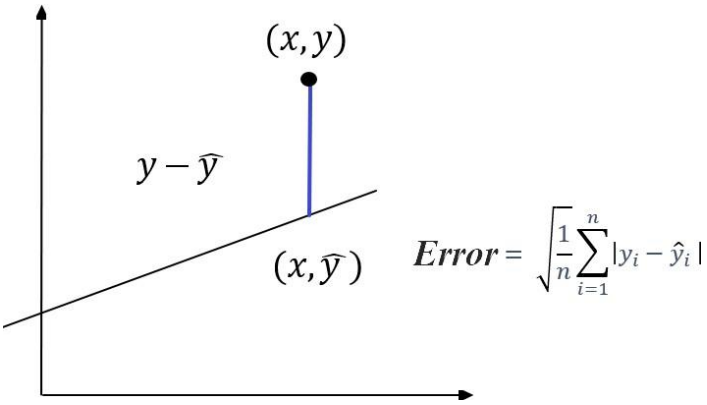
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Propiedades:

- Escala: Se expresa en las mismas unidades que la variable objetivo
- Un valor más bajo indica un mejor desempeño.
- Interpretación: "En promedio, nuestras predicciones se desvían en X unidades"
- Sensibilidad a valores atípicos: Moderada, menos sensible que RMSE
- Derivabilidad: No es diferenciable en todos los puntos (en $y_i = \hat{y}_i$)

Figura 5

Representación Gráfica MAE



Nota. Adaptado de Géron, A. (2022)

Aplicación práctica:

El MAE es preferible cuando los valores atípicos no deben tener un impacto desproporcionado en la evaluación del modelo. Es particularmente útil en contextos donde la interpretabilidad directa es importante.

MSE (Error Cuadrático Medio)

Según (Albon, 2018), MSE (Mean Squared Error) es una de las métricas de evaluación más comunes para modelos de regresión. Formalmente, se define así:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

donde:

- n es el número total de observaciones
- y_i es el valor real de la observación i
- $\hat{y}_i$ es el valor predicho por el modelo para esa observación.

El MSE mide la suma de los errores al cuadrado entre los valores reales y los predichos. Cuanto mayor es el MSE, mayor es la diferencia entre lo que el modelo predice y la realidad, y, por lo tanto, peor es su desempeño.

El uso del cuadrado del error tiene ventajas matemáticas, como asegurar que todos los errores sean positivos. Sin embargo, también tiene una desventaja: amplifica más unos pocos errores grandes que muchos errores pequeños. Es decir, dos modelos con el mismo total de errores pueden tener diferentes MSE.

RMSE (Raíz del Error Cuadrático Medio)

El RMSE amplifica y penaliza severamente los errores grandes debido a la operación cuadrática, proporcionando una medida de la magnitud promedio del error que da mayor peso a las diferencias grandes.

Fórmula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Propiedades:

- Escala: Se expresa en las mismas unidades que la variable objetivo
- Interpretación: Similar al MAE, pero con mayor penalización para errores grandes
- Sensibilidad a valores atípicos: Alta, debido a la operación cuadrática

- Derivabilidad: Diferenciable en todos los puntos, facilitando su uso en optimización

Relación con la desviación estándar:

El RMSE puede interpretarse como la desviación estándar de los residuos (errores de predicción). Está matemáticamente relacionado con el MSE (Error Cuadrático Medio):

$$RMSE = \sqrt{MSE}$$

2.2.10. Validación Cruzada Temporal

Según Guerra, Jorge. (2022), la validación cruzada constituye un método fundamental para evaluar la calidad predictiva de modelos de machine learning, superando las limitaciones de las métricas tradicionales que solo evalúan el ajuste a los datos de entrenamiento. Este enfoque permite determinar la verdadera capacidad predictiva de un modelo al evaluar su rendimiento en datos no utilizados durante la fase de ajuste.

El procedimiento estándar consiste en dividir el conjunto de datos en dos subconjuntos: entrenamiento (para ajustar el modelo) y validación (para evaluar los errores predictivos). Aunque se suele recomendar que el conjunto de validación represente aproximadamente el 20% de la muestra total, este porcentaje puede variar según las características particulares de la serie temporal y el horizonte de predicción deseado.

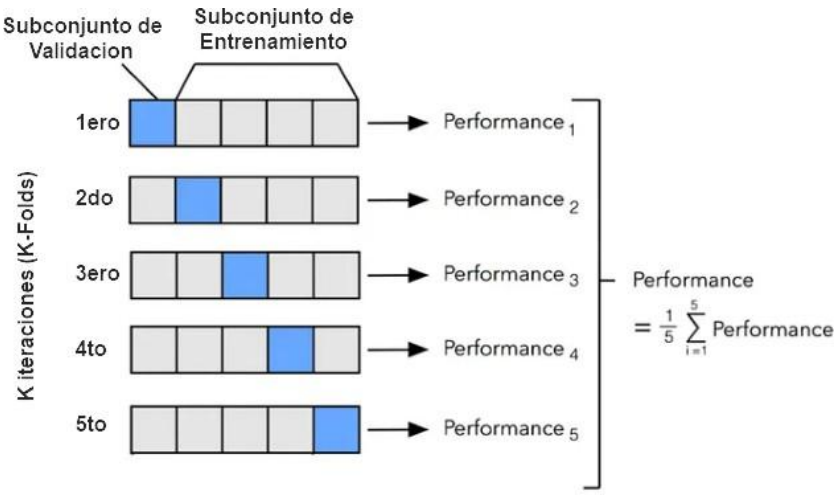
La validación cruzada temporal se implementa típicamente mediante dos métodos:

a) **Validación cruzada T-m+1**: En este enfoque iterativo, inicialmente se designa la m-ésima observación como punto de validación, mientras todas las observaciones anteriores constituyen el conjunto de entrenamiento. En las iteraciones subsiguientes, el punto de validación avanza cronológicamente,

incorporando la observación anteriormente validada al conjunto de entrenamiento.

Figura 6

Validación cruzada temporal



Nota. Adaptado de Guerra, Jorge. (2022)

b) **Validación cruzada de k pasos**: Esta variante mantiene una separación de k-1 observaciones entre los conjuntos de entrenamiento y validación, creando múltiples iteraciones de evaluación.

Para implementar estos métodos, el parámetro m debe ser suficientemente grande para permitir un entrenamiento adecuado del modelo con m-1 datos. La validación cruzada evalúa la fiabilidad predictiva utilizando las métricas de error discutidas previamente, aplicadas a cada conjunto de validación o iteración.

Estos métodos resultan superiores a las medidas tradicionales porque evalúan el rendimiento del modelo en datos nuevos, reflejando con mayor precisión su capacidad para predecir valores futuros desconocidos.

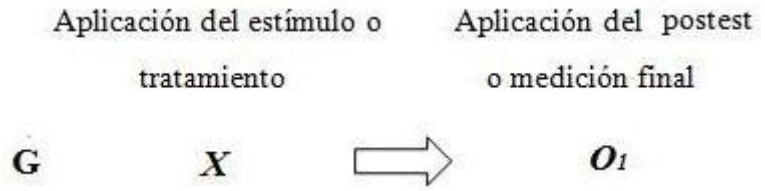
Los pasos de la comparación de modelos con validación cruzada temporal son los siguientes:

- División de datos: Se separan los datos en múltiples segmentos temporales consecutivos.
- Entrenamiento progresivo: Para cada segmento:
  - ✓ Se entrena el modelo con datos hasta un punto específico en el tiempo
  - ✓ Se evalúa su rendimiento prediciendo el siguiente período
  - ✓ Se avanza la ventana temporal
- Evaluación de múltiples modelos: Se aplica este proceso a diferentes algoritmos (arima, regresión, etc.)
- Métricas de comparación: Se calculan errores de predicción (RMSE, MAE y MSE) para cada modelo
- Selección del mejor modelo: Se elige el algoritmo con menor error promedio en las predicciones

CAPITULO III: METODOLOGIA

3.1. DISEÑO DE LA INVESTIGACIÓN

Por su naturaleza, la investigación pertenece a un diseño preexperimental, con "Diseño de grupo único, con solo postest"



Dónde:

G = Grupo Único (Datos de calificaciones de los estudiantes de Primer año de Secundaria)

X = Modelo Machine Learning

O1 = Métricas de precisión

Grupo Único: En un mismo grupo de datos se utiliza para entrenar y evaluar modelos de Machine Learning. (entrenamiento y prueba), ambos subconjuntos provienen del mismo grupo original de datos. La división se realiza para evaluar la efectividad del modelo.

3.2. POBLACIÓN Y MUESTRA

3.2.1. Población

Según (Hernández Sampieri et al., 2014), "la población se refiere al conjunto completo de elementos que comparten características comunes y que son de interés para el estudio. Puede estar compuesta por personas, animales, objetos, eventos o cualquier otro tipo de elemento. Debe estar bien definida y delimitada para que el estudio sea preciso y confiable. Finalmente, la población puede ser finita o infinita". En la presente investigación se consideró la siguiente población tomada de dos colegiosemblemáticos con una gran población de estudiantes:

- Alumnos matriculados en 1er año de Secundaria

En la presente investigación se consideró la siguiente población para cada indicador:

Tabla 2:

Población por cada indicador

Indicador	Población
Error Absoluto Medio (MAE)	Número de registros de calificaciones de los estudiantes utilizados para calcular el MAE.
Error Cuadrático Medio (MSE)	Número de registros de calificaciones de los estudiantes utilizados para calcular el MSE.
Raíz del Error Cuadrático Medio (RMSE)	Número de registros de calificaciones de los estudiantes utilizados para calcular el RMSE.
Comparación de métricas (MSE, MAE y RMSE)	Número de registros de calificaciones de los estudiantes utilizados para comparar las métricas.

Nota. Elaboración propia.

3.2.2. Muestra

De acuerdo la naturaleza del presente proyecto se trabajó con la población completa de registros de calificaciones de los estudiantes de primer año de secundaria, lo que permite obtener resultados más precisos y evitar errores de muestreo que podrían afectar la representatividad del análisis. El tamaño del conjunto de datos es manejable y los recursos computacionales disponibles permiten procesar la población completa, lo que asegura un análisis objetivo. El análisis se realizó sobre la población completa para garantizar que las métricas calculadas reflejen con precisión el desempeño del modelo en el contexto del estudio, evitando posibles sesgos derivados del muestreo.

.

3.3. Operacionalización de Variables

Tabla 3:

Variables y sus Indicadores

Variable	Definición Conceptual	Definición Operacional	Dimensiones	Indicadores	Escala de medición
Independiente: Modelo predictivo	Conjunto de algoritmos y técnicas del aprendizaje automático utilizados para identificar patrones y predecir resultados basados en datos históricos.	Técnicas y algoritmos de machine learning aplicados al conjunto de datos académicos, incluyendo selección de características, entrenamiento del modelo, validación cruzada y evaluación de desempeño.	Algoritmos de aprendizaje	Tipo de algoritmo implementado (Regresión múltiple, Árboles de decisión, Random Forest,)	Nominal
			Preprocesamiento de datos	Técnicas de limpieza y transformación aplicadas	Nominal
			Selección de características	Variables predictoras incluidas en el modelo	Nominal
			Validación del modelo	Método de validación empleado (validación cruzada, hold-out)	Nominal
			Ajuste del modelo	Hiperparámetros optimizados	Nominal/Razón
Dependiente: Predicción del rendimiento académico	Capacidad del modelo para anticipar el desempeño escolar en una asignatura, expresado generalmente en calificaciones o niveles de logro.	Resultado generado por el modelo predictivo al estimar la calificación o nivel de logro de los estudiantes, basado en sus datos históricos y académicos previos.	Métricas de precisión	Error Absoluto Medio (MAE) = Promedio de |valor real - valor predicho|	Razón
				Error Cuadrático Medio (MSE) = Promedio de (valor real - valor predicho) ^{2}	Razón
				Raíz del Error Cuadrático Medio (RMSE) = $\sqrt{MSE}$	Razón

Nota. Elaboración propia.

3.4. Técnicas e Instrumentos de recolección de datos

3.4.1. Técnicas de recolección de datos

- **Investigación bibliográfica**: Se recurrió a realizar una revisión sistemática de investigaciones previas sobre predictores del rendimiento académico en estudiantes de primer año de secundaria. Se consulto a bases de datos científicas (Scopus, Google Scholar) para obtener artículos relevantes sobre machine learning aplicado a educación. Se revisó los marcos normativos educativos como el análisis de los currículos y estándares de evaluación vigentes para comprender el contexto educativo formal.
- **Análisis de Datos Académicos Históricos:** Se revisó sistemáticamente los registros académicos para analizar calificaciones previas, que puedan establecer patrones de desempeño. Se complementa con un seguimiento longitudinal que permite recolectar datos de rendimiento a través del tiempo, identificando trayectorias y tendencias significativas en el desarrollo académico del estudiante.

CAPITULO IV: RESULTADOS Y DISCUSION

4.1. Análisis de Resultados:

El análisis de los resultados de esta presente investigación se fundamenta en la metodología propuesta por Géron, A. (2022). Esta metodología estructurada permitió abordar el problema de manera sistemática y obtener resultados robustos y confiables, siguiendo las mejores prácticas en ciencia de datos y aprendizaje automático.

4.1.1. Comprensión del panorama general

Objetivo del negocio y contexto

El proyecto se orienta a resolver uno de los problemas recurrentes en el ámbito educativo: anticipar la distribución del rendimiento académico a futuro en función de datos históricos. La finalidad es que los directores, profesores y responsables de la formulación de políticas educativas (UGEL) puedan disponer de información predictiva que permita identificar de manera temprana áreas críticas y, de esta forma, asignar recursos, diseñar intervenciones o implementar estrategias para mejorar el rendimiento global.

Según el (Ministerio de Educación [MINEDU], 2020), la escala de calificaciones es literal con la siguiente escala:

Tabla 4:

Variables y sus Indicadores

Escala	Nivel de Logro	Descripción
AD	LOGRO DESTACADO	Cuando el estudiante evidencia un nivel superior a lo esperado respecto a la competencia. Esto quiere decir que demuestra aprendizajes que van más allá del nivel esperado.
A	LOGRO ESPERADO	Cuando el estudiante evidencia el nivel esperado respecto a la competencia, demostrando manejo satisfactorio en todas las tareas propuestas y en el tiempo programado.
B	EN PROCESO	Cuando el estudiante está próximo o cerca al nivel esperado respecto a la competencia, para lo cual requiere acompañamiento durante un tiempo razonable para lograrlo.

C	EN INICIO	Cuando el estudiante muestra un progreso mínimo en una competencia de acuerdo al nivel esperado. Evidencia con frecuencia dificultades en el desarrollo de las tareas, por lo que necesita mayor tiempo de acompañamiento e intervención del docente.

Nota. Adaptado de (MINEDU, 2020).

Este esquema de calificación se implementó en todo el Perú a partir del año 2019 con el propósito fortalecer la evaluación formativa y enfocarse en el desarrollo de competencias, fomentando el progreso del estudiante.

Aunque los datos (calificaciones literales) son eficientes para representar el progreso del estudiante, no es tan eficiente para realizar predicciones de rendimiento académico, porque los algoritmos de **Machine Learning** funcionan mejor cuando trabajan con datos **cuantificables**. La mayoría de los modelos estadísticos y de aprendizaje automático requieren **valores numéricos** para calcular relaciones, patrones y realizar predicciones.

También se tiene data del año 2018 que está en escala vigesimal (0 -20) que representa un desafío a la hora de integrar toda la información con los años subsiguientes que están en otro formato.

El (Ministerio de Educación [MINEDU], 2024), para efectos de realizar otros procesos donde se necesita una escala numérica, presentó la conversión literal a escala decimal y vigesimal asignando pesos a las notas literales, tal como se muestra en la siguiente tabla:

Tabla 5:

Conversión de escalas

Literal	Puntaje	Escala 10	Escala 20
AD	**4**	**10**	**20.00**
A	**3**	**7.5**	**13.33**
B	**2.5**	**6.25**	**10.00**
C	**1**	**2.5**	**0.00**

Nota. Adaptado de (MINEDU, 2024).

Esto de acuerdo con la siguiente formula:

D = (3V / 8) + 2,5

Donde V = Nota Vigesimal

Tabla 6:

Conversión de escalas

Nota en la escala vigesimal (V)	Resultado en la nueva escala (D)
20	10.000
19	9.625
18	9.250
17	8.875
16	8.500
15	8.125
14	7.750
13	7.375
12	7.000
11	6.625
10	6.250
09	5.875
08	5.500
07	5.125
06	4.750
05	4.375
04	4.000
03	3.625
02	3.250
01	2.875
00	2.500

Nota. Adaptado de (MINEDU, 2024).

Bajo este contexto se procedió a realizar la conversión de escala literal y escala vigesimal a la nueva escala decimal, por los motivos antes mencionados.

Ahora podremos modelar la evolución de los desempeños académicos mediante patrones y tendencias temporales. La capacidad de ML para identificar relaciones complejas en conjuntos de datos amplios y heterogéneos facilita la predicción de la distribución porcentual en cada logro de aprendizaje, haciendo posible tomar decisiones basadas en la anticipación de condiciones futuras del rendimiento educativo.

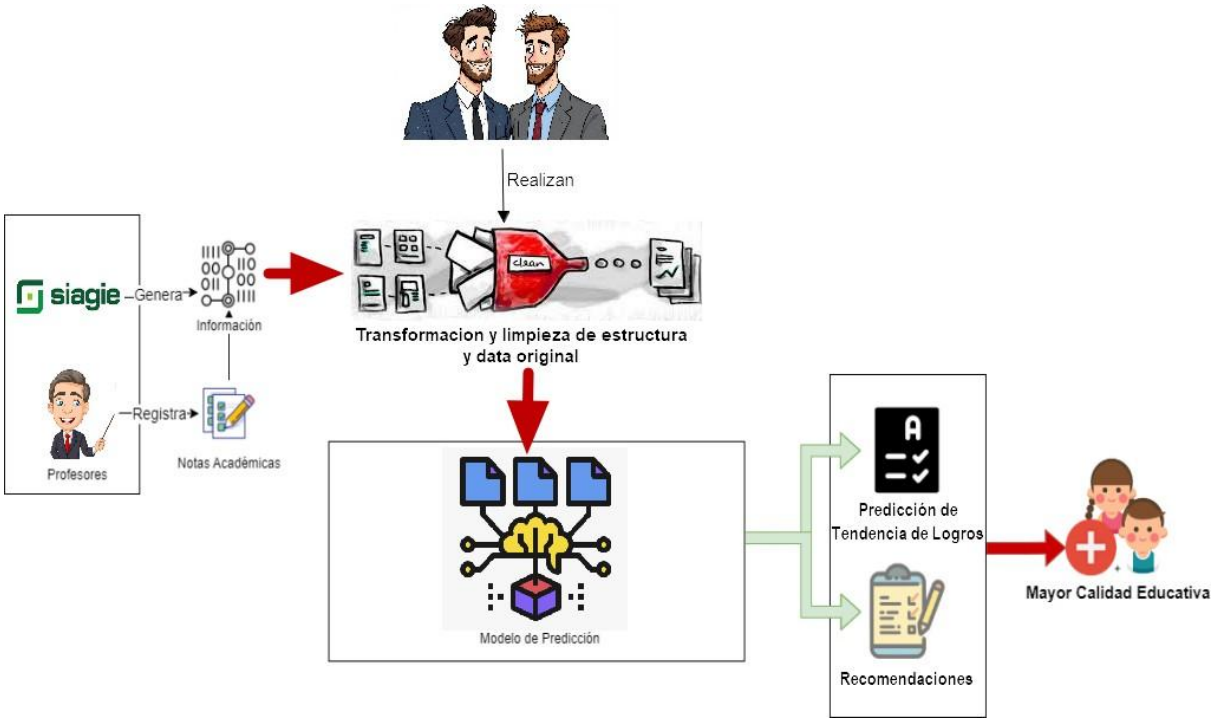
Esta visión integral permite que el modelo no solo actúe como una herramienta predictiva, sino como un sistema de apoyo a la toma de decisiones, enfocándose en intervenciones dirigidas a mejorar los procesos educativos y en la optimización de la asignación de recursos.

Definición del tipo de problema que se va a abordar

El problema para abordar se encuadra dentro del paradigma de aprendizaje supervisado. Contamos con datos históricos generados por el sistema SIAGEI del Ministerio de Educación, en la que la tarea consiste en predecir, sobre la base de patrones observados en estos datos, la distribución porcentual de estudiantes en los logros definidos para un año futuro.

Figura 7

Validación cruzada temporal



Nota. Elaboración propia.

Adicionalmente, el problema es de análisis temporal, puesto que se deben identificar y modelar tendencias a lo largo de diversos períodos (años). En este sentido, se combinan aspectos de series temporales y regresión, lo que exige la *adopción de enfoques que puedan capturar tanto la evolución en el tiempo como las particularidades de cada logro de rendimiento.*

Consideraciones de Granularidad y Agregación

A diferencia de modelos que se centran en predicciones a nivel individual, este proyecto requiere la agregación de datos a nivel de año y curso. Esto implica que el modelo debe aprender patrones agregados y distribuir los resultados en forma porcentual para cada una de las categorías. La conexión entre estos niveles de agregación y la evolución temporal exige el desarrollo de técnicas que integren análisis de series temporales, identificando la tendencia global del rendimiento y permitiendo segmentar los resultados según la variable.

Esta particularidad del problema influye directamente en la selección de algoritmos y técnicas de validación: se debe asegurar que la modelación respecta la dependencia temporal sin perder la capacidad predictiva en la agregación de los datos.

Naturaleza del problema:

Este es un problema de regresión multivariante y a la vez de predicción de distribuciones porcentuales (problema de regresión aplicado a logros).

No es clasificación por estudiante, porque el modelo no predice si un alumno tendrá "en inicio" o "destacado", sino que estima porcentajes de estudiantes en cada categoría.

