UNIVERSIDAD NACIONAL DE SAN MARTIN

RCSI
https://revistas.unsm.edu.pe/index.php/rcsi
e-ISSN: 2709-992X

Artículo Original
Original Article
Ene-Jun, 2026

Análisis y modelado predictivo del rendimiento académico mediante técnicas de aprendizaje automático en una institución de educación secundaria

Predictive analysis and modeling of academic performance using machine learning techniques in a secondary education institution

Zalasar, Alejandro Miguel[1]
Aramayo, Ramón[2]
Martínez, Cristian Alejandro[1*]

[1]Universidad Nacional de Salta, Salta - Capital, Argentina
[2]Escuela de Educación Técnica 3100 República de la India, Salta - Capital, Argentina

Recibido: 10 Oct. 2025 | **Aceptado:** 25 Dic. 2025 | **Publicado:** 20 Ene. 2026

Autor de correspondencia*: cmartinez@di.unsa.edu.ar

Cómo citar este artículo: Zalasar, A. M., Aramayo, R. & Martínez, C. A. (2026). Análisis y modelado predictivo del rendimiento académico mediante técnicas de aprendizaje automático en una institución de educación secundaria. *Revista Científica de Sistemas e Informática*, *6*(1), e1212. https://doi.org/10.51252/rcsi.v6i1.1212

RESUMEN

El rendimiento académico es un indicador clave para evaluar la calidad educativa y detectar áreas de mejora en los procesos de enseñanza y aprendizaje. En este trabajo se analizó un conjunto de datos de estudiantes de primer año del ciclo básico de una institución secundaria de la provincia de Salta, Argentina, con el objetivo de identificar variables que influyen en el desempeño estudiantil y apoyar la toma de decisiones orientadas a reducir el bajo rendimiento académico. Siguiendo la metodología CRISP-DM, se realizó un análisis exploratorio para identificar patrones relevantes en las calificaciones, se aplicaron modelos de aprendizaje no supervisado para detectar perfiles de estudiantes y, finalmente, modelos supervisados para predecir la aprobación del año a partir de las calificaciones del segundo trimestre. El mejor modelo alcanzó un *F1-Score* de 0,80 en la clase minoritaria y un *accuracy* del 89%. Los resultados permiten anticipar situaciones de riesgo académico y segmentar perfiles estudiantiles, aportando información útil para intervenciones pedagógicas más efectivas.

Keywords: aprendizaje automático; ciencia de datos; educación secundaria; python

ABSTRACT

Academic performance is a key indicator for evaluating educational quality and identifying areas for improvement in teaching and learning processes. This study analyzes a dataset of first-year lower secondary students from an educational institution in the province of Salta, Argentina, with the aim of identifying variables that influence student performance and supporting decision-making to mitigate low academic achievement. Following the CRISP-DM methodology, an exploratory analysis was conducted to identify relevant patterns in grades, unsupervised learning models were applied to detect student profiles, and supervised models were used to predict year completion based on second-term grades. The best-performing model achieved an *F1-score* of 0.80 for the minority class and an overall *accuracy* of 89%. The results enable early identification of students at academic risk and the segmentation of student profiles, providing valuable insights for more effective pedagogical interventions.

Palabras clave: machine learning; data science; secondary education; python

1. INTRODUCCIÓN

El análisis del rendimiento académico es actualmente uno de los temas fundamentales y de mayor preocupación que deben abordar las instituciones educativas. A lo largo del tiempo se ha estudiado en base a dos aspectos: los datos relacionados con la escuela como sistema educativo y en base a las características que los alumnos presentan a partir de su contexto social. Sin embargo, no se ha logrado identificar y comprender completamente cuáles son las variables que influyen en el rendimiento académico.

En la última década, la investigación sobre el rendimiento académico y el abandono de los estudios ha crecido significativamente en Argentina. La mayoría de estos estudios se centraron en el nivel superior, sin realizar hasta el presente un análisis integrador que permita sacar conclusiones generales sobre el estado del conocimiento en el campo. Gran parte de la producción académica sobre este problema ha abordado indirectamente el tema del rendimiento académico, estudiando las variables asociadas tanto al abandono como al desempeño de los estudiantes, pero sin ofrecer una visión global que integre ambos aspectos (García, 2014).

No obstante, gran parte de estas investigaciones se han centrado en el nivel superior, quedando menos explorado el nivel medio. Resulta entonces necesario entender también las dinámicas propias de la educación secundaria, especialmente en contextos regionales como Salta, donde las condiciones socioeconómicas y escolares presentan grandes desafíos.

En este contexto, la Minería de Datos (Ibarra, 2020) emerge como una herramienta prometedora para analizar el rendimiento académico con el propósito de resolver diferentes tipos de problemas como por ejemplo, el rendimiento académico de los alumnos, la deserción y el desgranamiento. En estos últimos años, la Minería de Datos Educativa (EDM) demostró un creciente impacto en el análisis del rendimiento académico y la identificación de estudiantes en riesgo, de acuerdo con una revisión reciente (Romero & Ventura, 2020). EDM se ha convertido en una herramienta útil para descubrir patrones relevantes a partir de datos educativos.

Un ejemplo de ello, es el estudio realizado en la Universidad Técnica de Manabí (Saltos-Mero & Cruz-Felipe, 2024), donde se aplicó la metodología CRISP-DM para analizar el rendimiento académico de estudiantes de las carreras de “Gastronomía y Turismo” y “Economía”, utilizando métodos de aprendizaje supervisado como Árboles de Decisión, Random Forests, Redes Neuronales y Máquinas de Soporte Vectorial (SVM). Los modelos fueron implementados en Python y, tras un proceso de evaluación comparativa, se concluyó que el modelo basado en Random Forests ofreció el mejor desempeño, alcanzando valores de *Accuracy* de 83% y 86% respectivamente.

Asimismo, Guanin-Fajardo et al. (2024) aplicaron la metodología CRISP-DM para predecir el rendimiento académico de estudiantes universitarios, utilizando un conjunto de datos de 6690 registros con variables académicas y socioeconómicas. Entre los métodos evaluados, XGBoost fue el que obtuvo mejores resultados, alcanzando un valor de *AUC* de 87,75%, lo que evidencia su alta capacidad predictiva. Además, el modelo permitió extraer reglas interpretables a partir de árboles de decisión, facilitando su aplicación práctica. El estudio resalta la importancia de implementar modelos predictivos tempranos para fortalecer las estrategias de retención estudiantil. La metodología es replicable en otros contextos académicos y demuestra el valor de combinar precisión con explicabilidad.

Por último, Bellaj et al. (2024) han desarrollado modelos de aprendizaje supervisado para predecir el rendimiento académico basados en las técnicas SVM, Random Forests, XGBoost, K-NN y Naïve Bayes. El mejor modelo fue XGBoost, seguido por un clasificador de votación (EVC). Los autores destacan la importancia de optimizar hiperparámetros para mejorar la precisión de los modelos predictivos. Además, remarcan que variables como el rendimiento previo, la interacción en plataformas virtuales y factores sociodemográficos influyen notablemente en las predicciones. El trabajo contribuye al desarrollo de sistemas de alerta temprana en educación superior.

El propósito de este estudio es identificar tendencias y patrones en el rendimiento académico a través de un Análisis Exploratorio de Datos (EDA) (Tukey, 1977) así como predecir el desempeño estudiantil mediante modelos de aprendizaje supervisado basados en técnicas como Random Forests (Breiman, 2001), XGBoost (Chen & Guestrin, 2016) y Extreme Learning Machine (Huang et al., 2006; Wang et al., 2022). Además, mediante modelos de aprendizaje no supervisado basados en K-Means (MacQueen, 1967) y BIRCH (Zhang et al., 1996) se pudo identificar grupos de estudiantes con características similares, facilitando medidas correctivas y la segmentación para intervenciones oportunas.

2. MATERIALES Y MÉTODOS

2.1. Caso de estudio

El presente estudio se llevó a cabo en la Escuela de Educación Técnica N° 3100 "República de la India", ubicada en la provincia de Salta, Argentina. La investigación fue tipo descriptiva y correlacional, con un diseño no experimental.

2.2. Dataset

Se trabajó con un conjunto de datos correspondientes a la población total de estudiantes del primer año del ciclo básico de la institución. El conjunto incluyó 787 registros de estudiantes, con 13 materias (Artística, Cs. Biológicas, Lengua I, Historia I, Tecnología I, Dibujo Técnico, Fisicoquímica, Lengua extranjera I, F.E. y C. I, Geografía I, Matemática I, Taller Prep. I, Ed. Física). Los registros abarcan los ciclos lectivos 2017, 2018, 2019, 2022 y 2023.

En la Tabla 1 se presentan los atributos relacionados con los registros académicos de la institución. Estos fueron considerados para analizar el rendimiento académico y detectar patrones en las calificaciones. Los datos se obtuvieron de fuentes oficiales de la institución, los cuales constituyeron el instrumento primario de recolección. Cada registro estuvo asociado a un estudiante mediante nombre, apellido e identificador único.

Tabla 1. Atributos del Dataset

Atributo	Tipo	Descripción
Curso	Categórica	Año de cursado y división del estudiante (ej. 1°1°, 1°2°, etc.)
Turno	Categórica	TM (Turno Mañana) o TT (Turno Tarde)
Sexo	Categórica	M (Masculino) o F (Femenino)
Año	Numérica	Año académico (2017–2019, 2022-2023)
Calificaciones por materia	Numérica (Entera)	Notas trimestrales obtenidas en 13 materias. Rango (1-10).
Estado Final	Categórica	Indica si el estudiante fue promovido o no al año siguiente (valores posibles: Promovido y Repite)

Cada materia se encuentra desagregada en tres columnas distintas, correspondientes a los trimestres del ciclo lectivo, e identificadas mediante los sufijos "_1t", "_2t" y "_3t" (por ejemplo: Matemática_1t, Matemática_2t, etc.).

