UNIVERSIDAD NACIONAL DE SAN MARTIN

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

Original article
Artículo original
Jan-Jun, 2026

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

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

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

Received: 10 Oct. 2025 | **Accepted:** 25 Dec. 2025 | **Published:** 20 Jan. 2026

Corresponding author*: cmartinez@di.unsa.edu.ar

How to cite this article: Zalasar, A. M., Aramayo, R. & Martínez, C. A. (2026). Predictive analysis and modeling of academic performance using machine learning techniques in a secondary education institution. *Revista Científica de Sistemas e Informática*, *6*(1), e1212. https://doi.org/10.51252/rcsi.v6i1.1212

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.

Keywords: machine learning; data science; secondary education; python

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.

Palabras clave: aprendizaje automático; ciencia de datos; educación secundaria; python

© The authors. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

The analysis of academic performance is currently one of the fundamental issues and a major concern that educational institutions must address. Over time, it has been studied from two main perspectives: data related to the school as an educational system and the characteristics that students exhibit based on their social context. However, it has not yet been possible to fully identify and understand the variables that influence academic performance.

In the last decade, research on academic performance and student dropout has grown significantly in Argentina. Most of these studies have focused on higher education, without providing, to date, an integrative analysis that allows for general conclusions about the state of knowledge in the field. A large part of the academic production on this issue has addressed academic performance indirectly by studying variables associated with both dropout and student achievement, but without offering a comprehensive view that integrates both aspects (García, 2014).

Nevertheless, much of this research has concentrated on higher education, leaving secondary education less explored. It is therefore necessary to also understand the specific dynamics of secondary education, especially in regional contexts such as Salta, where socioeconomic and school conditions pose significant challenges.

In this context, Data Mining (Ibarra, 2020) emerges as a promising tool for analyzing academic performance with the aim of addressing various problems, such as student achievement, dropout, and attrition. In recent years, Educational Data Mining (EDM) has shown a growing impact on the analysis of academic performance and the identification of at-risk students, according to a recent review (Romero & Ventura, 2020). EDM has become a useful tool for discovering relevant patterns from educational data.

An example of this is the study conducted at the Universidad Técnica de Manabí (Saltos-Mero & Cruz-Felipe, 2024), where the CRISP-DM methodology was applied to analyze the academic performance of students in the "Gastronomy and Tourism" and "Economics" degree programs, using supervised learning methods such as Decision Trees, Random Forests, Neural Networks, and Support Vector Machines (SVM). The models were implemented in Python, and after a comparative evaluation process, the Random Forest–based model was found to deliver the best performance, achieving *accuracy* values of 83% and 86%, respectively.

Similarly, Guanin-Fajardo et al. (2024) applied the CRISP-DM methodology to predict the academic performance of university students using a dataset of 6.690 records with academic and socioeconomic variables. Among the evaluated methods, XGBoost achieved the best results, reaching an *AUC* value of 87.75%, demonstrating high predictive capability. In addition, the model enabled the extraction of interpretable rules from decision trees, facilitating practical application. The study highlights the importance of implementing early predictive models to strengthen student retention strategies. The methodology is replicable in other academic contexts and demonstrates the value of combining accuracy with interpretability.

Finally, Bellaj et al., 2024 developed supervised learning models to predict academic performance based on techniques such as SVM, Random Forests, XGBoost, K-NN, and Naïve Bayes. The best-performing model was XGBoost, followed by an Ensemble Voting Classifier (EVC). The authors emphasize the importance of hyperparameter optimization to improve the accuracy of predictive models. They also note that variables such as prior academic performance, interaction with virtual

platforms, and sociodemographic factors significantly influence predictions. This work contributes to the development of early warning systems in higher education.

The purpose of this study is to identify trends and patterns in academic performance through Exploratory Data Analysis (EDA) (Tukey, 1977) as well as to predict student performance using supervised learning models based on techniques such as Random Forests (Breiman, 2001), XGBoost (Chen & Guestrin, 2016), and Extreme Learning Machine (Huang et al., 2006; Wang et al., 2022). In addition, unsupervised learning models based on K-Means (MacQueen, 1967) and BIRCH (Zhang et al., 1996) were used to identify groups of students with similar characteristics, facilitating corrective measures and segmentation for timely interventions.

2. MATERIALS AND METHODS

2.1. Case Study

The present study was conducted at the Technical Secondary School No. 3100 "República de la India", located in the province of Salta, Argentina. The research followed a descriptive and correlational approach, with a non-experimental design.

2.2. Dataset

The study was based on a dataset corresponding to the entire population of first-year students in the basic cycle of the institution. The dataset included 787 student records covering 13 subjects (Art, Biological Sciences, Spanish Language I, History I, Technology I, Technical Drawing, Physical Chemistry, Foreign Language I, Ethics and Citizenship I, Geography I, Mathematics I, Preparatory Workshop I, and Physical Education). The records span the academic years 2017, 2018, 2019, 2022, and 2023.

Table 1 presents the attributes related to the institution's academic records, which were considered for analyzing academic performance and detecting patterns in student grades. The data were obtained from official institutional sources and constituted the primary data collection instrument. Each record was associated with a student through their first name, last name, and a unique identifier.

Table 1. Dataset Attributes

Attribute	Type	Description
Class	Categorical	Student's grade level and class group (e.g., 1st A, 1st B, etc.)
Shift	Categorical	Morning shift (TM) or afternoon shift (TT)
Gender	Categorical	Male (M) or Female (F)
Academic Year	Numerical	Academic year (2017–2019, 2022–2023)
Subject Grades	Numerical (Integer)	Quarterly grades obtained in 13 subjects. Range: 1–10
Final Status	Categorical	Indicates whether the student was promoted to the next academic year (possible values: Promoted or Repeating)

Each subject is disaggregated into three separate columns corresponding to the three terms of the academic year, identified by the suffixes "_1t", "_2t", and "_3t" (e.g., Mathematics_1t, Mathematics_2t, etc.).

Appendix A1 describes the dataset related to second-year basic cycle students.

2.3. Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was adopted as the methodological framework for this study (Chapman, 2000), due to its structured and flexible nature for Data Mining projects. This approach proposes a cyclical process composed of different phases that enable the transformation of domain-specific objectives into analytical models with practical value for decision-making.

First, the *business understanding* phase made it possible to define the research problem and establish as the main objective the analysis of students' academic performance, with the aim of identifying relevant patterns in their grades. Based on these objectives, the study proceeded to the *data understanding* phase, in which the available records were collected and an initial exploration was conducted to describe their general characteristics, assess data quality, and detect potential inconsistencies, missing values, or outliers.

Subsequently, during the *data preparation* phase, the final dataset to be used in the analysis was constructed through the selection of relevant attributes and the application of data cleaning and transformation processes, ensuring its suitability for modeling. In the *modeling* phase, analytical techniques aimed at both the description and prediction of academic performance were applied, selecting and tuning the most appropriate algorithms according to the defined objectives.

