Education Science and Management
https://www.acadlore.com/journals/ESM

Comparative Analysis of Feature Selection Techniques in Predictive Modeling of Mathematics Performance: An Ecuadorian Case Study

Nadia N. Sánchez-Pozo[1]*, Liliana M. Chamorro-Hernández[1], Jorge Mina[1], Javier Montalvo Márquez[2]

[1] Postgraduate Center, Carchi State Polytechnic University, 040101 Tulcán, Ecuador
[2] Department of Logistics and Transportation, Carchi State Polytechnic University, 040101 Tulcán, Ecuador

* Correspondence: Nadia N. Sánchez-Pozo (nadia.sanchez@upec.edu.ec)

Received: 07-27-2023 **Revised:** 08-31-2023 **Accepted:** 09-10-2023

Citation: Sá nchez-Pozo,N. N., Chamorro-Hernández, L. M., Mina, J., & Márquez, J. M. (2023). Comparative analysis of feature selection techniques in predictive modeling of mathematics performance: An Ecuadorian case study. *Educ. Sci. Manag., 1*(2), 111-121. https://doi.org/10.56578/esm010205.

Abstract: The field of educational research increasingly emphasizes predictive modeling of academic performance, focusing on identifying determinants of student success and crafting models to forecast future achievements. This investigation evaluates the efficacy of different feature selection techniques in predicting mathematics performance among Ecuadorian students, based on data from the 2021-2022 cycle of the Ser Estudiante test. Employing supervised logistic regression for classification, the study compares three feature selection methods: selection based on the highest k-scores, recursive feature elimination with cross-validation (RFECV), and recursive feature elimination (RFE). The assessment reveals that both the highest k-scores and RFECV methods are highly effective in isolating the most pertinent features for predicting mathematical prowess. These methods achieved prediction accuracies exceeding 90%, with k-scores attaining 96% and RFECV 92%. Furthermore, they demonstrated remarkable recall (94% and 97%, respectively) and F1-Score (96% and 91%, respectively). Additionally, both methods presented Receiver Operating Characteristic (ROC) curves with an area under the curve (AUC) of 99%, signifying superior discriminatory capability. The findings illuminate the critical role of judicious feature selection in enhancing the precision of predictive models for academic performance, particularly in mathematics. The results advocate for the application of these techniques in pinpointing key factors contributing to student success. This study not only contributes to the methodological discourse in educational data analysis but also provides practical insights for the Ecuadorian education system in leveraging data-driven approaches to enhance student outcomes.

Keywords: High school; Academic performance; Prediction; Mathematics; Machine learning

1. Introduction

The challenge of low academic achievement in mathematics within secondary education institutions has emerged as a critical concern. This issue, highlighted by Malak et al. (2022), is closely linked to elevated dropout rates. Academic success, a key contributor to student self-esteem and motivation, is vitally important. Bravo et al. (2021) note that underperformance in assessments can lead to academic disengagement and dropout. The exploration of factors influencing academic performance has thus become a focal point in educational research (Hellas et al., 2018).

Academic achievement, a complex and multifaceted concept, is defined variably by different scholars. It is perceived either as the level of knowledge assessed by educators or as the realization of educational objectives set by students and teachers within a specific timeframe (Llamas-Díaz et al., 2022). The assessment of academic performance involves a spectrum of evaluation methods, capturing student progress in various coursework aspects, ranging from classroom practices to quizzes and examinations (Contreras et al., 2020). Evaluations of educational system effectiveness often consider critical indicators like annual dropout and graduation rates, acknowledging the multifaceted nature of academic achievement influenced by a diverse array of social, economic, and demographic factors (Muñoz Gualán et al., 2023).

Mathematics, a cornerstone discipline, equips students with vital skills for solving complex problems across numerous fields, including engineering and finance (Hwang & Tu, 2021). However, its challenging nature often leads to suboptimal achievement levels among students, a concern highlighted by UNESCO (2021). In response, Ecuador's Ministry of Education and the National Institute for Educational Evaluation (INEVAL) have undertaken initiatives such as the "Being a Student" assessment to gauge students' proficiency in various subjects (Instituto Nacional de Evaluación Educativa, 2016). The 2021-2022 assessment cycle results indicate that the national average in mathematics for the Unified General Baccalaureate (BGU) falls below the desired performance level, underscoring the necessity to address barriers hindering student advancement in STEM fields (Instituto Nacional de Evaluación Educativa, 2022).

A primary obstacle in mathematics education is identified as the lack of student engagement and motivation, often due to the subject's perceived difficulty (Hung et al., 2020). This challenge is compounded by insufficient resources, a scarcity of qualified educators, and inadequate pedagogical approaches (Rocha Feregrino et al., 2020). It is suggested that education policymakers should prioritize enhancing teaching quality through the adoption of innovative methods, development of interactive learning resources, and provision of professional development for teachers (Tarik et al., 2021).

The advent of technology has revolutionized data accessibility, enabling the use of sophisticated techniques, notably machine learning, to distill relevant information. The integration of machine learning in predicting students' mathematical performance in various Ecuadorian educational institutions could significantly advance national education. Machine learning is posited to refine decision-making processes and bolster educational progress.