En el *Anexo A1* se describe el conjunto de datos referido a los estudiantes de segundo año de ciclo básico.

2.3. Metodología de Trabajo

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) fue adoptada como marco metodológico para el desarrollo del presente estudio (Chapman, 2000), debido a su carácter estructurado y flexible para proyectos de Minería de Datos. Este enfoque propone un proceso cíclico compuesto por distintas fases que permiten transformar los objetivos del dominio de aplicación en modelos analíticos con valor práctico para la toma de decisiones.

En primer lugar, la fase de comprensión del negocio permitió definir el problema de estudio y establecer como objetivo principal el análisis del rendimiento académico de los estudiantes, con el fin de identificar patrones relevantes en sus calificaciones. A partir de estos objetivos, se avanzó hacia la comprensión de los datos, etapa en la cual se recopilaron los registros disponibles y se realizó una exploración inicial para describir sus características generales, evaluar su calidad y detectar posibles inconsistencias, valores faltantes o atípicos.

Posteriormente, en la fase de preparación de los datos, se construyó el conjunto final de información a utilizar en el análisis, mediante la selección de atributos relevantes y la aplicación de procesos de limpieza y transformación, asegurando su adecuación para el modelado. En la etapa de modelado se aplicaron técnicas analíticas orientadas tanto a la descripción como a la predicción del rendimiento académico, seleccionando y ajustando los algoritmos más apropiados de acuerdo con los objetivos planteados.

Finalmente, la fase de evaluación permitió analizar el desempeño y la utilidad de los modelos obtenidos, verificando su coherencia con los objetivos definidos y su aporte como soporte para la toma de decisiones en el ámbito educativo. De este modo, CRISP-DM proporcionó un marco metodológico integral que guió de manera sistemática el desarrollo del estudio y la generación de conocimiento relevante.

Las tareas de procesamiento, visualización y análisis de datos como el desarrollo de modelos de aprendizaje automático se desarrollaron mediante scripts en Python usando librerías como Pandas, NumPy, Scikit-Learn y Matplotlib. Este enfoque permitió estructurar el estudio de manera secuencial y guiada por CRISP-DM.

2.4. Análisis Exploratorio de Datos

Este estudio se apoyó en un conjunto de técnicas fundamentales para el análisis de datos. En primer lugar, se realizaron tareas de limpieza y transformación de los datos, con el fin de preparar la información para su posterior análisis, garantizando su consistencia y adecuación al contexto de estudio.

Posteriormente, el Análisis Exploratorio de Datos (EDA) proporcionó una primera aproximación a la estructura y calidad de los registros, permitiendo la detección de irregularidades y la comprensión general del comportamiento de las variables del dataset.

El EDA constituye una etapa transversal previa a la implementación de los pipelines de aprendizaje no supervisado y supervisado, y tiene como objetivo comprender la estructura de los datos, detectar valores atípicos y orientar las decisiones de preprocesamiento y modelado.

2.5. Aprendizaje No Supervisado

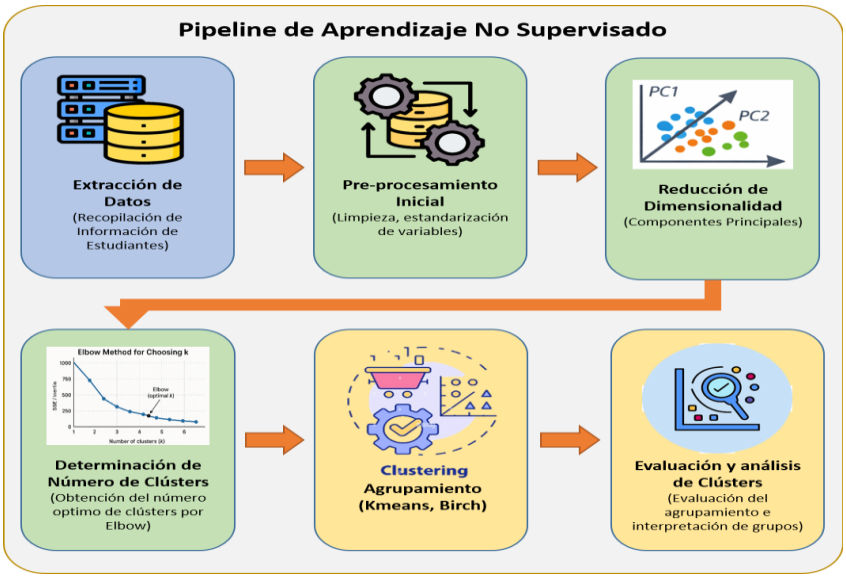

Figura 1. Pipeline de Aprendizaje No Supervisado propuesto

El aprendizaje no supervisado (Ghahramani, 2003) se empleó para identificar patrones y agrupar observaciones con características similares, mediante el desarrollo de modelos basados en técnicas como K-Means y BIRCH. Adicionalmente, el Análisis de Componentes Principales (PCA) (Yang et al., 2018) se utilizó como técnica de reducción de dimensión.

El pipeline completo de aprendizaje no supervisado se presenta en la Figura 1, mostrando de manera secuencial las etapas del proceso desde la extracción de datos hasta la evaluación de los clusters. Como se observa en la figura, el proceso inicia con la extracción y preprocesamiento de los datos, seguido de la reducción de dimensionalidad para agrupamiento y visualización, determinación del número óptimo de clusters mediante el método del codo (Elbow), aplicación de las técnicas de agrupamiento y evaluación de los clusters obtenidos.

2.5.1. Objetivo del Aprendizaje No Supervisado

En esta etapa del estudio, se llevó a cabo un análisis con el objetivo de identificar grupos de estudiantes con características similares en su rendimiento académico. El análisis busca generar conocimiento de valor para la toma temprana de decisiones pedagógicas, permitiendo implementar acciones correctivas basadas en las calificaciones del segundo trimestre, antes de que concluya el ciclo lectivo.

2.5.2. Conjunto de datos y variables utilizadas

Para este análisis, se trabajó con un conjunto de 13 atributos correspondientes a las calificaciones numéricas obtenidas por estudiantes de primer año del ciclo básico en distintas asignaturas del segundo trimestre, considerando los períodos 2017, 2018, 2019, 2022 y 2023. Los atributos considerados son: Artística_2t, Cs. Biológicas_2t, Lengua I_2t, Historia I_2t, Tecnología I_2t, Dibujo Técnico_2t, Fisicoquímica_2t, Lengua Extranjera I_2t, F.E. y C. I_2t, Geografía I_2t, Matemática I_2t, Taller Prep. I_2t y Ed. Física_2t.

2.5.3. Preprocesamiento de datos

Se realizó un ajuste de preprocesamiento de los datos que consistió en la estandarización de los mismos y la aplicación de PCA.

La estandarización se utilizó con el fin de atenuar el impacto de valores extremos, evitando que atributos con mayor escala dominaran el proceso de agrupamiento, sin eliminar observaciones reales del conjunto de datos. Posteriormente, se aplicó PCA sobre los datos estandarizados para que todos los atributos tuvieran la misma importancia en la proyección. Se seleccionaron las primeras 5 (cinco) componentes principales que explican el 71% de la variabilidad de los datos (CP1 = 0,448; CP2 = 0,079; CP3 = 0,065; CP4 = 0,061; CP5 = 0,05) y se utilizaron las primeras 2 (dos) para la visualización en un espacio bidimensional. Esto permitió manejar de manera sintética la información contenida en los 13 atributos originales y facilitar la visualización como la interpretación gráfica de los clusters obtenidos.

2.5.4. Selección del número de clusters

Para determinar el número óptimo de clusters se aplicó el método de Elbow (Thorndike, 1953; Syakur et al., 2018). Como se observa en la Figura 2, la curva presenta una disminución pronunciada de la inercia (WCSS) hasta $k = 3$, a partir de la cual la pendiente se atenúa, indicando un punto de inflexión. En función de este criterio, se seleccionó k=3 como el número adecuado de clusters.