En términos de aprendizaje automático, podríamos plantearlo como:

- **Input:** Datos históricos de desempeño académico segmentados por curso y año, junto con variables adicionales que puedan influir (si las hay, como cantidad de estudiantes, etc.).
- **Output:** Estimación de 4 valores continuos (los porcentajes de cada logro: Destacado, Esperado, En proceso, en Inicio) para cada curso en el año a predecir.

Tipo de problema:

- **Regresión multivariante** (ya que queremos predecir varias salidas continuas a la vez: los 4 porcentajes). En este caso se podría considerar: ***XGBoost o Random Forest.***
- ***Regresión lineal múltiple***. como modelo sencillo como usando el año como una variable predictora más, junto con otras características del curso

<u>Selección de métricas de rendimiento adecuadas para evaluar el modelo</u>

Dado que predeciremos porcentajes, es clave medir qué tan lejos están las predicciones de los valores reales. Algunas métricas adecuadas serían:

MAE (Mean Absolute Error)

Mide el error promedio absoluto entre los porcentajes reales y los predichos para cada categoría. Es fácil de interpretar en términos de "cuántos puntos porcentuales de diferencia".

MSE (Mean Squared Error)

El Error Cuadrático Medio es una medida utilizada para evaluar la calidad de un modelo de regresión. Se calcula como el promedio de los cuadrados de los errores, donde el error es la diferencia entre los valores predichos por el modelo y los valores reales. Penaliza más fuertemente los errores grandes debido al cuadrado de las diferencias. Esto lo hace útil para identificar modelos que pueden tener problemas con predicciones extremas.

RMSE (Root Mean Squared Error)

Penaliza más los errores grandes. Útil si queremos ser sensibles a desviaciones grandes en alguna categoría.

4.1.2. Obtención de datos

a) Fuentes de Datos Necesarias

Datos históricos académicos (2018-2024): Registros académicos por curso que contengan: Calificaciones por curso, género y año académico, datos agrupados por periodos académicos anuales.

Tabla 7:

Datos originales obtenidos de los registros académicos.

Categoría	Variable	Valores/Formato
Variables Demográficas y Contextuales	Sexo	H/M
	Grado	1er año
	Sección	A, B, C, etc.
	Turno	Mañana/Tarde
	DNI/Código	Número de identificación
Variables Académicas	Desarrollo Personal, Ciudadanía y Cívica	A, AD, B, C
	Ciencias Sociales	A, AD, B, C
	Educación Religiosa	A, AD, B, C
	Educación para el Trabajo	A, AD, B, C
	Educación Física	A, AD, B, C
	Comunicación	A, AD, B, C
	Arte y Cultura	A, AD, B, C
	Castellano como Segunda Lengua	A, AD, B, C
	Inglés	A, AD, B, C
	Matemática	A, AD, B, C
	Ciencia y Tecnología	A, AD, B, C
Variables de Rendimiento	Calificaciones por área	A, AD, B, C
	Situación final	Promovido/Repitente
	Competencias transversales	A, AD, B, C

Nota. Elaboración propia.

b) Formato Adecuado para el Procesamiento

Para conseguir el formato de procesamiento se realizó una sucesión de transformaciones a la estructura y data original con el fin de obtener la ideal para poder iniciar el entrenamiento. A continuación, se describe esas transformaciones a partir de un fragmento de los registros originales:

1er. Paso: Formatear la estructura original: Se considero solamente las variables imprescindibles para la transformación de la data y posterior entrenamiento.

Tabla 8:

Estructura modificada

Variable	Valores/Formato
Año	Entero
Nro_Alumno	Entero
Ciencias Sociales	A, AD, B, C
Educación Religiosa	A, AD, B, C
Educación para el Trabajo	A, AD, B, C
Educación Física	A, AD, B, C
Comunicación	A, AD, B, C
Arte y Cultura	A, AD, B, C
Castellano como Segunda Lengua	A, AD, B, C
Inglés	A, AD, B, C
Matemática	A, AD, B, C
Ciencia y Tecnología	A, AD, B, C

Nota. Elaboración propia.

En la siguiente tabla se muestra un fragmento de la estructura modificada con las notas literales por asignatura y competencias.

Tabla 9:

Fragmento de la estructura modificada con calificaciones literales.

Año	Desarrollo personal ciudadanía y Cívica		Ciencias Sociales			Educación para el Trabajo		Educación Física			Comunicación		
	Construye su identidad	Convive y participa democráticamente en la búsqueda del bien común	Construye interpretaciones históricas	Gestiona responsablemente el espacio	Gestiona responsablemente los recursos económicos	Gestiona proyectos de emprendimiento económico	Espacio(s)	Se desenvuelve de manera autónoma a través de su motricidad	Asume una vida saludable	Interactúa a través de sus habilidades socio motrices	Se comunica oralmente en su lengua materna	Lee diversos tipos de textos escritos en su lengua materna	Escribe diversos tipos de textos en su lengua materna
2021	A	B	A	A	B	B	A	AD	A	A	A	B	B
2021	AD	A	A	A	B	B	A	AD	A	A	A	A	A
2021	B	B	A	A	B	B	A	A	A	B	A	A	A
2021	B	B	B	AD	B	B	B	A	A	A	B	B	B
2021	AD	AD	A	B	A	A	C	AD	A	A	A	A	A

Nota. Elaboración propia.

2do. Paso: Se reemplazó las calificaciones literales basadas en logros de acuerdo a la siguiente información:

Literal	AD	A	B	C
Puntaje	4	3	2.5	1

El reemplazo se puede observar en la siguiente tabla.

Tabla 10:

Fragmento de la estructura modificada con calificaciones numéricas.

Año	Desarrollo personal ciudadanía y Cívica		Ciencias Sociales			Educación para el Trabajo		Educación Física			Comunicación		
	Construye su identidad	Convive y participa democráticamente en la búsqueda del bien común	Construye interpretaciones históricas	Gestiona responsablemente el espacio y el ambiente	Gestiona responsablemente los recursos económicos	Gestiona proyectos de emprendimiento económico	Espacio(s)	Se desenvuelve de manera autónoma a	Asume una vida saludable	Interactúa a través de sus habilidades socio motrices	Se comunica oralmente en su lengua materna	Lee diversos tipos de textos escritos en su lengua materna	Escribe diversos tipos de textos en su lengua materna
2021	3	2.5	3	3	2.5	2.5	3	4	3	3	3	2.5	2.5
2021	4	3	3	3	2.5	2.5	3	4	3	3	3	3	3
2021	2.5	2.5	3	3	2.5	2.5	3	3	3	2.5	3	3	3
2021	2.5	2.5	2.5	4	2.5	2.5	2.5	3	3	3	2.5	2.5	2.5
2021	4	4	3	2.5	3	3	1	4	3	3	3	3	3

Nota. Elaboración propia.

3er. Paso: Se procedió a obtener el promedio en base decimal de las competencias por cada asignatura de acuerdo con la siguiente formula:

$$PROMEDIO = \frac{\sum Ci}{n\ x\ 4}\ x\ 10$$

Donde:

- Ci = Calificación de cada competencia (donde A=3, B=2.5, AD=4, C=1)
- n = Número de competencias de cada asignatura

Tabla 11:

Fragmento de la estructura modificada con calificaciones en base decimal.

Nro Alumno	Año	Desarrollo personal ciudadanía y Cívica	Ciencias Sociales	Educación para el Trabajo	Educación Física	Comunicación
1	2021	6.88	7.08	6.88	8.33	6.67
2	2021	8.75	7.08	6.88	8.33	7.50
3	2021	6.25	7.08	6.88	7.08	7.50
4	2021	6.25	7.50	6.25	7.50	6.25
5	2021	10.00	7.08	5.00	8.33	7.50

Nota. Elaboración propia.

Figura 8

Validación cruzada temporal

	Anio	Nro_Alumno	Civica	Sociales	Religion	Comunicacion	Matematica	Tecnologia	Educ_Trabajo	Educ_Fisica	Arte	Ingles
0	2017	1	7.375	6.625000	8.125	6.625000	6.625	6.625000	8.125	8.125	7.375	7.375000
1	2017	2	7.375	7.375000	7.750	7.750000	7.000	6.625000	8.875	8.125	7.375	7.750000
2	2017	3	7.000	7.750000	7.000	8.125000	7.375	7.000000	8.500	7.375	7.000	8.500000
3	2017	4	8.125	7.750000	8.125	7.750000	7.375	7.000000	8.500	8.125	7.750	7.375000
4	2017	5	7.750	7.750000	8.500	7.375000	7.750	7.375000	8.500	8.125	8.125	7.750000
...	...	...	...	...	...	...	...	...	...	...	...	...
3009	2022	384	7.500	7.083333	7.500	6.250000	6.250	7.083333	7.500	7.500	7.500	7.083333
3010	2022	385	6.875	6.250000	6.875	7.083333	7.500	7.500000	7.500	7.500	7.500	8.333333
3011	2022	386	6.250	7.500000	6.875	7.500000	8.750	7.500000	7.500	7.500	7.500	10.000000
3012	2022	387	7.500	7.500000	6.875	7.083333	6.250	7.500000	7.500	7.500	7.500	7.500000
3013	2022	388	6.250	6.250000	6.875	6.250000	7.500	6.250000	7.500	7.500	7.500	7.083333

3014 rows × 13 columns

Nota. Elaboración propia.

4to. Paso: Finalmente se procede a la agrupación por logros de acuerdo con la siguiente escala:

Tabla 12:

Escala de Logros

Valores	Escala Decimal
AD	**10**
A	**[7.5 – 10.0>**
B	**[6.25 - 7.5>**
C	**[2.5 - 6.25>**

Nota. Elaboración propia.

En este paso se generaría ya la estructura final para su entrenamiento con estas características:

- Tabla estructurada con formato CSV o Excel conteniendo:
  - Columnas para cada año académico (2018-2024)
  - Columnas para identificación de cursos
  - En las columnas de cursos se generarán las calificaciones numéricas para cada logro de rendimiento de acuerdo con la siguiente escala:

Tabla 13:

Estructura final propuesta para el entrenamiento.

Variable	Tipo de Dato	Descripción	Rango/ Valores	Observaciones
Anio	Entero	Año académico	2017-2024	8 años
Curso	Categórico	Asignatura	10 Logros	['Arte', 'Cívica', 'Comunicación', 'Educ. Física', 'Educ. Trabajo', 'Ingles', 'Matemática', 'Religión', 'Sociales', 'Ciencia y Tecnología']
AD	Flotante	% logro destacado	0 - 100%	
A	Flotante	% logro esperado	0 - 100%	
B	Flotante	% en proceso	0 - 100%	
C	Flotante	% en inicio	0 - 100%	

Nota. Elaboración propia.

Tabla *14*:

Variables Independientes (Predictoras)

Variable	Tipo	Descripción	Valores Reales
Año Académico	Numérica Discreta	Año del registro académico	2017-2024 (8 años)
Asignatura	Categórica Nominal	Curso específico	Cívica, Sociales, Religión, Comunicación, Matemática, Ciencia y Tecnología, Educación para el trabajo, Educación Física, Arte, Inglés (10 asignaturas)

Nota. Elaboración propia.

Tabla 15:

Variables Dependientes (Resultados)

Variable	Tipo	Descripción	Rango
Porcentaje AD	Numérica Continua	Porcentaje de logro destacado	0-100%
Porcentaje A	Numérica Continua	Porcentaje de logro esperado	0-100%
Porcentaje B	Numérica Continua	Porcentaje de estudiantes en proceso	0-100%
Porcentaje C	Numérica Continua	Porcentaje de estudiantes en inicio	0-100%

Nota. Elaboración propia.

Tabla 16:

Data y estructura de entrenamiento inicial

Año	Curso	AD	A	B	C
2017	Cívica	0	0.622739	0.377261	0
2017	Sociales	0	0.576227	0.418605	0.005168
2017	Religión	0	0.713178	0.273902	0.01292
2017	Comunicación	0	0.426357	0.547804	0.02584
2017	Matemática	0	0.286822	0.664083	0.049096
2017	Ciencia y Tecnología	0	0.374677	0.599483	0.02584
2017	Educ_Trabajo	0	0.816537	0.178295	0.005168
2017	Educ. Física	0.007752	0.945736	0.046512	0
2017	Arte	0.002584	0.715762	0.27907	0.002584
2017	Ingles	0	0.395349	0.578811	0.02584
2018	Ingles	0	0.413613	0.586387	0
2018	Arte	0	0.876963	0.096859	0
2018	Educ. Física	0	0.774869	0.225131	0
2018	Educ_Trabajo	0	0.772251	0.227749	0
2018	Tecnología	0	0.557592	0.429319	0.013089
2018	Sociales	0	0.691099	0.308901	0
2018	Comunicación	0	0.434555	0.549738	0.015707
2018	Religión	0	0.526178	0.434555	0.039267
2018	Cívica	0	0.615183	0.384817	0
2018	Matemática	0	0.34555	0.604712	0.049738
2019	Educ_Trabajo	0.057895	0.610526	0.321053	0.010526
2019	Ingles	0.065789	0.718421	0.194737	0.021053
2019	Educ. Física	0.094737	0.771053	0.131579	0.002632
2019	Tecnología	0.107895	0.663158	0.218421	0.010526
2019	Arte	0.139474	0.602632	0.257895	0
2019	Comunicación	0.165789	0.452632	0.381579	0
2019	Religión	0.107895	0.623684	0.223684	0.044737
2019	Sociales	0.134211	0.568421	0.297368	0
2019	Cívica	0.073684	0.521053	0.405263	0
2019	Matemática	0.078947	0.426316	0.318421	0.176316
2020	Tecnología	0	0.05641	0.94359	0

2020	Matemática	0	0.125641	0.874359	0
2020	Comunicación	0.002564	0.225641	0.771795	0
2020	Religión	0.089744	0.4	0.510256	0
2020	Sociales	0.041026	0.479487	0.479487	0
2020	Cívica	0.033333	0.823077	0.14359	0
2020	Arte	0.248718	0.279487	0.471795	0
2020	Ingles	0	0.197436	0.802564	0
2020	Educ. Física	0.038462	0.176923	0.784615	0
2020	Educ_Trabajo	0.020513	0.407692	0.571795	0
2021	Arte	0.06015	0.466165	0.473684	0
2021	Ingles	0.002506	0.348371	0.649123	0
2021	Cívica	0.152882	0.343358	0.503759	0
2021	Sociales	0.067669	0.378446	0.553885	0
2021	Religión	0.080201	0.318296	0.601504	0
2021	Educ_Trabajo	0.117794	0.408521	0.473684	0
2021	Matemática	0.005013	0.243108	0.75188	0
2021	Tecnología	0.085213	0.280702	0.634085	0
2021	Educ. Física	0	0.283208	0.716792	0
2021	Comunicación	0.035088	0.426065	0.538847	0
2022	Cívica	0.095361	0.551546	0.353093	0
2022	Religión	0.069588	0.440722	0.489691	0
2022	Comunicación	0.025773	0.502577	0.471649	0
2022	Matemática	0	0.342784	0.657216	0
2022	Tecnología	0.033505	0.322165	0.64433	0
2022	Educ_Trabajo	0.090206	0.778351	0.131443	0
2022	Educ. Física	0.015464	0.610825	0.373711	0
2022	Sociales	0.015464	0.420103	0.564433	0
2022	Ingles	0.002577	0.425258	0.572165	0
2022	Arte	0.012887	0.969072	0.018041	0
2023	Arte	0.024259	0.568733	0.393531	0.013477
2023	Cívica	0.008086	0.838275	0.153639	0
2023	Sociales	0.005391	0.58221	0.385445	0.026954
2023	Religión	0.002695	0.54717	0.336927	0.113208
2023	Comunicación	0.010782	0.363881	0.493261	0.132075
2023	Matemática	0.010782	0.16442	0.455526	0.369272
2023	Tecnología	0.026954	0.226415	0.501348	0.245283
2023	Educ_Trabajo	0	0.002695	0.121294	0.876011
2023	Educ. Física	0.018868	0.41779	0.54717	0.016173
2023	Ingles	0.008086	0.296496	0.506739	0.188679
2024	Cívica	0.037855	0.498423	0.422713	0.041009
2024	Sociales	0.012618	0.397476	0.485804	0.104101
2024	Religión	0.078864	0.520505	0.321767	0.078864
2024	Comunicación	0.006309	0.280757	0.529968	0.182965
2024	Matemática	0.022082	0.205047	0.432177	0.340694
2024	Tecnología	0.069401	0.55205	0.356467	0.022082
2024	Educ_Trabajo	0.091483	0.690852	0.217666	0

2024	Educ. Física	0	0.55836	0.438486	0.003155
2024	Ingles	0.015773	0.176656	0.463722	0.343849
2024	Arte	0.006309	0.769716	0.217666	0.006309

Nota. Elaboración propia.

4.1.3. Exploración y visualización de los datos

En esta etapa se analizará y comprenderá de forma profunda los datos antes de aplicar algoritmos complejos. Esta fase permite identificar patrones, tendencias, correlaciones y anomalías mediante análisis exploratorio y visualizaciones. Al examinar distribuciones, relaciones entre variables podemos detectar problemas de calidad (valores atípicos, datos faltantes, inconsistencias), formular hipótesis iniciales y determinar qué técnicas de preprocesamiento serán necesarias. Además, a partir de esta exploración nos dará una idea más clara de que modelo de machine learning aplicar.

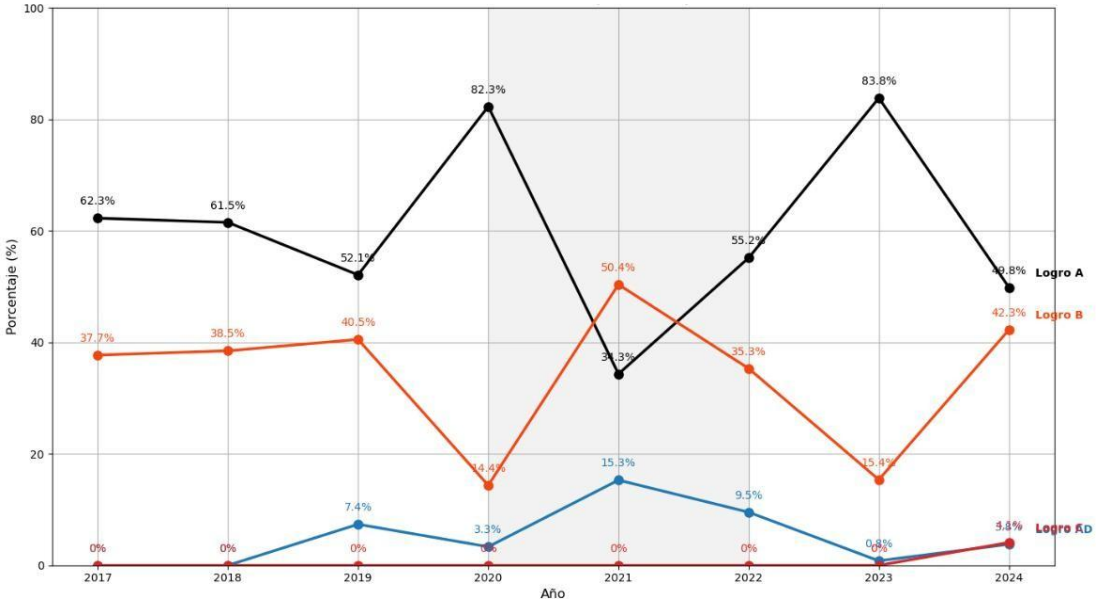
Identificación de los tipos de datos

- Datos estructurados:
  - ✓ Registros académicos históricos (2017-2024)
  - ✓ Distribuciones porcentuales por logro (destacado (AD), esperado (A), en proceso (B), en inicio (C))
  - ✓ Variables temporales (años académicos, periodos)
- Datos continuos:
  - ✓ Porcentajes de estudiantes en cada Logro

a) Técnicas de visualización y análisis estadístico de los datos

Figura 9

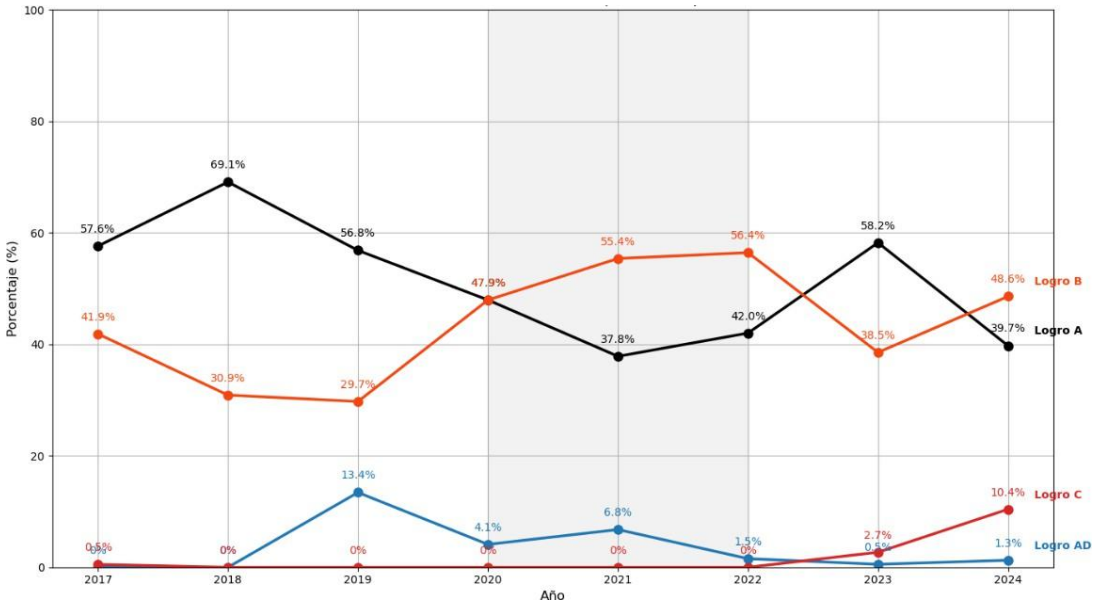
Análisis del rendimiento del curso Educación Cívica



Nota. Elaboración propia.