Finally, the *evaluation* phase allowed for the analysis of the performance and usefulness of the obtained models, verifying their consistency with the established objectives and their contribution as decision-support tools in the educational context. In this way, CRISP-DM provided a comprehensive methodological framework that systematically guided the development of the study and the generation of relevant knowledge.

Data processing, visualization, and analysis tasks, as well as the development of machine learning models, were carried out using Python scripts and libraries such as Pandas, NumPy, Scikit-Learn, and Matplotlib. This approach enabled a sequential and CRISP-DM–guided structure for the study.

2.4. Exploratory Data Analysis

This study relied on a set of fundamental techniques for data analysis. First, data cleaning and transformation tasks were performed in order to prepare the information for subsequent analysis, ensuring its consistency and suitability for the study context.

Subsequently, Exploratory Data Analysis (EDA) provided an initial approach to the structure and quality of the records, enabling the detection of irregularities and a general understanding of the behavior of the dataset variables.

EDA constitutes a cross-cutting stage prior to the implementation of unsupervised and supervised learning pipelines, and its objective is to understand the data structure, detect outliers, and guide preprocessing and modeling decisions.

2.5. Unsupervised Learning

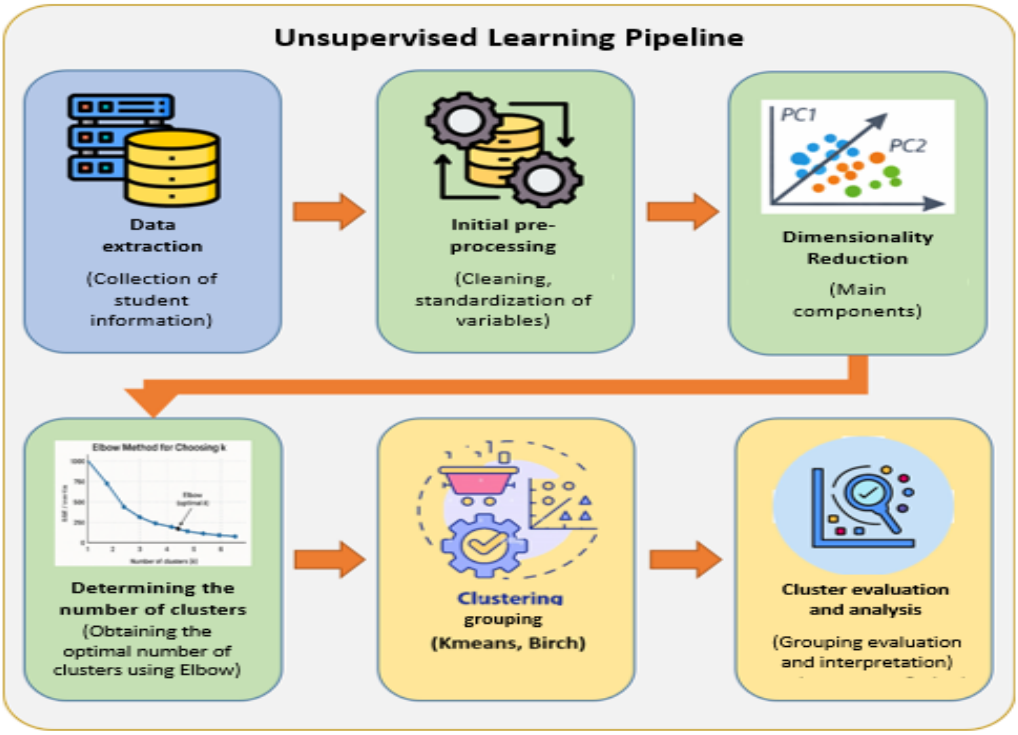

Figure 1. Proposed Unsupervised Learning Pipeline

Unsupervised learning (Ghahramani, 2003) was employed to identify patterns and group observations with similar characteristics through the development of models based on techniques such as K-Means and BIRCH. In addition, Principal Component Analysis (PCA) (Yang et al., 2018) was used as a dimensionality reduction technique.

The complete unsupervised learning pipeline is presented in Figure 1, sequentially illustrating the stages of the process from data extraction to cluster evaluation. As shown in the figure, the process begins with data extraction and preprocessing, followed by dimensionality reduction for clustering and visualization, determination of the optimal number of clusters using the Elbow method, application of clustering techniques, and evaluation of the resulting clusters.

2.5.1. Objective of Unsupervised Learning

At this stage of the study, an analysis was conducted with the objective of identifying groups of students with similar characteristics in their academic performance. The analysis aims to generate valuable knowledge for early pedagogical decision-making, enabling the implementation of corrective actions based on second-term grades before the end of the academic year.

2.5.2. Dataset and Variables Used

For this analysis, a set of 13 attributes corresponding to the numerical grades obtained by first-year basic cycle students in different subjects during the second term was used, considering the academic years 2017, 2018, 2019, 2022, and 2023. The attributes considered were: Art_2t, Biological Sciences_2t, Spanish Language I_2t, History I_2t, Technology I_2t, Technical Drawing_2t, Physical Chemistry_2t, Foreign Language I_2t, Ethics and Citizenship I_2t, Geography I_2t, Mathematics I_2t, Preparatory Workshop I_2t, and Physical Education_2t.

2.5.3. Data Preprocessing

A data preprocessing adjustment was performed, consisting of data standardization and the application of PCA.

Standardization was applied in order to mitigate the impact of extreme values, preventing attributes with larger scales from dominating the clustering process, without removing real observations from the dataset. Subsequently, PCA was applied to the standardized data so that all attributes had equal importance in the projection. The first five principal components, which explain 71% of the data variability (PC1 = 0.448; PC2 = 0.079; PC3 = 0.065; PC4 = 0.061; PC5 = 0.050), were selected, and the first two were used for visualization in a two-dimensional space. This approach allowed a concise handling of the information contained in the 13 original attributes and facilitated both the visualization and graphical interpretation of the resulting clusters.

2.5.4. Selection of the Number of Clusters

To determine the optimal number of clusters, the Elbow method was applied (Thorndike, 1953; Syakur et al., 2018). As shown in Figure 2, the curve exhibits a pronounced decrease in inertia (WCSS) up to k = 3, after which the slope becomes less steep, indicating an inflection point. Based on this criterion, k = 3 was selected as the appropriate number of clusters.

