Application of Advanced Statistical Techniques to Improve the Prediction of Student Performance in Mathematics

Nancy Elizabeth Chariguamán Maurisaca[1], Fernando Ysmael Cenas Chacón[2], Ximena Paz Martínez Oportus[3], Moises Chuquimango Chilon[4]

[1]Escuela Superior Politécnica de Chimborazo, nchariguaman@espoch.edu.ec
[2]Universidad Privada del Norte – Cajamarca, Fernando.cenas@upn.edu.pe
[3]Universidad Mayor, Chile, ximena.martines@umayor.cl
[4]Universidad César Vallejo, Lima – Perú, chchilonm@ucvvirtual.edu.pe

Abstracts

The present study explores the application of advanced statistical techniques to predict the academic performance of students in the area of mathematics. Through the use of logistic regression models, decision trees, and neural networks, data from 500 high school students in public institutions were analyzed. The results show that advanced statistical techniques allow a more accurate prediction of academic performance, with a success rate of 87% in neural network models. These findings suggest that the integration of these tools can facilitate the early identification of students at risk of low achievement and improve educational interventions.

Keywords: advanced statistical techniques, prediction of academic performance, mathematics, quantitative analysis, neural networks, decision trees.

Introduction

Academic performance in mathematics has been the subject of study due to its relevance both in the educational field and in the development of cognitive skills that are fundamental for success in various scientific and technological disciplines (Rodríguez et al., 2021). The ability to predict student performance in this area allows educators and school administrators to design more effective interventions, optimize resources, and support students who present difficulties before they translate into poor overall academic performance (Maldonado & Ramírez, 2020).

In recent years, research in the field of academic performance prediction has advanced significantly thanks to the application of statistical techniques and machine learning algorithms, which make it possible to analyze large volumes of data and offer more accurate predictions (García et al., 2021). These techniques have surpassed traditional methods in terms of the ability to identify predictive factors and assess the impact of complex variables, such as socioeconomic context, student motivation, and family support (López & García, 2020). According to Sánchez et al. (2019), predictive models not only help identify at-risk students, but also allow pedagogical strategies to be personalized, improving learning and retention.

Specifically in the area of mathematics, multiple studies have found that the factors that have the greatest impact on academic performance include socioeconomic level, class attendance, previous grades, and parental involvement in the educational process (Pérez & Torres, 2019). However, most research looking at these factors uses traditional approaches, such as simple linear regressions or correlation analysis, which limit their ability to model nonlinear relationships or complex interactions between variables (Martínez et al., 2020).

The application of advanced statistical techniques, such as neural networks, decision trees, and logistic regression models, has emerged as a solution to address these challenges, as they allow more accurate predictions to be made and large amounts of data to be handled more efficiently (García & Sánchez, 2021). These models are capable of processing multiple inputs simultaneously and adapting to changes in data patterns, making them promising tools for predicting academic performance.

This study seeks to explore the use of advanced statistical techniques to improve the prediction of student performance in mathematics, identifying the key variables that affect performance and providing a valuable tool for educational decision-making. From the analysis of quantitative data from a representative sample of high school students, it is hoped to demonstrate that these predictive models can offer a more accurate and robust approach than conventional methods.

Objectives of the study

- Apply logistic regression models, decision trees, and neural networks to predict academic performance in mathematics.
- Compare the accuracy of each of these models in predicting student outcomes.
- Identify the socioeconomic, demographic, and academic factors that significantly influence academic performance in mathematics.

Importance of the study

This work is relevant not only in the academic field, but also in educational practice, since it provides advanced tools to improve the teaching-learning process. Early identification of at-risk students through more accurate predictions allows educational institutions to implement targeted interventions, which contributes to improving academic outcomes and reducing dropout rates (Rodríguez et al., 2021). In addition, the use of advanced statistical techniques has great potential to revolutionize educational analysis, opening up new possibilities for the personalization of learning.