Figura 10

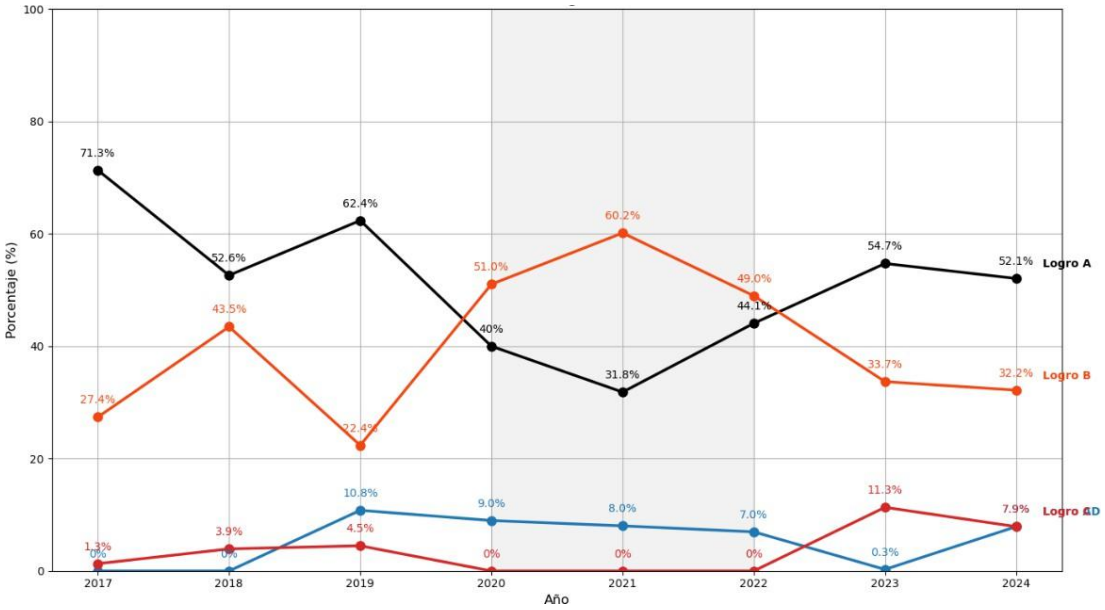
Análisis del rendimiento del curso de Ciencias Sociales



Nota. Elaboración propia.

Figura 11

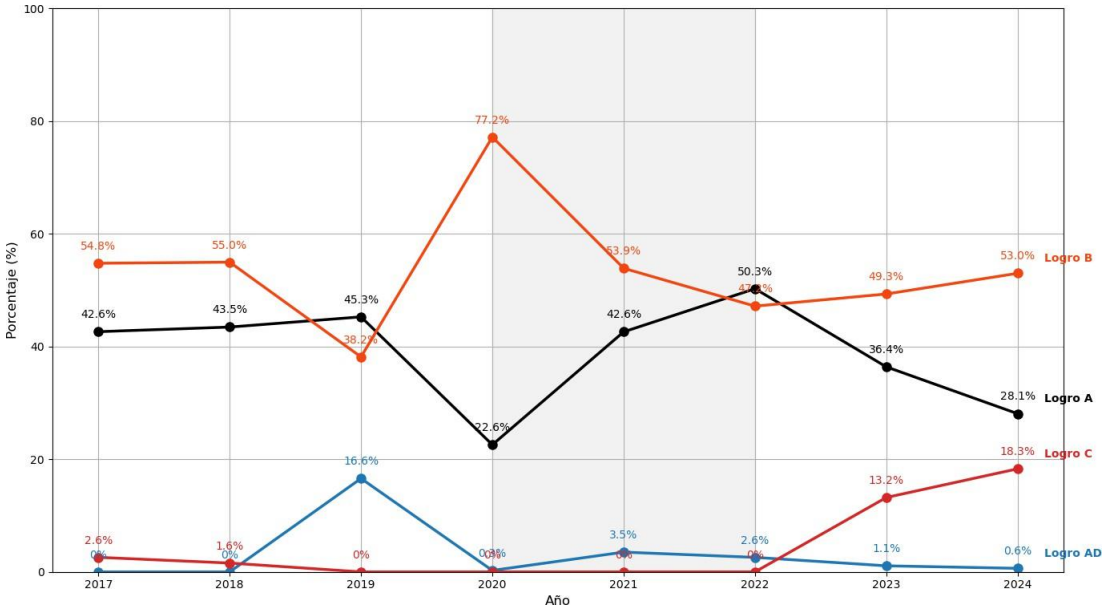
Análisis del rendimiento del curso de Religión



Nota. Elaboración propia.

Figura 12

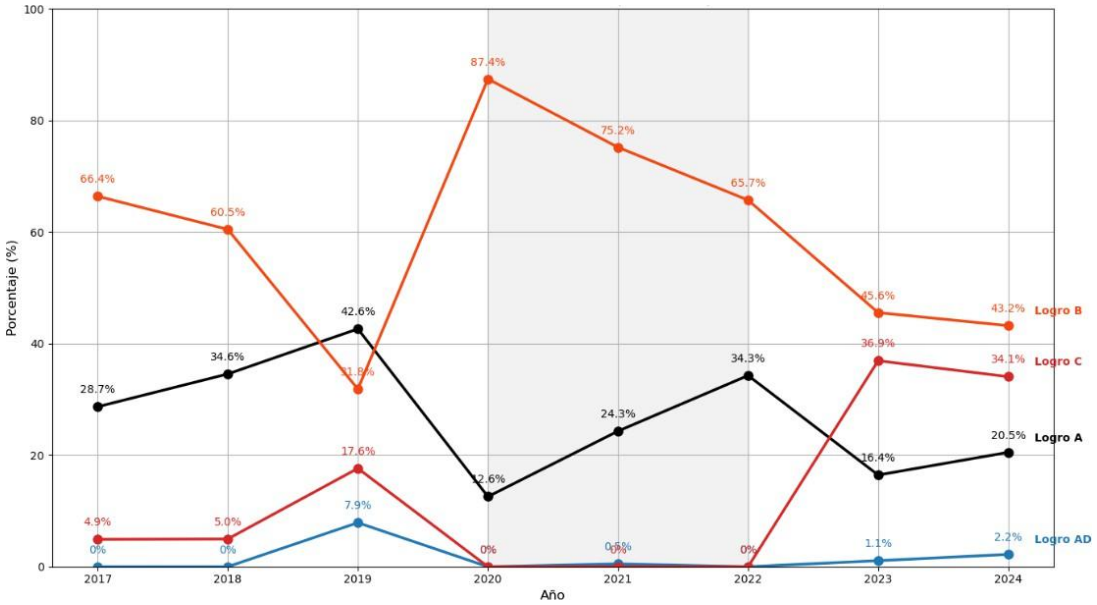
Análisis del rendimiento del curso de Comunicación



Nota. Elaboración propia.

Figura 13

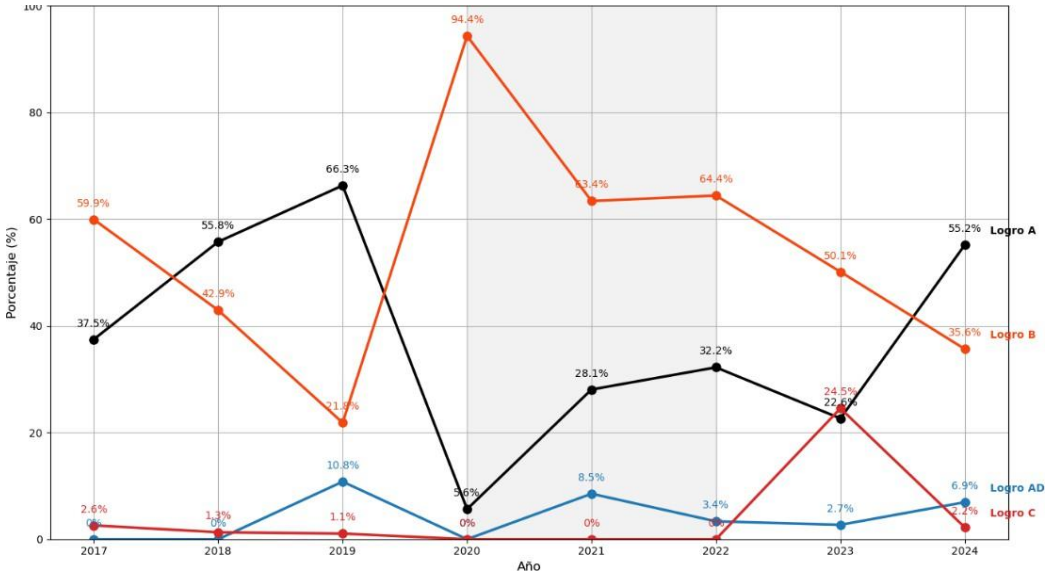
Análisis del rendimiento del curso de Matemática



Nota. Elaboración propia.

Figura 14

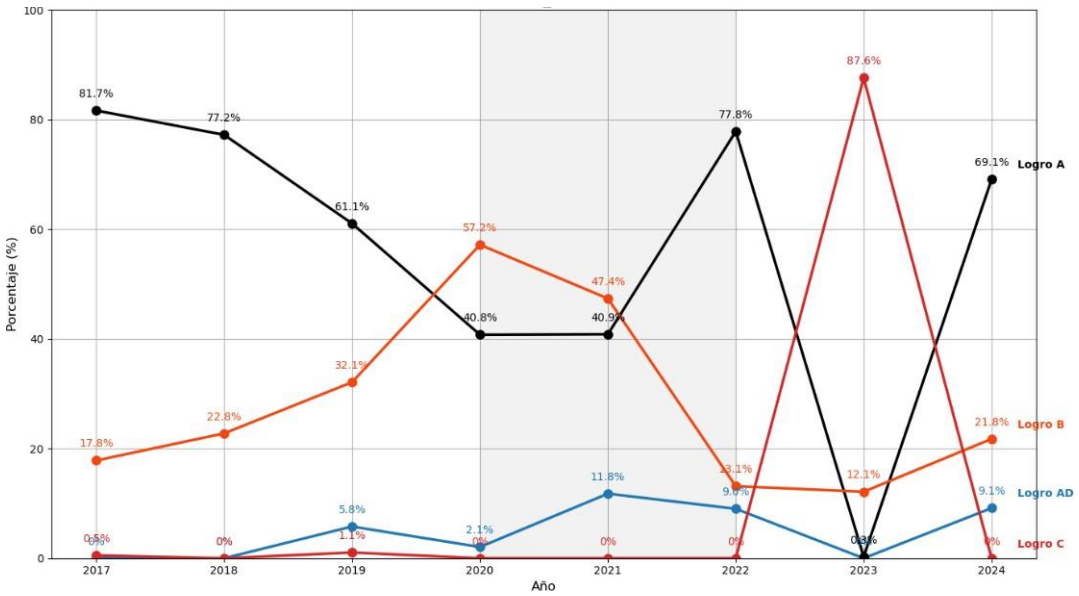
Análisis del rendimiento del curso de Ciencia y Tecnología



Nota. Elaboración propia.

Figura 15

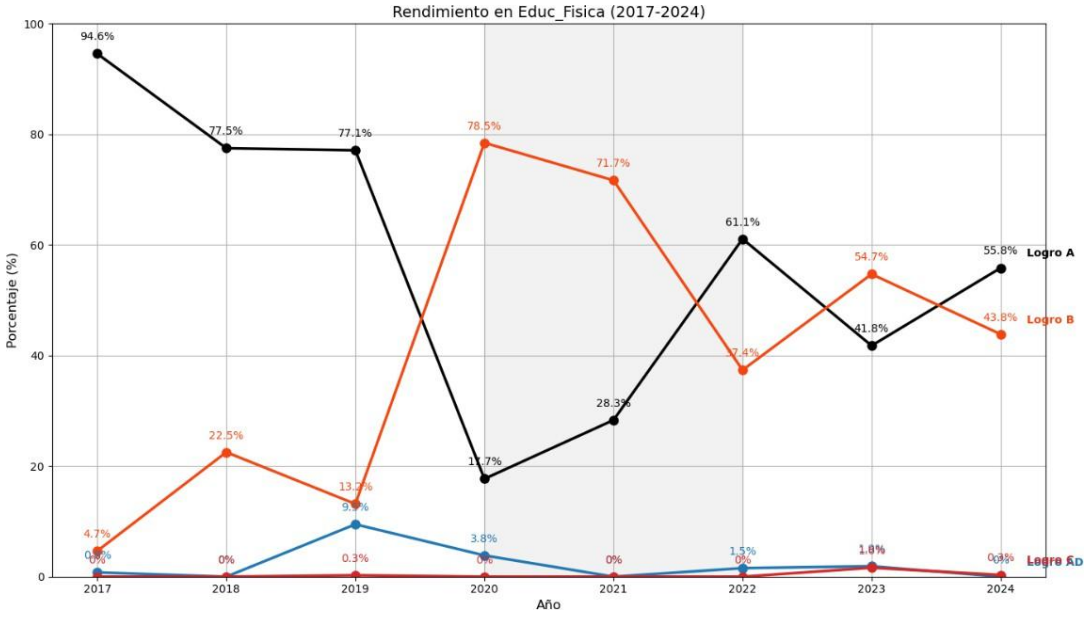
Análisis del rendimiento del curso de Educación para el trabajo



Nota. Elaboración propia.

Figura 16

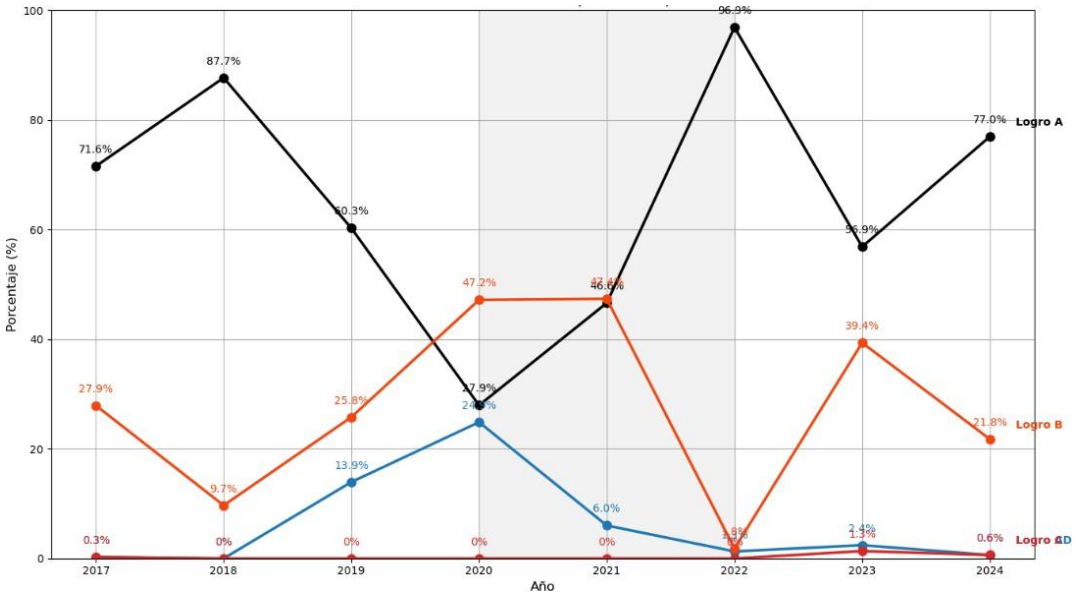
Análisis del rendimiento del curso de Educación Física



Nota. Elaboración propia.

Figura 17

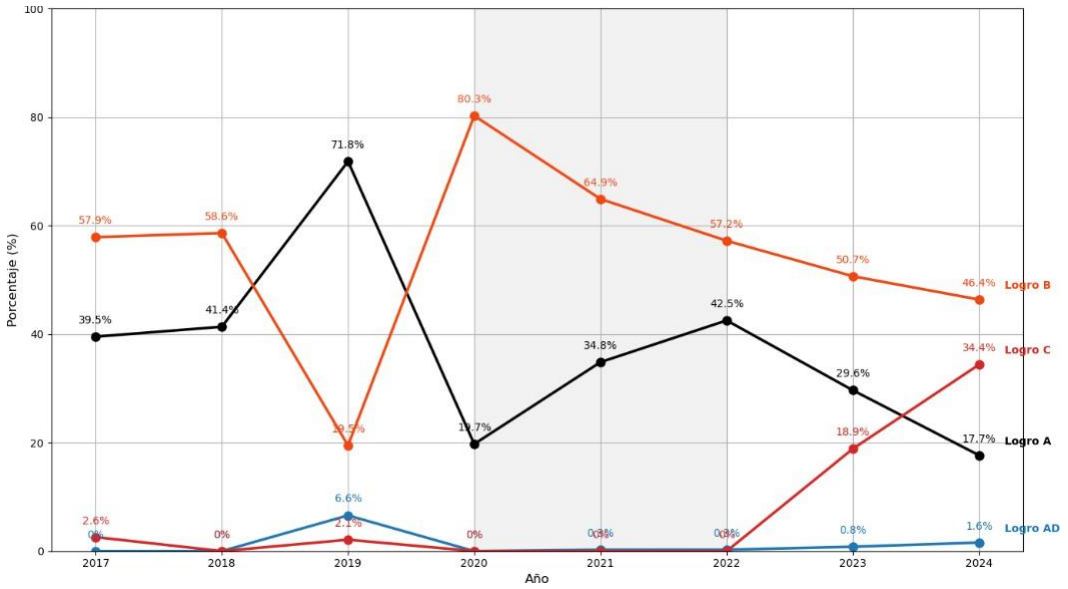
Análisis del rendimiento del curso de Arte y Cultura



Nota. Elaboración propia.

Figura 18

Análisis del rendimiento del curso de Ingles



Nota. Elaboración propia.

Tabla 17:

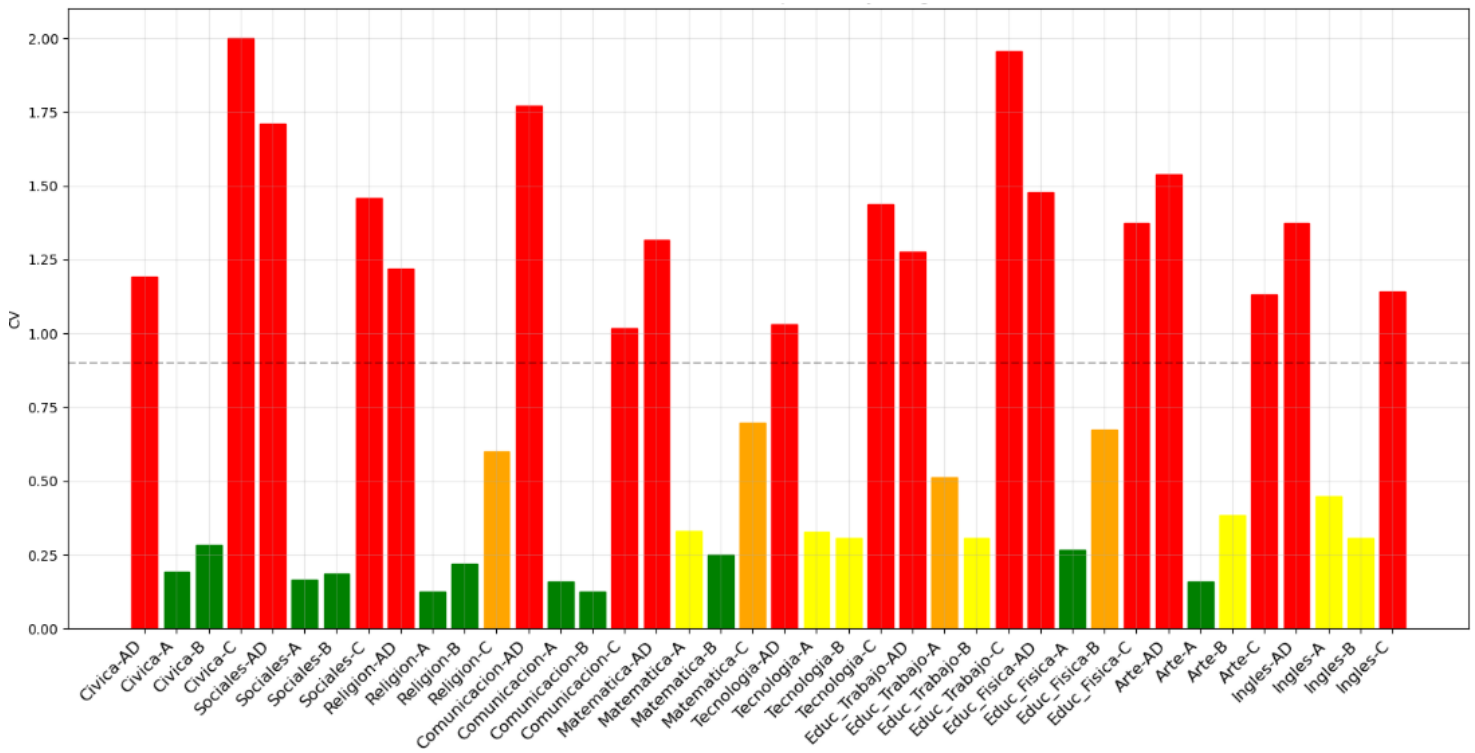
Coeficiente de variación y por Cursos y logros

Curso	AD	A	B	C
Arte	1.5396	0.1587	0.3852	1.131
Cívica	1.1919	0.1941	0.2834	2.000
Comunicación	1.7699	0.1608	0.1255	1.0189
Educ. Física	1.4793	0.2663	0.6755	1.3737
Educ_Trabajo	1.2753	0.5124	0.3078	1.9561
Ingles	1.3741	0.4501	0.3071	1.1435
Matemática	1.3169	0.3305	0.2512	0.6971
Religión	1.2203	0.1253	0.2213	0.6014
Sociales	1.711	0.1673	0.1849	1.4570
Tecnología	1.0294	0.3263	0.3071	1.4382

Nota. Elaboración propia.