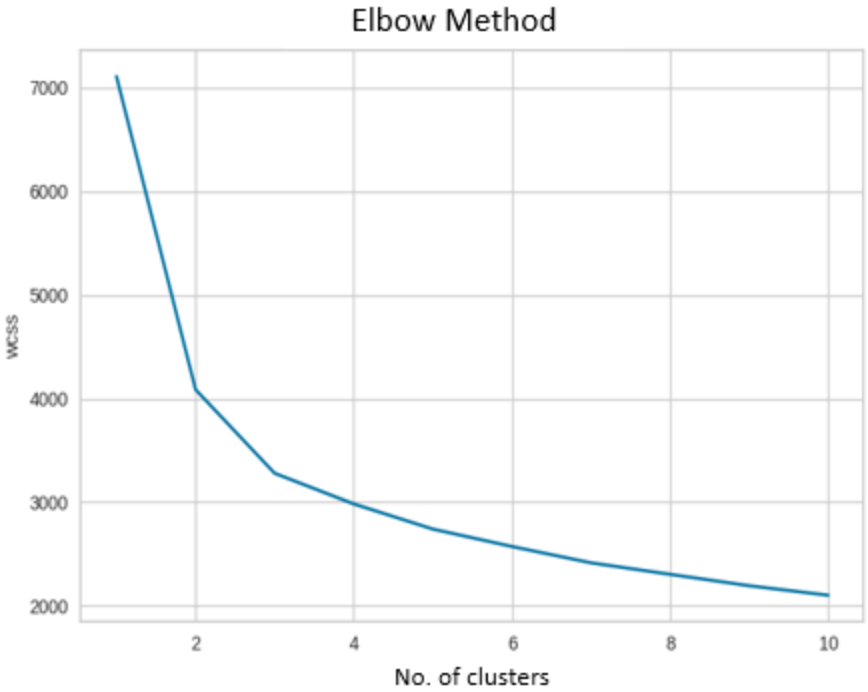

Figure 2. Application of the Elbow Method to the Dataset

2.5.5. Data Clustering

Unsupervised learning models based on the K-Means and BIRCH techniques were developed, allowing the comparison of different methodological approaches. Both techniques were selected due to their conceptual differences in the clustering process. K-Means is a partition-based clustering technique widely used for numerical and standardized datasets, which produces compact and well-defined clusters. In contrast, BIRCH is a hierarchical clustering technique that can capture more flexible data structures, making it suitable for contrasting the results obtained with K-Means. The comparison between both models made it possible to assess the stability and coherence of the groupings under different methodological assumptions.

Table 2 shows the selected configuration for both models.

Table 2. Configuration of Unsupervised Models

Model	Parameters
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. Evaluation and Interpretation

The performance of the unsupervised models was assessed using three specific clustering evaluation metrics: the Calinski–Harabasz index (Calinski & Harabasz, 1974), the Silhouette coefficient (Rousseeuw, 1987), and the Davies–Bouldin index (Ros et al., 2023). The Calinski–Harabasz index measures the ratio between inter-cluster dispersion and intra-cluster dispersion, where higher values indicate better clustering quality. In contrast, the Silhouette coefficient evaluates the internal cohesion of clusters and their separation from other groups, with values close to 1 indicating a well-defined clustering structure. Finally, the Davies–Bouldin index analyzes the relationship between the internal dispersion of each cluster and the distance to the nearest cluster, where lower values represent more compact and better-separated partitions. As a general rule, values close to 0 represent very compact and well-separated clusters, values between 1 and 2 indicate moderate cohesion and separation, and values greater than 2 reflect poorly defined clusters or significant overlap.

The methodological procedure developed made it possible to apply unsupervised learning techniques for the identification of patterns in academic performance. The results derived from this analysis are presented in Section 3.

2.6. Supervised Learning

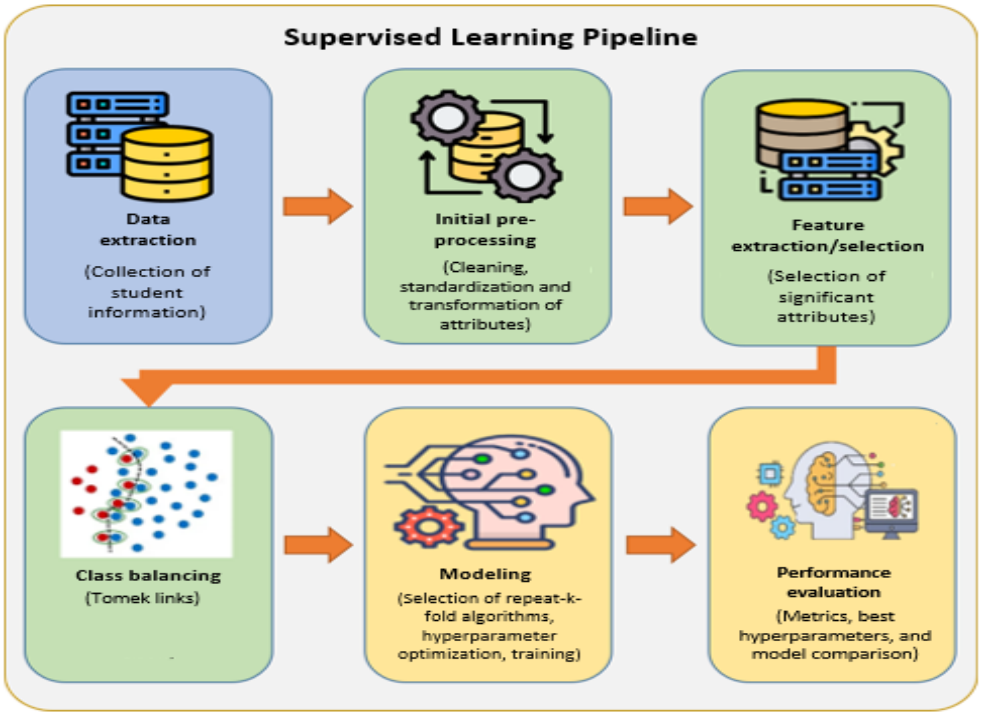

Figure 3. Proposed Supervised Learning Pipeline

Supervised learning techniques such as Random Forests, XGBoost, and Extreme Learning Machine were employed for the development of predictive models (Plathottam et al., 2023). The complete supervised learning pipeline is presented in Figure 3, where the stages of the process are described sequentially.

As shown in the figure, the process begins with data extraction and initial preprocessing. Subsequently, feature extraction and selection are performed, followed by class balancing. Next, the modeling stage is carried out using the selected supervised learning techniques, and finally, model performance is evaluated using appropriate metrics, providing objective criteria to assess their effectiveness and consistency with the defined objectives.

2.6.1. Objective of Supervised Learning

The objective of this stage is to predict whether a student will be promoted to the next academic year using supervised learning models. To this end, predictive models based on Random Forests (RF), XGBoost, and Extreme Learning Machine (ELM) were developed and evaluated in order to identify students at risk at an early stage and support pedagogical decision-making.

2.6.2. Selection of attributes or features

For model construction, the attributes Gender, Academic Year, and the grades from the first and second terms corresponding to the academic years 2017, 2018, 2019, 2022, and 2023 for first-year basic cycle students were initially considered. These attributes were previously standardized, as they represent a critical period for implementing interventions before the end of the academic year.

The selected significant features correspond to second-term grades from six subjects: History I_2T, Spanish Language I_2T, Technical Drawing_2T, Technology I_2T, Mathematics I_2T, and Foreign Language I_2T, for the academic years 2017, 2018, 2019, 2022, and 2023. These features were selected using the Feature Importance technique (Breiman, 2001), obtained from a Random Forest model trained with 250 decision trees. Importance values were computed for all available attributes, and only the six second-term subjects with the highest contribution to academic performance prediction were retained. The selection focused on the second term, as it provides more recent information about student performance and enables predictive decision-making before the beginning of the third term. Figure 4 illustrates the application of this technique and the relative contribution of each attribute.