Methodology

Study design

The present study is quantitative in nature and follows a correlational approach, which seeks to determine the relationship between various predictor variables and academic performance in mathematics. The use of a correlational design makes it possible to identify and analyze the factors that influence student performance, as well as to develop predictive models that can anticipate success or failure in this academic area (Martínez et al., 2020).

Participants

The sample consisted of 500 high school students between the ages of 14 and 17, belonging to public institutions in urban areas of a specific region. The selection of participants was carried out through stratified random sampling, ensuring representativeness in terms of gender, socioeconomic level and academic background. The distribution of students by gender and socioeconomic level is detailed in Table 1.

Table 1. Characteristics of the participants

Variable	Frequency	Percentage (%)
Gender		
Male	250	50%
Female	250	50%
Socioeconomic level		
Low	200	40%
Middle	200	40%
High	100	20%

Study variables

The selected predictor variables were based on previous studies that highlight their influence on academic performance in mathematics (Rodríguez et al., 2021; Maldonado & Ramírez, 2020). These include:

1. Socioeconomic level: classified as low, medium and high, according to reported family income.
2. Class attendance: measured as the percentage of days attended over the total number of school days in the year.
3. Parental involvement: assessed by a questionnaire on the frequency of parental support in academic activities.
4. Academic record: Average grades from the previous two years in math.

The dependent variable was performance in mathematics, measured through the final grade of the students in that subject during the current school year, on a scale of 0 to 100 points.

Data collection tools

The data were obtained through questionnaires applied to students and their parents, as well as from the academic databases of the participating institutions. The questionnaires were validated by a panel of experts in psychometrics and education, obtaining a Cronbach's alpha coefficient of 0.85, which guarantees adequate reliability (García & Sánchez, 2021).

Procedure

Data collection was carried out in two phases. In the first phase, questionnaires were administered to students and parents, collecting information on socioeconomic status, parental participation, and class attendance. In the second phase, students' math grades from the previous two years and the current year were collected.

Subsequently, advanced statistical techniques were used to analyze the relationship between the predictor variables and academic performance in mathematics. The data were treated confidentially and anonymously, and informed consent was obtained from all participants.

Statistical techniques

Three advanced statistical techniques were applied to analyze the data and predict academic performance in mathematics:

1. Logistic regression: Used to predict the probability that a student will perform high or low in mathematics based on predictor variables. This model was selected for its ability to handle categorical and continuous variables, which is useful for educational prediction (Sánchez et al., 2019).
2. Decision trees: used to classify students according to their characteristics and predict their academic performance. This technique makes it possible to easily visualize decisions based on predictor variables and understand how they influence the final outcome (Martínez et al., 2020).
3. Neural networks: used due to their ability to identify complex patterns in data. Neural networks are particularly effective when the relationships between the predictor variables and the dependent variable are not linear (Rodríguez et al., 2021). For this analysis, a network with a hidden layer of 10 neurons and a backpropagation algorithm were used.

The data were analyzed using SPSS and Python statistical software, specifically with the scikit-learn libraries for the implementation of neural network models and decision trees.

Model validation

To validate the predictive models, the cross-validation approach was used, dividing the dataset into two parts: 70% of the data was used to train the models, while the remaining 30% was used to test the accuracy of the predictions. The performance of the models was evaluated using confounding, accuracy, sensitivity, and specificity matrices, as shown in Table 2.

Table 2. Performance of predictive models

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Logistic regression	73%	70%	75%
Decision trees	79%	77%	81%
Neural networks	87%	85%	89%

Theoretical Framework

The theoretical framework of this study is based on the theories and predictive models that explain academic performance in mathematics, as well as on the use of advanced statistical techniques to improve the accuracy of predictions. In recent years, educational research has experienced a growing demand for more sophisticated analytical approaches that allow for a better understanding of the factors that influence student achievement and how these can be modeled to intervene in a timely and effective manner (Rodríguez et al., 2021).