Figura *19*

Coeficiente de variación y por Cursos y logros



Nota. Elaboración propia.

De acuerdo con el análisis, existen logros con sus asignaturas que tienen un alto coeficiente de variación, por lo que se procederá a aplicar la técnica de suavizado a los que tienen un coeficiente de variación > 0.5

Tabla 18:

Evaluación del coeficiente de variación y por Cursos y logros

Curso	Logro	CV	Rango	Nivel CV	Requiere Suavizado
Arte	A	0.1587	CV < 0.3	Baja	No ✓
Arte	AD	1.5396	CV ≥ 1	Muy Alta	Sí ⚠
Arte	B	0.3852	0.3 ≤ CV < 0.5	Moderada	No ✓
Arte	C	1.131	CV ≥ 1	Muy Alta	Sí ⚠
Cívica	A	0.1941	CV < 0.3	Baja	No ✓
Cívica	AD	1.1919	CV ≥ 1	Muy Alta	Sí ⚠
Cívica	B	0.2834	CV < 0.3	Baja	No ✓
Cívica	C	2	CV ≥ 1	Muy Alta	Sí ⚠
Comunicación	A	0.1608	CV < 0.3	Baja	No ✓
Comunicación	AD	1.7699	CV ≥ 1	Muy Alta	Sí ⚠
Comunicación	B	0.1255	CV < 0.3	Baja	No ✓
Comunicación	C	1.0189	CV ≥ 1	Muy Alta	Sí ⚠
Educ. Física	AD	1.4793	CV ≥ 1	Muy Alta	Sí ⚠
Educ. Física	B	0.6755	0.5 ≤ CV < 1	Alta	Sí ⚠
Educ. Física	C	1.3737	CV ≥ 1	Muy Alta	Sí ⚠
Educ. Trabajo	A	0.5124	0.5 ≤ CV < 1	Alta	Sí ⚠
Educ. Trabajo	AD	1.2753	CV ≥ 1	Muy Alta	Sí ⚠
Educ. Trabajo	B	0.3078	0.3 ≤ CV < 0.5	Moderada	No ✓
Educ. Trabajo	C	1.9561	CV ≥ 1	Muy Alta	Sí ⚠
Inglés	A	0.4501	0.3 ≤ CV < 0.5	Moderada	No ✓
Inglés	AD	1.3741	CV ≥ 1	Muy Alta	Sí ⚠
Inglés	B	0.3071	0.3 ≤ CV < 0.5	Moderada	No ✓
Inglés	C	1.1435	CV ≥ 1	Muy Alta	Sí ⚠
Matemática	A	0.3305	0.3 ≤ CV < 0.5	Moderada	No ✓
Matemática	AD	1.3169	CV ≥ 1	Muy Alta	Sí ⚠
Matemática	B	0.2512	CV < 0.3	Baja	No ✓
Matemática	C	0.6971	0.5 ≤ CV < 1	Alta	Sí ⚠
Religión	A	0.1253	CV < 0.3	Baja	No ✓
Religión	AD	1.2203	CV ≥ 1	Muy Alta	Sí ⚠
Religión	B	0.2213	CV < 0.3	Baja	No ✓
Religión	C	0.6014	0.5 ≤ CV < 1	Alta	Sí ⚠
Sociales	A	0.1673	CV < 0.3	Baja	No ✓
Sociales	AD	1.711	CV ≥ 1	Muy Alta	Sí ⚠
Sociales	B	0.1849	CV < 0.3	Baja	No ✓
Sociales	C	1.457	CV ≥ 1	Muy Alta	Sí ⚠
Tecnología	A	0.3263	0.3 ≤ CV < 0.5	Moderada	No ✓
Tecnología	AD	1.0294	CV ≥ 1	Muy Alta	Sí ⚠
Tecnología	B	0.3071	0.3 ≤ CV < 0.5	Moderada	No ✓
Tecnología	C	1.4382	CV ≥ 1	Muy Alta	Sí ⚠

Nota. Elaboración propia

4.1.4. Preparar los datos para los algoritmos de Machine Learning:

Se observo datos atípicos entre los años 2020-2022 por lo que se excluyó por estos motivos:

- Las calificaciones de este período no reflejan el rendimiento real de los estudiantes sino una política del gobierno excepcional por pandemia.
- Las predicciones estarían distorsionadas, con estos datos se generaría un modelo con predicciones sesgadas, ya que el algoritmo generaría patrones que no corresponden a condiciones normales de evaluación.
- Representan una anomalía o discontinuidad estadística que no sigue los patrones naturales de rendimiento académico.
- Crearían una falsa impresión de mejora seguida de un deterioro abrupto, cuando en realidad se trata de un cambio metodológico temporal.

4.1.4.1. Suavizar los datos atípicos.

El suavizado es una técnica de preprocesamiento que se aplica antes del entrenamiento del modelo para: reducir la variabilidad excesiva, mejorar la estabilidad de las predicciones, manejar valores atípicos y ruido en los datos, y preparar características más robustas para el modelo. Se aplicará el Suavizado Exponencial (EMA) que es una técnica que asigna mayor importancia a los datos recientes mientras mantiene la influencia de datos históricos, utilizando un factor de suavizado (α) que determina el balance entre ambos. Para datos educativos, se recomienda un α entre 0.2 y 0.3, lo que significa que aproximadamente el 30% del nuevo valor proviene del dato actual y el 70% de la historia previa.

- Un valor de α=0.3 significa que el 30% del valor de la EMA en el tiempo *t* proviene de la observación más reciente *Xt*.

- El restante 70% proviene del valor de la EMA en el tiempo anterior EMAt−1.

Esta técnica es especialmente útil porque reduce la variabilidad excesiva sin perder las tendencias importantes en el rendimiento académico.

Tabla 19:

Evaluación del suavizado a los datos originales

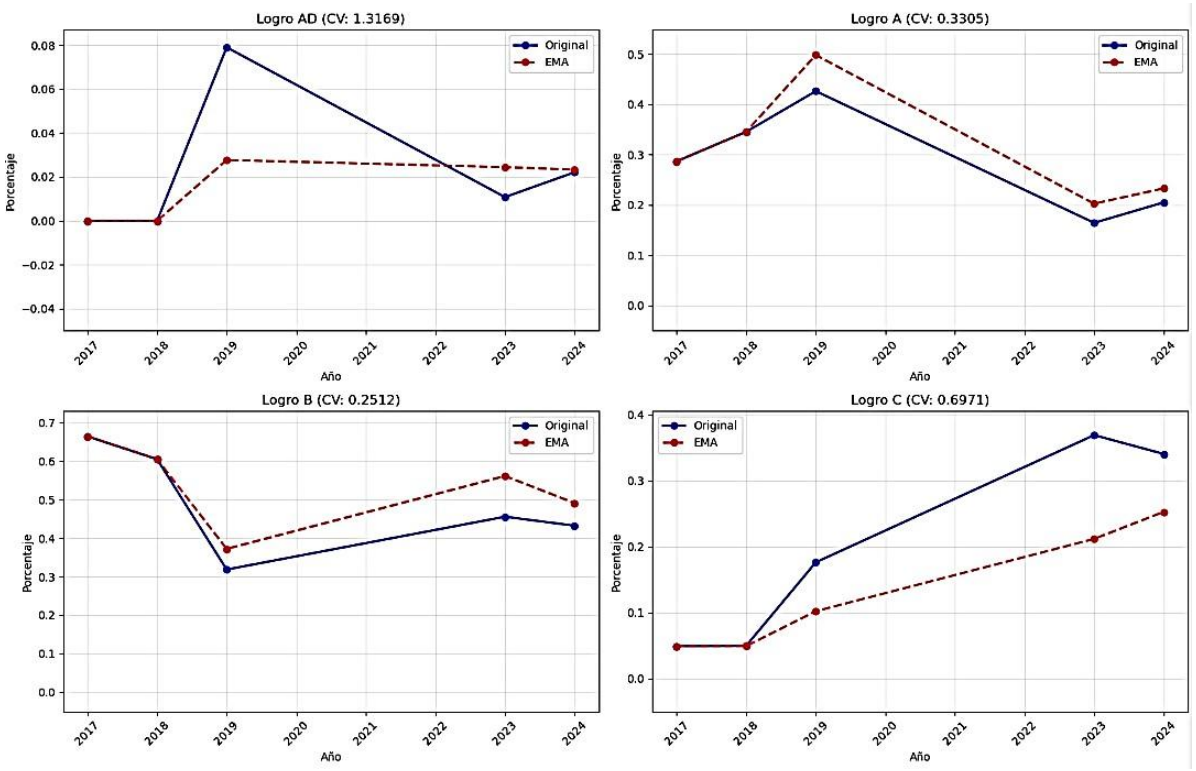
Curso	Logro	CV Original	CV Suavizado	Requirió Suavizado	% Reducción
Arte	A	0.1587	0.1511	No	4.77
Arte	AD	1.5396	0.7876	Sí	48.85
Arte	B	0.3852	0.3791	No	1.6
Arte	C	1.131	0.4948	Sí	56.25
Cívica	A	0.1941	0.1727	No	11.04
Cívica	AD	1.1919	0.8364	Sí	29.82
Cívica	B	0.2834	0.2947	No	-4
Cívica	C	2	2	Sí	0
Comunicación	A	0.1608	0.1608	No	0
Comunicación	AD	1.7699	0.8736	Sí	50.64
Comunicación	B	0.1255	0.0983	No	21.71
Comunicación	C	1.0189	0.6835	Sí	32.92
Educ. Física	A	0.2663	0.176	No	33.9
Educ. Física	AD	1.4793	0.6235	Sí	57.85
Educ. Física	B	0.6755	0.6479	Sí	4.08
Educ. Física	C	1.3737	1.1388	Sí	17.11
Educ. Trabajo	A	0.5124	0.1524	Sí	70.26
Educ. Trabajo	AD	1.2753	1.0104	Sí	20.77
Educ. Trabajo	B	0.3078	0.2586	No	15.97
Educ. Trabajo	C	1.9561	1.2152	Sí	37.88
Ingles	A	0.4501	0.4255	No	5.47
Ingles	AD	1.3741	0.8193	Sí	40.38
Ingles	B	0.3071	0.2957	No	3.73
Ingles	C	1.1435	0.9842	Sí	13.93
Matemática	A	0.3305	0.3337	No	-0.97
Matemática	AD	1.3169	0.8221	Sí	37.57
Matemática	B	0.2512	0.1869	No	25.61
Matemática	C	0.6971	0.634	Sí	9.05
Religión	A	0.1253	0.1204	No	3.91
Religión	AD	1.2203	0.8636	Sí	29.23
Religión	B	0.2213	0.2066	No	6.65
Religión	C	0.6014	0.5378	Sí	10.57
Sociales	A	0.1673	0.1537	No	8.14
Sociales	AD	1.711	0.8721	Sí	49.03

Sociales	B	0.1849	0.1878	No	-1.54
Sociales	C	1.457	1.1513	Sí	20.98
Tecnología	A	0.3263	0.3139	No	3.8
Tecnología	AD	1.0294	0.8227	Sí	20.08
Tecnología	B	0.3071	0.3203	No	-4.3
Tecnología	C	1.4382	0.6865	Sí	52.27

Nota. Elaboración propia

Figura 20

Análisis de la validación Cruzada para el curso de Arte y Cultura



Nota. Elaboración propia

4.1.5. Selección del modelo.

Como se visualiza en el presente *estudio existe una alta variabilidad en los datos* del rendimiento académico, por el cual se propone los siguientes algoritmos y enfoques:

- **Algoritmos Robustos a Valores atípicos (Valores Atípicos) y Variabilidad**
  - Modelo de regresión múltiple: Más sencillo

- Árboles de Decisión y Random Forest: Menos sensibles a valores atípicos y alta variabilidad.
- Gradient Boosting (XGBoost): Manejan bien datos con alta varianza mediante el aprendizaje secuencial

Para seleccionar el algoritmo más adecuado para cada caso *(Logros por cada Curso)* se procedió a evaluar los algoritmos propuestos anteriormente (Regresión Múltiple, Random Forest, Xgboost) con las métricas MAE, MSE y RMSE mediante la *validación cruzada temporal* que es una técnica de evaluación de modelos usada en problemas donde los datos tienen una estructura temporal. A diferencia de la validación cruzada tradicional, que asume que los datos son independientes e intercambiables, la validación cruzada temporal respeta el orden cronológico de los datos para evitar fugas de información del futuro al pasado. En la siguiente tabla se muestra un ejemplo con el mecanismo de cálculo para el modelo Random Forest aplicado al curso de **Matemática** en el logro **AD**. Posteriormente se realizará el mismo calculo para todas las combinaciones del modelo, curso y logro.

Tabla *20*:

Determinación de validación cruzada: Random Forest | Matemática | AD

Split	Años Entrenamiento	Años Prueba	MSE	RMSE	MAE
1	2017, 2018	2019	0.006233	0.078947	0.078947
2	2017, 2018, 2019	2023	0.001708	0.041324	0.041324
3	2017, 2018, 2019, 2023	2024	0.000009	0.002941	0.002941
Promedio	-	-	**0.002650**	**0.041071**	**0.041071**

Nota. Elaboración propia.

Tabla 21:

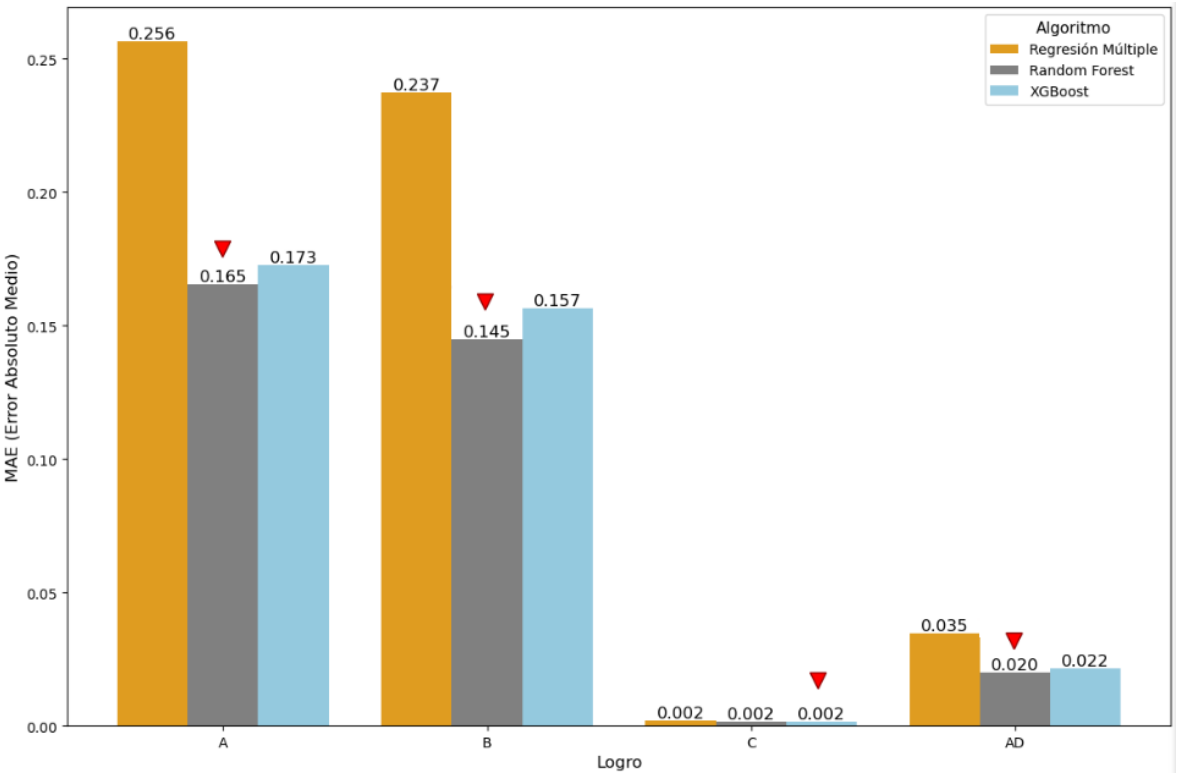
Validación Cruzada para el curso de Arte y Cultura

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Arte	A	Regresión Múltiple	0.256409609	0.078751408	0.25641
Arte	A	Random Forest	0.16541126	0.027653282	0.165411
Arte	A	XGBoost	0.172675507	0.032709963	0.172676
Arte	B	Regresión Múltiple	0.237353163	0.064642591	0.237353
Arte	B	Random Forest	0.144852978	0.021114981	0.144853
Arte	B	XGBoost	0.156666227	0.025755232	0.156666
Arte	C	Regresión Múltiple	0.001782995	6.04073E-06	0.001783
Arte	C	Random Forest	0.001757509	4.38039E-06	0.001758
Arte	C	XGBoost	0.001725952	3.7989E-06	0.001726
Arte	AD	Regresión Múltiple	0.034641635	0.001279103	0.034642
Arte	AD	Random Forest	0.020104261	0.000732315	0.020104
Arte	AD	XGBoost	0.021654592	0.000751505	0.021655

Nota. Elaboración propia.

Figura 21

Análisis de la validación Cruzada para el curso de Arte y Cultura



Nota. Elaboración propia

Tabla 22:

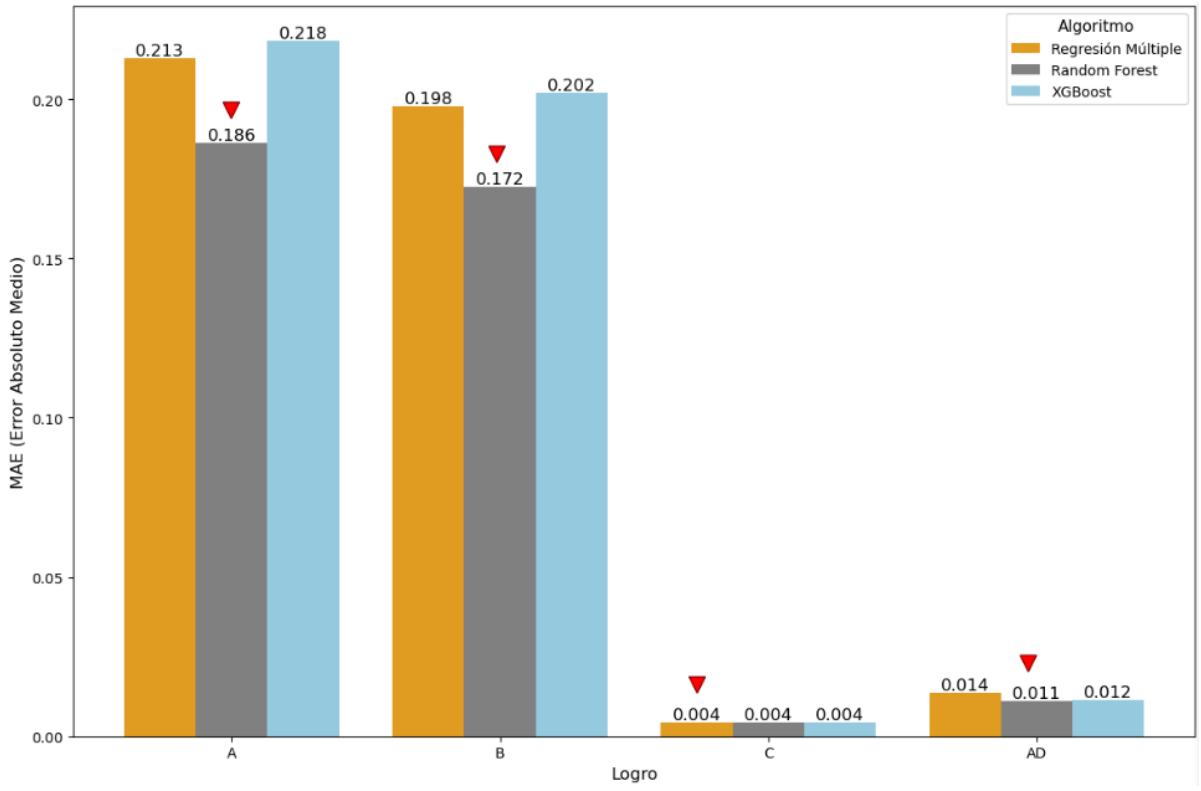
Validación Cruzada para el curso de Ciudadanía y Cívica

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Cívica	A	Regresión Múltiple	0.212919337	0.057500758	0.212919
Cívica	A	Random Forest	0.18614273	0.041763357	0.186143
Cívica	A	XGBoost	0.218320129	0.059300271	0.21832
Cívica	B	Regresión Múltiple	0.197892238	0.052589275	0.197892
Cívica	B	Random Forest	0.17245802	0.03831463	0.172458
Cívica	B	XGBoost	0.201957914	0.053397698	0.201958
Cívica	C	Regresión Múltiple	0.004283757	5.50517E-05	0.004284
Cívica	C	Random Forest	0.004283757	5.50517E-05	0.004284
Cívica	C	XGBoost	0.004283757	5.50517E-05	0.004284
Cívica	AD	Regresión Múltiple	0.013695595	0.000247032	0.013696
Cívica	AD	Random Forest	0.010962757	0.000200372	0.010963
Civica	AD	XGBoost	0.011508	0.000203	0.011508

Nota. Elaboración propia.

Figura 22

Análisis de la validación Cruzada para el curso de Ciudadanía y Cívica



Nota. Elaboración propia.