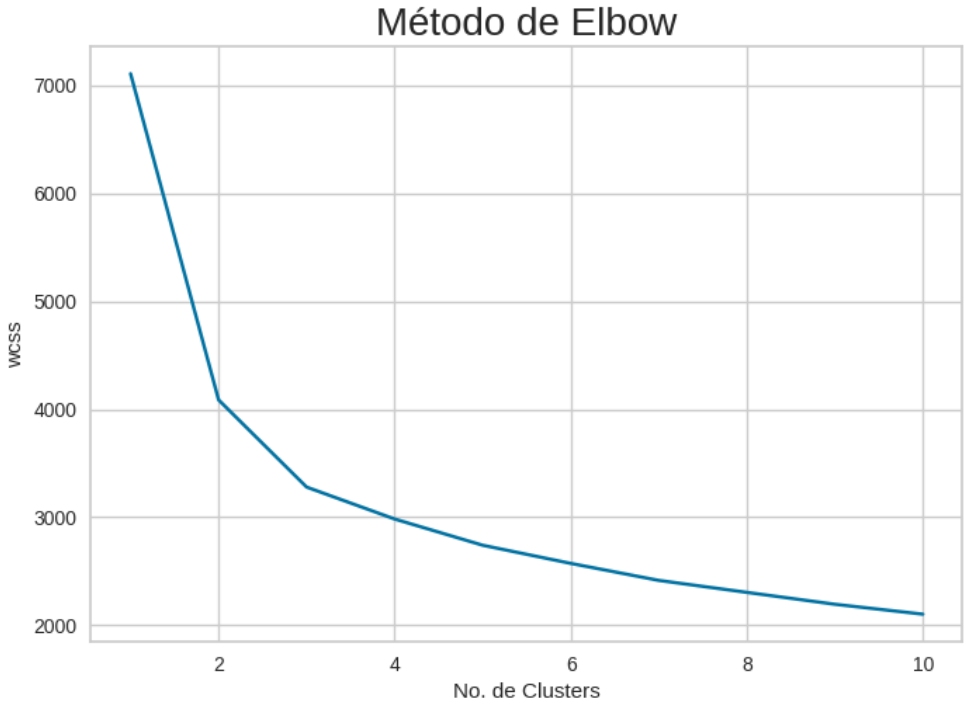

Figura 2. Aplicación del método de Elbow sobre el conjunto de datos

2.5.5. Agrupamiento de datos

Se desarrollaron modelos de aprendizaje no supervisado basados en las técnicas K-Means y BIRCH lo que permitió contrastar enfoques metodológicos distintos. Ambas técnicas fueron seleccionadas debido a sus diferencias conceptuales en el proceso de agrupamiento de datos. K-Means es una técnica de agrupamiento particional, ampliamente utilizada en conjuntos de datos numéricos y estandarizados, la cual obtiene clusters compactos y definidos. Por su parte, BIRCH es una técnica de agrupamiento jerárquico que permite capturar estructuras más flexibles en los datos, resultando adecuado para contrastar los resultados obtenidos mediante K-Means. La comparación entre ambos modelos permitió evaluar la estabilidad y coherencia de los agrupamientos bajo supuestos metodológicos diferentes.

En la Tabla 2 se muestra la configuración elegida para ambos modelos.

Tabla 2. Configuración de los modelos no supervisados

Modelo	Parámetros
K-Means	n_clusters = 3 max_iter = 300 n_init = 10 random_state = 0
BIRCH	n_clusters = 3 threshold = 0,05; 0,1; 0,2; 0,3; 0,4 branching_factor = 3; 5; 10; 15; 20; 30

2.5.6. Evaluación e interpretación

La performance de los modelos no supervisados se midió mediante 3 (tres) métricas específicas de evaluación de clustering: el índice de Calinski-Harabasz (Calinski & Harabasz, 1974), el coeficiente Silhouette (Rousseeuw, 1987) y el índice de Davies–Bouldin (Ros et al., 2023). El índice de Calinski–Harabasz mide la relación entre la dispersión inter-cluster y la dispersión intra-cluster, donde los valores más altos indican una mejor calidad de agrupamiento. En cambio, el coeficiente Silhouette evalúa la cohesión interna de los clusters y su separación respecto de otros grupos, donde valores cercanos a 1 indican una estructura de clustering bien definida. Finalmente, Davies-Bouldin analiza la relación entre la dispersión interna de cada cluster y la distancia al cluster más cercano, donde valores más bajos

representan particiones más compactas y mejor separadas. Como regla general, valores cercanos a 0 representan clusters muy compactos y bien separados, valores entre 1 y 2 indican una cohesión y separación moderadas, y valores mayores a 2 reflejan clusters mal definidos o con solapamiento significativo.

El procedimiento metodológico desarrollado permitió aplicar técnicas de aprendizaje no supervisado para la identificación de patrones en el rendimiento académico. Los resultados derivados de este análisis se presentan en la Sección 3.

2.6. Aprendizaje Supervisado

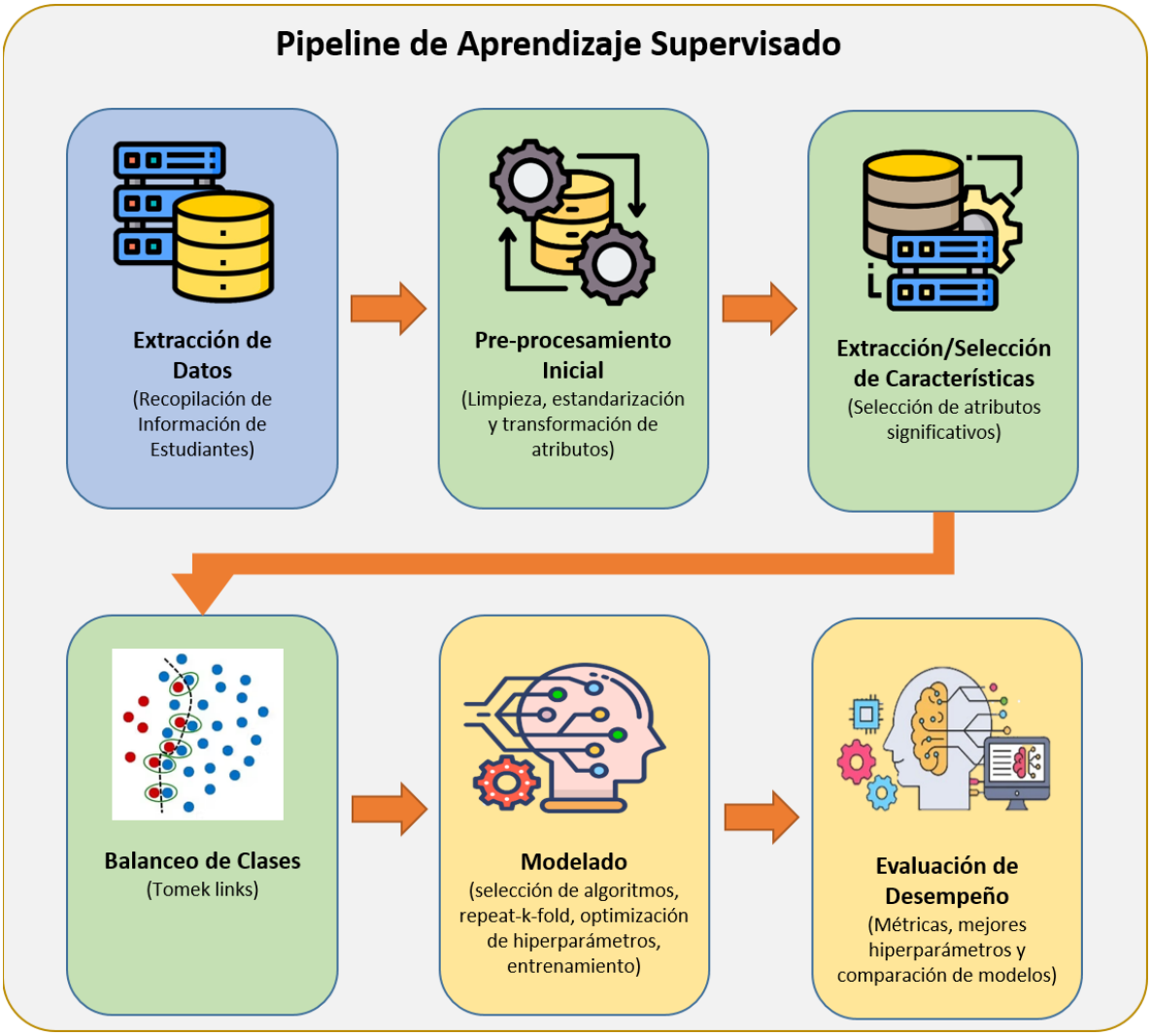

Figure 3. Pipeline de Aprendizaje Supervisado propuesto

Se emplearon técnicas de aprendizaje supervisado como Random Forests, XGBoost y Extreme Learning Machine para el desarrollo de modelos predictivos (Plathottam et al., 2023). El pipeline completo de aprendizaje supervisado se presenta en la Figura 3, donde se describen de manera secuencial las etapas del proceso.

Como se observa en dicha figura, el proceso inicia con la extracción de los datos y su preprocesamiento inicial. Posteriormente, se realiza la extracción y selección de atributos significativos, seguida del balanceo de clases. A continuación, se lleva a cabo la etapa de modelado mediante técnicas de aprendizaje supervisado seleccionadas y finalmente, se evalúa el desempeño de los modelos mediante métricas apropiadas, proporcionando criterios objetivos para valorar su eficacia y consistencia con los objetivos planteados.

2.6.1. Objetivo del Aprendizaje Supervisado

El objetivo de esta etapa es predecir si un estudiante será promovido al año siguiente, utilizando modelos de aprendizaje supervisado. Para ello, se desarrollaron y evaluaron modelos predictivos basados en Random Forests (RF), XGBoost y Extreme Learning Machine (ELM), con el fin de identificar tempranamente a los alumnos en riesgo y apoyar la toma de decisiones pedagógicas.

2.6.2. Selección de atributos o características

Para la construcción de los modelos, se consideraron inicialmente los atributos Sexo, Año y las calificaciones del primer y segundo trimestre de los períodos 2017, 2018, 2019, 2022 y 2023, correspondientes al 1er año de ciclo básico, previamente estandarizadas, ya que estos representan un periodo clave para realizar intervenciones antes del cierre del ciclo lectivo.

Los atributos significativos elegidos corresponden a calificaciones de 6 materias del segundo trimestre, Historia I_2T, Lengua I_2T, Dibujo Técnico_2T, Tecnología I_2T, Matemática I_2T, Lengua Extranjera I_2T, de los periodos 2017, 2018, 2019, 2022 y 2023. Estos fueron elegidos mediante la técnica de selección de características Feature Importance (Breiman, 2001), obtenida mediante un modelo Random Forests entrenado con 250 árboles de decisión. Se calcularon los valores de importancia de todos los atributos disponibles y se conservaron únicamente las 6 materias del segundo trimestre con mayor contribución a la predicción del rendimiento académico. Esta selección se centró en el segundo trimestre, ya que proporciona información más reciente sobre el desempeño del alumno y permite tomar decisiones predictivas antes del inicio del tercero. La Figura 4 muestra la aplicación de la técnica y la contribución relativa de cada atributo.