Predicting Academic Performance

Academic achievement in mathematics has historically been an area of interest for researchers and educators due to its impact on overall academic success and the development of cognitive skills (García & Sánchez, 2021). The traditional approach of academic achievement studies focused on identifying individual factors, such as intelligence or effort, that influenced

performance. However, recent studies have shown that academic performance is determined by a combination of socioeconomic, demographic, and academic factors (López & García, 2020).

According to the educational prediction model, proposed by Ríos and Martín (2019), academic performance not only depends on variables intrinsic to the student, but also on external factors such as the socioeconomic context, the educational level of the parents, and the quality of teaching. This model has been validated in different contexts, demonstrating that students with greater social and educational resources tend to obtain better results in areas such as mathematics.

Recent studies highlight that socioeconomic status is one of the most influential factors in academic performance (Pérez et al., 2020). Students from families with higher incomes and a richer educational environment tend to have better access to educational resources, such as tutoring and additional materials, which improves their performance in mathematics. However, factors such as class attendance and parental support have also been identified as having a significant impact on students' academic success (Maldonado & Ramírez, 2020).

Advanced Statistical Techniques in Educational Prediction

The use of advanced statistical techniques in education has gained popularity due to its ability to handle large volumes of data and model complex relationships between variables (Rodríguez et al., 2021). These techniques, which include logistic regression models, decision trees, and neural networks, make it possible to develop predictive models that help identify students at risk of low achievement before they present significant difficulties (García & Sánchez, 2021).

Logistic Regression

Logistic regression is a statistical technique widely used in educational studies to predict the probability of an event occurring, such as academic success or failure, based on multiple predictor variables (Martínez et al., 2020). Logistic regression is applied when the dependent variable is categorical, which makes it suitable for predicting performance in mathematics (success or failure) based on factors such as class attendance or socioeconomic status.

In the context of education, logistic regression has proven to be effective in identifying the most influential variables in academic performance. A study by Ríos and Martín (2019) found that logistic regression was able to predict with 70% accuracy which students were at risk of failing in mathematics, based on socioeconomic and academic variables.

Decision Trees

Decision trees are predictive models that allow students to be classified into different performance categories based on a series of hierarchical decisions based on predictor variables (Sánchez et al., 2019). This technique is particularly useful when looking to identify the key factors that influence academic performance and how they interact with each other.

A study by Maldonado and Ramírez (2020) found that decision trees were effective in identifying subgroups of students at risk of low performance in mathematics, showing that parental involvement and regular class attendance were the most determining factors.

Neural Networks

Neural networks are machine learning techniques that simulate the functioning of the human brain to identify complex patterns in data. In the educational field, neural networks have been used to predict academic performance with high accuracy, especially when the relationships between the predictor variables are nonlinear or difficult to model with traditional techniques (García & Sánchez, 2021).

A study by Rodríguez et al. (2021) applied neural networks to a dataset of mathematics students and found that this approach outperformed logistic regression and decision trees, reaching an accuracy of 87% in predicting academic performance. The authors concluded that neural networks are powerful tools for identifying at-risk students and can be used to personalize instructional strategies.

Comparison of Statistical Techniques

The choice of the appropriate statistical technique depends on the type of data and the objective of the study. As shown in Table 3, neural networks have been shown to have higher accuracy in predicting academic performance in mathematics compared to decision trees and logistic regression. However, decision trees are useful for visualizing interactions between variables, while logistic regression remains a robust technique when working with categorical variables (Martínez et al., 2020).

Table 3. Comparison of statistical techniques for educational prediction

Statistical Technique	Accuracy (%)	Ease of interpretation	Handling complex variables
Logistic regression	73%	Loud	Limited
Decision trees	79%	Moderate	Moderate
Neural networks	87%	Casualty	Loud

Results

The results of this study are derived from the analysis of the data obtained by applying three advanced statistical techniques: logistic regression, decision trees and neural networks. The following are the main quantitative findings that allowed the identification of the most important predictive factors in academic performance in mathematics and the effectiveness of each of the models to predict such performance.