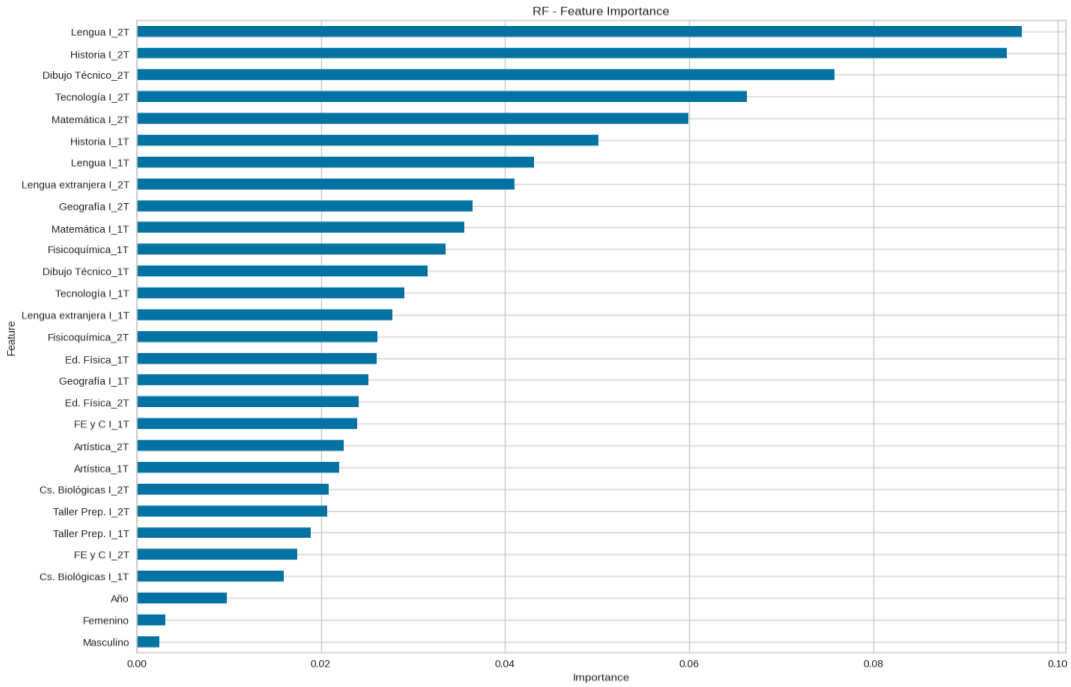

Figure 4. Application of Feature Importance

2.6.3. Class Balancing

Given that the dataset exhibited a strong class imbalance, with a significant majority of promoted students compared to a smaller proportion of retained students, the Tomek Links technique was applied (Leng et al., 2024). This method identifies pairs of instances from different classes that are mutual nearest neighbors and removes those belonging to the majority class (in this case, promoted students).

As a result, the majority class was reduced from 574 to 554 instances, while the minority class remained unchanged. This slight reduction helped mitigate class imbalance, promoting more balanced learning in the supervised models and contributing to improved predictive performance for the minority class, without removing critical information from the majority class.

Data augmentation techniques were not employed, as only real academic data were used and generating synthetic observations was deemed inappropriate. The creation of synthetic data could introduce artificial grade values, potentially affecting both the pedagogical validity of the results and the reliability of the model predictions.

2.6.4. Data Splitting (Train/Test Split)

The data were split by allocating 75% of the dataset for training and 25% for testing the predictive models.

2.6.5. Validation Strategy (Repeated K-Fold)

In order to ensure robust validation and minimize the risk of overfitting, the Repeated K-Fold Cross-Validation technique was applied (Kohavi, 2001) with the following configuration:

- n_splits: 10 (number of splits)
- n_repeats: 10 (number of repetitions)
- random_state: 42 (seed value to ensure reproducibility)

This technique was chosen over alternatives such as simple K-Fold or Stratified K-Fold because it provides a more stable estimation of model performance by reducing metric variability across different dataset partitions. Each observation participates in multiple training and validation sets, ensuring a more comprehensive analysis of model behavior across the entire dataset and improving the reliability of hyperparameter selection and the reported final performance. The Repeated K-Fold configuration is suitable for a dataset with fewer than 1000 samples, balancing the stability of supervised model performance metrics and the computational cost of model training.

2.6.6. Hyperparameter Optimization

In order to obtain the best possible models, hyperparameter optimization techniques were employed to improve performance and adaptability. Specifically, Grid Search (Belete & Huchaiah, 2022; Ogunsanya et al. 2023), Randomized Search (Breiman, 2001) and BayesSearch (Snoek et al., 2012) were used, as these approaches offer different strategies for exploring the configuration space and selecting those that maximize model quality.

Table 3 presents the specific values tested for each hyperparameter.

Table 3. Techniques and Hyperparameter Values (Martínez et al., 2025)

ML Model	Technique	Hyperparameters	Values
ELM	Grid Search	Number of neurons in hidden layers	1000; 2000; 3000; 4000; 5000; 6000; 7000
		Activation function	sigmoid; relu; sin; leaky_relu; tanh
		C (regularization parameter)	0.001; 0.01; 0.1; 0.3; 0.5; 0.7; 0.9; 1; 1.3; 1.5; 2
		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

Table 4 presents the optimal hyperparameters used for training the predictive models. The selection of these configurations was based on both global and class-level performance metrics during the optimization process, considering *Accuracy* as the global metric and *Precision*, *Recall*, and *F1-Score* at the class level. Additionally, the CPU time associated with each optimal configuration is reported.

Table 4. Best Hyperparameters of the Supervised Models

Model	Best Hyperparameters	CPU Time
Random Forests (Bayes Search)	• max_depth: 30 • max_features: sqrt • min_samples_leaf: 5 • min_samples_split: 10 • n_estimators: 100	18 minutes, 32 seconds
XGBoost	• subsample: 0.7 • n_estimators: 500	2 minutes, 56 seconds

(Randomized Search)	• max_depth: 11 • learning_rate: 0.01 • gamma: 3	
ELM (Grid Search)	• Hidden_Units = 1000 • Activación = Sigmoid • C = 0.1 • random_type: normal	1 hour, 38 minutes y 27 seconds

With the established methodology, the final supervised models were trained to predict academic performance. The results obtained are presented and analyzed in Section 3.

3. RESULTS AND DISCUSSION

3.1. Results of the Exploratory Data Analysis (EDA)

Distribution of Promoted and Non-Promoted Students

Table 5 shows the distribution of promoted and non-promoted first-year basic cycle students for the academic years 2017–2019 and 2022–2023. The objective is to provide an overall view of academic performance, without segmentation by subject or term, as a starting point for more specific analyses.

It can be observed that the proportion of students who were promoted to the next academic year is higher than that of those who were not, reflecting a generally positive overall performance.