Research has demonstrated that feature selection techniques can enhance the precision and efficiency of predictive models in academic performance. Jalil et al. (2019) utilized a feature selection method based on correlation and variance, markedly improving the accuracy of a predictive model for college students' academic performance (Jalil et al., 2019). In a similar vein, Shreem et al. (2022) employed a feature selection technique based on feature importance, which showed enhanced accuracy in predicting high school students' academic performance compared to using all available features (Shreem et al., 2022). Chen et al. (2020) implemented feature selection using Least Absolute Shrinkage and Selection Operator (LASSO) regression to predict elementary school students' math achievement, resulting in increased accuracy and reduced model complexity. Fan et al. (2019) underlined the critical role of appropriate feature selection in developing reliable educational models, highlighting the beneficial impact of integrating machine learning in teaching and assessment (Fan et al., 2019).

This study aims to refine the predictive model for high school students' academic achievement in mathematics in Ecuador. It will employ feature selection techniques within a logistic regression classification framework. A comprehensive evaluation and comparison of three feature selection methodologies, highest k-scores, RFECV, and RFE, will be conducted, using data from the Ser Estudiante test. The impact of each feature selection technique on the model's accuracy and interpretability will be assessed.

The overarching goal is to enhance the prediction model's accuracy and efficiency, contributing to the understanding and forecasting of academic performance in mathematics. By identifying key factors influencing mathematics achievement, the study seeks to provide insights for educational policymakers and practitioners to develop targeted interventions and improve student outcomes.

1.1 Academic Performance

Academic performance, a multifaceted construct, has been extensively examined by various scholars. It encompasses the assimilation of knowledge and competencies by students, typically evaluated through grades, test scores, and overall achievements within educational settings. Influenced by numerous factors, including motivation, study habits, resource availability, support systems, and the quality of the educational environment, academic performance is a crucial determinant of future academic and professional success. Students demonstrating high academic performance are observed to have enhanced prospects in subsequent educational pursuits and their future careers (Findiana et al., 2020).

1.2 Factors Influencing Academic Performance

Academic performance is subject to a range of influences that vary across disciplines, educational levels, and socioeconomic contexts. Key factors include:

1. Emotional factors. Stress, anxiety, and external personal and emotional challenges significantly impact students' concentration and learning.
2. Sociological Factors. These factors, closely linked to a student's family financial status, environment, neighborhood, parents' education level, and child labor, play a pivotal role in academic outcomes. Economically disadvantaged parents may struggle to finance their children's education, leading to demotivation and potential academic underperformance (Pérez Gutiérrez, 2020).

3. Psychological factors. These relate to students' adjustment, social acceptance, and emotional states, exerting significant effects on their academic development (Llamas-Díaz et al., 2022).
4. Pedagogical factors. Pedagogical efficacy, the availability of educational resources (including technological access and instructional materials), and the teaching methods employed influence students' learning motivation and academic achievements (Lazo Salcedo & Meza Paucar, 2022).

The interaction among these factors results in complex and often unpredictable effects on academic performance. Consequently, an effective educational approach should incorporate a diverse array of tools and strategies, customized to meet the distinct requirements of individual students and optimize their academic potential.

1.3 Ser Estudiante Evaluation (SEST)

SEST, formulated by INEVAL, is an instrumental tool for assessing educational achievement within the Ecuadorian National School System. Its primary objective is to discern students' competencies and developmental needs, facilitating informed decision-making for educational enhancement. INEVAL adopts the Context, Input, Process, and Product (CIPP) model, a comprehensive framework encapsulating the multifarious elements that impact academic performance in Ecuador. This model integrates critical domains such as teachers and training, students and families, and schools, offering a thorough and multifaceted examination of the factors influencing academic success. Complementarily, the Associated Factors Survey, part of the SEST, probes additional aspects related to learning, including socio-emotional skills that may affect academic outcomes.

The SEST model is a comprehensive approach to evaluating student knowledge in diverse educational settings. This approach incorporates three main assessment tools: structured tests, analytic rubrics, and checklists. These tools are strategically combined to assess students' cognitive abilities and proficiencies within a competency-based framework. Students' responses are categorized into four performance levels: Proficient, Satisfactory, Basic, and Below Proficient, based on their depth of understanding and skill demonstration. The SEST model, aligning with both national and international standards, integrates benchmarks and core components to maintain consistency with large-scale assessments (Instituto Nacional de Evaluación Educativa, 2022).

1.4 Machine Learning

Machine learning plays a pivotal role in analyzing and modeling educational data, offering insights into student performance and other variables. Machine learning's application in education is geared towards developing predictive models that personalize learning experiences, identify students at risk, and enhance instructional quality (Buenaño-Fernández et al., 2019). Additionally, machine learning techniques analyze various student data, including demographics, standardized test performance, and interactions with educational technology platforms (Alyahyan & Düştegör, 2020).

These algorithms are capable of developing predictive models that estimate student outcomes, such as dropout probabilities or success likelihood in specific courses (Ghorbani & Ghousi, 2020). Machine learning algorithms are classified into several types, including supervised, unsupervised, and reinforcement learning (Hung et al., 2020). Within these, supervised learning, further divided into regression and classification, is particularly notable. Among the plethora of machine learning algorithms, decision trees, logistic regression, naive Bayes, random forest, and k-Nearest Neighbor (KNN) are widely recognized and utilized in educational research (Sánchez-Pozo et al., 2021).