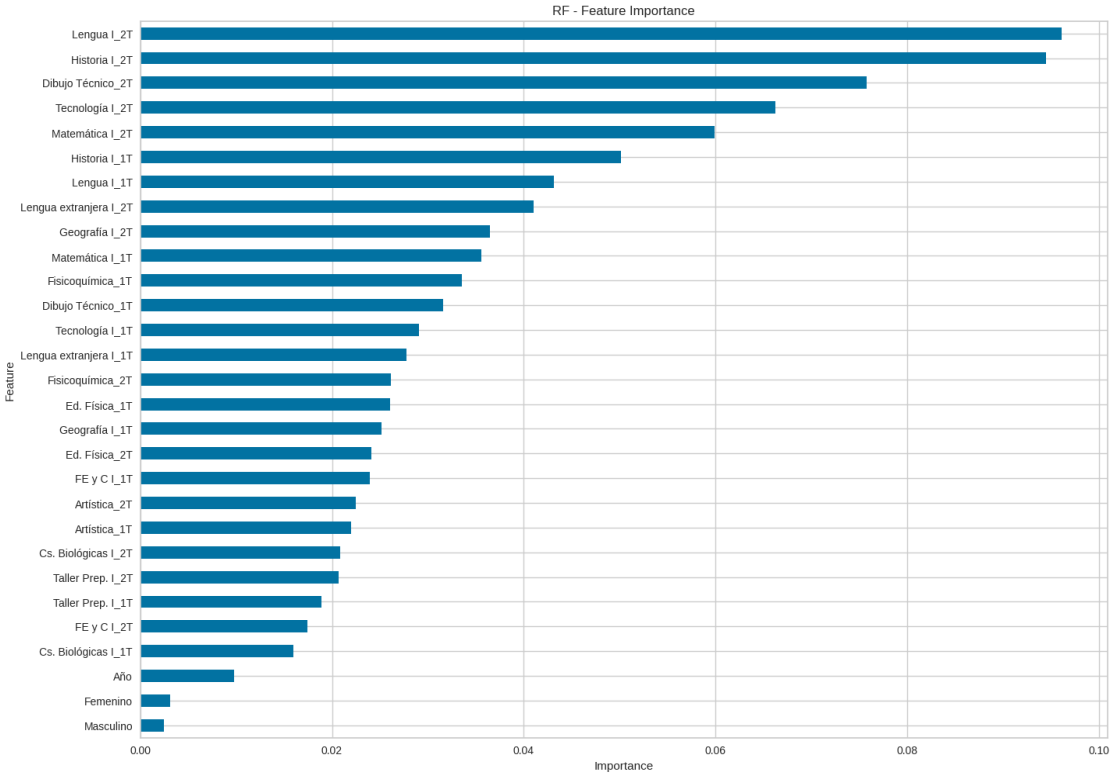

Figure 4. Aplicación de Feature Importance

2.6.3. Balanceo de clases

Dado que el conjunto de datos presentaba un fuerte desbalance de clases, con una mayoría significativa de alumnos promovidos frente a una menor proporción de repitentes, se aplicó la técnica Tomek Links (Leng et al., 2024). Esta identifica pares de instancias de diferentes clases que son mutuamente las más cercanas entre sí, y elimina aquellas pertenecientes a la clase mayoritaria (en este caso, los promovidos).

Como resultado, la clase mayoritaria se redujo de 574 a 554 instancias, mientras que la clase minoritaria permaneció sin cambios. Esta ligera reducción ayudó a atenuar el desbalance, favoreciendo un aprendizaje más equilibrado de los modelos supervisados y contribuyendo a mejorar la capacidad de predicción de la clase minoritaria, sin eliminar información crítica de la clase mayoritaria.

No se emplearon técnicas de *data augmentation*, dado que se trabajó únicamente con datos académicos reales y no se consideró adecuado generar observaciones sintéticas. La generación de observaciones sintéticas podría introducir valores artificiales en las calificaciones, afectando tanto la validez pedagógica de los resultados como la confiabilidad de las predicciones del modelo.

2.6.4. División de datos (Train/Test Split)

El split de los datos se realizó tomando el 75% de los datos para entrenamiento y el 25% para prueba de los modelos predictivos.

2.6.5. Estrategia de validación (Repeated K-Fold)

Con el fin de garantizar una validación robusta y minimizar el riesgo de sobreajuste, se aplicó la técnica de Repeated K-Fold Cross Validation (Kohavi, 2001) con la configuración:

- n_splits: 10 (número de splits).
- n_repeats: 10 (número de repeticiones).
- random_state: 42 (valor de la semilla, para asegurar reproducibilidad)

Esta técnica fue elegida frente a alternativas como K-Fold simple o Stratified K-Fold, ya que permite obtener una medición más estable del desempeño del modelo, reduciendo la variabilidad de las métricas entre diferentes particiones del dataset. Cada observación participa en múltiples conjuntos de entrenamiento y validación, lo que asegura un análisis más completo del comportamiento del modelo en todo el conjunto de datos y mejora la confiabilidad de la selección de hiperparámetros y del rendimiento final reportado. La configuración de Repeated K-Fold es adecuada para un dataset de menos de 1000 muestras, balanceando estabilidad de las métricas de performance de los modelos supervisados y costo computacional de entrenamiento de los mismos.

2.6.6. Optimización de hiperparámetros

Con el objetivo de obtener los mejores modelos posibles, se emplearon técnicas de optimización de hiperparámetros que permitieron mejorar su desempeño y adaptabilidad. En particular, se recurrió a Grid Search (Belete & Huchaiah 2022; Ogunsanya et al. 2023), Randomized Search (Breiman, 2001) y BayesSearch (Snoek et al., 2012), enfoques que ofrecen distintas estrategias para explorar el espacio de configuraciones y seleccionar aquellas que maximizan la calidad de los modelos.

En la Tabla 3 se encuentran los valores específicos probados para cada hiperparámetro.

Tabla 3. Técnicas y valores de hiperparámetros (Martínez et al., 2025)

Modelo ML	Técnica	Hiperparametros	Valores
ELM	Grid Search	número de neuronas en capas ocultas	1000; 2000; 3000; 4000; 5000; 6000; 7000
		función de activación	sigmoid; relu; sin; leaky_relu; tanh
		C (parámetro de regularización)	0,001; 0,01; 0.1; 0,3; 0,5; 0,7; 0,9; 1; 1,3, 1,5, 2
		tipo de aleatoriedad (Random Type)	uniform; normal
		include	False
RF	Bayes Search	n_estimators	100; 300
		max_depth	5; 30
		min_samples_split	2; 10
		min_samples_leaf	1; 5
		max_features	sqrt; log2

XGBoost	Randomized Search	subsample	0,7; 0,8; 0,85; 0,9
		max_depth	5; 7; 9; 10; 11
		learning rate	0,001; 0,01; 0,05; 0,1
		gamma	0; 0,1; 1; 3; 5
		n_estimators	500; 900

La Tabla 4 presenta los hiperparámetros óptimos utilizados para el entrenamiento de los modelos predictivos. La selección de estas configuraciones se basó en métricas de desempeño global y por clase durante el proceso de optimización, considerando *Accuracy* como métrica global y *Precision*, *Recall* y *F1-Score* a nivel de clase. Asimismo, se reporta el tiempo de CPU asociado a cada configuración óptima.

Tabla 4. Mejores hiperparámetros de los modelos supervisados

Modelo	Mejores hiperparámetros	Tiempo de CPU
Random Forests (Bayes Search)	• max_depth: 30 • max_features: sqrt • min_samples_leaf: 5 • min_samples_split: 10 • n_estimators: 100	18 minutos, 32 segundos
XGBoost (Randomized Search)	• subsample: 0,7 • n_estimators: 500 • max_depth: 11 • learning_rate: 0,01 • gamma: 3	2 minutos, 56 segundos
ELM (Grid Search)	• Hidden_Units = 1000 • Activación = Sigmoid • C = 0,1 • random_type: normal	1 hora, 38 minutos y 27 segundos

Con la metodología establecida, se entrenaron los modelos supervisados finales para la predicción del rendimiento académico. Los resultados alcanzados se presentan y analizan en la Sección 3.

3. RESULTADOS Y DISCUSIÓN

3.1. Resultados del Análisis Exploratorio (EDA)

Distribución de alumnos aprobados y desaprobados

En la Tabla 5 se muestra la distribución de alumnos aprobados y desaprobados de primer año de ciclo básico en los periodos 2017-2019, 2022-2023. El objetivo es brindar una visión general del desempeño académico, sin segmentaciones por materia, trimestre, como punto de partida para análisis más específicos.

Se puede observar que la proporción de estudiantes que lograron ser promovidos al siguiente año es superior a la de aquellos que no lo hicieron, lo que refleja un desempeño global positivo.

Tabla 5. Cantidad y porcentaje de alumnos promovidos y repitentes

Estado del alumno	Cantidad	Porcentaje (%)
Promovidos	560	74,5
Repitentes	192	25,5

Histogramas por Materias del Primer y Segundo Trimestre

Para analizar el desempeño de los estudiantes en cada materia, se presentan en la Figura 5 y 6 los histogramas que muestran la distribución de notas del primer y segundo trimestre de los periodos 2017, 2018, 2019, 2022 y 2023.