Table 5. Number and Percentage of Promoted and Retained Students

Student Status	Count	Percentage (%)
Promoted	560	74.5
Retained	192	25.5

Histograms by Subject for the First and Second Terms

To analyze student performance in each subject, Figs. 5 and 6 present histograms showing the distribution of grades for the first and second terms of the academic years 2017, 2018, 2019, 2022, and 2023.

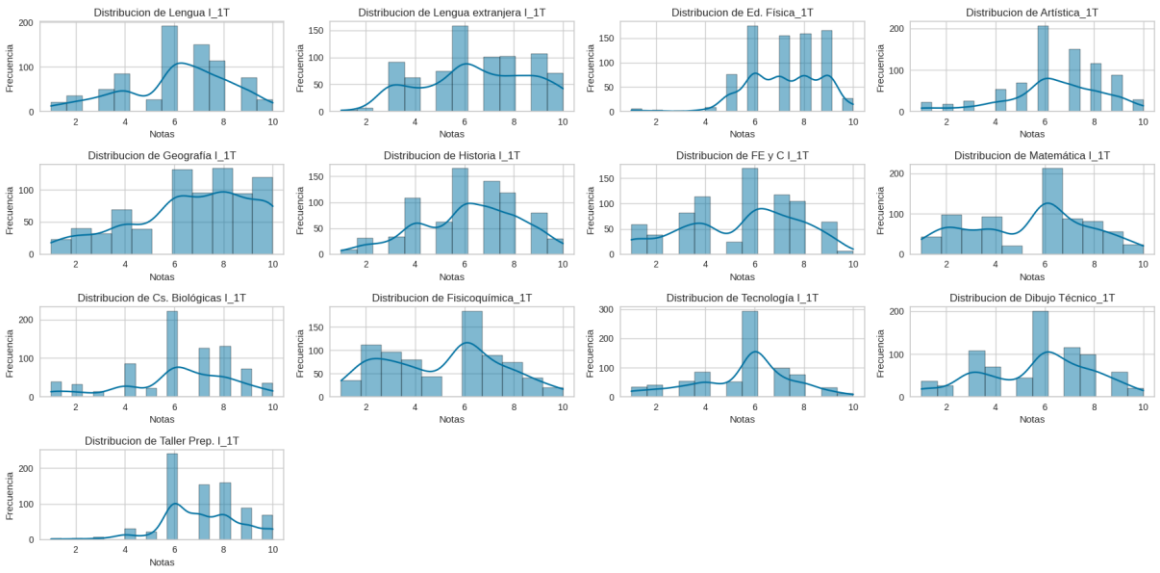

Figure 5. Distribution of First-Term Grades

In the first term, different patterns were identified in the distribution of the data. Subjects such as Spanish Language, History, Foreign Language, Ethics and Citizenship, Biological Sciences, and Workshop exhibit slightly symmetric or centered distributions, with means between 6.16 and 7.13 and medians around 6 or 7, concentrating most grades within the [6–8] range.

In contrast, subjects such as Mathematics, Physical Chemistry, Technology, and Technical Drawing show positively skewed distributions, characterized by lower means (between 5.13 and 5.71) and first quartiles between 3 and 4, with a higher concentration of low grades (between 3 and 6), reflecting greater academic difficulties, particularly in Mathematics and Physical Chemistry.

Conversely, subjects such as Art, Geography, and Physical Education exhibit negatively skewed distributions, with means above 6.4, third quartiles between 8 and 9, and maximum values close to 10, indicating generally favorable performance in these subjects.

Finally, in some cases such as Physical Education, Workshop, and Technical Drawing, multimodal distributions are observed, suggesting the presence of subgroups with differentiated performance within the classroom.

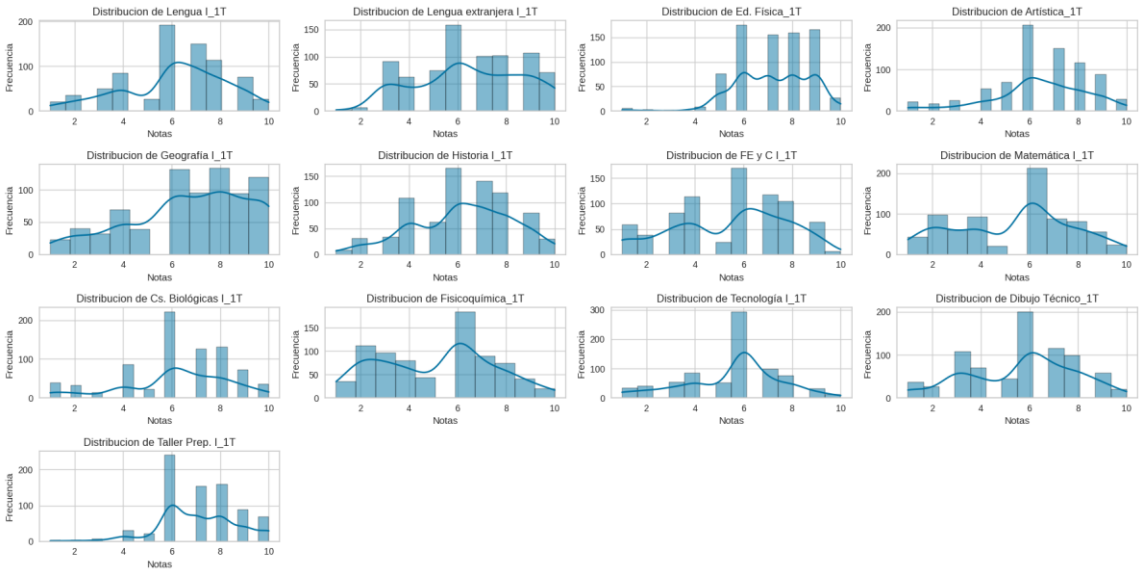

Figure 6. Distribution of Second-Term Grades

Based on the observations of the second-term grade distributions, it can be highlighted that the grades once again reflect a variety of distributions and asymmetries, revealing diverse academic performance depending on the subject. Physical Education, Arts, Ethics and Citizenship Education, Geography, and Workshop exhibit negative skewness, with mean values between 6.33 and 7.30, medians close to 7–8, and third quartiles around 8 or 9, indicating a predominance of high grades and a low failure rate. Other subjects such as Language, History, Foreign Language, and Biological Sciences show centered distributions, with means close to 6, medians equal to 6, and interquartile ranges concentrated between 5 and 8, suggesting intermediate and relatively stable academic performance. In contrast, Mathematics, Physicochemistry, Technology, and Technical Drawing display positive skewness, with mean values below 5.75, first quartiles between 3 and 5, and greater dispersion of grades. In particular, Physicochemistry (mean = 5.15) and Technical Drawing (mean = 5.38) show a higher proportion of students with low academic performance. Furthermore, bimodality is observed in the histograms of these subjects, suggesting a clear

differentiation between students who successfully understand the content and those who experience greater learning difficulties.