Strengths and limitations of using machine learning in predictive modeling of academic performance

Machine learning has been increasingly recognized as a pivotal tool for deriving insights from extensive data sets and developing accurate predictive models, particularly in educational contexts:

- Data-driven insights. It has been observed that machine learning algorithms are adept at uncovering patterns and relationships within large-scale educational data. This capability provides educators with crucial insights into student performance, including strengths, weaknesses, and learning tendencies. Consequently, such insights facilitate the tailoring of instruction and intervention strategies to meet individual student needs.
- Predictive capabilities. The predictive nature of machine learning models enables the anticipation of student outcomes, such as academic performance and dropout rates, with a notable degree of accuracy. These capabilities permit educators to proactively identify students at risk and offer timely support, potentially enhancing educational outcomes.
- Personalization of learning. Machine learning's ability to discern individual student profiles allows for the personalization of learning experiences. By identifying specific strengths and weaknesses, machine learning

can recommend customized learning paths, adaptive materials, and targeted support strategies, thereby optimizing the learning experience for each student.

However, the application of machine learning in educational settings is not without limitations:

- Data dependency. The effectiveness of machine learning models is intrinsically linked to the quality of the underlying data. Issues with data bias or incompleteness can result in skewed predictions and exacerbate existing educational disparities. Ensuring data quality and representativeness is therefore critical in the educational application of machine learning models.
- Interpretability challenges. The complexity of machine learning models often renders them opaque, complicating the understanding of their predictive reasoning. This lack of transparency raises concerns regarding fairness and accountability in educational decisions. It is imperative for educators to comprehend the logic behind machine learning predictions to prevent the inadvertent introduction of biases.
- Ethical considerations. The utilization of machine learning in education comes with ethical challenges, including concerns about privacy, algorithmic bias, and potential discrimination. It is essential to address these ethical considerations to ensure machine learning's responsible and equitable use in educational contexts. Implementing robust privacy measures to safeguard student data and rigorously evaluating machine learning algorithms to mitigate bias are fundamental to achieving fair and equitable student outcomes.

2. Methodology

The study employs an experimental methodology, articulated through five distinct phases: data collection, data preprocessing, feature selection, classification, and a comparative analysis of feature selection approaches. Figure 1 delineates the workflow, illustrating the interconnections among these phases and highlighting the specific algorithms utilized at each stage.

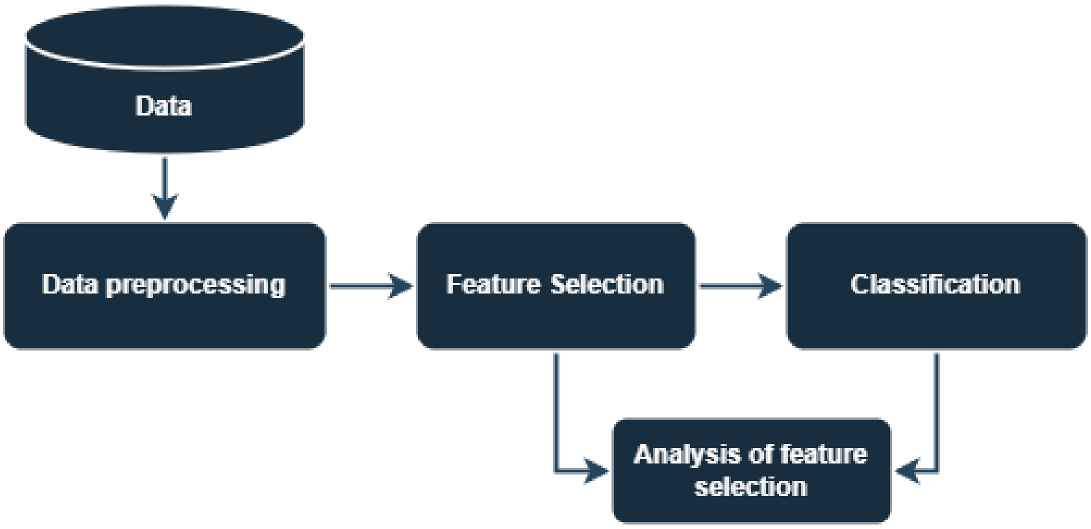

Figure 1. Experimental methodology

2.1 Data Collection

Data for this research was sourced from the publicly available Ser Estudiante Cycle 2021-2022 database, hosted on the official INEVAL website. This database comprises two key files: the Microfile and the Associated Student Factors file. The Microfile encompasses extensive details on academic performance, inclusive of scores across various subjects, socio-economic indicators, and school-level factors. The Associated Student Factors file enriches this data with insights into personal characteristics, interests, and home environments of the students.

Non-probability convenience sampling was adopted for selecting the study's sample, focusing on readily available participants. This approach led to the selection of third-year students of the General Unified High School from the Ser Estudiante Cycle 2021-2022 database, who also completed the Associated Factors survey. The sample represents approximately 24.14% of Ecuador's total third-year high school student population, with an average age of 16 years and a gender distribution of 51.7% female and 48.3% male (Figure 2), reflecting the national demographic distribution.