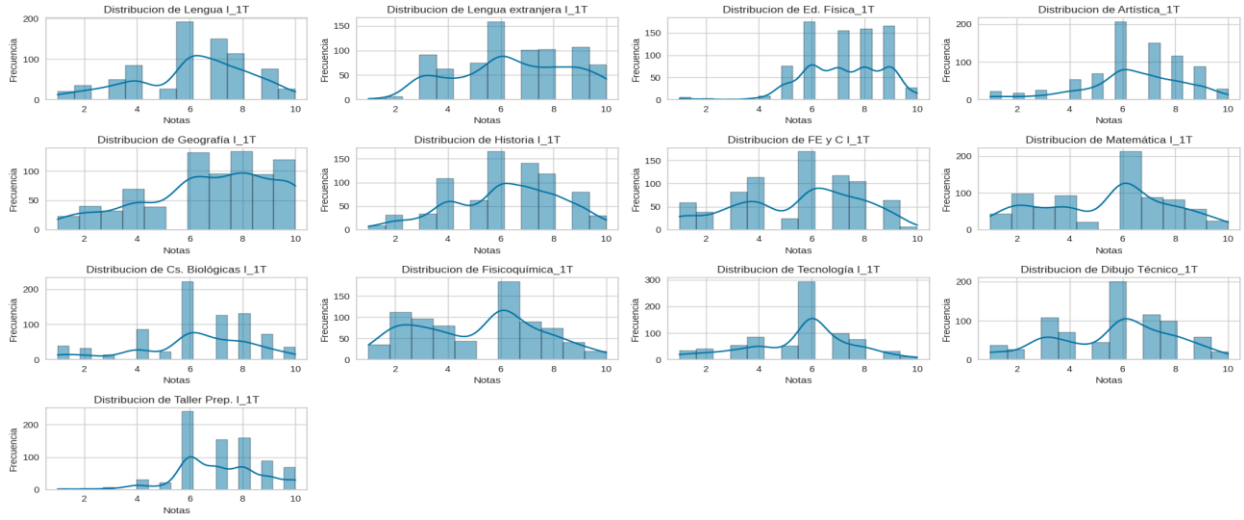

Figura 5. Distribución de notas del primer trimestre

En el primer trimestre se identificaron distintos patrones en la distribución de los datos. Algunas materias como Lengua, Historia, Lengua Extranjera, Formación Ética y Ciudadana, Ciencias Biológicas y Taller, presentan distribuciones levemente simétricas o centradas, con medias entre 6,16 y 7,13 y medianas en torno a 6 o 7, concentrando la mayor parte de las calificaciones en el rango [6–8]. Por otro lado, materias como Matemática, Fisicoquímica, Tecnología y Dibujo Técnico muestran distribuciones asimétricas positivas, caracterizadas por medias más bajas (entre 5,13 y 5,71) y primer cuartil ubicado entre 3 y 4, con mayor concentración de calificaciones bajas (entre 3 y 6), lo que refleja mayores dificultades académicas, especialmente en las áreas de Matemática y Fisicoquímica. En contraste, materias como Artística, Geografía y Educación Física exhiben distribuciones asimétricas negativas, con medias superiores a 6,4, tercer cuartil entre 8 y 9, y valores máximos cercanos a 10, lo que evidencia un desempeño generalmente favorable en estas asignaturas. Finalmente, en algunos casos como Educación Física, Taller y Dibujo Técnico se observan distribuciones multimodales, lo que sugiere la presencia de subgrupos con desempeños diferenciados dentro del aula.

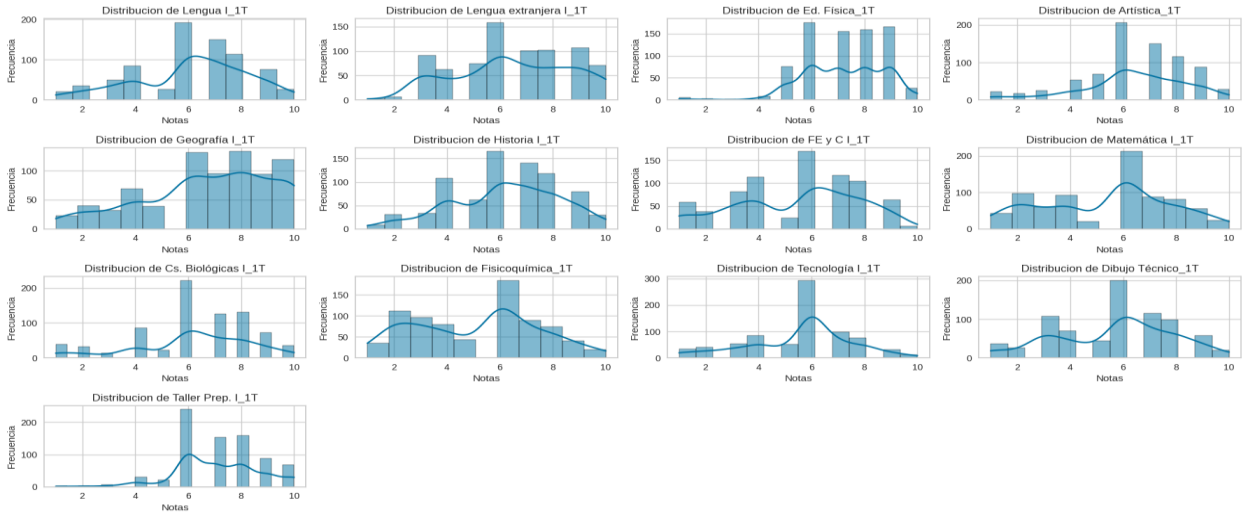

Figura 6. Distribución de notas del segundo trimestre

A partir de las observaciones de las distribuciones de notas del segundo trimestre, podemos destacar que las calificaciones reflejan nuevamente una variedad de distribuciones y asimetrías, lo que da cuenta de un desempeño académico diverso según la materia. Educación Física, Artística, Formación Ética y Ciudadana, Geografía y Taller, presentan asimetría negativa, con medias entre 6,33 y 7,30, medianas cercanas 7 - 8 y tercer cuartil cercano a 8 o 9, lo que evidencia un predominio de calificaciones altas y bajo nivel de

desaprobación. Otras asignaturas como Lengua, Historia, Lengua Extranjera y Ciencias Biológicas muestran distribuciones centradas, con medias cercanas a 6, medianas iguales a 6 y rangos intercuartílicos concentrados entre 5 y 8, indicando un rendimiento académico intermedio y relativamente estable. En cambio, Matemática, Fisicoquímica, Tecnología y Dibujo Técnico, evidencian asimetría positiva con medias inferiores a 5,75, primer cuartil entre 3 y 5 y mayor dispersión de las calificaciones. En particular, Fisicoquímica (media = 5,15) y Dibujo Técnico (media = 5,38) reflejan un mayor porcentaje de estudiantes con bajo rendimiento. Además, en estas materias se observa bimodalidad en los histogramas, lo que sugiere una clara diferenciación entre estudiantes que logran comprender los contenidos y aquellos que presentan mayores dificultades.

Comparación con el primer trimestre

En comparación con el primer trimestre, se observa una leve mejora general en varias materias, aunque persisten ciertas asimetrías y dispersión en asignaturas como Matemática, Fisicoquímica y Tecnología. En líneas generales, las notas siguen concentrándose entre 6 y 8, con algunas materias mostrando mejor desempeño general. Las materias más críticas que requieren atención especial siguen siendo Matemática, Fisicoquímica, Tecnología por su dispersión y cantidad de alumnos con rendimiento bajo en riesgo académico. Estos patrones sugieren un rendimiento general positivo, incluso superior al del primer trimestre, aunque con una variabilidad característica.

Análisis multivariado de datos

En este análisis se examina la relación entre las calificaciones de las distintas materias mediante una matriz de correlación, utilizando datos acumulados de los años 2017-2019 y 2022-2023. El objetivo es identificar dependencias y patrones relevantes que ayuden a comprender mejor el rendimiento académico de los estudiantes. A continuación, se presenta la matriz de correlación correspondiente al segundo trimestre (Figura 7).

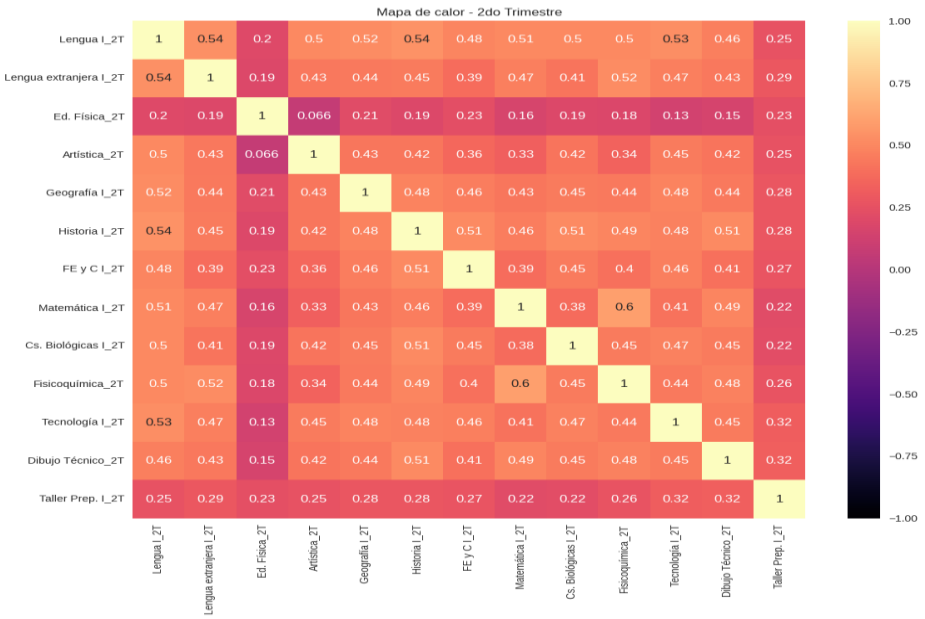

Figura 7. Matriz de correlación - 2do trimestre

Se observan correlaciones moderadamente altas y positivas entre Lengua, Lengua Extranjera, Historia, Matemática, Fisicoquímica y Tecnología, tanto entre estas asignaturas como con otras materias, aunque con correlaciones más débiles y moderadas. Este patrón sugiere que los estudiantes que obtienen buenos resultados en una de estas materias tienden a desempeñarse bien en las demás, posiblemente debido a la presencia de habilidades cognitivas compartidas como la comprensión lectora, el razonamiento lógico y la capacidad de análisis.