Tabla 23:

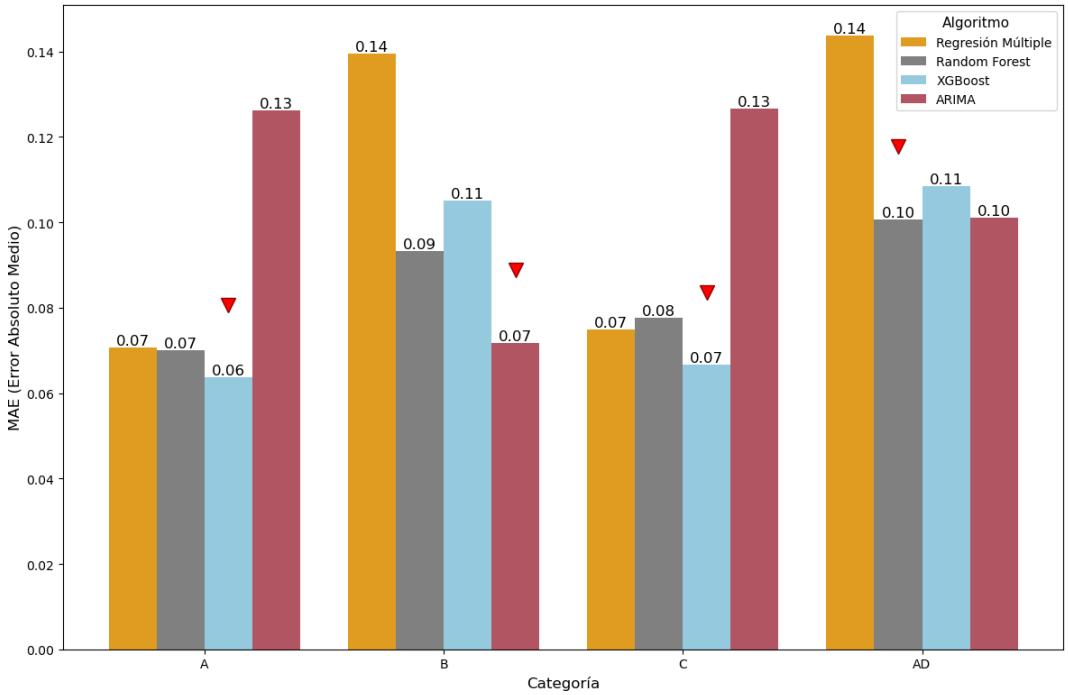
Validación Cruzada para el curso de Comunicación

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Comunicación	A	Regresión Múltiple	0.110964797	0.013409488	0.110965
Comunicación	A	Random Forest	0.092851345	0.008877393	0.092851
Comunicación	A	XGBoost	0.091136281	0.008667502	0.091136
Comunicación	B	Regresión Múltiple	0.123243299	0.015332726	0.123243
Comunicación	B	Random Forest	0.082188629	0.007587288	0.082189
Comunicación	B	XGBoost	0.089428199	0.00892728	0.089428
Comunicación	C	Regresión Múltiple	0.029691479	0.00128368	0.029691
Comunicación	C	Random Forest	0.031181278	0.001359649	0.031181
Comunicación	C	XGBoost	0.028516374	0.001084811	0.028516
Comunicación	AD	Regresión Múltiple	0.041969981	0.001852195	0.04197
Comunicación	AD	Random Forest	0.023015479	0.00105644	0.023015
Comunicación	AD	XGBoost	0.026433622	0.001116971	0.026434

Nota. Elaboración propia.

Figura 23

Análisis de la validación Cruzada para el curso de Comunicación



Nota. Elaboración propia.

Tabla 24:

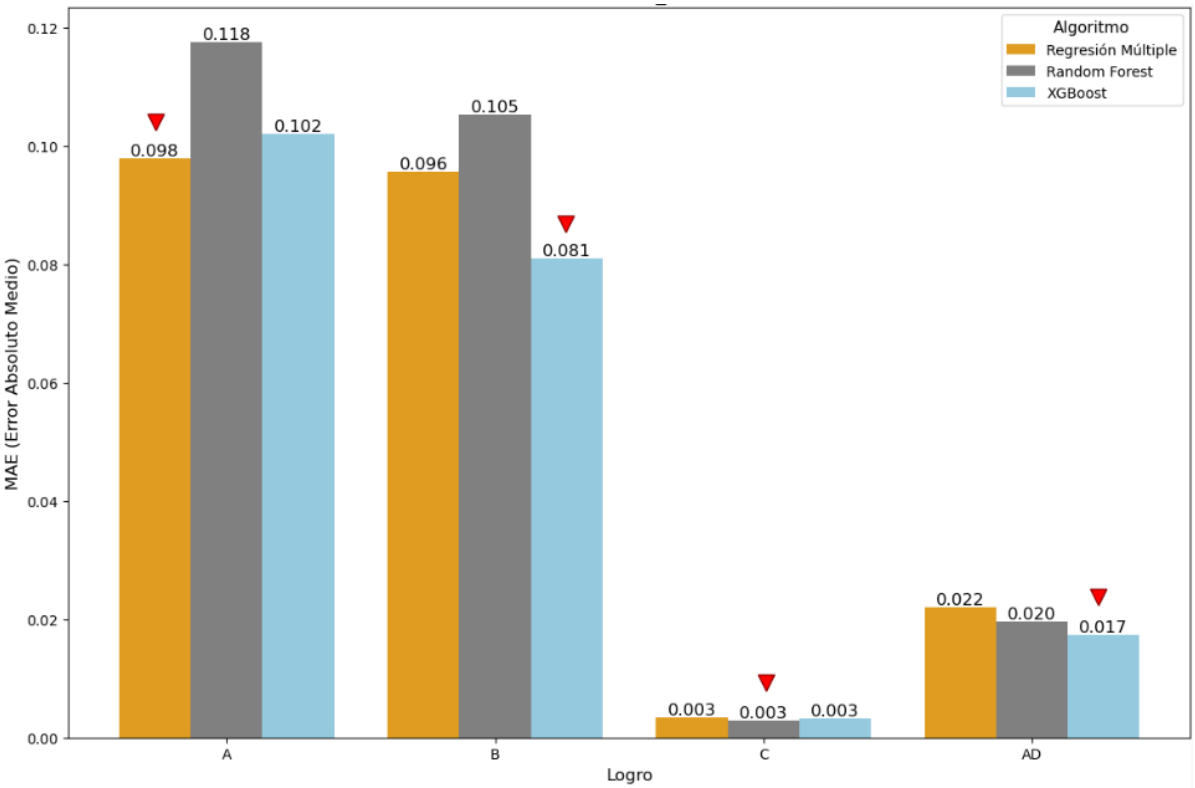
Validación Cruzada para el curso de Educación Física

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Educ. Física	A	Regresión Múltiple	0.097929524	0.013771635	0.09793
Educ. Física	A	Random Forest	0.117554293	0.023420329	0.117554
Educ. Física	A	XGBoost	0.1020954	0.020151304	0.102095
Educ. Física	B	Regresión Múltiple	0.095603182	0.012842642	0.095603
Educ. Física	B	Random Forest	0.10523418	0.019499141	0.105234
Educ. Física	B	XGBoost	0.081069704	0.017521461	0.08107
Educ. Física	C	Regresión Múltiple	0.003467312	1.77819E-05	0.003467
Educ. Física	C	Random Forest	0.002825921	1.77533E-05	0.002826
Educ. Física	C	XGBoost	0.003277783	1.97954E-05	0.003278
Educ. Física	AD	Regresión Múltiple	0.021996423	0.000661776	0.021996
Educ. Física	AD	Random Forest	0.019585697	0.000425092	0.019586
Educ. Física	AD	XGBoost	0.017324052	0.000379982	0.017324

Nota. Elaboración propia.

Figura 24

Análisis de la validación Cruzada para el curso de Educación Física



Nota. Elaboración propia.

Tabla 25:

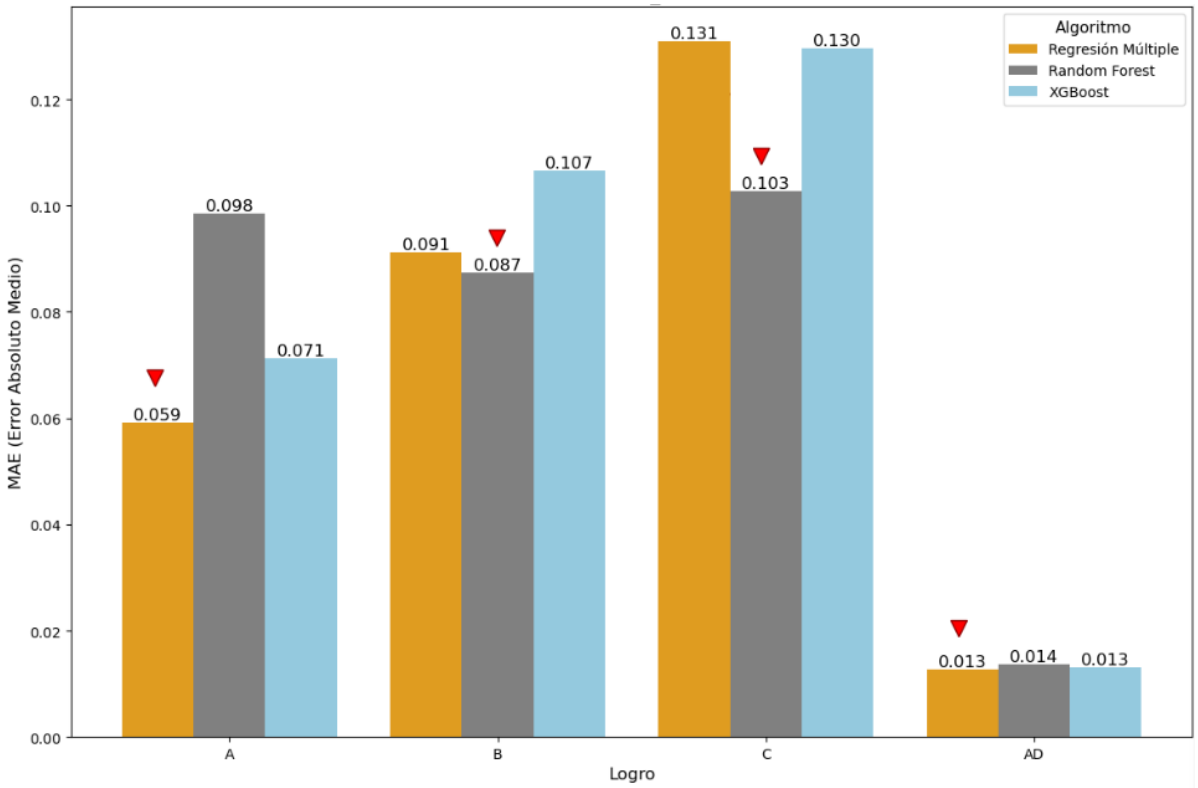
Validación Cruzada para el curso de Educación para el trabajo

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Educ_Trabajo	A	Regresión Múltiple	0.059252438	0.003541005	0.059252
Educ_Trabajo	A	Random Forest	0.098445269	0.011708267	0.098445
Educ_Trabajo	A	XGBoost	0.071325524	0.007603678	0.071326
Educ_Trabajo	B	Regresión Múltiple	0.091276133	0.016114865	0.091276
Educ_Trabajo	B	Random Forest	0.08747838	0.009145367	0.087478
Educ_Trabajo	B	XGBoost	0.10665248	0.012924826	0.106652
Educ_Trabajo	C	Regresión Múltiple	0.130883039	0.030630296	0.130883
Educ_Trabajo	C	Random Forest	0.102679305	0.027122927	0.102679
Educ_Trabajo	C	XGBoost	0.129683046	0.030557301	0.129683
Educ_Trabajo	AD	Regresión Múltiple	0.012784836	0.000174949	0.012785
Educ_Trabajo	AD	Random Forest	0.013737134	0.000257564	0.013737
Educ_Trabajo	AD	XGBoost	0.013164318	0.000244632	0.013164

Nota. Elaboración propia.

Figura 25

Análisis de la validación Cruzada para el curso de Educ. para el trabajo



Nota. Elaboración propia.

Tabla 26:

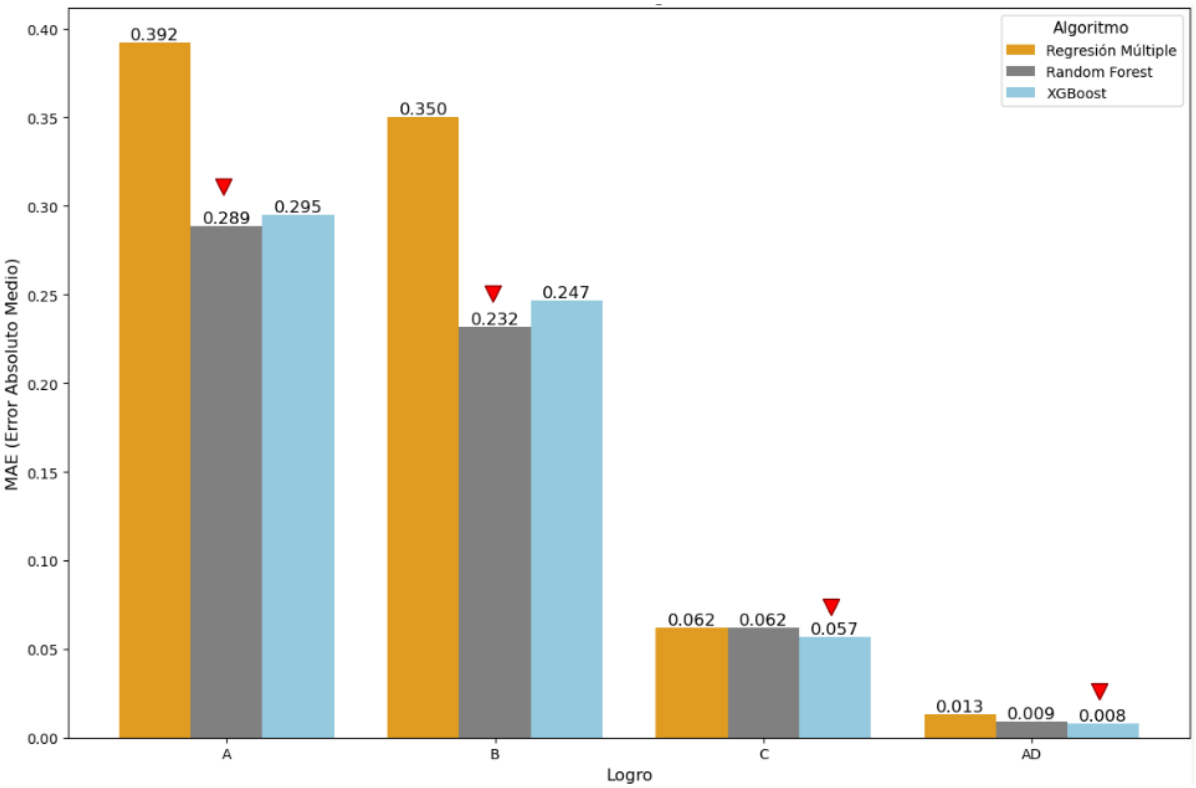
Validación Cruzada para el curso de Inglés

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Ingles	A	Regresión Múltiple	0.392321141	0.165848991	0.392321
Ingles	A	Random Forest	0.288784682	0.086691035	0.288785
Ingles	A	XGBoost	0.295239497	0.104081686	0.295239
Ingles	B	Regresión Múltiple	0.350095042	0.137739334	0.350095
Ingles	B	Random Forest	0.23151909	0.067496906	0.231519
Ingles	B	XGBoost	0.24696361	0.090396574	0.246964
Ingles	C	Regresión Múltiple	0.062000229	0.005586014	0.062
Ingles	C	Random Forest	0.061903768	0.00650717	0.061904
Ingles	C	XGBoost	0.056947614	0.005235497	0.056948
Ingles	AD	Regresión Múltiple	0.01294506	0.000197989	0.012945
Ingles	AD	Random Forest	0.009232501	0.000152191	0.009233
Ingles	AD	XGBoost	0.007798456	0.000144511	0.007798

Nota. Elaboración propia.

Figura 26

Análisis de la validación Cruzada para el curso de Inglés



Nota. Elaboración propia.

Tabla 27:

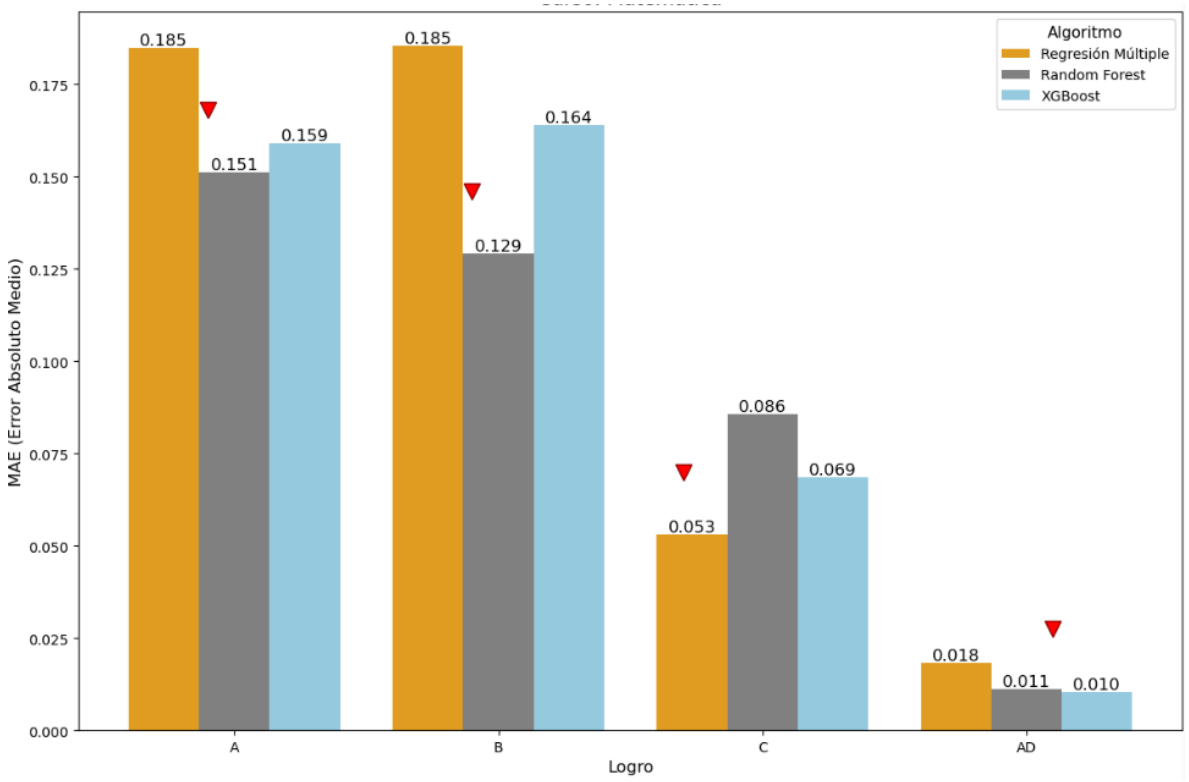
Validación Cruzada para el curso de Matemática

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Matemática	A	Regresión Múltiple	0.184828026	0.054381107	0.184828
Matemática	A	Random Forest	0.151076169	0.029374184	0.151076
Matemática	A	XGBoost	0.159060699	0.037050713	0.159061
Matemática	B	Regresión Múltiple	0.185247616	0.043228791	0.185248
Matemática	B	Random Forest	0.128986165	0.024697622	0.128986
Matemática	B	XGBoost	0.163911638	0.031649357	0.163912
Matemática	C	Regresión Múltiple	0.053020423	0.003812067	0.05302
Matemática	C	Random Forest	0.085691488	0.008318259	0.085691
Matemática	C	XGBoost	0.068559861	0.005611309	0.06856
Matemática	AD	Regresión Múltiple	0.018376639	0.000381946	0.018377
Matemática	AD	Random Forest	0.011318893	0.00026788	0.011319
Matemática	AD	XGBoost	0.010493447	0.000257755	0.010493

Nota. Elaboración propia.

Figura 27

Análisis de la validación Cruzada para el curso de Matemática



Nota. Elaboración propia.

Tabla 28:

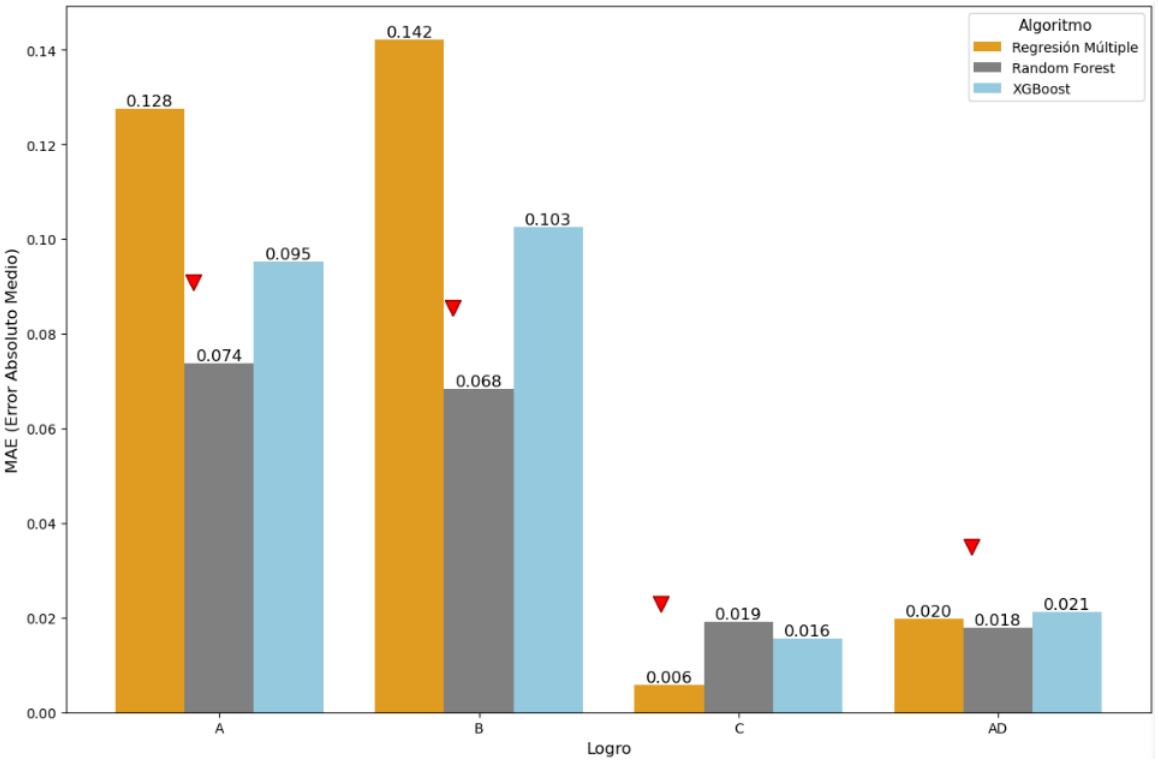
Validación Cruzada para el curso de Educación Religiosa

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Religión	A	Regresión Múltiple	0.127582249	0.036730619	0.127582
Religión	A	Random Forest	0.073822446	0.006128098	0.073822
Religión	A	XGBoost	0.095235455	0.012197428	0.095235
Religión	B	Regresión Múltiple	0.142119383	0.04552686	0.142119
Religión	B	Random Forest	0.06839192	0.00845892	0.068392
Religión	B	XGBoost	0.102509386	0.016294987	0.102509
Religión	C	Regresión Múltiple	0.005866959	8.74715E-05	0.005867
Religión	C	Random Forest	0.019111113	0.00041448	0.019111
Religión	C	XGBoost	0.01552918	0.000293944	0.015529
Religión	AD	Regresión Múltiple	0.01967613	0.000602177	0.019676
Religión	AD	Random Forest	0.017948503	0.000523933	0.017949
Religión	AD	XGBoost	0.021157713	0.000561691	0.021158

Nota. Elaboración propia.