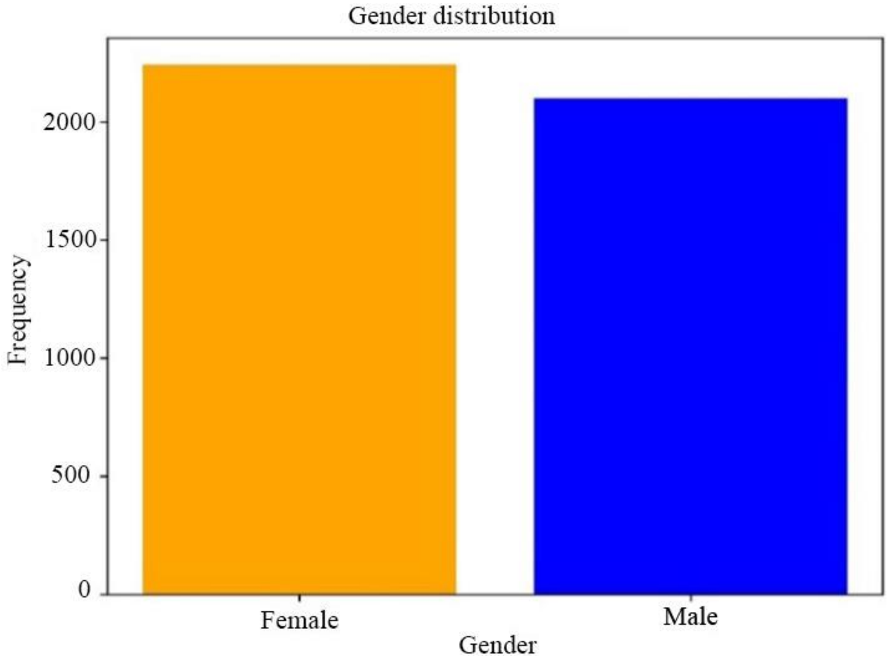

Figure 2. Gender distribution of the selected sample

2.2 Rationale for Sampling Strategy

The adoption of a non-probabilistic convenience sampling method in this study was informed by several pivotal considerations. Primarily, the availability of the Ser Estudiante Cycle 2021-2022 database offered an accessible and comprehensive data source. This availability obviated the necessity for extensive recruitment efforts, thereby facilitating the study's execution within a practical and efficient timeframe.

Furthermore, the extensive scope of the database, encompassing both academic and personal data, aligned seamlessly with the study's research objectives. The database provided intricate details regarding students' academic performance, socioeconomic backgrounds, and personal characteristics. This wealth of information enabled a thorough investigation into the various factors influencing mathematics achievement.

The utilization of a non-probabilistic convenience sampling method facilitated a focused analysis on a specific subset of the population: third-year high school students who participated in the SEST and completed the Associated Factors survey. This approach enabled a tailored examination pertinent to the relevant age group, ensuring the direct relevance of the data to the research questions posed.

However, it is imperative to acknowledge the inherent limitations of non-probabilistic convenience sampling. The sample may not comprehensively represent the entire population of high school seniors in Ecuador, potentially excluding students from varied socioeconomic backgrounds or geographical regions. Additionally, a risk of selection bias exists, as the sample might disproportionately include students more inclined to participate in the SEST.

To address these limitations, careful consideration was given to the potential for bias in the analysis of the results. Measures were taken to ensure that the sample size was sufficiently large to yield meaningful insights. Furthermore, the results were interpreted with due caution, given the nature of the sampling method employed.

2.3 Data Preprocessing

In the initial phase, the data underwent cleansing, involving the elimination of variables with over 20% missing values and outliers. The primary variable, students' average scores in mathematics, was discretized into four categories: elementary (600-699 points), satisfactory (700-799 points), excellent (800-1000 points), and insufficient (less than 599 points).

To prepare the data for machine learning algorithms, z-score normalization was implemented. This process standardizes the data by calculating the number of standard deviations each observation lies from the mean, as per Eq. (1).

$$z = \frac{x - u}{\sigma} \tag{1}$$

2.4 Feature Selection

Feature selection, a crucial step for enhancing prediction accuracy and simplifying decision-making, was undertaken to identify the most relevant student characteristics impacting mathematical performance. Post data preprocessing, 4,344 records with 192 characteristics were available for analysis. These characteristics spanned a range of domains, including motivation, family and social environment, and other factors pertinent to academic achievement.

Three feature selection techniques were employed:

1. Select K Best, which selects features based on their relationship with the target variable, quantified by a specific score (Chen et al., 2020).
2. RFECV, which begins with all features, iteratively eliminating the least significant ones. It utilizes cross-validation to assess the model's performance at each step of elimination (Mustaqim et al., 2021).
3. RFE, which assesses each feature's importance using a model, iteratively removing the least significant features (Alwarthan et al., 2022).