Durante el análisis exploratorio, se identificaron ciertos valores extremos en las calificaciones y algunos registros incompletos.

Los valores extremos fueron conservados, ya que corresponden a resultados académicos reales y su eliminación podría introducir sesgos en el análisis. Por otro lado, los registros incompletos fueron eliminados del conjunto de datos, debido a la presencia de valores faltantes que impedían su correcta utilización en las técnicas de análisis y modelado aplicadas.

Por otra parte, el atributo objetivo del estudio corresponde al Estado Final del alumno, indicando si fue promovido al siguiente año o si repitió la cursada. Para su uso en modelos supervisados, este atributo categórico se codificó de manera binaria, asignando 1 a "Promovido" y 0 a "Repite".

3.2. Resultados del Aprendizaje No Supervisado

En esta sección se presenta la evaluación de los modelos K-Means y BIRCH mediante las métricas de Silhouette, Calinski–Harabasz y Davies–Bouldin. Mediante comparación numérica, se selecciona el modelo con mejor desempeño, cuya representación gráfica se analiza para observar la distribución de los estudiantes en los clusters identificados.

En la Tabla 6 se muestran los valores de las métricas para ambos modelos.

Tabla 6. Resultados de las métricas de agrupamiento

Modelo	Silhouette Score	Calinski–Harabasz	Davies Bouldin
K-Means	0,2518	450,236	1,32
BIRCH	0,2265	408,79	1,36

Los valores listados en la Tabla 6 evidencian que ambos modelos logran identificar una estructura de agrupamiento interpretable sobre el conjunto de datos. En particular, K-Means obtiene valores superiores en el coeficiente de Silhouette (0,2518) y en el índice de Calinski–Harabasz (450,236), lo que sugiere una mayor cohesión interna de los clusters y una mejor separación entre los grupos formados en comparación con BIRCH. Asimismo, el índice de Davies–Bouldin presenta un valor menor para K-Means (1,32), indicando una menor dispersión interna relativa respecto al cluster más cercano. Considerando de manera conjunta las tres métricas, se concluye que el modelo basado en K-Means ofrece una calidad de agrupamiento adecuada y consistente para el análisis realizado.

Discusión de resultados

La Figura 8 muestra a los estudiantes agrupados utilizando el modelo basado en K-Means usando las 2 (dos) primeras Componentes Principales (PCA). Los puntos en el gráfico representan las calificaciones de los estudiantes, previamente estandarizadas.

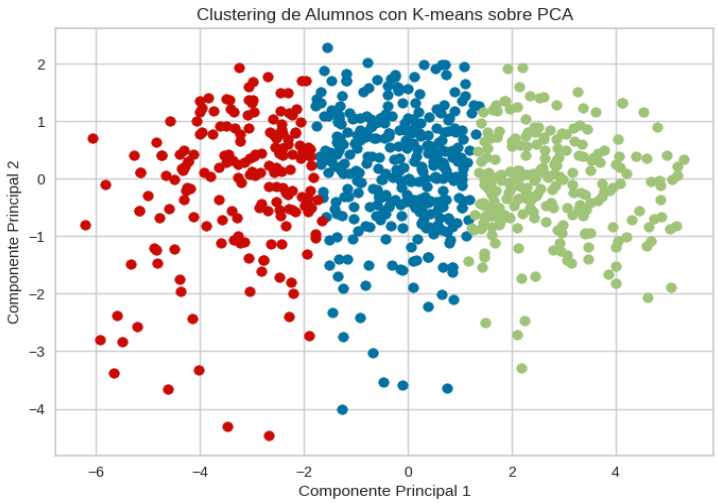

Figure 8. Agrupamiento de datos según modelo K-Means

A partir del gráfico de clusters en el espacio bidimensional, se identificaron tres grupos de estudiantes:

- Clase 0 (Azul) – Rendimiento Medio - (337 elementos): Los puntos correspondientes a esta clase se concentran en torno a valores intermedios de la CP1, que está principalmente influenciada por materias teóricas como Lengua I (10,36), Historia I (9,62), Fisicoquímica (8,91), Tecnología I (8,84), Geografía I (8,69) y Lengua Extranjera I (8,61). La mayor concentración de puntos sobre CP2, dominada por materias prácticas como Ed. Física (66,07), Taller Preparación (23,73) y Artística (5,18), se relaciona con estudiantes con un rendimiento académico promedio.
- Clase 1 (Verde) – Rendimiento Alto - (247 elementos): Los puntos de esta clase se ubican predominantemente en valores positivos de la CP1 y CP2, presentando una distribución relativamente compacta. La CP1 positiva sugiere un excelente desempeño en materias teóricas (Lengua, Historia, Fisicoquímica, Tecnología, Geografía), mientras que los valores positivos en CP2 reflejan también un buen desempeño en materias prácticas o físicas (Ed. Física, Taller, Artística). Este grupo corresponde a estudiantes con un rendimiento académico superior y perfiles más homogéneos en las materias evaluadas.
- Clase 2 (Rojo) – Bajo Rendimiento - (191 elementos): Los puntos asociados a esta clase se concentran en valores negativos de la CP1, lo que evidencia desempeño bajo en materias teóricas, con una mayor dispersión en la CP2 lo que indica variaciones en el rendimiento en materias prácticas. Este comportamiento sugiere un grupo de estudiantes con desempeño académico bajo.

En este estudio, la aplicación de K-Means permitió identificar tres clusters de rendimiento académico (bajo, medio y alto) utilizando el método del codo (Elbow), con un coeficiente de Silhouette de 0,2518, un índice de Calinski-Harabasz de 450,236 y Davies–Bouldin de 1,32. La distribución en los clusters fue de 191, 337 y 247 estudiantes respectivamente, lo que garantiza una segmentación equilibrada y representativa.

Comparando con la literatura reciente, Mohamed Nafuri et al. (2022) identificaron cinco clusters utilizando K-Means, pero los valores de Silhouette (0,16–0,192) y Calinski-Harabasz (17,358–24,946) fueron considerablemente menores, indicando una menor separación y densidad interna de los grupos. Por su parte, Amalia et al. (2021) reportó clusters óptimos también en 3 grupos con métricas de validación que variaron entre 0,340–0,514 para Silhouette y 27,174–84,529 para Calinski-Harabasz; sin embargo, sus datasets eran pequeños (20–50 muestras por modelo), lo que limita la representatividad de los clusters. Estos resultados sugieren que la segmentación en tres grupos de nuestro estudio ofrece mayor claridad y aplicabilidad educativa, combinando una adecuada separación entre los clusters con tamaños de muestra suficientes para un análisis confiable.

Además, la estructura obtenida confirma el potencial del análisis para apoyar decisiones pedagógicas tempranas, ya que permite reconocer distintos niveles de rendimiento académico y orientar medidas correctivas antes del cierre del ciclo lectivo.

En los *Anexos A2* y *A3* se presentan resultados del aprendizaje no supervisado usando el conjunto de datos de segundo año de ciclo básico.

3.3. Resultados del Aprendizaje Supervisado

En esta sección se presentan los resultados obtenidos por los modelos propuestos para la predicción del rendimiento académico.

Se emplearon las métricas de clasificación *F1-Score* (Rainio et al., 2024), *Precision*, *Recall* y *Accuracy* (Shobha & Rangaswamy, 2018) para evaluar el desempeño de los modelos desarrollados. Estas permiten obtener una visión general de la calidad de las predicciones, teniendo en cuenta los aciertos totales y el equilibrio entre los errores de clasificación. Las Tablas 7 y 8 muestran la performance de los mejores modelos supervisados, a nivel general y por clases (Clase 0: Repite; Clase 1: Promovido).

Tabla 7. Resultados de métrica de clasificación a nivel de modelo

Modelo	Accuracy (%)
XGBoost	89
ELM	87
RF	87

Tabla 8. Resultados de métricas de clasificación a nivel de clase

Modelo	Clase	Precision (%)	F1-Score (%)	Recall (%)
XGBoost	Repite (0)	84	80	76
	Promovido (1)	91	92	94
ELM	Repite (0)	82	78	74
	Promovido (1)	90	92	93
RF	Repite (0)	84	76	69
	Promovido (1)	88	91	95

De acuerdo a los resultados de las Tablas 7 y 8, se puede apreciar que el modelo XGBoost presentó un desempeño superior en comparación a Random Forests y ELM. Más allá de la capacidad general del modelo, si analizamos el rendimiento por clases, en específico considerando los valores de *F1-Score*, se aprecia claramente que el modelo XGBoost presenta un desempeño equilibrado, con un valor de 80% para la clase "Repite" y 92% para la clase "Promovido". Esto indica que identifica correctamente a los estudiantes de ambas clases, con un rendimiento particularmente sólido en la clase mayoritaria. Comparado con ELM (78% y 92%) y RF (76% y 91%), XGBoost muestra un mejor equilibrio y capacidad de clasificación, especialmente para la clase minoritaria. El modelo XGBoost se destaca como el más eficiente, logrando el mejor equilibrio entre tiempo de cómputo y rendimiento. De esta manera se posiciona como la opción más sólida para predecir si un estudiante es promovido o no al siguiente año.