Comparison with the First Term

Compared to the first term, a slight overall improvement is observed in several subjects, although certain asymmetries and dispersion persist in subjects such as Mathematics, Physicochemistry, and Technology. In general, grades remain concentrated between 6 and 8, with some subjects showing improved overall performance. The most critical subjects that continue to require special attention are Mathematics, Physicochemistry, and Technology, due to their higher dispersion and the number of students with low performance at academic risk. These patterns suggest a generally positive performance, even higher than that of the first term, although with characteristic variability.

Multivariate Data Analysis

This analysis examines the relationships among grades across different subjects by means of a correlation matrix, using aggregated data from the periods 2017–2019 and 2022–2023. The objective is to identify dependencies and relevant patterns that contribute to a better understanding of students' academic performance. Below, the correlation matrix corresponding to the second quarter resented (Figure 7).

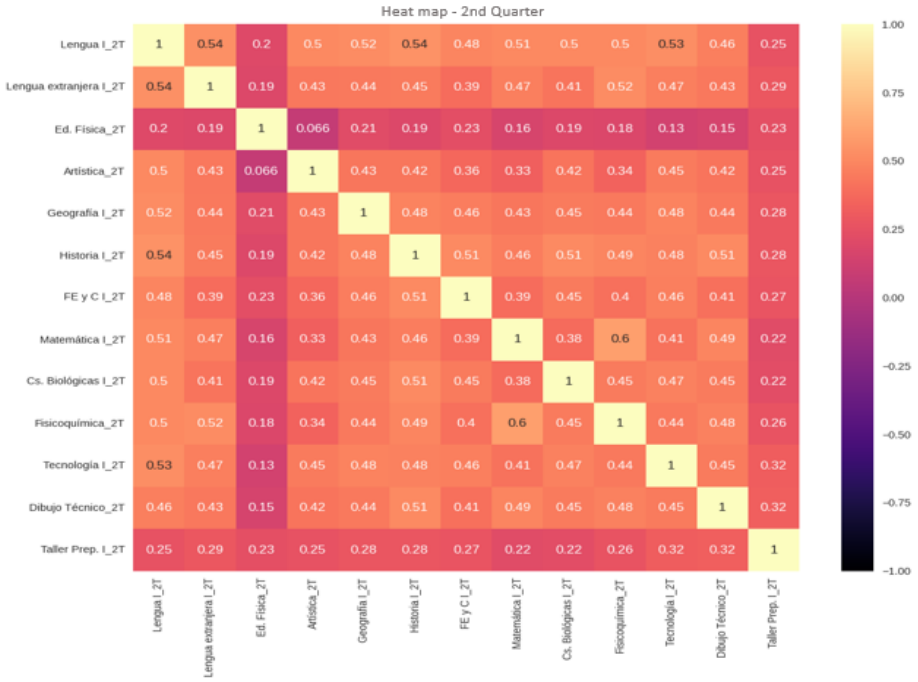

Figure 7. Correlation matrix – Second quarter

Moderately high and positive correlations are observed among Language, Foreign Language, History, Mathematics, Physical Chemistry, and Technology, both within these subjects and with other subjects, albeit with weaker and moderate correlations. This pattern suggests that students who perform well in one of these subjects tend to perform well in the others, possibly due to shared cognitive skills such as reading comprehension, logical reasoning, and analytical ability.

During the exploratory analysis, certain outlier grades and some incomplete records were identified. Outliers were retained, as they correspond to actual academic results, and their removal could introduce bias into the analysis. In contrast, incomplete records were removed from the

dataset due to missing values that prevented their proper use in the applied analysis and modeling techniques.

Furthermore, the target attribute of the study corresponds to the student's Final Status, indicating whether the student was promoted to the next year or had to repeat the course. For use in supervised models, this categorical attribute was binary-encoded, assigning 1 to "Promoted" and 0 to "Retained".

3.2. Results of Unsupervised Learning

This section presents the evaluation of the K-Means and BIRCH models using the Silhouette, Calinski–Harabasz, and Davies–Bouldin metrics. Through numerical comparison, the model with the best performance is selected, and its graphical representation is analyzed to observe the distribution of students within the identified clusters.

Table 6 shows the metric values for both models.

Table 6. Clustering Metric Results

Model	Silhouette Score	Calinski–Harabasz	Davies Bouldin
K-Means	0.2518	450.236	1.32
BIRCH	0.2265	408.79	1.36

The values listed in Table 6 indicate that both models are able to identify an interpretable clustering structure within the dataset. In particular, K-Means achieves higher values in the Silhouette coefficient (0.2518) and the Calinski–Harabasz index (450.236), suggesting greater internal cohesion of the clusters and better separation between the formed groups compared to BIRCH. Additionally, the Davies–Bouldin index shows a lower value for K-Means (1.32), indicating lower relative internal dispersion with respect to the nearest cluster. Considering all three metrics together, it is concluded that the K-Means–based model provides adequate and consistent clustering quality for the performed analysis.

Discussion of Results

Figure 8 shows the students grouped using the K-Means–based model, employing the first two Principal Components (PCA). The points in the plot represent the students' grades, previously standardized.

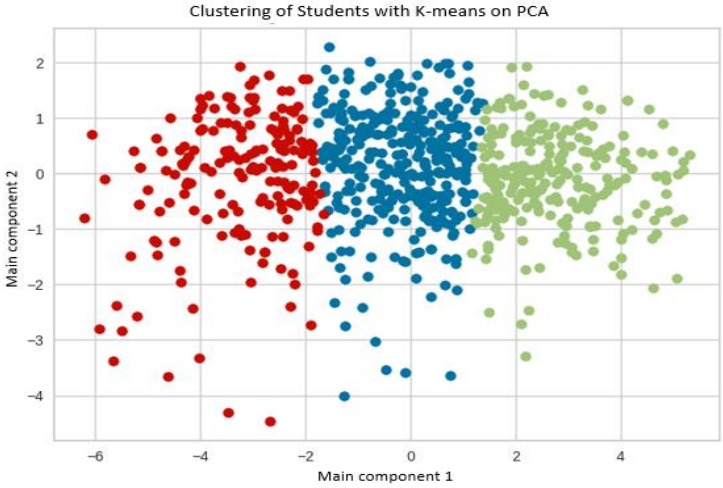

Figure 8. Data Clustering Using the K-Means Model

Based on the cluster plot in the two-dimensional space, three groups of students were identified:

- Class 0 (Blue) – Medium Performance – 337 students: The points corresponding to this class are concentrated around intermediate values of PC1, which is mainly influenced by theoretical subjects such as Language I (10.36), History I (9.62), Physical Chemistry (8.91), Technology I (8.84), Geography I (8.69), and Foreign Language I (8.61). The higher concentration of points along PC2, dominated by practical subjects such as Physical Education (66.07), Workshop Preparation (23.73), and Arts (5.18), corresponds to students with average academic performance.
- Class 1 (Green) – High Performance – 247 students: The points of this class are predominantly located in positive values of both PC1 and PC2, showing a relatively compact distribution. Positive PC1 values suggest excellent performance in theoretical subjects (Language, History, Physical Chemistry, Technology, Geography), while positive values in PC2 also reflect good performance in practical or physical subjects (Physical Education, Workshop, Arts). This group corresponds to students with superior academic performance and more homogeneous profiles across the evaluated subjects.
- Class 2 (Red) – Low Performance – 191 students: The points associated with this class are concentrated in negative PC1 values, indicating low performance in theoretical subjects, with greater dispersion along PC2, reflecting variability in practical subjects. This pattern suggests a group of students with low academic achievement.