Figura 28

Análisis de la validación Cruzada para el curso de Educación Religiosa



Nota. Elaboración propia

Tabla 29:

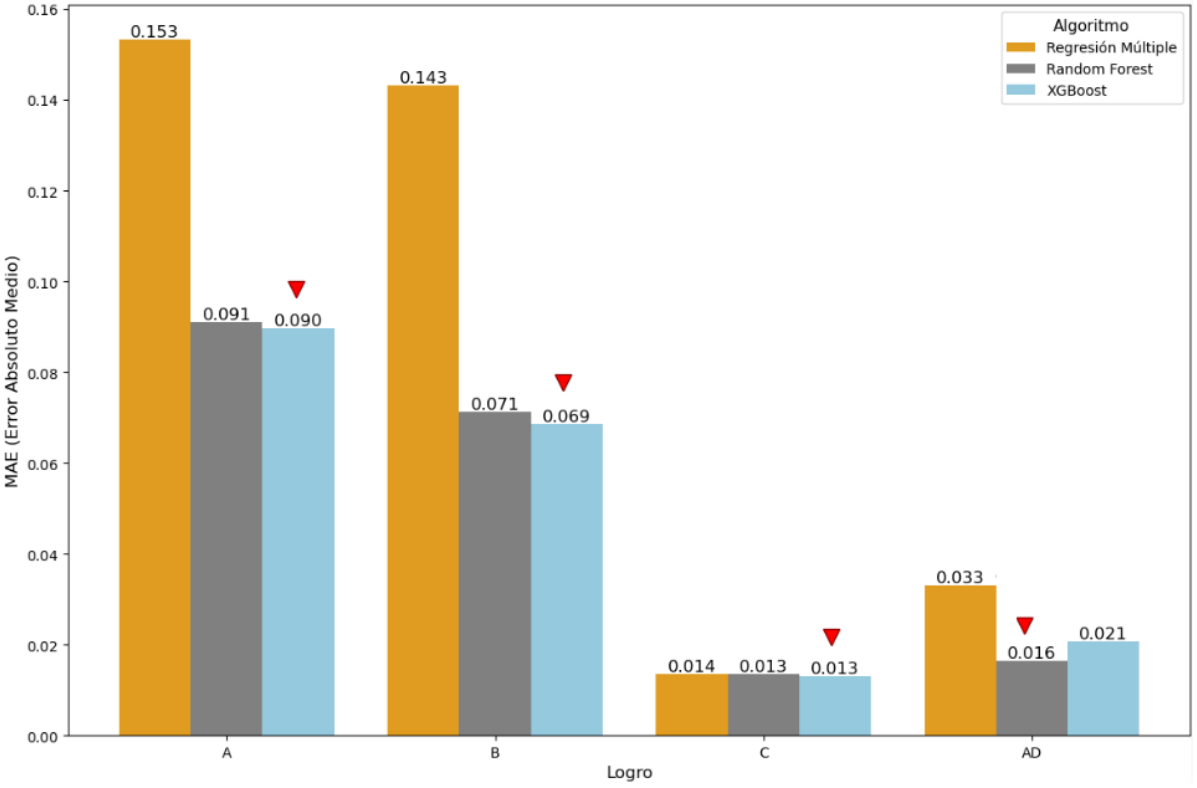
Validación Cruzada para el curso de Ciencias Sociales

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Sociales	A	Regresión Múltiple	0.153154617	0.024796346	0.153155
Sociales	A	Random Forest	0.091090789	0.012054391	0.091091
Sociales	A	XGBoost	0.089712552	0.010477709	0.089713
Sociales	B	Regresión Múltiple	0.143190266	0.02106835	0.14319
Sociales	B	Random Forest	0.071345453	0.008419673	0.071345
Sociales	B	XGBoost	0.068643395	0.006920111	0.068643
Sociales	C	Regresión Múltiple	0.013589909	0.000358381	0.01359
Sociales	C	Random Forest	0.013453071	0.000367106	0.013453
Sociales	C	XGBoost	0.013103943	0.000345783	0.013104
Sociales	AD	Regresión Múltiple	0.033168316	0.001164671	0.033168
Sociales	AD	Random Forest	0.016451518	0.000661991	0.016452
Sociales	AD	XGBoost	0.020691062	0.000722566	0.020691

Nota. Elaboración propia.

Figura 29

Análisis de la validación Cruzada para el curso de Ciencias Sociales



Nota. Elaboración propia.

Tabla 30:

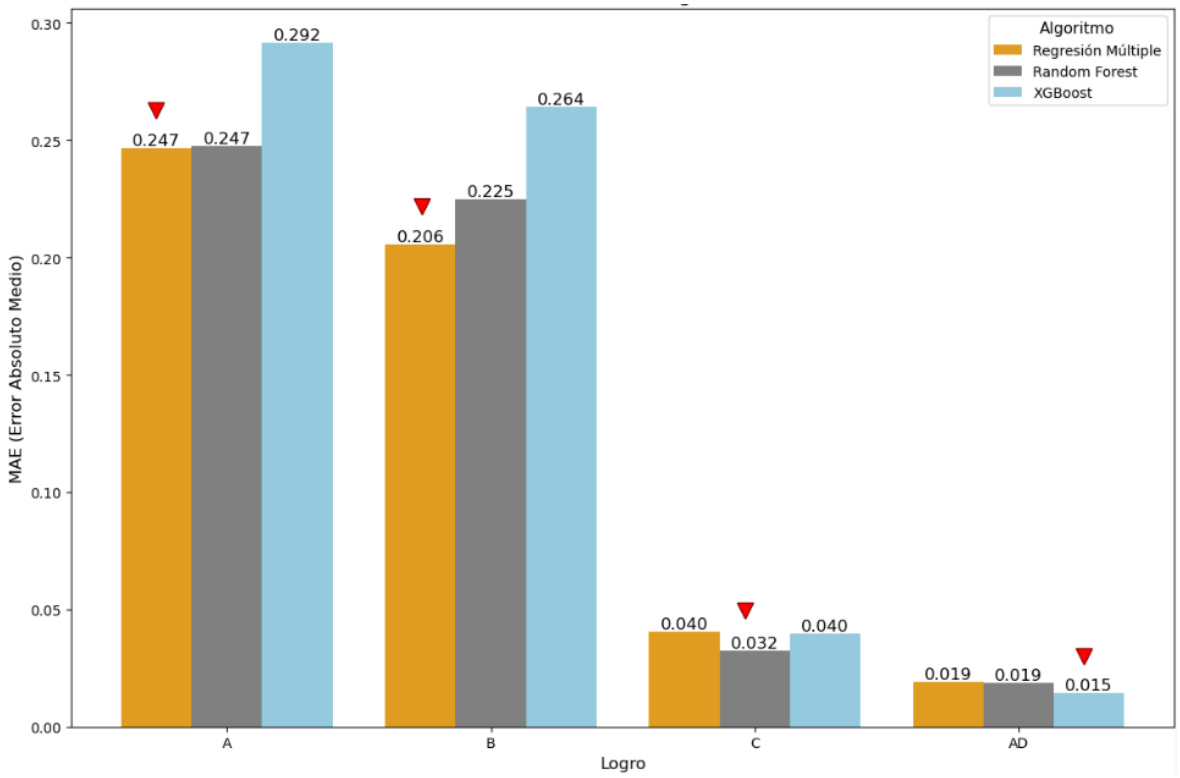
Validación Cruzada para el curso de Ciencia y Tecnología

Curso	Logro	Algoritmo	MAE	MSE	RMSE
Tecnología	A	Regresión Múltiple	0.246764377	0.129836564	0.246764
Tecnología	A	Random Forest	0.247363598	0.070426275	0.247364
Tecnología	A	XGBoost	0.29164336	0.098557431	0.291643
Tecnología	B	Regresión Múltiple	0.205526233	0.098073584	0.205526
Tecnología	B	Random Forest	0.224732007	0.05351739	0.224732
Tecnología	B	XGBoost	0.264144341	0.074594429	0.264144
Tecnología	C	Regresión Múltiple	0.040473969	0.002831691	0.040474
Tecnología	C	Random Forest	0.032419255	0.002280469	0.032419
Tecnología	C	XGBoost	0.039875535	0.002634025	0.039876
Tecnología	AD	Regresión Múltiple	0.018919581	0.000483473	0.01892
Tecnología	AD	Random Forest	0.01876805	0.000484093	0.018768
Tecnología	AD	XGBoost	0.014515452	0.000416753	0.014515

Nota. Elaboración propia.

Figura 30

Análisis de la validación Cruzada para el curso de Ciencia y Tecnología



Nota. Elaboración propia.

Una vez realizada la validación cruzada la selección del mejor modelo se basa en analizar las métricas que consiste en:

- Comparar MSE, RMSE y MAE entre splits
- *Identificar el modelo con valores más bajos*
- Priorizar la métrica más relevante.

Tabla 31:

Modelos finales seleccionados por Curso y Logro

Curso	Logro	Algoritmo	MAE	MSE	RMSE	PRECISION
Arte	A	Random Forest	0.165411	0.027653	0.165411	83.45887
Arte	AD	Random Forest	0.020104	0.000732	0.020104	97.98957
Arte	B	Random Forest	0.144853	0.021115	0.144853	85.5147
Arte	C	XGBoost	0.001726	3.8E-06	0.001726	99.8274
Cívica	A	Random Forest	0.186143	0.041763	0.186143	81.38573
Cívica	AD	Random Forest	0.010963	0.0002	0.010963	98.90372
Cívica	B	Random Forest	0.172458	0.038315	0.172458	82.7542
Cívica	C	Regresión Múltiple	0.004284	5.51E-05	0.004284	99.57162
Comunicación	A	XGBoost	0.091136	0.008668	0.091136	90.88637
Comunicación	AD	Random Forest	0.023015	0.001056	0.023015	97.69845
Comunicación	B	Random Forest	0.082189	0.007587	0.082189	91.78114
Comunicación	C	XGBoost	0.028516	0.001085	0.028516	97.14836
Educación Física	A	Regresión Múltiple	0.09793	0.013772	0.09793	90.20705
Educación Física	AD	XGBoost	0.017324	0.00038	0.017324	98.26759
Educación Física	B	XGBoost	0.08107	0.017521	0.08107	91.89303
Educación Física	C	Random Forest	0.002826	1.78E-05	0.002826	99.71741
Educación para el Trabajo	A	Regresión Múltiple	0.059252	0.003541	0.059252	94.07476
Educación para el Trabajo	AD	Regresión Múltiple	0.012785	0.000175	0.012785	98.72152
Educación para el Trabajo	B	Random Forest	0.087478	0.009145	0.087478	91.25216
Educación para el Trabajo	C	Random Forest	0.102679	0.027123	0.102679	89.73207
Ingles	A	Random Forest	0.288785	0.086691	0.288785	71.12153
Ingles	AD	XGBoost	0.007798	0.000145	0.007798	99.22015
Ingles	B	Random Forest	0.231519	0.067497	0.231519	76.84809
Ingles	C	XGBoost	0.056948	0.005235	0.056948	94.30524
Matemática	A	Random Forest	0.151076	0.029374	0.151076	84.89238
Matemática	AD	XGBoost	0.010493	0.000258	0.010493	98.95066
Matemática	B	Random Forest	0.128986	0.024698	0.128986	87.10138
Matemática	C	Regresión Múltiple	0.05302	0.003812	0.05302	94.69796
Religión	A	Random Forest	0.073822	0.006128	0.073822	92.61776
Religión	AD	Random Forest	0.017949	0.000524	0.017949	98.20515
Religión	B	Random Forest	0.068392	0.008459	0.068392	93.16081
Religión	C	Regresión Múltiple	0.005867	8.75E-05	0.005867	99.4133
Ciencias Sociales	A	XGBoost	0.089713	0.010478	0.089713	91.02874
Ciencias Sociales	AD	Random Forest	0.016452	0.000662	0.016452	98.35485
Ciencias Sociales	B	XGBoost	0.068643	0.00692	0.068643	93.13566
Ciencias Sociales	C	XGBoost	0.013104	0.000346	0.013104	98.68961

Ciencia y Tecnología	A	Regresión Múltiple	0.246764	0.129837	0.246764	75.32356
Ciencia y Tecnología	AD	XGBoost	0.014515	0.000417	0.014515	98.54845
Ciencia y Tecnología	B	Regresión Múltiple	0.205526	0.098074	0.205526	79.44738
Ciencia y Tecnología	C	Random Forest	0.032419	0.00228	0.032419	96.75807

Nota. Elaboración propia.

4.1.6. Entrenamiento del modelo.

Se procedió a entrenar el modelo de acuerdo con los algoritmos más eficientes identificados anteriormente. Los resultados al 2025 se pueden visualizar en la tabla a continuación:

Tabla 32:

Resultados del entrenamiento

Curso	Logro	Modelo	2017	2018	2019	2023	2024	2025
Arte	A	Random Forest	0.7158	0.8972	0.666	0.5661	0.754	0.7037
Arte	AD	Random Forest	0.0026	0.0019	0.0476	0.0373	0.0275	0.0309
Arte	B	Random Forest	0.2791	0.0991	0.285	0.3917	0.2132	0.2606
Arte	C	XGBoost	0.0026	0.0019	0.0014	0.0049	0.0052	0.0045
Cívica	A	Random Forest	0.6227	0.6152	0.5494	0.8301	0.5206	0.6019
Cívica	AD	Random Forest	0	0	0.0233	0.0177	0.025	0.0225
Cívica	B	Random Forest	0.3773	0.3848	0.4273	0.1521	0.4416	0.3671
Cívica	C	Regresión múltiple	0	0	0	0	0.0129	0.0086
Comunicación	A	XGBoost	0.4264	0.4315	0.503	0.3847	0.302	0.3032
Comunicación	AD	Random Forest	0	0	0.0553	0.0402	0.0307	0.0342
Comunicación	B	Random Forest	0.5478	0.5459	0.424	0.5214	0.57	0.5472
Comunicación	C	XGBoost	0.0258	0.0226	0.0177	0.0537	0.0973	0.0961
Educación Física	A	Regresión múltiple	0.9457	0.8801	0.844	0.6035	0.6324	0.5483
Educación Física	AD	XGBoost	0.0078	0.0062	0.0353	0.0408	0.0224	0.0229
Educación Física	B	XGBoost	0.0465	0.1137	0.1199	0.3479	0.3399	0.3403
Educación Física	C	Random Forest	0	0	0.0009	0.0078	0.0054	0.0055
Educ. Trabajo	A	Regresión múltiple	0.8165	0.7764	0.6842	0.5663	0.5655	0.5076
Educ. Trabajo	AD	Regresión múltiple	0	0	0.0159	0.0132	0.0355	0.0318
Educ. Trabajo	B	Random Forest	0.1783	0.2201	0.2947	0.1314	0.2148	0.1996
Educ. Trabajo	C	Random Forest	0.0052	0.0035	0.0052	0.2891	0.1843	0.1943

Ingles	A	Random Forest	0.3953	0.4063	0.7547	0.3334	0.2185	0.2885
Ingles	AD	XGBoost	0	0	0.0207	0.0183	0.0199	0.0193
Ingles	B	Random Forest	0.5788	0.576	0.2046	0.5698	0.5735	0.5468
Ingles	C	XGBoost	0.0258	0.0178	0.0199	0.0786	0.1881	0.1869
Matemática	A	Random Forest	0.2868	0.3457	0.4981	0.2026	0.2329	0.2462
Matemática	AD	XGBoost	0	0	0.0277	0.0244	0.0233	0.0237
Matemática	B	Random Forest	0.6641	0.605	0.3721	0.5612	0.4909	0.5025
Matemática	C	Regresión múltiple	0.0491	0.0493	0.1021	0.2118	0.2528	0.2773
Religión	A	Random Forest	0.7132	0.5361	0.6871	0.5693	0.5517	0.5652
Religión	AD	Random Forest	0	0	0.0357	0.0244	0.0425	0.0366
Religión	B	Random Forest	0.2739	0.4427	0.2464	0.3506	0.341	0.3388
Religión	C	Regresión múltiple	0.0129	0.0212	0.0308	0.0557	0.0648	0.0714
Ciencias Sociales	A	XGBoost	0.5762	0.6886	0.6256	0.578	0.4201	0.4213
Ciencias Sociales	AD	Random Forest	0	0	0.0443	0.0296	0.0261	0.0277
Ciencias Sociales	B	XGBoost	0.4186	0.3078	0.3273	0.3826	0.5135	0.5123
Ciencias Sociales	C	XGBoost	0.0052	0.0036	0.0028	0.0098	0.0403	0.0391
Ciencia y Tecn.	A	Regresión múltiple	0.3747	0.5527	0.7111	0.2679	0.5422	0.4331
Ciencia y Tecn.	AD	XGBoost	0	0	0.0347	0.0364	0.0416	0.0404
Ciencia y Tecn.	B	Regresión múltiple	0.5995	0.4255	0.2342	0.5932	0.3501	0.4227
Ciencia y Tecn.	C	Random Forest	0.0258	0.0218	0.0199	0.1025	0.066	0.071

Nota. Elaboración propia.

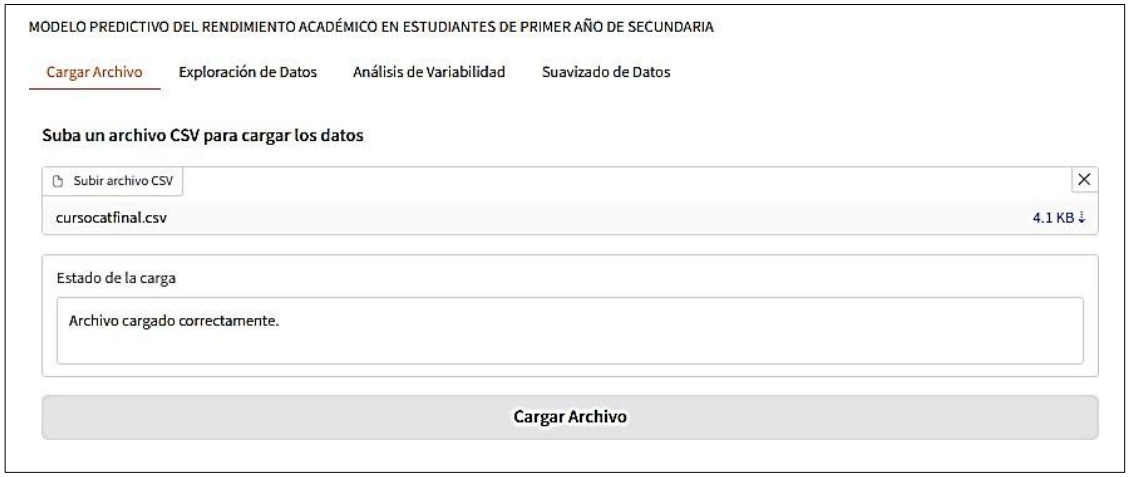
4.1.7. Presentación e implementación de la solución

Se implemento el modelo en Python, tanto para la interfaz gráfica con el paquete de código abierto *GRADIO,* como también para el desarrollo del modelo predictivo.

A continuación, se muestran las interfaces graficas del modelo propuesto.