The selection of feature selection techniques in this study was predicated on their efficacy in extracting pertinent features from extensive datasets, their capacity to manage high-dimensional data, and their computational efficiency. The Select K Best is characterized by its straightforward and computationally efficient nature. It ranks features based on their individual importance scores and selects the top k features, providing an uncomplicated yet effective means of reducing dimensionality. This method's computational efficiency is particularly advantageous for handling large datasets, such as the Ser Estudiante Cycle 2021-2022 database utilized in this study. RFECV stands out for its robustness and efficacy in feature selection. By iteratively discarding features and assessing the performance of the remaining set using cross-validation, RFECV ensures that the chosen features are not only individually significant but also collectively enhance the model's predictive accuracy. This technique is particularly adept at isolating the most relevant features from datasets that are noisy or high-dimensional.

RFE offers a relatively simple and efficient approach to feature selection. It systematically removes features, selecting the one with the least impact on model performance in each iteration. This process effectively eliminates redundant or irrelevant features. The simplicity of RFE facilitates ease of implementation and interpretation, while its efficiency renders it suitable for application to large datasets. The decision to employ these techniques was guided by a consideration of effectiveness, computational efficiency, and interpretability. The Select K Best excels in simplicity and computational speed. RFECV is notable for its robustness and precision in feature selection. RFE, conversely, strikes a balance between ease of use, efficiency, and clarity in interpretability.

In the context of this research, the selection of feature selection techniques was critically aligned with the extensive size and high dimensionality of the Ser Estudiante Cycle 2021-2022 database. The computational efficiency of the Select K Best and RFE techniques was paramount for managing the substantial dataset. Concurrently, the robustness of RFECV was instrumental in extracting the most pertinent features from the data's high-dimensional nature.

The Select K Best is lauded for its simplicity and computational speed, making it adept for large datasets. However, it may fall short in pinpointing the most crucial features and is vulnerable to correlations among features. Conversely, RFECV is distinguished by its robustness and precision, particularly in handling high-dimensional data, though it is computationally more demanding and may not always isolate the most optimal feature set. RFE maintains a balance between ease of use, computational efficiency, and interpretability but may not always identify the most crucial features due to its sensitivity to the order in which features are removed.

The principal characteristics identified through these methods encompassed various aspects of the students' educational and personal environment:

- Evaluation regime, which determined whether the student belonged to the Costa - Galapagos or Sierra - Amazon regime.
- Location of educational institution, which distinguished between urban and rural settings.
- Access to technology, which assessed whether the student had access to a computer and the Internet at home.
- Attendance, which evaluated the average number of days per week that students attended classes.
- Grades. Which analyzed students' grades in diverse subjects, including language arts, biology, physics, chemistry, history, civics, and philosophy.
- Motivation, which gauged the students' motivation to learn and interest in school.
- Study habits, which scrutinized study habits, such as time spent studying, homework completion, and reading habits.
- Relationship with parents, which investigated the quality of communication and relationship with parents.
- Relationship with teachers, which examined the students' interactions and relationships with their teachers.

2.5 Classification

In this study, classification, a process where machine learning algorithms automate the labeling of data, is utilized. The objective is to construct a model capable of discerning patterns within the data, thereby predicting the category to which unknown data points might belong (Küchemann et al., 2020). The application of machine learning for classification offers the advantage of analyzing large datasets more rapidly than manual methods. Furthermore, these algorithms can detect patterns not immediately apparent to human observers, enhancing the precision and efficiency of the classification (Cachero et al., 2023).

However, challenges arise in utilizing machine learning for classification, notably the reliance on high-quality, ample training data to develop accurate models (Al-Emran et al., 2022). An additional concern is the model's decisions' lack of explainability due to the automated nature of machine learning.

For the classification of students into high and low academic achievement groups in mathematics, logistic regression was employed. This technique accommodates both numerical and categorical variables and offers relative ease of interpretation. Logistic regression has been demonstrated to yield satisfactory results in terms of precision and accuracy in educational predictive models (Albreiki et al., 2021). In a study by Zhang et al. (2020), logistic regression achieved an accuracy of 83.4% in predicting high school students' mathematical performance.

Nevertheless, logistic regression is not without limitations and assumptions. The assumption of linearity between independent and dependent variables, sensitivity to outliers, challenges with high-dimensional data, potential for overfitting, and class imbalance are pertinent considerations. The assumptions of independence of observations, homoscedasticity, and the absence of multicollinearity must also be acknowledged. Despite these factors, when data align with these assumptions and relationships are approximately linear, logistic regression can be effectively utilized, albeit with careful interpretation.

Three distinct scenarios were explored in this study:

Scenario 1: The k-highest score selection was used, selecting the k features with the greatest weight in the model.

Scenario 2: Employed the RFECV technique, evaluating the performance of each subset of features and selecting the optimal set using a logistic regression model.

Scenario 3: Implemented the RFE technique, identifying the subset of features that best fit the logistic regression model.

3. Results

This section delineates the evaluation metrics utilized and presents the outcomes of the three classification scenarios developed using Google Colab.

3.1 Performance Measures

The effectiveness of the classification models was appraised using accuracy, sensitivity, and F1-Score.