Adicionalmente, la Figura 9 presenta la matriz de confusión (Menacho Chiok, 2017) del modelo XGBoost, seleccionado como el de mejor desempeño. Esta representación permite visualizar de manera directa la distribución de la clasificación observada (real) y la predecida por el clasificador para las distintas categorías de la variable clase, reforzando la interpretación del comportamiento del modelo en la identificación de estudiantes promovidos y en riesgo de repitencia. La matriz de confusión es una herramienta ampliamente utilizada en la literatura para evaluar la calidad de clasificación de modelos predictivos.

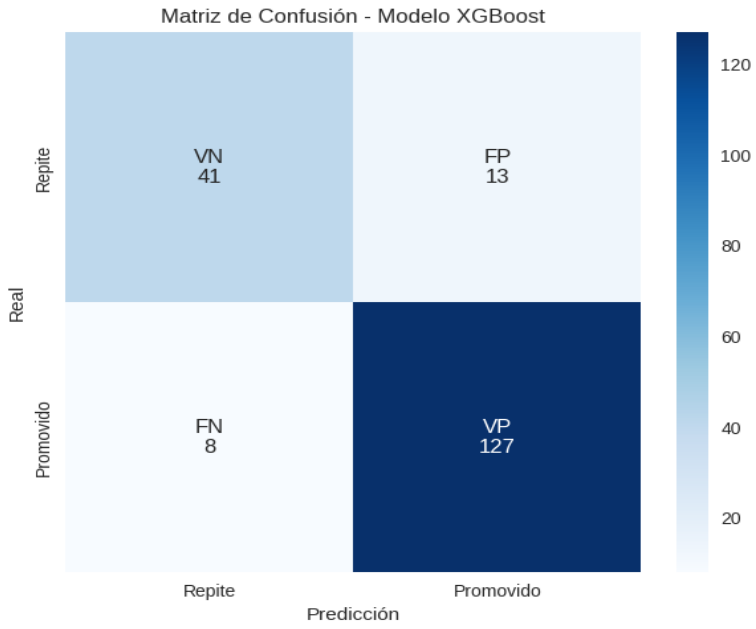

Figura 9. Matriz de confusión del modelo XGBoost

Discusión de resultados

Desde una perspectiva pedagógica, un valor de *F1-Score* de 0,80 en la clase *"Repite"* permite utilizar el modelo como un sistema de alerta temprana, útil para orientar intervenciones pedagógicas antes de que se concrete la repitencia, manteniendo un balance razonable entre omisiones y clasificaciones erróneas. En el ámbito institucional, esta información puede emplearse para orientar intervenciones educativas focalizadas, tales como tutorías, acompañamiento pedagógico o seguimiento personalizado, contribuyendo a una asignación más eficiente de los recursos disponibles. En este sentido, el modelo no reemplaza la evaluación docente, sino que actúa como un soporte para la toma de decisiones, fortaleciendo las estrategias de prevención del abandono y la repitencia escolar.

Los resultados obtenidos son consistentes con la literatura revisada. Por ejemplo, el estudio de Saltos-Mero & Cruz-Felipe (2024) abordó un problema de clasificación binaria ("Aprobar" / "Desaprobar") y reportó que su modelo basado en Random Forests superó a otros propuestos (Árbol de decisión, Redes Neuronales, SVM), alcanzando valores de *Accuracy* de 0,86 en Economía y 0,83 en Turismo, confirmando así la eficacia de los métodos basados en árboles para problemas de clasificación binaria.

De manera similar, Guanin-Fajardo et al. (2024) evaluaron XGBoost y RF sobre un dataset de 6690 registros con balanceo mediante EasyEnsemble y clasificación multiclase (Aprobado, Cambio, Abandono). Allí, XGBoost obtuvo valores de *Accuracy*=0,7949, *F1-Score*= 0,8306, *Precision*=0,8214 y *Recall*=0,8425, mientras que RF presentó resultados comparables. Aunque los valores reportados son ligeramente menores a los alcanzados en nuestro estudio, los autores abordaron un problema de mayor complejidad y tamaño.

Finalmente, Bellaj et al. (2024) analizaron un conjunto de datos de 480 casos, 16 atributos y tres niveles de rendimiento académico (Bajo, Medio, Alto), aplicando CRISP-DM, optimización de hiperparámetros y validación cruzada estratificada de 10 folds con GridSearchCV, lo que permitió evaluar la robustez de los modelos frente a diferentes particiones del dataset. Sus mejores modelos (Voting, XGBoost y RF con HPO) alcanzaron valores de *Accuracy*=0,84–0,86, *F1-Score*=0,85–0,87, *Precision*=0,84 y *Recall*=0,84, confirmando la efectividad de los métodos basados en árboles incluso en clasificación multiclase.

En conjunto, la comparación evidencia que aunque los datasets y la complejidad de las clases varían, los modelos basados en árboles, especialmente XGBoost, consistentemente muestran un desempeño superior. Además, las técnicas de preprocesamiento y optimización implementadas en nuestro estudio, como selección de características, manejo de desbalanceo y Repeated K-Fold, contribuyen a aumentar la robustez y confiabilidad de las predicciones en un contexto binario, de manera análoga a cómo la validación cruzada mejora la confianza en las métricas reportadas en estudios multiclase.

En los *Anexos A3, A4 y A5* se presentan resultados del aprendizaje supervisado usando el conjunto de datos de segundo año de ciclo básico.

CONCLUSIONES

A partir del análisis de los datos académicos de estudiantes de primer año del ciclo básico y siguiendo la metodología CRISP-DM, se identificaron patrones consistentes de rendimiento que permiten caracterizar perfiles académicos diferenciados a través de clusters. Desde una perspectiva educativa, estos perfiles pueden interpretarse como distintos niveles de riesgo y estabilidad académica, lo que posibilita el diseño de estrategias de acompañamiento pedagógico diferenciadas según las necesidades de cada grupo de estudiantes.

Estos hallazgos, en concordancia con los objetivos planteados, confirman la utilidad de aplicar técnicas de minería de datos y aprendizaje automático en contextos escolares para anticipar situaciones de riesgo académico mediante predicciones y orientar intervenciones pedagógicas más efectivas y oportunas. En

particular, la identificación temprana de estudiantes con mayor probabilidad de repitencia permite actuar antes de que se consoliden trayectorias escolares desfavorables.

Como principal aporte, los resultados ofrecen una base sólida para el diagnóstico institucional y la mejora en la toma de decisiones, además de evidenciar que técnicas de aprendizaje automático como K-Means y XGBoost resultan adecuados para clasificar y predecir el desempeño estudiantil. En este sentido, los modelos desarrollados no deben entenderse como herramientas de decisión automática, sino como un soporte analítico que complementa la mirada pedagógica y la experiencia docente. Asimismo, permiten integrar información de distintos niveles de análisis, desde perfiles generales de rendimiento hasta predicciones individuales, facilitando la planificación de estrategias de seguimiento personalizado, la asignación eficiente de recursos educativos y la identificación temprana de estudiantes que podrían beneficiarse de intervenciones focalizadas. De este modo, los modelos propuestos contribuyen a consolidar un enfoque más sistemático y fundamentado en evidencia para la gestión académica, fortaleciendo la capacidad de la institución para anticipar problemas, evaluar resultados y diseñar políticas educativas más efectivas.

Se identifican oportunidades para que futuras investigaciones amplíen el tamaño de la muestra, incorporen variables socioemocionales y factores contextuales, y analicen el impacto de estrategias de intervención basadas en los perfiles obtenidos. Asimismo, se propone como línea de trabajo futuro el desarrollo de una aplicación web que integre gráficos estadísticos junto con los modelos supervisados y no supervisados desarrollados, de manera que la institución pueda contar con una herramienta interactiva que facilite el seguimiento, análisis y gestión del rendimiento académico, contribuyendo así a la construcción de un sistema educativo más inclusivo, personalizado y efectivo.

Una limitación del estudio es la ausencia de datos correspondientes a los años 2020 y 2021, período afectado por la pandemia del COVID-19, lo que pudo haber influido en los patrones de rendimiento observados. Se podría subsanar esto una vez puesta en producción la solución desarrollada y usando las notas de las cohortes actuales para predicción y reentrenamiento permanente de los modelos supervisados y no supervisados.

AGRADECIMIENTO

Se agradece especialmente a las autoridades de la Escuela de Educación Técnica 3100 por haber dispuesto y facilitado el conjunto de datos utilizado en este estudio, cuya colaboración resultó fundamental para el desarrollo y validación de los modelos presentados.

FINANCIAMIENTO

El trabajo fue parcialmente financiado por el Proyecto CIUNSa 2735 dependiente del Consejo de Investigación de la Universidad Nacional de Salta (Argentina).

CONFLICTO DE INTERESES

Los autores declaran no tener ningún tipo de conflicto de interés relacionado con el desarrollo del estudio.

CONTRIBUCIÓN DE LOS AUTORES

Conceptualización: Martinez, C. Curación de datos y análisis formal: Zalasar, A. Adquisición de fondos: Martinez, C. Investigación: Zalasar, A. Metodología y administración del proyecto: Martinez, C. Recursos: Aramayo, R. Software: Zalasar, A. Supervisión: Martinez, C. y Aramayo, R. Validación: Aramayo, R. Visualización: Zalasar, A. Redacción - borrador original: Aramayo, R. y Zalasar, A. Redacción - revisión y edición: Martínez, C. y Zalasar, A.