Figura 31

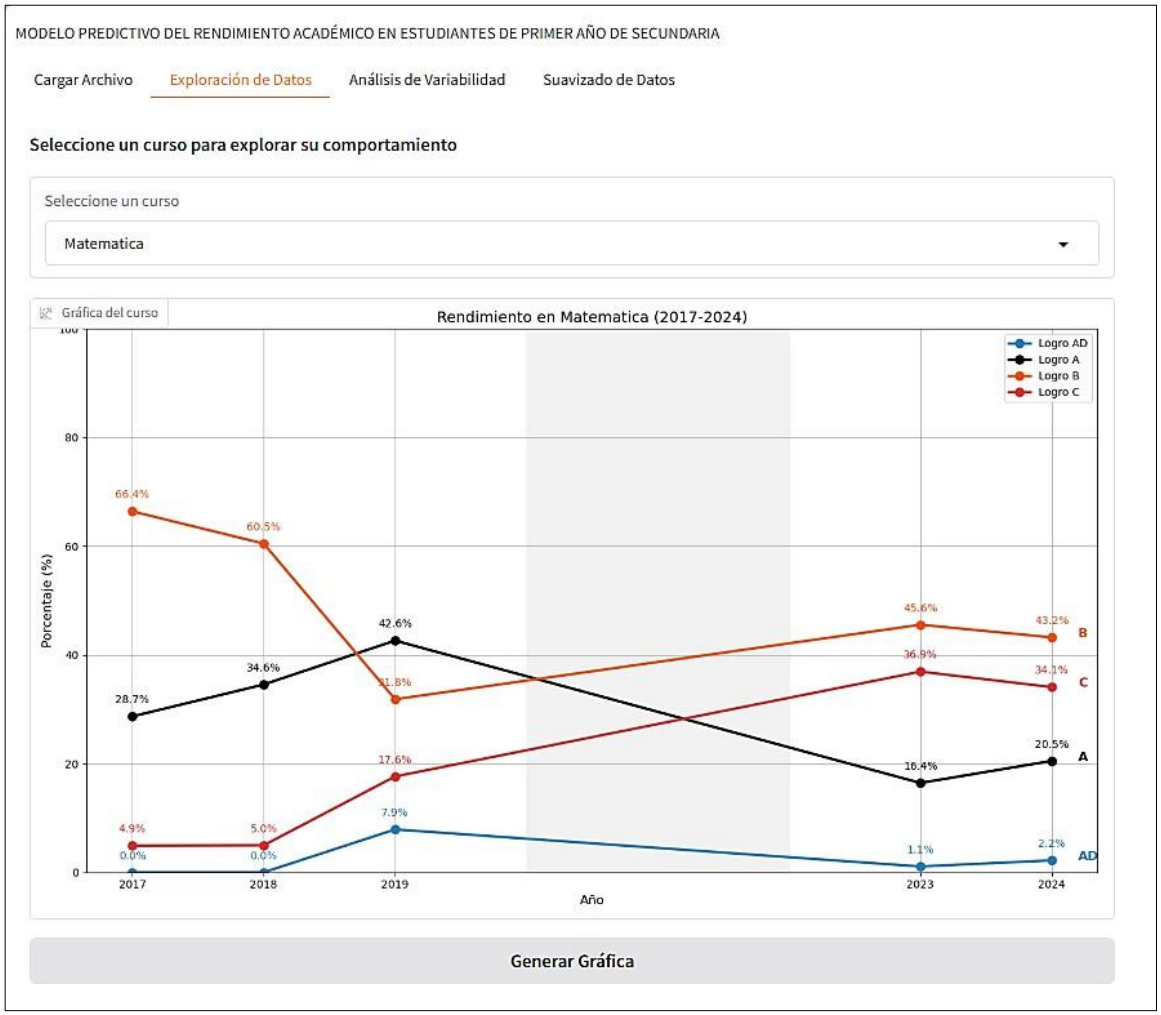
Interfaz gráfica del modelo propuesto: Carga de Archivo



Nota. Elaboración propia.

Figura 32

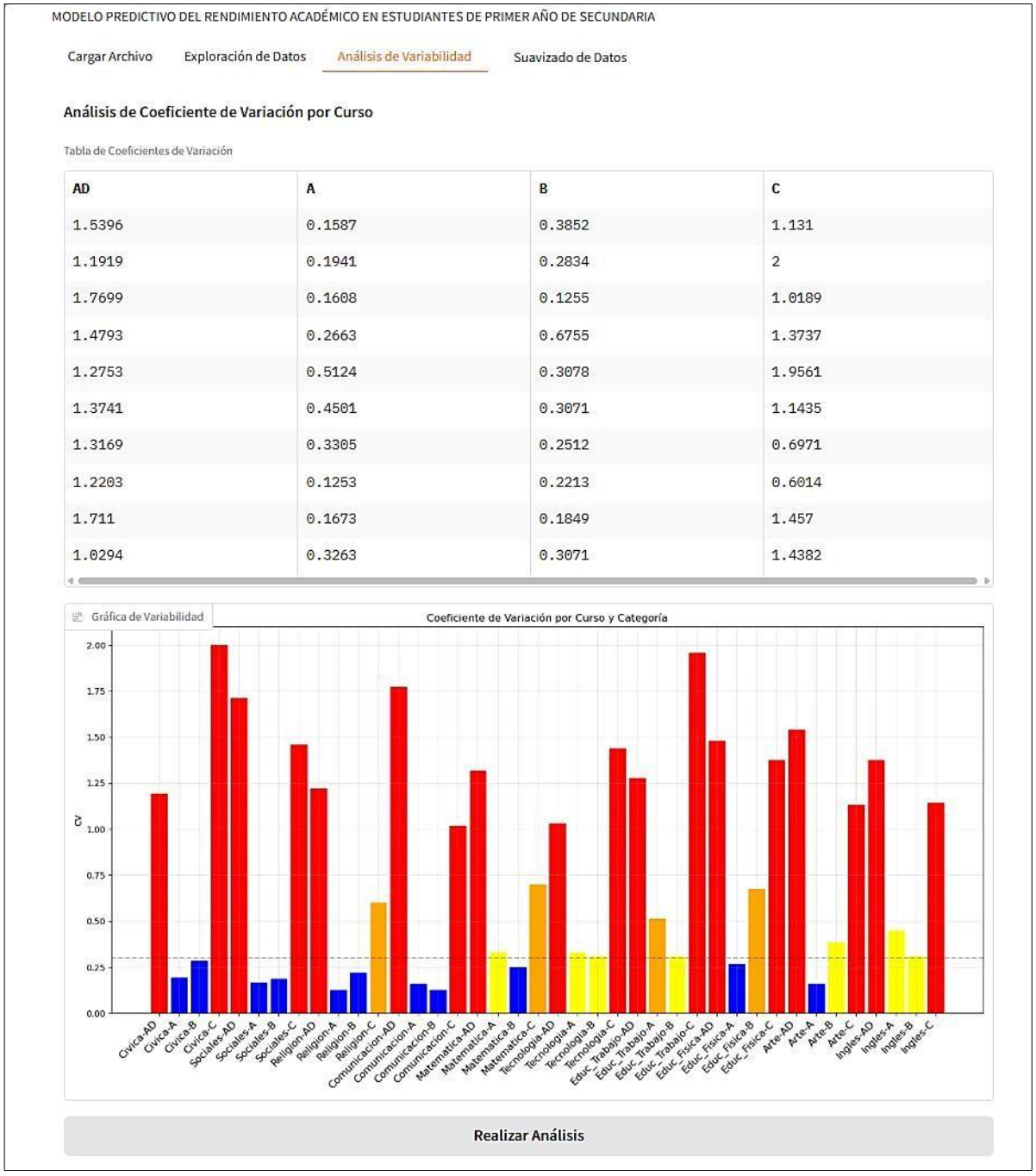
Interfaz gráfica del modelo propuesto: Exploración de datos



Nota. Elaboración propia.

Figura 33

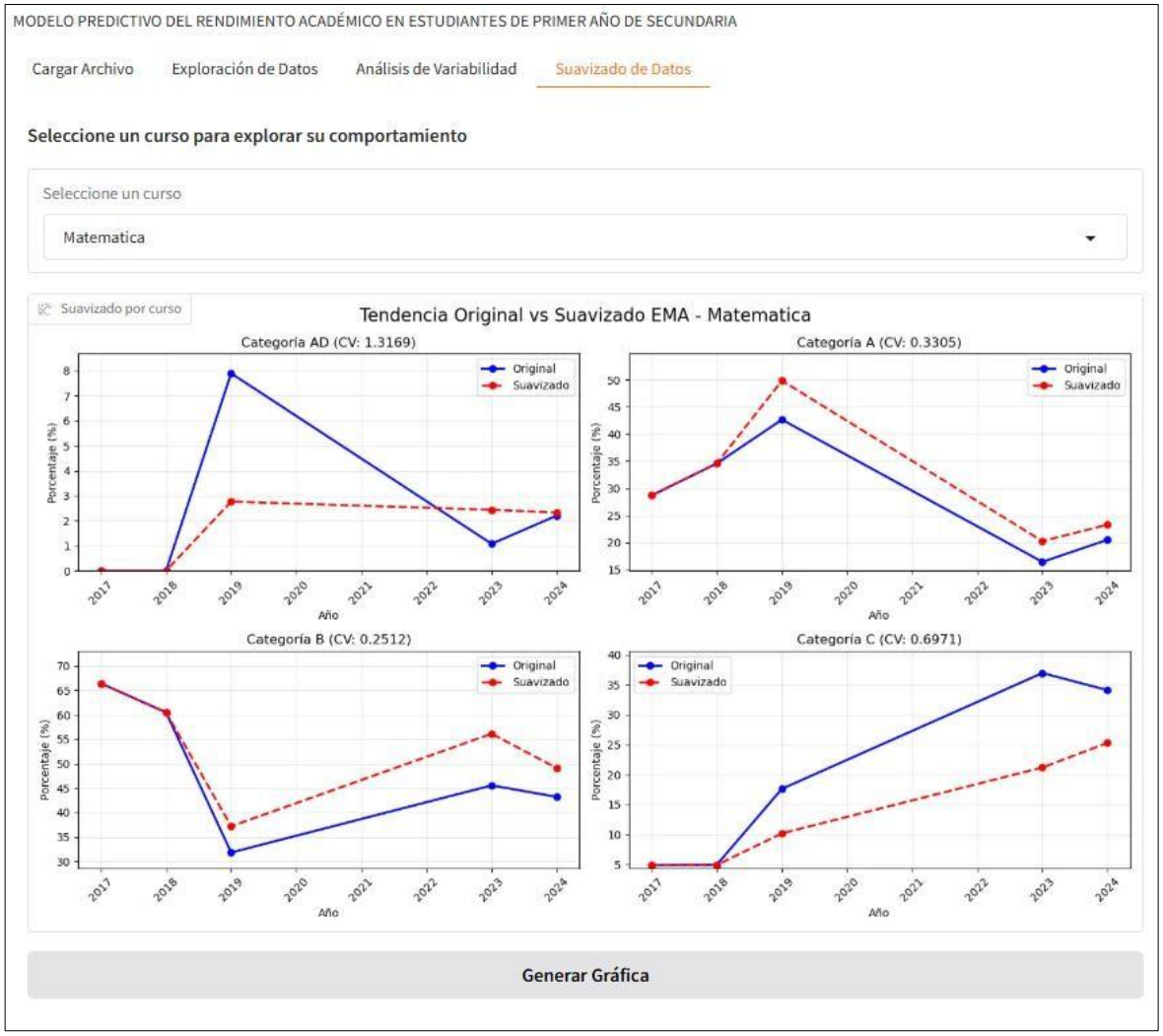
Interfaz gráfica: Análisis de variabilidad



Nota. Elaboración propia.

Figura 34

Interfaz gráfica: Suavizado de Datos



Nota. Elaboración propia.

4.2. Discusión

- La presente investigación se centró en desarrollar modelos de aprendizaje automático para predecir el rendimiento académico en estudiantes de primer año de secundaria.
- Uno de los aspectos importantes para este tipo de investigación es la calidad de los datos utilizados para realizar las predicciones. Fueron necesarias varias trasformaciones y ajustes menores a los datos para las diferentes técnicas aplicadas. La alta variabilidad encontrada en los datos de rendimiento académico representó un desafío significativo, lo que motivó la aplicación de algoritmos para capturar adecuadamente los patrones subyacentes en cada curso y nivel de logro específico.
- Se destacó la necesidad de abordar la alta variabilidad en los datos mediante la aplicación de diversos algoritmos y la validación cruzada temporal para asegurar la generalización de los modelos.
- El algoritmo Random Forest demostró ser el más efectivo, predominando en el 50% de las combinaciones analizadas. Su éxito se observó particularmente en áreas como Arte, Cívica y Religión, lo que sugiere su capacidad para capturar interacciones complejas entre variables en estas disciplinas.
- XGBoost demostró ser el segundo algoritmo más efectivo, representando el 30% de las combinaciones de casos óptimas. Su fortaleza radica en la identificación de interacciones no lineales complejas, lo que lo hizo valioso en asignaturas como Ciencias Sociales, Comunicación, Educación Física e Inglés.
- La Regresión Lineal Múltiple, aunque menos frecuente, resultó óptima en el 20% de los casos, especialmente en Educación para el Trabajo y Tecnología. Esto indica que, en ciertas situaciones, las relaciones lineales entre las variables

predictoras y el rendimiento académico son suficientes para obtener predicciones precisas.

- El análisis de la precisión muestra que **XGBoost** es el algoritmo más eficiente, alcanzando una precisión promedio del **95.99%**, lo que lo convierte en la mejor opción para predecir logros académicos con alta fiabilidad. Le sigue la **Regresión Múltiple** con un **91.43%**, ofreciendo un buen equilibrio entre precisión e interpretabilidad. Finalmente, **Random Forest** obtiene una precisión promedio de **89.96%**, ligeramente menor pero aún adecuada. En conjunto, todos los algoritmos presentan desempeños aceptables, destacando XGBoost como el más preciso.
- En general, este estudio ha logrado demostrar la importancia de un enfoque diferenciado en la aplicación de técnicas de aprendizaje automático para predecir el rendimiento académico. La variabilidad encontrada en el rendimiento óptimo de los algoritmos según asignaturas y niveles de logro sugiere que los factores que influyen en el éxito académico son distintos según la naturaleza de cada materia, lo que abre nuevas perspectivas para capturar la complejidad inherente al proceso de transición de primaria a secundaria y su impacto en el rendimiento académico.

CAPITULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

- Los resultados del estudio evidencian que no existe un único modelo predictivo óptimo para todas las asignaturas. Random Forest demostró ser el algoritmo más efectivo, abarcando el 50% de los casos, destacándose en materias como Arte, Cívica y Religión. XGBoost mostró un desempeño significativo con un 30% de efectividad, especialmente en Ciencias Sociales, Comunicación y Educación Física, mientras que la Regresión Lineal Múltiple tuvo una presencia moderada con un 20% de los casos, siendo particularmente útil en Educación para el Trabajo y Tecnología.
- La validación cruzada temporal desempeñó un papel fundamental en la selección de los modelos más robustos, permitiendo conservar la secuencia cronológica de los datos y garantizando una capacidad predictiva sostenida. Las métricas de rendimiento revelaron que XGBoost logró el mejor MAE promedio (0.040), mientras que la Regresión Múltiple alcanzó la mejor precisión predictiva promedio (62.9%).
- El estudio confirma que los datos educativos presentan una alta variabilidad, reflejo de la naturaleza multifactorial del rendimiento académico. Para abordar esta variabilidad, se implementó el Suavizado Exponencial (EMA), que permitió reducir el ruido en los datos y capturar tendencias más claras en el rendimiento académico. Esta técnica, junto con los algoritmos empleados, permitió adaptarse mejor a las distintas tendencias observadas en las asignaturas, como lo demuestra la distribución de efectividad entre los tres modelos principales.
- Los resultados sugieren que la efectividad de los modelos predictivos está estrechamente relacionada con la naturaleza pedagógica de cada asignatura. Esto refuerza la necesidad de considerar las características específicas de cada materia

al momento de seleccionar y aplicar modelos de predicción del rendimiento académico.

- Los resultados confirman la hipótesis. El modelo **XGBoost** alcanzó una precisión promedio de **95.99%**, seguido por Regresión Múltiple con **91.43%** y Random Forest con **89.96%**. Estas cifras demuestran que las técnicas de aprendizaje automático seleccionadas permiten predecir con **alta precisión** el rendimiento académico de estudiantes de primer año de secundaria.

5.2. RECOMENDACIONES

- Los hallazgos de este estudio ponen en evidencia la necesidad de ampliar la recolección de datos con el objetivo de mejorar la capacidad predictiva de los modelos de aprendizaje automático aplicados al rendimiento académico de estudiantes de primer año de secundaria. La variabilidad en el desempeño de los algoritmos por asignatura sugiere la presencia de factores específicos vinculados a la naturaleza de cada materia, los cuales no fueron completamente capturados en esta investigación. Identificar y analizar estos factores permitiría alcanzar una comprensión más integral de los determinantes del éxito académico en esta etapa clave de transición educativa.
- En consecuencia, se recomienda extender la muestra a instituciones públicas y de diversos contextos socioeconómicos. La inclusión de estudiantes con distintas condiciones de acceso a recursos y entornos de estudio podría revelar patrones diferentes a los observados hasta ahora. Evaluar el desempeño de los algoritmos en escenarios marcados por mayores niveles de desigualdad contribuiría a validar su eficacia predictiva y, en su caso, identificar la necesidad de ajustes específicos según el contexto institucional.
- Asimismo, resulta importante incorporar nuevas variables que potencien la precisión de los modelos utilizados. Aspectos de índole psicológica y socioemocional y económicas pueden tener una influencia significativa en el rendimiento escolar durante la transición a secundaria, pero no fueron considerados en este análisis. Igualmente, sería pertinente incluir información sobre las estrategias de adaptación al nuevo nivel educativo, la participación en programas de tutoría o acompañamiento, así como la existencia de responsabilidades familia.

- También deben considerarse variables como el apoyo familiar en tareas escolares y la detección temprana de dificultades de aprendizaje. Estos factores podrían aportar significativamente a la mejora de los modelos predictivos, particularmente al identificar combinaciones entre asignatura y nivel de logro.
- Otro aspecto relevante es la percepción del estudiante sobre la metodología docente y el nivel de exigencia de los cursos. La satisfacción con el profesorado, la dificultad percibida de las materias y las estrategias pedagógicas empleadas inciden directamente en la motivación y el rendimiento. Evaluar el acceso a recursos tecnológicos y materiales complementarios también resulta fundamental, especialmente en aquellas asignaturas donde modelos como XGBoost y Random Forest mostraron un rendimiento superior.
- La implementación de un sistema integrado de monitoreo, que combine los algoritmos más adecuados para cada asignatura, permitiría diseñar intervenciones pedagógicas diferenciadas y personalizadas, maximizando el potencial de los modelos predictivos para mejorar el rendimiento académico durante la transición a la educación secundaria.
- La inclusión de estas nuevas variables, junto con una muestra más amplia y diversa, no solo incrementaría la precisión de los modelos por asignatura, sino que también permitiría una visión más completa y profunda de los múltiples factores que inciden en el desempeño escolar, facilitando así el diseño de estrategias educativas basadas en evidencia sólida.

REFERENCIAS BIBLIOGRAFICAS

Albon, C. (2018). *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*. O'Reilly Media.

Anderson D., Sweeney D. y Williams T. (2008), "Estadística para Administración y Economía". 10º edición. Ed. Thomson. México.

Cahuana, J., (2021). Factores determinantes asociados al rendimiento académico mediante machine learning en estudiantes de la asignatura de matemática I, UNASAM – 2019 [Universidad Nacional Santiago Antúnez de Mayolo]. http://repositorio.unasam.edu.pe/handle/UNASAM/5054.

Candia, D. (2019). Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático. Repositorio Institucional - UNSAAC: https://repositorio.unsaac.edu.pe/handle/20.500.12918/4120

Caselli, H. (2021). *Modelo predictivo basado en Machine Learning como soporte para el seguimiento académico del estudiante universitario.* Repositorio Institucional Universidad Nacional del Santa: https://repositorio.uns.edu.pe/handle/20.500.14278/3804

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. 785-794.

Espinoza, G., León, E. (2020). Modelo de Machine Learning para la clasificación de estudiantes de acuerdo a su rendimiento académico en el Centro de Idiomas de la Universidad Nacional del Santa [Tesis, Universidad Nacional del Santa]. https://hdl.handle.net/20.500.14278/3588.

Garcia, J., (2021). Machine learning para predecir el rendimiento académico de los estudiantes universitarios [Universidad César Vallejo]. https://hdl.handle.net/20.500.12692/83442

Genuer, R., & Poggi, J. M. (2020). Random Forests with R. Springer International Publishing.

Géron, A. (2022). Hands-on machine learning with scikit-learn, Keras, and TensorFlow. O'Reilly Media.

Gómez Barrantes, M. (2012). Elementos de estadística descriptiva (5ª ed.). Costa Rica: EUNED. ISBN: 9789968482400

Goodfellow, I., Bengio, Y., & Courville, A. (2020). *Deep Learning*. MIT Press. pp. 96-161.

Guamán, S., Mullo, H., Marcatoma, J. (2023) Comparación entre Modelos de Regresión Lineal Múltiple Vs Redes Neuronales Artificiales Supervisadas en la Predicción de Calificaciones Ser Bachiller 2018-2019 del Ecuador. Revista Iberoamericana De educación, 7(2).

Guerra, Jorge. (2022). Fundamentos y variantes de los modelos ARIMA para el análisis de series temporales. Aplicación a la estadística universitaria.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Henríquez Cabezas, N., & Vargas Escobar, D. (2022). Modelos predictivos de rendimiento y deserción académica en estudiantes de primer año de una universidad pública chilena. Revista de Estudios y Experiencias en Educación, 21(45), 299-316. https://doi.org/10.21703/0718-5162.v21.n45.2022.015.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). Metodología de la investigación (6a ed.). McGraw-Hill Education.

Himmel, E. (2018). Modelos de análisis de la deserción estudiantil en la educación superior. Calidad en la Educación, (17), 91-108. https://doi.org/10.31619/caledu.n17.409.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

Kotsiantis, S. B., Pierrakeas, C. J., Zaharakis, I. D., & Pintelas, P. E. (2003). Eficacia de las técnicas de aprendizaje automático en la predicción del rendimiento de los estudiantes en sistemas de aprendizaje a distancia. In Proceedings of the International Conference on Artificial Intelligence in Education (pp. 269-276). IOS Press.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

Lourdes, López-García., Carlos, Lino-Ramírez Guamán Luna, S. V., Mullo Guaminga, H. S., & Marcatoma Tixi, J. A. (2023). Comparison between Multiple Linear Regression models vs supervised Artificial Neural Networks in the prediction of Ecuadorian Ser Bachiller 2018-2019 grades. Revista Iberoamericana de la Educación, 7(2). https://doi.org/10.31876/ie.v7i2.249

Ministerio de Educación. (2020). Resolución Viceministerial N.º 094-2020-MINEDU que aprueba el documento normativo denominado "Disposiciones para el desarrollo del año escolar 2020". Ministerio de Educación del Perú.