In this study, the application of K-Means allowed the identification of three academic performance clusters (low, medium, and high) using the Elbow method, with a Silhouette coefficient of 0.2518, a Calinski-Harabasz index of 450.236, and a Davies–Bouldin index of 1.32. The distribution across clusters was 191, 337, and 247 students, respectively, ensuring a balanced and representative segmentation.

Compared to recent literature, Mohamed Nafuri et al. (2022) identified five clusters using K-Means, but the Silhouette values (0.16–0.192) and Calinski-Harabasz indices (17.358–24.946) were considerably lower, indicating less separation and internal density among the groups. Meanwhile, Amalia et al. (2021) also reported three optimal clusters with validation metrics ranging from 0.340–0.514 for Silhouette and 27.174–84.529 for Calinski-Harabasz; however, their datasets were small (20–50 samples per model), limiting cluster representativeness. These results suggest that the three-group segmentation in our study provides greater clarity and educational applicability, combining adequate separation between clusters with sufficiently large sample sizes for reliable analysis.

Furthermore, the obtained structure confirms the potential of this analysis to support early pedagogical decisions, as it allows recognition of different academic performance levels and guides corrective actions before the end of the academic year.

Appendices A2 and A3 present results of unsupervised learning using the second-year basic cycle dataset.

3.3. Results of Supervised Learning

This section presents the results obtained by the proposed models for predicting academic performance.

Classification metrics including *F1-Score* (Rainio et al., 2024), *Precision*, *Recall*, and *Accuracy* (Shobha & Rangaswamy, 2018) were employed to evaluate the performance of the developed models. These metrics provide an overall view of the prediction quality, taking into account total correct predictions and the balance between classification errors.

Tables 7 and 8 show the performance of the best supervised models, both at the overall level and by class (Class 0: Retained; Class 1: Promoted).

Table 7. Classification Metric Results at the Model Level

Model	Accuracy (%)
XGBoost	89
ELM	87
RF	87

Table 8. Classification Metric Results by Class

Model	Class	Precision (%)	F1-Score (%)	Recall (%)
XGBoost	Retained (0)	84	80	76
	Promoted (1)	91	92	94
ELM	Retained (0)	82	78	74
	Promoted (1)	90	92	93
RF	Retained (0)	84	76	69
	Promoted (1)	88	91	95

According to the results presented in Tables 7 and 8, the XGBoost model outperformed both Random Forests and ELM. Beyond the overall model capability, analyzing performance by class—specifically considering F1-Score values—clearly shows that XGBoost achieves a balanced performance, with an F1-Score of 80% for the "Retained" class and 92% for the "Promoted" class. This indicates that it correctly identifies students in both classes, with particularly strong performance for the majority class.

Compared to ELM (78% and 92%) and RF (76% and 91%), XGBoost demonstrates better balance and classification ability, especially for the minority class. The XGBoost model stands out as the most efficient, achieving the best trade-off between computational time and performance. Thus, it is the strongest option for predicting whether a student will be promoted to the next year.

Additionally, Figure 9 shows the confusion matrix (Menacho Chiok, 2017) of the XGBoost model, selected as the best-performing model. This representation allows direct visualization of the distribution between observed (actual) and predicted classifications for the different categories of the target class, reinforcing the interpretation of the model's behavior in identifying promoted students and those at risk of retention. The confusion matrix is a widely used tool in the literature to evaluate the classification quality of predictive models.

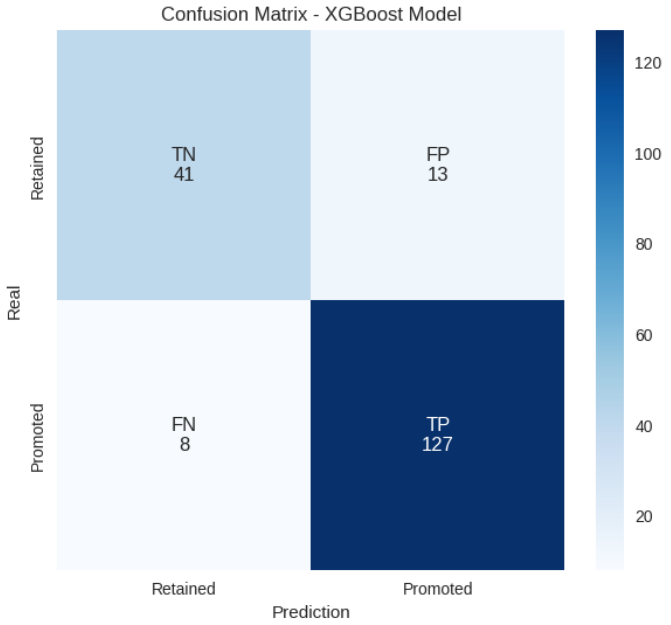

Figure 9. Confusion Matrix of the XGBoost Model

Discussion of results

From a pedagogical perspective, an *F1-Score* of 0.80 for the *"Retained"* class allows the model to be used as an early warning system, useful for guiding educational interventions before retention occurs, maintaining a reasonable balance between omissions and misclassifications. At the institutional level, this information can support targeted educational interventions, such as tutoring, pedagogical support, or personalized follow-up, contributing to a more efficient allocation of available resources. In this sense, the model does not replace teacher evaluation but serves as a decision-support tool, strengthening strategies to prevent school dropout and retention.

The results obtained are consistent with the reviewed literature. For example, the study by Saltos-Mero & Cruz-Felipe (2024) addressed a binary classification problem ("Pass" / "Fail") and reported that their Random Forests–based model outperformed other proposed models (Decision Trees, Neural Networks, SVM), achieving *Accuracy* values of 0.86 in Economics and 0.83 in Tourism, confirming the effectiveness of tree-based methods for binary classification problems.

Similarly, Guanin-Fajardo et al. (2024) evaluated XGBoost and RF on a dataset of 6,690 records with class balancing using EasyEnsemble and multiclass classification (Passed, Changed, Dropped Out). There, XGBoost achieved *Accuracy* = 0.7949, *F1-Score* = 0.8306, *Precision* = 0.8214, and *Recall* = 0.8425, while RF showed comparable results. Although the reported values are slightly lower than those in our study, the authors addressed a more complex problem with a larger dataset.