3.1.1 Accuracy

Accuracy is defined as the proportion of true positives among the total classified cases, both true and false positives. It serves as a crucial metric for minimizing the occurrence of false positives.

$$Accuracy = \frac{VP}{VP + FP} \tag{2}$$

3.1.2 Sensitivity

Sensitivity, also known as recall, measures the proportion of actual positive cases accurately identified as positive. It should be noted that a subset of true positives may be incorrectly classified as negative, known as false negatives.

$$Recall = \frac{VP}{VP + FN} \tag{3}$$

3.1.3 F1-Score

The F1-Score is a composite metric, amalgamating precision and recall. It is computed as the harmonic mean of these two measures, providing a balanced evaluation of the model's performance.

$$f1 - score = \frac{2 \, * \, (Precisión * Recall)}{Precisión \, + \, Recall} \tag{4}$$

3.1.4 ROC curve

The ROC curve is a statistical tool for assessing binary classification model performance. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR), indicating the proportion of correctly identified positive cases and the proportion of negative cases incorrectly classified as positive.

Scenario 1: Highest k-scores technique

In the first scenario, the highest k-scores technique was employed, a method that prioritizes features based on their individual significance scores, selecting the top k features. This approach, notable for its simplicity and efficiency, is particularly suitable for analyzing large datasets. Within the framework of this study, the highest k-scores technique was utilized to ascertain the most salient features for predicting mathematics achievement, considering their individual contributions to the model's performance. The results, as elucidated in Table 1, revealed that the accuracy for class 0 was 94%, with sensitivity at 100% and an F1-Score of 97%. Conversely, for class 1, the accuracy stood at 100%, sensitivity at 88%, and F1-Score at 93%. The model exhibited an overall accuracy of 97%, a sensitivity of 94%, and an F1-Score of 95%, proficiently predicting mathematics achievement across both high and low achievement groups.

Table 1. Results of the first scenario

Class	Accuracy	Recall	F1-Score
0	94%	100%	97%
1	100%	88%	93%

Scenario 2: RFECV

The second scenario implemented the RFECV technique. This method involves the systematic elimination of features, assessing the performance of the remaining set through cross-validation, thus ensuring the retained features consistently augment the model's accuracy across different data subsets. As depicted in Table 2, for class 0, the model achieved an accuracy of 90%, a sensitivity of 97%, and an F1-Score of 93%. For class 1, accuracy was 100%, sensitivity 97%, and F1-Score 89%. Overall, the model demonstrated an impressive accuracy of 95%, with a recall of 97% and an F1-Score of 91%.

Table 2. Results of second scenario

Class	Accuracy	Recall	F1-Score
0	90%	97%	93%
1	100%	97%	89%

Scenario 3: RFE

The third scenario applied the RFE technique, which iteratively removes features, selecting those with the least impact on the model's performance. This process aims to identify a concise set of features while maintaining the model's predictive accuracy. The results, outlined in Table 3, indicated uniform accuracy, sensitivity, and F1-Score of 60% for both classes.

Table 3. Results of the third scenario

Class	Accuracy	Recall	F1-Score
0	60%	60%	60%
1	60%	60%	60%

3.2 Comparative Analysis

This analysis aims to juxtapose the efficacy of distinct feature selection techniques in predicting academic performance in mathematics, utilizing the logistic regression algorithm for classification. Through this evaluation, the most salient features for prediction were discerned, offering insights for educators to enhance strategies in this pivotal subject. Three feature selection models were appraised. Initially, the k highest scores selection was implemented, achieving an accuracy of 97%. Subsequently, the RFECV technique was employed, yielding an accuracy of 95%. The final approach RFE resulted in a markedly lower accuracy of 60%. This outcome underscores the impact of including irrelevant features in the model on its predictive capacity.

Table 4 presents the specific results of these scenarios. Notably, a heightened sensitivity was observed in the RFECV scenario, while the RFE scenario's accuracy was significantly inferior to the other models. These findings have profound implications for enhancing mathematical academic performance.

In addition to accuracy measures, ROC curve analysis was utilized, depicted in Figure 3, to assess the models' discriminatory ability. This ability, crucial for distinguishing between varying academic capabilities, was evaluated using AUC.

Table 4. Comparative results of classification scenarios

Scenario	Accuracy	Sensitivity	F1-Score
1	97%	94%	95%
2	95%	97%	91%
3	60%	60%	60%

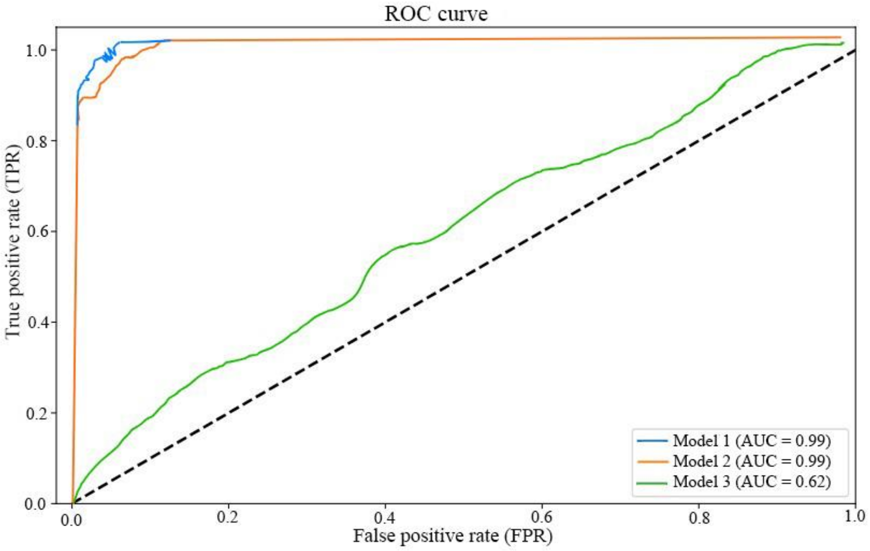

Figure 3. ROC curve analysis of the three proposed models

The analysis yielded diverse AUC values. Both the first and second models achieved an AUC of 0.99, indicating their exceptional capability in discriminating between high and low academic achievers. Conversely, the third model exhibited an AUC of 0.62, suggesting limitations in accurately predicting the performance of certain student groups. This lower AUC value implies challenges in effectively distinguishing between specific student categories, potentially leading to less precise predictions.