Ministerio de Educación. (2024). Resolución Viceministerial N.º 048-2024-MINEDU que regula la evaluación de las competencias de los estudiantes de la Educación Básica ". Ministerio de Educación del Perú.

Mueller, J. & Massaron, L. (2016). Machine Learning for Dummies. John Wiley & Sons.

Oscuvilca Tapia, A. L. (2018). Factores influyentes del bajo rendimiento académico en los estudiantes de secundaria en la institución educativa Túpac Amaru de Chilca.

Posada Hernández, G. J. (2016). *Elementos básicos de estadística descriptiva para el análisis de datos* [Recurso electrónico]. Fundación Universitaria Luis Amigó.

Raschka, S. (2022). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 3, 4th Edition. Packt Publishing Ltd.

Roback, P., & Legler, J. (2020). *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R*. Routledge

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. https://doi.org/10.1037/BUL0000098

Theobald, O. (2017). Machine Learning for Absolute Beginners: A Plain English Introduction (2nd ed.). Independently published.

Touron, J. (1985). La predicción del rendimiento académico: procedimientos, resultados e implicaciones. Revista Española de Pedagogía, no 169-170, pp. 473-495. https://dadun.unav.edu/handle/10171/18774

Vargas, I. M., & Vergara, N. B. C. (2012) Análisis del Problema y propuestas de alternativas de la Educación Peruana.

Víctor, Manuel, Zamudio-Rodríguez., Josué, Del, Valle-, Hernández. (2022). Predictive model for the analysis of academic performance and preventing student dropout using machine learning techniques. Revista de educación técnica, 1-5. doi: 10.35429/jote.2022.16.6.1.5

Vega, J. (2019). Modelo de pronóstico de rendimiento académico de alumnos en los cursos del programa de estudios básicos de la Universidad Ricardo Palma usando algoritmos de Machine Learning. Repositorio Institucional Universidad Ricardo Palma: https://repositorio.urp.edu.pe/handle/20.500.14138/2914

Wu, X., Gao, Y. y Jiao, D. (2019). Clasificación multietiqueta basada en un algoritmo de bosque aleatorio para un sistema de monitorización de carga no intrusivo. *Processes* , *7* (6), 337. https://doi.org/10.3390/pr7060337

ANEXOS

Anexo A: *Formato oficial de registros de calificaciones*

Datos de la Instancia de Gestión Educativa Descentralizada (UGEL) (1)		Datos de la Institución Educativa o Programa Educativo						Período Lectivo (8)	Inicio	15/03/2021	Fin	17/12/2021
		Número y/o Nombre										
Código	0 2 0 0 1 8	Código Modular - Anexo										
Nombre de UGEL	UGEL Santa	Resolución de Creación N°										
		Modalidad (3)	EBR	Grado (5)	1	Turno (7)	M					
		Gestión (4)	P	Sección(6)	F							

N° de Orden	D.N.I. / Código del Estudiante (2)	Apellidos y Nombres (Orden Alfabético)	Sexo H/M	ÁREAS															N° Áreas y Talleres que no alcanzan el calificativo mínimo exigido (10)	Situación final (11)	Motivo de Retiro (12)
				DESARROLLO PERSONAL CIUDADANÍA Y CÍVICA (A)	CIENCIAS SOCIALES (B)			EDUCACIÓN PARA EL TRABAJO (C)		EDUCACIÓN FÍSICA (D)			COMUNICACIÓN (E)			ARTE Y CULTURA (F)					
				Construye su identidad	Convive y participa democráticamente en la búsqueda del bien común	Construye interpretaciones históricas	Gestiona responsablemente el espacio y el ambiente	Gestiona responsablemente los recursos económicos	Gestiona proyectos de emprendimiento económico o social	Especialidad(14)	Se desenvuelve de manera autónoma a través de su motricidad	Asume una vida saludable	Interactúa a través de sus habilidades sociomotrices	Se comunica oralmente en su lengua materna	Lee diversos tipos de textos escritos en su lengua materna	Escribe diversos tipos de textos en su lengua materna	Aprecia de manera crítica manifestaciones artístico-culturales	Crea proyectos desde los lenguajes artísticos			
1	D N I		M	B	A	B	A	A	B	A	AD	A	A	A	B	B	A	A	0	PRO	
2	D N I		M	A	AD	A	A	B	A	B	AD	A	A	A	A	A	A	A	0	PRO	
3	D N I		H	A	B	B	A	A	B	A	A	A	B	A	A	A	A	A	0	PRO	
4	D N I		M	A	B	B	B	B	B	B	A	A	B	B	A	A	A	A	0	PG	
5	D N I		H	B	B	B	B	AD	B	B	A	A	A	B	B	B	A	A	0	PRO	
6	D N I		H	A	AD	AD	A	A	A	01	A	A	A	A	A	A	A	A	0	PRO	
7	D N I		M	A	AD	AD	A	B	A	01	AD	A	A	A	A	A	A	A	0	PRO	

Anexo B: Script de transformación de la Data Original

```
#  CARGA Y CLASIFICACION DE LA DATA ORIGINAL
# Crear un DataFrame para almacenar los resultados
results = []

# Procesar los datos por año y curso
for year in data['Anio'].unique():
  for course in courses:
    # Filtrar los datos por año y curso
    course_data = data[data['Anio'] == year][course]

    # Calcular los porcentajes de cada categoría
    total = len(course_data)
    a = len(course_data[(course_data >= 7.5) & (course_data < 10)]) / total
    ad = len(course_data[course_data == 10]) / total
    b = len(course_data[(course_data >= 6.25) & (course_data < 7.5)]) / total
    c = len(course_data[(course_data >= 2.5) & (course_data < 6.25)]) / total

    # Agregar los resultados al DataFrame
    results.append({'Anio': year, 'Curso': course, 'AD': ad, 'A': a, 'B': b, 'C': c})

# Convertir los resultados a un DataFrame
results_df = pd.DataFrame(results)
# Ordenar el DataFrame por año
results_df = results_df.sort_values(by='Anio')
results_df.to_csv('cursocategoria.csv', index=False)
results_df.to_excel('cursocategoria.xlsx', index=False)

# Mostrar los resultados
results_df
```

Anexo C: Script del análisis de rendimiento por curso y logro

```
# ANALISIS DEL RENDIMIENTO POR CADA CURSO Y LOGRO
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# Cargar los datos
df = pd.read_csv('cursocategoria.csv')

# Multiplicar las columnas de categorías por 100
categorias = ['AD', 'A', 'B', 'C']
for cat in categorias:
  df[cat] = df[cat] * 100

# Obtener la lista de cursos únicos
cursos = df['Curso'].unique()

# Crear un gráfico para cada curso
for curso in cursos:
  # Filtrar datos para el curso actual
  curso_df = df[df['Curso'] == curso]

  # Crear el gráfico
  plt.figure(figsize=(14, 8))

  # Definir colores similares a la imagen de referencia
  colors = {
    'AD': '#1f77b4', # Azul
    'A': '#000000', # Negro
    'B': '#f44611', # Naranja
    'C': '#d62728'  # Rojo
  }

  # Graficar cada categoría
  for category in ['AD', 'A', 'B', 'C']:
    line, = plt.plot(curso_df['Anio'], curso_df[category], 'o-',
         color=colors[category],
         linewidth=2.5,
         markersize=8,
         label=f'Logro {category}')

    # Añadir etiqueta al final de la línea
    last_year = curso_df['Anio'].iloc[-1]
    last_value = curso_df[curso_df['Anio'] == last_year][category].values[0]

    # Colocar la etiqueta de categoría encima de la línea al final
    plt.annotate(f'Logro {category}',
          xy=(last_year, last_value),
          xytext=(last_year + 0.2, last_value + 2),
          color=colors[category],
          fontweight='bold')
```

```
        # Añadir el porcentaje real en cada punto
        for year in curso_df['Anio']:
            value = curso_df[curso_df['Anio'] == year][category].values[0]
            # Formatear el valor para mostrar solo un decimal si es necesario
            if value == int(value):
                value_str = f"{int(value)}%"
            else:
                value_str = f"{value:.1f}%"

            plt.annotate(value_str,
                         xy=(year, value),
                         xytext=(year, value + 2),  # Colocar ligeramente arriba
                         ha='center',
                         va='bottom',
                         color=colors[category],
                         fontsize=10)

    # Añadir área sombreada para pandemia 2020-2022
    plt.axvspan(2020, 2022, alpha=0.1, color='gray')

    # Configurar el gráfico
    plt.grid(True)
    plt.xlabel('Año', fontsize=12)
    plt.ylabel('Porcentaje (%)', fontsize=12)
    plt.xticks(curso_df['Anio'])
    plt.ylim(0, 100)  # Ajustado a 100% para mostrar todo el rango
    plt.title(f'Rendimiento en {curso} (2017-2024)', fontsize=14)

    # No añadir leyenda ya que tenemos etiquetas en las líneas

    # Ajustar diseño
    plt.tight_layout()

    # Guardar
    plt.savefig(f'rendimiento_{curso}_porcentajes.png', dpi=300, bbox_inches='tight')

# Crear un archivo combinado con todos los gráficos
plt.figure(figsize=(20, 30))

for i, curso in enumerate(cursos):
    # Filtrar datos para el curso actual
    curso_df = df[df['Curso'] == curso]

    # Crear subplot
    plt.subplot(5, 2, i+1)

    # Graficar cada categoría
    for category in ['AD', 'A', 'B', 'C']:
        line, = plt.plot(curso_df['Anio'], curso_df[category], 'o-',
                 color=colors[category],
                 linewidth=2.5,
```

```
                markersize=8,
                label=f'Categoría {category}')

        # Añadir etiqueta al final de la línea solo para el gráfico combinado
        if i < 10: # Solo para los primeros 10 gráficos
            last_year = curso_df['Anio'].iloc[-1]
            last_value = curso_df[curso_df['Anio'] == last_year][category].values[0]

            # Usar etiquetas más cortas para el gráfico combinado
            plt.annotate(f'{category}',
                         xy=(last_year, last_value),
                         xytext=(last_year + 0.05, last_value + 1),
                         color=colors[category],
                         fontweight='bold',
                         fontsize=8)

    # Añadir área sombreada para 2019-2023
    plt.axvspan(2019, 2023, alpha=0.1, color='gray')

    # Configurar el gráfico
    plt.grid(True)
    plt.xlabel('Año', fontsize=10)
    plt.ylabel('Porcentaje (%)', fontsize=10)
    plt.xticks(curso_df['Anio'], fontsize=8)
    plt.ylim(0, 100)
    plt.title(f'Rendimiento en {curso}', fontsize=12)

plt.tight_layout()
plt.savefig('rendimiento_todos_cursos_etiquetado.png', dpi=300, bbox_inches='tight')

print("Gráficos generados con etiquetas y porcentajes para cada curso")
```

Anexo D: Script del análisis de variabilidad

```
# ANALISIS DE VARIABILIDAD
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

def calcular_cv(data):
    resultados = []

    # Para mantener el orden original para la tabla
    for curso in data['Curso'].unique():
        for categoria in ['AD', 'A', 'B', 'C']:
            mask = data['Curso'] == curso
            valores = data.loc[mask, categoria].values

            if len(valores) > 0 and np.mean(valores) != 0:
                cv = np.std(valores) / np.mean(valores)
                resultados.append({
                    'Curso': curso,
                    'Categoria': categoria,
                    'CV': cv
                })

    # DataFrame para la tabla
    df_tabla = pd.DataFrame(resultados)

    # Crear tabla pivote para el formato deseado
    tabla_formato = df_tabla.pivot(index='Curso', columns='Categoria', values='CV')
    tabla_formato = tabla_formato[['AD', 'A', 'B', 'C']]  # Reordenar columnas
    tabla_formato = tabla_formato.round(4)  # Redondear a 4 decimales

    # DataFrame para el gráfico (mantener orden original)
    df_grafico = df_tabla.copy()

    return tabla_formato, df_grafico

# Cargar datos
data = pd.read_csv('cursocatfinal.csv')

# Calcular CV
tabla_formato, df_grafico = calcular_cv(data)

# Mostrar resultados en formato de tabla
print("\nCoeficiente de Variación por Curso y Categoría:")
print("="*80)
print(tabla_formato.to_string())

# Exportar a Excel
tabla_formato.to_excel('coeficiente_variacion.xlsx')

# Crear gráfico de barras con el orden original
```

```
plt.figure(figsize=(15, 8))
bars = plt.bar(range(len(df_grafico)), df_grafico['CV'])

# Colorear barras según nivel de variabilidad
for i, bar in enumerate(bars):
  cv = df_grafico['CV'].iloc[i]
  if cv < 0.3:
    bar.set_color('blue')
  elif cv < 0.5:
    bar.set_color('yellow')
  elif cv < 1:
    bar.set_color('orange')
  else:
    bar.set_color('red')

plt.xticks(range(len(df_grafico)),
      [f"{row['Curso']}-{row['Categoria']}" for _, row in df_grafico.iterrows()],
      rotation=45, ha='right', fontsize=12)
plt.axhline(y=0.3, color='black', linestyle='--', alpha=0.3)
plt.title('Coeficiente de Variación por Curso y Categoría')
plt.ylabel('CV')
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()
```

Anexo E: Script del suavizado de datos con EMA

```
# SUAVIZADO
import pandas as pd
import numpy as np

def aplicar_ema_normalizado(data, curso, alpha=0.3):
  # Obtener datos del curso
  curso_data = data[data['Curso'] == curso].copy()
  categorias = ['AD', 'A', 'B', 'C']

  # Aplicar EMA a cada categoría que lo necesite
  for categoria in categorias:
    valores = curso_data[categoria].values
    cv = np.std(valores) / np.mean(valores) if np.mean(valores) != 0 else 0

    if cv >= 0.5:
      # Aplicar EMA
      ema = [valores[0]]
      for n in range(1, len(valores)):
        ema.append(alpha * valores[n] + (1-alpha) * ema[n-1])
      curso_data[categoria] = ema

  # Normalizar filas para que sumen 1
  for idx in curso_data.index:
    row_sum = curso_data.loc[idx, categorias].sum()
    if row_sum != 0:
      curso_data.loc[idx, categorias] = curso_data.loc[idx, categorias] / row_sum

  return curso_data

# Cargar datos originales
data_original = pd.read_csv('cursocatfinal.csv')
data_suavizada = pd.DataFrame()

# Procesar cada curso
for curso in data_original['Curso'].unique():
  curso_suavizado = aplicar_ema_normalizado(data_original, curso)
  data_suavizada = pd.concat([data_suavizada, curso_suavizado])

# Ordenar por Año y Curso como en el original
data_suavizada = data_suavizada.sort_values(['Anio', 'Curso']).reset_index(drop=True)

# Guardar archivo suavizado
data_suavizada.to_csv('cursologrosuavizado.csv', index=False)
```

```
# VALIDACION CRUZADA TEMPORAL:  MAE MSE RMSE PRECISION
import pandas as pd
from sklearn.model_selection import TimeSeriesSplit
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error
import numpy as np
import warnings

# Ignorar todos los warnings
warnings.filterwarnings("ignore")

# Cargar el archivo CSV
data  =  pd.read_csv('cursologrosuavizado.csv')

# Configurar la validación cruzada temporal
ts_split = TimeSeriesSplit(n_splits=3)

# Inicializar los modelos
models  = {
   'Regresión Múltiple': LinearRegression(),
   'Random Forest': RandomForestRegressor(n_estimators=100, random_state=42),
   'XGBoost': XGBRegressor(n_estimators=100, random_state=42),
}

# Crear un DataFrame para almacenar los resultados finales
final_results = []

# Iterar por cada curso y categoría
for curso in data['Curso'].unique():
   for categoria in ['A', 'B', 'C', 'AD']:
      # Filtrar los datos por curso y categoría
      subset = data[data['Curso'] == curso]
      if categoria not in subset.columns:
         continue

      X = subset.index.values.reshape(-1, 1)  # Índices como características
      y = subset[categoria]

      # Verificar si hay suficientes datos para la validación cruzada
      if len(y) < 4:  # Se necesitan al menos 4 datos para 3 splits
         continue

      # Validación cruzada temporal para los modelos basados en machine learning
      for model_name, model in models.items():
         partial_results = []
         for train_index, test_index in ts_split.split(X):
            X_train, X_test = X[train_index], X[test_index]
            y_train,  y_test  =  y.iloc[train_index],  y.iloc[test_index]
```

```
            # Entrenar el modelo
            model.fit(X_train, y_train)
            # Predecir
            y_pred = model.predict(X_test)
            # Calcular métricas
            mae = mean_absolute_error(y_test, y_pred)
            mse = mean_squared_error(y_test, y_pred)
            rmse = np.sqrt(mse)
            # Guardar resultados parciales (sin calcular precisión aquí)
            partial_results.append({'MAE': mae, 'MSE': mse, 'RMSE': rmse})

        # Promediar los resultados de error
        avg_results = {metric: np.mean([fold[metric] for fold in partial_results])
                for metric in ['MAE', 'MSE', 'RMSE']}

        # Calcular precisión una sola vez con el RMSE promedio
        precision = 100 * (1 - avg_results['RMSE'])

        # Guardar los resultados finales
        final_results.append({
            'Curso': curso,
            'Categoria': categoria,
            'Algoritmo': model_name,
            'MAE': avg_results['MAE'],
            'MSE': avg_results['MSE'],
            'RMSE': avg_results['RMSE'],
            'Precisión': precision
        })

# Convertir los resultados finales a un DataFrame
final_results_df = pd.DataFrame(final_results)

# Determinar el mejor modelo por curso y categoría basado en RMSE
best_models = final_results_df.loc[final_results_df.groupby(['Curso',
'Categoria'])['RMSE'].idxmin()]
best_models = best_models.sort_values(['Curso', 'Categoria'])

# Exportar los resultados finales y los mejores modelos a archivos CSV y Excel
final_results_df.to_csv('comparacion_algoritmos.csv', index=False)
final_results_df.to_excel('comparacion_algoritmos.xlsx', index=False)
best_models.to_csv('mejores_modelos.csv', index=False)
best_models.to_excel('mejores_modelos.xlsx', index=False)

# Imprimir resultados
print("\nResumen de todos los resultados:")
print("=" * 100)
print(final_results_df)

print("\nMejores Modelos por Curso y Categoría:")
print("=" * 100)
for _, row in best_models.iterrows():
```

```
    print(f"Curso: {row['Curso']:<15} | Categoría: {row['Categoria']:<2} | "
          f"Mejor Modelo: {row['Algoritmo']:<20} | MAE: {row['MAE']:.4f} | "
          f"RMSE: {row['RMSE']:.4f} | Precisión: {row['Precisión']:.2f}%")

# Mostrar resumen general por algoritmo
print("\nResumen General de Rendimiento por Algoritmo:")
print("=" * 60)
summary = final_results_df.groupby('Algoritmo')[['MAE', 'MSE', 'RMSE', 'Precisión']].mean()
print(summary.round(4))
```

```
# PREDICCION FINAL 2025
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.linear_model import LinearRegression
import warnings

# Ignorar warnings
warnings.filterwarnings("ignore")

# Cargar los datos
data = pd.read_csv('cursologrosuavizado1.csv')
mejores_modelos = pd.read_excel('mejores_modelos.xlsx')

# Función para hacer predicciones
def predict_2025(curso, categoria, algoritmo, historical_data):
  years = np.array([2017, 2018, 2019, 2023, 2024]).reshape(-1, 1)
  y = historical_data.values


  if algoritmo == 'Random Forest':
    model = RandomForestRegressor(n_estimators=100, random_state=42)
    model.fit(years, y)
    return max(0, model.predict([[2025]])[0])

  elif algoritmo == 'XGBoost':
    model = XGBRegressor(n_estimators=100, random_state=42)
    model.fit(years, y)
    return max(0, model.predict([[2025]])[0])

  elif algoritmo == 'Regresión Múltiple':
    model = LinearRegression()
    model.fit(years, y)
    return max(0, model.predict([[2025]])[0])

  return 0

# Crear DataFrame para almacenar resultados
predictions = []

# Para cada combinación en mejores_modelos
for _, row in mejores_modelos.iterrows():
  curso = row['Curso']
  categoria = row['Categoria']
  algoritmo = row['Algoritmo']

  # Obtener datos históricos
  historical_values = {}
  for year in [2017, 2018, 2019, 2023, 2024]:
```

```
        year_data = data[data['Anio'] == year]
        if not year_data.empty:
            value = year_data[year_data['Curso'] == curso][categoria].iloc[0]
            historical_values[str(year)] = value
        else:
            historical_values[str(year)] = 0

    # Hacer predicción para 2025
    historical_series = pd.Series(list(historical_values.values()))
    pred_2025 = predict_2025(curso, categoria, algoritmo, historical_series)

    # Almacenar resultados
    predictions.append({
        'Curso': curso,
        'Categoria': categoria,
        'Modelo': algoritmo,
        '2017': historical_values['2017'],
        '2018': historical_values['2018'],
        '2019': historical_values['2019'],
        '2023': historical_values['2023'],
        '2024': historical_values['2024'],
        '2025': round(pred_2025, 4)
    })

# Convertir a DataFrame y ordenar
predictions_df = pd.DataFrame(predictions)
predictions_df = predictions_df.sort_values(['Curso', 'Categoria'])

# Redondear todos los valores numéricos a 4 decimales
for col in ['2017', '2018', '2019', '2023', '2024', '2025']:
    predictions_df[col] = predictions_df[col].round(4)

# Mostrar resultados
print("\nPredicciones históricas y 2025 por Modelo:")
print(predictions_df)

# Guardar resultados
predictions_df.to_csv('predicciones_historicas_y_2025.csv', index=False)
predictions_df.to_excel('predicciones_historicas_y_2025.xlsx', index=False)

predictions_df
```