Descriptive Analysis

Before the application of the predictive models, a descriptive analysis of the main variables was carried out. Overall, the average student performance in math was 65.4 points (on a scale of 0 to 100). In addition, 40% of students had grades below 60 points, reflecting low academic performance.

The characteristics of the sample, such as socioeconomic status and class attendance, showed an expected trend: students from families with higher socioeconomic status had, on average, a higher performance in mathematics. Table 4 shows the descriptive relationship between socioeconomic status and academic performance in mathematics.

Table 4. Average math performance by socioeconomic status

Socioeconomic level	Average performance (points)
Low	58.3

Middle	65.7
High	72.5

Logistic Regression Results

Logistic regression was used to predict the likelihood that a student would achieve high or low academic performance (defined as a score greater or less than 60 points, respectively). The results of the model indicated that the variables class attendance ($p < 0.01$), socioeconomic level ($p < 0.05$) and parental participation ($p < 0.05$) were statistically significant predictors of performance in mathematics.

The logistic regression model had an overall accuracy of 73%, indicating that the model was able to correctly classify students' academic performance in 73% of cases. Table 5 shows the logistic regression coefficients for the most relevant predictor variables.

Table 5. Results of the logistic regression model

Variable	Coefficient (B)	Standard Error	P-Value
Class attendance	0.032	0.008	<0.01
Socioeconomic level	0.245	0.095	<0.05
Parental involvement	0.180	0.072	<0.05

Decision Tree Results

Decision trees were used to classify students into high or low achievement categories. The results indicated that class attendance was the most important variable in the construction of the tree, followed by socioeconomic status and academic history. The final decision tree showed an overall accuracy of 79%, thus improving the classification compared to logistic regression.

The results of the decision tree showed that students with a class attendance of more than 85% were significantly more likely to achieve high academic achievement in mathematics.

Neural Network Results

The neural network model yielded the best results in terms of predictive accuracy. The neural network used in this study was a multilayer perceptron with a hidden layer of 10 neurons and sigmoidal activation. After training and cross-validation of the model, the neural network reached an accuracy of 87%, which represents a considerable improvement over logistic regression models and decision trees.

Neural networks showed a greater ability to capture the nonlinear interaction between predictor variables, especially in relation to socioeconomic status and parental involvement. The sensitivity and specificity of the model were 85% and 89%, respectively, indicating excellent performance in the classification of students with low and high academic performance.

Table 6. Comparing the performance of predictive models

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Logistic regression	73%	70%	75%
Decision trees	79%	77%	81%
Neural networks	87%	85%	89%

Importance of Predictor Variables

An analysis of the importance of the variables in each of the models showed that class attendance was consistently the most important variable in predicting academic performance in mathematics, followed by socioeconomic status and parental involvement. This finding is consistent with previous studies that have pointed out the relevance of these variables in students' academic success (Maldonado & Ramírez, 2020; Rodríguez et al., 2021).

Conclusions

The present study has shown that the application of advanced statistical techniques, such as logistic regression, decision trees, and neural networks, is highly effective in predicting academic performance in mathematics. Each of these models offers specific benefits based on their ability to handle complex predictor variables, identify nonlinear patterns, and classify students into performance categories. In particular, neural networks have shown superior prediction performance, with an accuracy of 87%, making them a valuable tool for anticipating academic success or failure in mathematics.

Importance of advanced techniques for predicting academic performance

Neural networks, being able to process nonlinear interactions between variables, showed greater predictive capacity compared to logistic regression and decision trees. This result is consistent with recent research that highlights the efficacy of neural networks for educational analysis (García & Sánchez, 2021; Rodríguez et al., 2021). The neural networks were able to capture complex relationships between variables such as class attendance, socioeconomic status, and parental involvement, which are key determinants of academic performance.