Finally, Bellaj et al. (2024) analyzed a dataset of 480 cases, 16 attributes, and three academic performance levels (Low, Medium, High), applying CRISP-DM, hyperparameter optimization, and 10-fold stratified cross-validation with GridSearchCV, which allowed them to evaluate model robustness across different dataset partitions. Their best models (Voting, XGBoost, and RF with HPO) achieved *Accuracy* = 0.84–0.86, *F1-Score* = 0.85–0.87, *Precision* = 0.84, and *Recall* = 0.84, confirming the effectiveness of tree-based methods even in multiclass classification.

Overall, the comparison indicates that, although datasets and class complexities vary, tree-based models, particularly XGBoost, consistently show superior performance. Moreover, the preprocessing and optimization techniques implemented in our study, such as feature selection, class imbalance handling, and Repeated K-Fold, contribute to increased robustness and reliability of predictions in a binary context, analogous to how cross-validation improves confidence in reported metrics in multiclass studies.

Appendices A3, A4 and A5 present results of supervised learning using the second year of basic cycle dataset.

CONCLUSIONS

Based on the analysis of academic data from first-year students in the basic cycle and following the CRISP-DM methodology, consistent performance patterns were identified, allowing the characterization of differentiated academic profiles through clusters. From an educational perspective, these profiles can be interpreted as distinct levels of academic risk and stability, enabling the design of tailored pedagogical support strategies according to the needs of each student group.

These findings, aligned with the study objectives, confirm the usefulness of applying data mining and machine learning techniques in school contexts to anticipate academic risk situations through predictions and guide more effective and timely pedagogical interventions. In particular, the early identification of students with a higher likelihood of retention allows for intervention before unfavorable academic trajectories are consolidated.

As a main contribution, the results provide a solid foundation for institutional diagnosis and improved decision-making, demonstrating that machine learning techniques such as K-Means and XGBoost are suitable for classifying and predicting student performance. In this sense, the developed models should not be understood as automatic decision-making tools but as analytical support that complements pedagogical insight and teacher expertise. Additionally, they allow the integration of information across different levels of analysis, from general performance profiles to individual predictions, facilitating the planning of personalized monitoring strategies, efficient allocation of educational resources, and early identification of students who could benefit from targeted interventions. In this way, the proposed models contribute to establishing a more systematic, evidence-based approach to academic management, strengthening the institution's capacity to anticipate problems, evaluate outcomes, and design more effective educational policies.

Opportunities for future research include expanding the sample size, incorporating socio-emotional and contextual variables, and analyzing the impact of intervention strategies based on the obtained profiles. Additionally, a future line of work is the development of a web application integrating statistical graphics with the supervised and unsupervised models developed, enabling the institution to have an interactive tool for monitoring, analyzing, and managing academic performance, thus contributing to the construction of a more inclusive, personalized, and effective educational system.

A limitation of this study is the absence of data for 2020 and 2021, a period affected by the COVID-19 pandemic, which may have influenced the observed performance patterns. This limitation could be addressed once the developed solution is in production, using current cohort grades for continuous prediction and retraining of both supervised and unsupervised models.

ACKNOWLEDGEMENTS

Special thanks are due to the authorities of Technical Education School 3100 for providing and facilitating the data set used in this study, whose collaboration was fundamental to the development and validation of the models presented.

FINANCING

This work was partially funded by Project CIUNSa 2735, under the Research Council of the National University of Salta (Argentina).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest related to the development of this study.

AUTHORSHIP CONTRIBUTION

Conceptualization: Martinez, C. Data curation and formal analysis: Zalasar, A. Funding acquisition: Martinez, C. Investigation: Zalasar, A. Methodology and project administration: Martinez, C. Resources: Aramayo, R. Software: Zalasar, A. Supervision: Martinez, C. and Aramayo, R. Validation: Aramayo, R. Visualization: Zalasar, A. Writing – original draft: Aramayo, R. and Zalasar, A. Writing – review and editing: Martinez, C. and Zalasar, A.

REFERENCES

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

APPENDICES

Second-Year Basic Cycle Dataset

Appendix A.1 presents the attributes related to the academic records of the second year of the basic cycle at the institution.

Appendix A.1. Study Attributes – 2nd Year BC

Variable	Type	Description
Course	Categorical	Year and division of the student (e.g., 2°1°, 2°2°, etc.)
Shift	Categorical	TM (Morning Shift) or TT (Afternoon Shift)
Gender	Categorical	M (Male) or F (Female)
Year	Numeric	Academic year (2017–2019, 2022–2023)
Grades by Subject	Numeric (Integer)	Quarterly grades obtained in 14 subjects. Range (1–10)
Final Status	Categorical	Indicates whether the student was promoted to the next year or not (possible values: Promoted and Repeats)

Unsupervised Learning Results for the 2nd Year of the Basic Cycle

A dataset of 14 attributes corresponding to student grades in the subjects of the second trimester of the 2nd year of the basic cycle was used, covering the periods 2017, 2018, 2019, 2022, and 2023.

Following the pipeline shown in Figure 1, the numerical attributes were standardized, and PCA was applied. The first six principal components, explaining 71% of the total data variability, were selected and used as input features for the unsupervised models. The configuration of the models and the metric results are presented in Appendix A.2 and A.3, respectively.

Appendix A.2. Model Configuration

Model	Parameters
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

Appendix A.3. Clustering Metrics Results

Model	Silhouette Score	Calinski–Harabasz	Davies Bouldin
K-Means	0.238	310.11	1.355
BIRCH	0.229	279.86	1.318

Supervised Learning Results for the 2nd Year of the Basic Cycle

For the construction of the models, grades from six subjects of the second trimester—Language II, Physics, Biological Sciences II, Art II, History II, and Preparatory Workshop II—corresponding to the years 2017, 2018, 2019, 2022, and 2023 of the second year of the basic cycle were used. These attributes were selected using the Feature Importance technique. The target variable to be predicted is Outcome, which takes the value 1 for Promoted and 0 for Repeats.

Appendix A.4 shows the hyperparameter configurations of the three predictive models developed. Additionally, Appendix A.5 and A.6 present their performance at the model level and at the class level, respectively.

Appendix A.4. Best Hyperparameters of the Supervised Models

Model	Best Hyperparameters	CPU Time
Random Forests (Bayes Search)	• max_depth: 5 • max_features: log2 • min_samples_leaf: 1 • min_samples_split: 10 • n_estimators: 100	15 minutes, 48 seconds
XGBoost (Randomized Search)	• subsample: 0.8 • n_estimators: 500 • max_depth: 9 • learning_rate: 0.01 • gamma: 5	2 minutes, 25 seconds
ELM (Grid Search)	• hidden_units = 1000 • activación = Sigmoid • C = 0.001 • random_type: normal	1 hour, 39 minutes

Appendix A.5. Classification Metric Results at the Model Level

Model	Accuracy (%)
XGBoost	87
ELM	88
RF	87

Appendix A.6. Classification Metric Results at the Class Level

Model	Class	Precision (%)	F1-Score (%)	Recall (%)
XGBoost	Retained (0)	84	76	70
	Promoted (1)	88	91	94
ELM	Retained (0)	83	75	68
	Promoted (1)	90	93	96
RF	Retained (0)	78	72	68
	Promoted (1)	90	92	96