In summary, the ROC curve analysis elucidates the discriminative prowess of the three predictive models. The first two models demonstrate remarkable ability in differentiating high and low achievers, while the third model exhibits a reduced capacity in this regard, hinting at potential constraints in uniformly predicting academic performance.

4. Conclusions

The application of feature selection and machine learning techniques in this study has substantiated their effectiveness in refining the accuracy of predictive models for academic performance, particularly in mathematics. It has been demonstrated that through these methods, educators can ascertain pivotal factors influencing student achievement, thereby enabling the development of bespoke tutoring programs aimed at enhancing the performance of students who are underachieving.

Academically, the implications of this study extend beyond mathematics education. The methodologies employed herein can significantly inform decision-making processes across various educational fields. The utility of machine learning in discerning distinct behavioral patterns that contribute to academic success is not confined to mathematics alone but is also applicable to disciplines such as science, engineering, and technology. Prospective research focusing on the efficacy of specific tutoring strategies could further guide educators in optimizing academic outcomes across these domains.

Furthermore, the identification of behavior patterns associated with academic success has broader ramifications in the educational sector and beyond. For instance, exploring the nexus between technological access and academic achievement in mathematics could inform curriculum design and pedagogical approaches within educational systems.

This study, while underscoring the transformative potential of feature selection and machine learning in education, particularly in predicting mathematics achievement, does recognize certain limitations. The specificity of the study to mathematics achievement and the reliance on a singular dataset may circumscribe its generalizability to other educational spheres or to diverse student demographics. Additionally, the employment of multiple machine learning algorithms and feature selection techniques introduces a level of complexity that may obfuscate the interpretation of results.

To surmount these limitations, future research endeavors should venture into the applicability of these methodologies across various subjects and educational levels, incorporating larger and more diverse datasets. Simplifying models to enhance interpretability is also advisable.

Informed Consent Statement

Not applicable.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Educ. Sci.*, *11*(9), 552. https://doi.org/10.3390/educsci11090552

Al-Emran, M., Al-Nuaimi, M. N., Arpaci, I., Al-Sharafi, M. A., & Anthony Jnr., B. (2022). Towards a wearable education: Understanding the determinants affecting students' adoption of wearable technologies using machine learning algorithms. *Educ. Info. Technol.*, *28*(3), 2727–2746. https://doi.org/10.1007/s10639-022-11294-z

Alwarthan, S., Aslam, N., & Khan, I. U. (2022). An explainable model for identifying at-risk student at higher education. *IEEE Access*, *10*, 107649-107668. https://doi.org/10.1109/ACCESS.2022.3211070

Alyahyan, E. & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *Int. J. Educational Technol. High. Educ.*, *17*(1), 1–21. https://doi.org/10.1186/s41239-020-0177-7

Bravo, L. E. C., López, H. J. F., & Rivas Trujilllo, E. R. (2021). Análisis del rendimiento académico mediante técnicas de aprendizaje automático con métodos de ensamble. *Rev. Boletín Redipe*, *10*(13), 171–190. https://doi.org/10.36260/rbr.v10i13.1737

Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of machine learning in predicting performance for computer engineering students: A case study. *Sustainability*, *11*(10), 2833. https://doi.org/10.3390/su11102833

Cachero, C., Rico-Juan, J. R., & Macià, H. (2023). Influence of personality and modality on peer assessment evaluation perceptions using machine learning techniques. *Expert Syst. Appl.*, *213*, 119150. https://doi.org/10.1016/j.eswa.2022.119150

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *J. Big Data*, *7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4

Contreras, L. E., Fuentes, H. J., & Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación Universitaria*, *13*(5), 233–246. https://doi.org/10.4067/s0718-50062020000500233

Fan, Y. M., Liu, Y., Chen, H. S., & Ma, J. L. (2019). Data Mining-based design and implementation of college physical education performance management and analysis system. *Int. J. Emerg. Technol. Learn.*, *14*(6), 87-97. https://doi.org/10.3991/ijet.v14i06.10159

Findiana, R., Yuniarno, E. M., & Endroyono. (2020). Classification of graduates student on entrance selection public higher education through report card grade path using support vector machine method. In *2020 3rd International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia*, pp. 7–11. https://doi.org/ 10.1109/ICOIACT50329.2020.9332072

Ghorbani, R. & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access*, *8*, 67899–67911. https://doi.org/10.1109/ACCESS.2020.2986809

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V, Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, Larnaca, Cyprus, pp. 175–199. https://doi.org/10.1145/3293881.3295783

Hung, H. C., Liu, I. F., Liang, C. T., & Su, Y. S. (2020). Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. *Symmetry*, *12*(2). 213. https://doi.org/10.3390/sym12020213

Hwang, G. J. & Tu, Y. F. (2021). Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review. *Math.*, *9*(6), 584. https://doi.org/10.3390/math9060584

Instituto Nacional de Evaluación Educativa. (2016). *Ser Estudiante*. http://evaluaciones.evaluacion.gob.ec/BI/ser-estudiante/.