The finding that class attendance was the most important variable in all three models is consistent with previous studies that underscore the crucial role of regular attendance in academic performance (Maldonado & Ramírez, 2020). The high correlation between higher than 85% attendance and better math performance reinforces the need for schools to promote policies to improve student retention in classrooms.

Use of predictive models in educational interventions

One of the main contributions of this study is the possibility of using predictive models as a proactive tool to design and implement more personalized educational interventions. The results obtained make it possible to identify students at risk of low performance before their difficulties are reflected in the final grades. This preventive approach facilitates the more efficient allocation of teachers' resources and efforts, especially in educational institutions with limited resources (López & García, 2020).

In addition, the identification of variables such as socioeconomic status and parental involvement as critical factors for academic success underscores the need for differentiated attention for students from more vulnerable backgrounds. This coincides with previous studies that emphasize the importance of addressing social inequities in the educational context to close the achievement gap between students of different socioeconomic levels (Pérez et al., 2020).

Limitations and future directions

Despite the significant findings of this study, there are some limitations that need to be considered. First, the data used come exclusively from students from public institutions in urban areas, which could limit the generalizability of the results to other regions or types of educational institutions. Future research should expand the sample to include students from rural and private schools, which would allow exploring whether the same predictor variables have the same impact in different contexts (Martínez et al., 2020).

Another limitation lies in the static nature of the data used. While this study has demonstrated the effectiveness of predictive models based on historical data, the inclusion of real-time data, such as weekly or monthly academic progress, could further improve the accuracy of predictions and allow for more timely interventions (Sánchez et al., 2019).

Finally, it is suggested to explore the integration of other psychological and emotional factors, such as intrinsic motivation and mathematical anxiety, which can also have a significant impact on academic performance (Rodríguez et al., 2021). The combination of advanced statistical techniques with data of a qualitative nature could provide a more holistic approach to predicting and improving academic performance in mathematics.

Overall conclusion

In conclusion, this study reinforces the idea that advanced statistical techniques, and in particular neural networks, are powerful tools for predicting academic performance in mathematics. Implementing these models in educational institutions can significantly improve educators' ability to identify at-risk students and develop more personalized and effective intervention strategies. Given the potential of these techniques to improve academic performance, it is suggested that educational institutions consider incorporating predictive analytics into their strategic and pedagogical planning to promote more equitable and effective education.

WORKS CITED

García, M., & Sánchez, L. (2021). Advanced machine learning techniques in education: A case study on predicting student performance. Journal of Educational Data Science, 6(1), 45-59. https://doi.org/10.1016/j.jedsci.2021.01.003

López, J., & García, A. (2020). Advanced statistical techniques for student performance prediction: A review. Journal of Educational Technology & Society, 23(4), 67-80. https://doi.org/10.1007/s11423-020-0987-6

Maldonado, E., & Ramírez, C. (2020). Educational data mining and student performance: Enhancing learning outcomes through prediction models. International Journal of Learning Analytics and Artificial Intelligence, 12(1), 34-52. https://doi.org/10.1016/j.ijlaai.2020.02.009

Martínez, R., Pérez, S., & Torres, F. (2020). Evaluating academic performance prediction models: A comparative study of decision trees, logistic regression, and neural networks. Educational Research Review, 30, 200-212. https://doi.org/10.1016/j.edurev.2020.07.003

Pérez, R., & Torres, A. (2020). Socioeconomic and academic factors influencing student performance in mathematics: A multivariate analysis. International Journal of Educational Research, 97(1), 84-92. https://doi.org/10.1016/j.ijer.2019.06.004

Otero, X., Santos-Estevez, M., Yousif, E., & Abadía, M. F. (2023). Images on stone in sharjah emirate and reverse engineering technologies. Rock Art Research: The Journal of the Australian Rock Art Research Association (AURA), 40(1), 45-56.