REFERENCIAS

Amalia, N. L. R., Supianto, A. A., Setiawan, N. Y., Zilvan, V., Yuliani, A. R., & Ramdan, A. (2021). Student Academic Mark Clustering Analysis and Usability Scoring on Dashboard Development Using K-Means Algorithm and System Usability Scale. *Jurnal Ilmu Komputer Dan Informasi*, *14*(2), 137–143. https://doi.org/10.21609/jiki.v14i2.980

Belete, D. M., & Huchaiah, M. D. (2022). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, *44*(9), 875–886. https://doi.org/10.1080/1206212X.2021.1974663

Bellaj, M., Ben Dahmane, A., Boudra, S., & Lamarti Sefian, M. (2024). Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance. *International Journal of Online and Biomedical Engineering (IJOE)*, *20*(03), 55–74. https://doi.org/10.3991/ijoe.v20i03.46287

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Chapman, P. (2000). *Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide*. https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman/54bad20bbc7938991bf34f86dde0babfbd2d5a72

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

García, A. M. (2014). Rendimiento académico y abandono universitario modelos, resultados y alcances de la producción académica en la Argentina. *Revista Argentina de Educación Superior*. http://hdl.handle.net/11336/35674

Ghahramani, Z. (2003). Unsupervised Learning. *ML Summer Schools*. https://doi.org/https://doi.org/10.1007/978-3-540-28650-9_5

Guanin-Fajardo, J. H., Guaña-Moya, J., & Casillas, J. (2024). Predicting Academic Success of College Students Using Machine Learning Techniques. *Data*, *9*(4), 60. https://doi.org/10.3390/data9040060

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, *70*(1–3), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

Ibarra, C. S. (2020). *TÉCNICAS DE DATA MINING APLICADAS A LA DESERCIÓN DE LOS ESTUDIANTES DE LA FACULTAD DE CIENCIAS EXACTAS* [Universidad del Norte Santo Tomás de Aquino]. https://doi.org/https://doi.org/10.13140/RG.2.2.29986.66244

Kohavi, R. (2001). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, *2*, 1137–1143. https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection

Leng, Q., Guo, J., Tao, J., Meng, X., & Wang, C. (2024). OBMI: oversampling borderline minority instances by a two-stage Tomek link-finding procedure for class imbalance problem. *Complex & Intelligent Systems*, *10*(4), 4775–4792. https://doi.org/10.1007/s40747-024-01399-y

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Berkeley Symp. on Math. Statist. and Prob.* University of California, Los Angeles.

Martínez, C. A., Hohl, D. M., Gutiérrez, M. de los A., Palmal, S., Faux, P., Adhikari, K., Gonzalez-Jose, R., Bortolini, M. C., Acuña-Alonzo, V., Gallo, C., Linares, A. R., Rothhammer, F., Catanesi, C. I., & Barrientos, R. J. (2025). DNA-based prediction of eye color in Latin American population applying Machine Learning models. *Computers in Biology and Medicine*, *194*, 110404. https://doi.org/10.1016/j.compbiomed.2025.110404

Menacho Chiok, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos*, *78*(1), 26. https://doi.org/10.21704/ac.v78i1.811

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering Analysis for Classifying Student Academic Performance in Higher Education. *Applied Sciences*, *12*(19), 9467. https://doi.org/10.3390/app12199467

Ogunsanya, M., Isichei, J., & Desai, S. (2023). Grid search hyperparameter tuning in additive manufacturing processes. *SME North American Manufacturing Research Conference*. https://doi.org/https://doi.org/10.1016/j.mfglet.2023.08.056

Plathottam, S. J., Rzonca, A., Lakhnori, R., & Iloeje, C. O. (2023). A review of artificial intelligence applications in manufacturing operations. *Journal of Advanced Manufacturing and Processing*, *5*(3). https://doi.org/10.1002/amp2.10159

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Wiley Interdisciplinary Reviews*, *10*(3). https://doi.org/https://doi.org/10.1002/widm.1355

Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*, *528*, 178–199. https://doi.org/10.1016/j.neucom.2023.01.043

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Saltos-Mero, J., & Cruz-Felipe, M. (2024). Análisis del rendimiento académico de estudiantes de las carreras Economía y Turismo con Power BI en los periodos (2021). *593 Digital Publisher CEIT*, *9*(1), 762–772. https://doi.org/10.33386/593dp.2024.1.2162

Shobha, G., & Rangaswamy, S. (2018). *Machine Learning* (pp. 197–228). https://doi.org/10.1016/bs.host.2018.07.004

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Cornell University*. https://doi.org/https://doi.org/10.48550/arXiv.1206.2944

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*, 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Thorndike, R. L. (1953). Who Belongs in the Family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

Tukey, J. W. (1977). *Exploratory Data Analysis, Volumen 2* (18th ed.). Addison-Wesley Publishing Company.

Wang, J., Lu, S., Wang, S.-H., & Zhang, Y.-D. (2022). A review on extreme learning machine. *Multimedia Tools and Applications*, *81*(29), 41611–41660. https://doi.org/10.1007/s11042-021-11007-7

Yang, S. J. H., Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., & Hiroaki Ogata, A. J. Q. L. (2018). Predicting

Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis. *J-Stage*, *26*, 170–176. https://doi.org/https://doi.org/10.2197/ipsjjip.26.170

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH. *ACM SIGMOD Record*, *25*(2), 103–114. https://doi.org/10.1145/235968.233324

ANEXOS

Conjunto de datos del segundo año de ciclo básico

En el Anexo A.1 se presentan los atributos relacionados con los registros académicos del segundo año de ciclo básico de la institución.

Anexo A.1. Atributos del estudio 2do año CB

Variable	Tipo	Descripción
Curso	Categórica	Año de cursado y división del estudiante (ej. 2°1°, 2°2°, etc.)
Turno	Categórica	TM (Turno Mañana) o TT (Turno Tarde).
Sexo	Categórica	M (Masculino) o F (Femenino).
Año	Numérica	Año académico (2017–2019, 2022-2023).
Calificaciones por materia	Numérica (Entera)	Notas trimestrales obtenidas en 14 materias. Rango (1-10)
Estado Final	Categórica	Indica si el estudiante fue promovido o no al año siguiente (valores posibles: Promovido y Repite)

Resultados del Aprendizaje No Supervisado del 2do año de ciclo básico

Se trabajó con un conjunto de 14 atributos correspondientes a las calificaciones de alumnos en las materias del segundo trimestre de 2do año de ciclo básico, de los periodos 2017, 2018, 2019, 2022 y 2023.

Siguiendo el pipeline de la Figura 1, se estandarizaron los atributos numéricos y se aplicó PCA sobre los mismos. Se seleccionaron las primeras seis componentes las cuales explican el 71% de la variabilidad total de los datos. Las mismas fueron utilizadas como atributos de entrada para los modelos no supervisados. La configuración de los mismos y los resultados de las métricas se presentan en los Anexos A.2 y A.3 respectivamente.

Anexo A.2. Configuración de modelos

Modelo	Parámetros
K-Means	n_clusters = 3 max_iter = 300 n_init = 10 random_state = 0
BIRCH	n_clusters = 3 threshold = 0,05; 0,1; 0,2; 0,3; 0,4 branching_factor = 3; 5; 10; 15; 20; 30

Anexo A.3. Resultados de las métricas de agrupamiento

Modelo	Silhouette Score	Calinski–Harabasz	Davies Bouldin
K-Means	0,238	310,11	1,355
BIRCH	0,229	279,86	1,318

Resultados del Aprendizaje Supervisado del 2do año de ciclo básico

Para la construcción de los modelos, se utilizaron las calificaciones de seis materias del segundo trimestre — Lengua II, Física, Ciencias Biológicas II, Artística II, Historia II y Taller de Preparación II — correspondientes a los años 2017, 2018, 2019, 2022 y 2023 de 2do año de ciclo básico. Estos atributos

fueron seleccionados mediante la técnica Feature Importance. Por otro lado, la variable objetivo a predecir es Resultado, la cual toma el valor 1 para "Promovido" y 0 para "Repite".

El anexo A.4 indica las configuraciones de hiperparámetros de los tres modelos predictivos desarrollados. Por otra parte, los anexos A.5 y A.6 presentan la performance de los mismos, a nivel de modelo y de clase.

Anexo A.4. Mejores hiperparámetros de los modelos supervisados

Modelo	Mejores Hiperparámetros	Tiempo de CPU
Random Forests (Bayes Search)	• max_depth: 5 • max_features: log2 • min_samples_leaf: 1 • min_samples_split: 10 • n_estimators: 100	15 minutos, 48 segundos
XGBoost (Randomized Search)	• subsample: 0,8 • n_estimators: 500 • max_depth: 9 • learning_rate: 0,01 • gamma: 5	2 minutos, 25 segundos
ELM (Grid Search)	• hidden_units = 1000 • activación = Sigmoid • C = 0,001 • random_type: normal	1 hora, 39 minutos

Anexo A.5. Resultados de métrica de clasificación a nivel de modelo

Modelo	Accuracy (%)
XGBoost	87
ELM	88
RF	87

Anexo A.6. Resultados de métricas de clasificación a nivel de clase

Modelo	Clase	Precision (%)	F1-Score (%)	Recall (%)
XGBoost	Repite (0)	84	76	70
	Promovido (1)	88	91	94
ELM	Repite (0)	83	75	68
	Promovido (1)	90	93	96
RF	Repite (0)	78	72	68
	Promovido (1)	90	92	96