Instituto Nacional de Evaluación Educativa. (2022). *Informe de resultados Ser Bachiller lectivo 2020-2021*. https://cloud.evaluacion.gob.ec/dagireportes/sbciclo19/distrito/03D02.pdf.

Jalil, N. A., Hwang, H. J., & Dawi, N. M. (2019). Machines learning trends, perspectives and prospects in education sector. In *Proceedings of the 2019 3rd International Conference on Education and Multimedia Technology*. pp. 201–205. https://doi.org/10.1145/3345120.3345147

Küchemann, S., Klein, P., Becker, S., Kumari, N., & Kuhn, J. (2020). Classification of students' conceptual understanding in stem education using their visual attention distributions: A comparison of three machine-learning approaches. In *Proceedings of the 12th International Conference on Computer Supported Education*. pp. 36–46. https://doi.org/10.5220/0009359400360046

Lazo Salcedo, C. A., & Meza Paucar, T. M. (2022). Medidas de política económica en el sector educación durante la pandemia de la COVID-19 en Huánuco. *Innovación Empresarial*, *2* (1), https://doi.org/10.37711/rcie.2022.2.1.12

Llamas-Díaz, D., Cabello, R., Megías-Robles, A., & Fernández-Berrocal, P. (2022). Systematic review and meta-analysis: The association between emotional intelligence and subjective well-being in adolescents. *J. Adolescence*, *94*(7), 925–938. https://doi.org/10.1002/jad.12075

Malak, M. Z., Shuhaiber, A. H., Al-amer, R. M., Abuadas, M. H., & Aburoomi, R. J. (2022). Correlation between psychological factors, academic performance and social media addiction: Model-based testing. *Behav. Inf. Technol.*, *41*(8), 1583–1595. https://doi.org/10.1080/0144929X.2021.1891460

Muñoz Gualán, G. G., Gualán, E. D. M., & Gualán, A. P. M. (2023). Determinants of admission in the academic performance of excellence in the career of military sciences. *Rev. Científica Sinapsis*, *1*(22), 173–190. https://doi.org/10.37117/s.v1i22.745

Mustaqim, A. Z., Adi, S., Pristyanto, Y., & Astuti, Y. (2021). The effect of recursive feature elimination with cross-validation (RFECV) feature selection algorithm toward classifier performance on credit card fraud detection. In *2021 International Conference on Artificial Intelligence and Computer Science Technology, Yogyakarta, Indonesia*, pp. 270–275. https://doi.org/10.1109/ICAICST53116.2021.9497842

Pérez Gutiérrez, B. R. (2020). Comparison of data mining techniques to identify signs of student desertion, based on academic performance. *Rev. UIS Ingenierías*, *19*(1), 193–204. http://repositorio.ufps.edu.co/handle/ufps/1592.

Rocha Feregrino, G., López, J. A. J., Gómez, O. L. F., & Méndez, G. R. (2020). El rendimiento académico y las actitudes hacia las matemáticas con un sistema tutor adaptativo. *PNA. Rev. de Investigación En Didáctica de La Matemática*, *14*(4), 271–294. https://doi.org/10.30827/pna.v14i4.15202

Sánchez-Pozo, N. N., Mejía-Ordóñez, J. S., Chamorro, D. C., Mayorca-Torres, D., & Peluffo-Ordóñez, D. H. (2021). Predicting high school students' academic performance: A comparative study of supervised machine learning techniques. In *2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop*, pp. 1–6. https://doi.org/10.1109/IEEECONF53024.2021.9733756

Shreem, S. S., Turabieh, H., Al Azwari, S., & Baothman, F. (2022). Enhanced binary genetic algorithm as a feature selection to predict student performance. *Soft Comput.*, *26*(4), 1811–1823. https://doi.org/10.1007/s00500-021-06424-7

Tarik, A., Aissa, H., & Yousef, F. (2021). Artificial intelligence and machine learning to predict student performance during the COVID-19. *Procedia Comput. Sci.*, *184*, 835–840. https://doi.org/10.1016/j.procs.2021.03.104

UNESCO. (2021). *Las Matemáticas, enseñanza e investigación para enfrentar los desafíos de estos tiempos*. https://es.unesco.org/news/matematicas-ensenanza-e-investigacion-enfrentar-desafios-estos-tiempos.

Zhang, L., Liu, Z., Ren, T. W., Liu, D. Y., Ma, Z., Tong, L., Zhang, C., Zhou, T. Y., Zhang, X. D., & Li, S. M. (2020). Identification of seed maize fields with high spatial resolution and multiple spectral remote sensing using random forest classifier. *Remote Sens.*, *12*(3), 362. https://doi.org/10.3390/rs12030362