OmniaScience

Journal of Technology and Science Education

JOTSE, 2025 – 15(2): 322-334 – Online ISSN: 2013-6374 – Print ISSN: 2014-5349

https://doi.org/10.3926/jotse.2945

MACHINE LEARNING MODEL FOR CLASSIFYING HIGH SCHOOL STUDENTS' ACADEMIC PERFORMANCE IN MATHEMATICS AMIDST THE COVID-19 CONTEXT

Gisella Luisa Elena Maquen-Niño[1*], Moisés Alvin Miguel-Flores[1], Leysi Yarely Aurich-Mio[1], Ivan Adrianzén-Olano[2], Percy Edwin de la Cruz Vélez de Villa[3], Diana Mercedes Castro-Cárdenas[1]

[1]Universidad Nacional Pedro Ruiz Gallo (Peru)

[2]Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (Peru)

[3]Universidad Nacional Mayor de San Marcos (Peru)

*Corresponding author: gmaquenn@unprg.edu.pe

mmiguel@unprg.edu.pe, laurich@unprg.edu.pe, ivan.adrianzen@untrm.edu.pe

pdelacruzv@unmsm.edu.pe, dcastroc@unprg.edu.pe

Received June 2024

Accepted April 2025

Abstract

There is uncertainty in classifying academic performance due to the large number of variables involved in its measurement within educational institutions. Therefore, this research aimed to develop a machine learning model for the classification of mathematical academic performance in students of a private Regular Educational Institution in the department of Lambayeque, Peru. Using the CRISP-DM methodology, a dataset was obtained through data collection instruments from 711 students in the first, second, and third years of high school during the period from 2019 to 2021 in the context of the COVID-19 pandemic. This dataset, which was validated by expert judgment, contained 34 input variables and 1 output variable that could have three possible values: 1: deficient, 2: improvable, and 3: optimal. The variables most related to the classification of academic performance were the student's self-perception of performance and grades from the previous year in mathematics subjects, as analyzed in the correlation matrix. The model was trained and subsequently evaluated, obtaining 91% accuracy. It was concluded that it is possible to classify academic performance using a machine learning model, with the K-nearest neighbors model being the most appropriate for this research because it works with categorical data and achieves an adequate level of certainty.

Keywords **–** Machine learning, Classification model, Academic performance, Mathematics, KNN algorithm.

To cite this article:

Maquen-Niño, G.L.E., Miguel-Flores, M.A., Aurich-Mio, L.Y., Adrianzén-Olano, I., de la Cruz-Vélez de Villa, P.E., & Castro-Cárdenas, D.M. (2025). Machine learning model for classifying high school students' academic performance in mathematics amidst the COVID-19 context. *Journal of Technology and Science Education,* 15(2), 322-334. https://doi.org/10.3926/jotse.2945

---

1. Introduction

In today's knowledge society, education plays an essential role in human development, with academic performance serving as a primary indicator of students' learning outcomes, commonly assessed through grades obtained in standardized tests. The PISA Report indicates that academic performance reflects students' ability to apply their knowledge and skills in real situations, providing a comparative measure of educational success between different countries and educational systems (OCDE, 2023).

Traditionally, academic performance has been evaluated through periodic tests, adjusting grades to the demands of each subject.

Mathematic, a discipline that requires high levels of logical reasoning and analytical thinking, and the practical application of concepts for problem solving, is particularly relevant in performance assessments. However, evaluating students' mathematical proficiency remains a challenge, especially when external factors influence learning conditions.

In this research analyzing the academic performance of students at a private Regular Educational Institution in the Lambayeque region, Peru in the period from 2019 to 2021, finding that overall course averages in mathematics were low during this period. The shift to virtual education significantly impacted students' academic performance, particularly in mathematics.

Disruptive factors such as student distraction, power outages, and slow internet speed affected students' learning experiences, leading to concerns from both teachers and parents regarding content comprehension and performance outcomes. A survey of students to assess the influence of internal and external factors on their academic performance corroborated the existence of these difficulties in the learning process.

Accurately classifying academic performance is essential for identifying at-risk students and implementing timely interventions. However, traditional methods of evaluation rely heavily on periodic tests, which may not fully capture the underlying factors influencing student performance.

Machine learning has become a powerful tool for analyzing large data sets to uncover hidden patterns and predict values more accurately (Maquen-Niño, Ayelen, Carrión-barco & Adrianzén-Olano, 2023; Maquen-Niño, Nuñez-Fernandez, Taquila-Calderon, Adrianzén-Olano, de la Cruz-VdV & Carrión-Barco, 2024; Yauri, Lagos, Vega-huerta, de la Cruz-VdV, Maquen-Niño & Condor-Tinoco, 2023). A classification model in machine learning seeks to determine the most accurate output based on the input variables, categorizing academic performance levels based on relevant predictors (Pérez-Mármol, Castro-Sánchez, Chacón-Cuberos & Gamarra-Vengoechea, 2023).

Therefore, this study aims to develop a machine learning application to classify the mathematical academic performance of high school students in the first, second, and third year at a private regular educational institution in the Lambayeque region of Peru. To achieve this objective, the research addresses the following questions:

Which predictor variables are important to include in the dataset for academic performance classification?

What is the accuracy level of the machine learning model?

2. Background

2.1. Academic Performance

In the technological era where education plays a fundamental role in human development, academic performance emerges as a preponderant variable in the educational field. Traditionally, academic performance has been assessed through periodic evaluations, where grades are adjusted to the demands of each subject. In mathematics, a subject that demands high levels of logical reasoning and analytical

thinking, the evaluation of performance is particularly relevant. Academic performance is an indicator of the level of learning achieved by students, measured through test scores (Orihuela-Maita, 2019).