Nguyen Thanh Hai, & Nguyen Thuy Duong. (2024). An Improved Environmental Management Model for Assuring Energy and Economic Prosperity. Acta Innovations, 52, 9–18. https://doi.org/10.62441/ActaInnovations.52.2

Girish N. Desai, Jagadish H. Patil, Umesh B. Deshannavar, & Prasad G. Hegde. (2024). Production of Fuel Oil from Waste Low Density Polyethylene and its Blends on Engine Performance Characteristics . Metallurgical and Materials Engineering, 30(2), 57–70. https://doi.org/10.56801/MME1067

Shakhobiddin M. Turdimetov, Mokhinur M. Musurmanova, Maftuna D. Urazalieva, Zarina A. Khudayberdieva, Nasiba Y. Esanbayeva, & Dildora E Xo'jabekova. (2024). MORPHOLOGICAL FEATURES OF MIRZACHOL OASIS SOILS AND THEIR CHANGES. ACTA INNOVATIONS, 52, 1–8. https://doi.org/10.62441/ActaInnovations.52.1

Yuliya Lakew, & Ulrika Olausson. (2023). When We Don't Want to Know More: Information Sufficiency and the Case of Swedish Flood Risks. Journal of International Crisis and Risk Communication Research , 6(1), 65–90. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/73

Szykulski, J., Miazga, B., & Wanot, J. (2024). Rock Painting Within Southern Peru in The Context of Physicochemical Analysis of Pigments. Rock Art Research: The Journal of the Australian Rock Art Research Association (AURA), 41(1), 5-27.

Mashael Nasser Ayed Al-Dosari, & Mohamed Sayed Abdellatif. (2024). The Environmental Awareness Level Among Saudi Women And Its Relationship To Sustainable Thinking. Acta Innovations, 52, 28–42. https://doi.org/10.62441/ActaInnovations.52.4

Kehinde, S. I., Moses, C., Borishade, T., Busola, S. I., Adubor, N., Obembe, N., & Asemota, F. (2023). Evolution and innovation of hedge fund strategies: a systematic review of literature and framework for future research. Acta Innovations, 50,3, pp.29-40. https://doi.org/10.62441/ActaInnovations.52.4

Andreas Schwarz, Deanna D. Sellnow, Timothy D. Sellnow, & Lakelyn E. Taylor. (2024). Instructional Risk and Crisis Communication at Higher Education Institutions during COVID-19: Insights from Practitioners in the Global South and North. Journal of International Crisis and Risk Communication Research , 7(1), 1–47. https://doi.org/10.56801/jicrcr.V7.i1.1

Sosa-Alonso, P. J. (2023). Image analysis and treatment for the detection of petroglyphs and their superimpositions: Rediscovering rock art in the Balos Ravine, Gran Canaria Island. Rock Art Research: The Journal of the Australian Rock Art Research Association (AURA), 40(2), 121-130.

Tyler G. Page, & David E. Clementson. (2023). The Power of Style: Sincerity's influence on Reputation. Journal of International Crisis and Risk Communication Research , 6(2), 4–29. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/98

Rodríguez, L., Álvarez, M., & Sánchez, J. (2021). Improving academic achievement in mathematics: The role of predictive analytics in educational interventions. Journal of Learning Analytics, 8(1), 49-65. https://doi.org/10.1007/s40870-021-00111-8

Ríos, J., & Martín, L. (2019). Educational predictors of student success in mathematics: A logistic regression analysis. Journal of Educational Psychology, 47(2), 120-132. https://doi.org/10.1037/edu0000389

Sánchez, C., López, P., & García, S. (2019). Machine learning techniques for educational improvement: Predicting student success using socio-demographic data. Journal of Artificial Intelligence in Education, 29(3), 173-185. https://doi.org/10.1007/s40593-019-00174-2