Studies such as those by Álvarez-Rodriguez (2020) and Orihuela-Maita (2019) have explored the application of machine learning models to predict academic performance, considering variables such as socioeconomic background, previous grades, and personal learning behaviors.

2.2. Mathematical Academic Performance

Mathematical academic performance has been extensively studied due to its important role in students' logical and comprehensive development.. Candia-Oviedo (2019) states that Logical-mathematical reasoning skills, along with processes such as problem-solving and the interpretation of mathematical language, are considered fundamental skills in student development. A solid mathematical foundation and its continuous development facilitates progress in science and technology and raises the educational level of society. Over time, the concept of academic performance has evolved, adopting new perspectives. Currently, performance is not limited solely to numerical grades but also focuses on soft skills and the student's environment, such as their interest and attention.

2.3. Factors Influencing Academic Performance

Study Modality: The COVID-19 pandemic significantly impacted education, forcing students to adapt to virtual learning environments. The effectiveness of learning in this context varies, especially in subjects like mathematics that require high levels of concentration. Studies highlight how different modalities influence students' ability to effectively grasp mathematical concepts.

Social Environment: Learning experiences are shaped by a student's social context. An individual's sociocultural environment significantly influences his or her educational as well as professional development (Quezada-Mora & Pardo-Frias, 2018). Family relationships: The role of parents is fundamental in students' academic performance. Higher parental education levels often imply greater expectations and support for students.

Economic: The family's economic income often determines whether a student may achieve better academic performance due to the ease of acquiring educational or technological tools, which significantly contributes to their development.

Self-Concept: Students' ability to self-assess their strengths and weaknesses correlates with their performance, especially in mathematics, considering dimensions such as personality, maturity, and confidence. These psychological and cognitive aspects influence how students approach problem-solving and academic challenges. Identifying these factors is decisive, as they can enhance predictive models for academic performance.

2.4. Machine Learning

Machine learning involves algorithms and models that identify the optimal outcome and the associated confidence level, in order to select the most suitable approach for implementation in the prediction process (Vega-Huerta, Pantoja-Pimentel, Quintanilla-Jaimes, Maquen-Niño, de la Cruz-VdV & Guerra-Grados, 2024). It is a branch of artificial intelligence dedicated to developing algorithms capable of learning progressively (Cubas & Niño, 2022; Zegarra-Vargas, 2020). In education, machine learning can be used to classify academic performance and identify at-risk students.

2.5. K-Nearest Neighbors (KNN)

A machine learning algorithm for supervised learning, employed within classification models is K-nearest neighbors (KNN), which assigns a label or category to an unknown data point based on the classification of its nearest neighbors (Abualhaj, Abu-Shareha, Shambour, Alsaaidah, Al-Khatib & Anbar, 2024; Moujahid, Inza & Larrañaga, 2008).

The effectiveness of KNN algorithm in classification tasks extends beyond academic performance prediction. Research introduces three classification models for efficiently predicting iris flower species. The proposed model utilizes Exploratory Data Analysis (EDA) to analyze and preprocess the dataset, employing three classification models: "Logistic Regression," "Support Vector Machine (SVM)," and "K-Nearest Neighbors (KNN)." All proposed models were tested using the Iris dataset and reached maximum accuracies of 96.43%, 98.21%, and 94.64%, respectively (Gupta, Arora, Rani, Jaiswal, Bansal & Dev, 2022).

Other research used weighted KNN classifiers to diagnose Alzheimer's disease, achieving recovery rates of 96.59% for weighted KNN and 93.68% for standard KNN classifiers (Kaur, Thacker, Goswami, Thamizhvani, Abdulrahman & Raj, 2023).

By leveraging this approach, it is possible to develop a predictive model that enhances academic performance classification, providing valuable insights for educators and policymakers.

2.6. Classification Models for Predicting Academic Performance

The use of machine learning in predicting academic performance has been extensively studied. Research used a feedforward multilayer neural network with backpropagation to predict student performance, achieving an accuracy of over 90%. However, the model's effectiveness decreased when assigning precise numerical grades (Álvarez-Rodriguez, 2020). Other research employed Logistic Regression and Random Forest models to predict academic performance among Systems Engineering students, achieving over 75% accuracy when incorporating socioeconomic and educational factors (Orihuela-Maita, 2019).

Morales-Hernández, González-Camacho, Robles-Vásquez, del Valle-Paniagua and Durán-Moreno (2022) conducted a study in which they implemented two machine learning classifiers, a multilayer neural network (MLP) and a gradient boosting (GB) model, to predict academic achievement in Spanish and mathematics in sixth grade and third grade high school students in Tlaxcala, Mexico. Using 13 contextual variables obtained from the National Exams of Academic Achievement in Schools (Enlace), they trained the models with data from 11,036 students. The results showed that the MLP performed better in Spanish, with an accuracy of 70.1% in 2008 and 61.1% in 2011, while the GB model was more effective in mathematics, reaching an accuracy of 68.8% in 2008 and 63.5% in 2011.

Contreras-Bravo, Fuentes and Rivas (2021) explored the use of machine learning techniques, specifically ensemble methods, to analyze students' academic performance. In their research, they highlighted that the combination of multiple models can significantly increase the accuracy of predictions. They also noted that this approach not only improves predictive ability, but also provides a more holistic view of the factors that affect students' academic performance, which facilitates decision-making in educational settings.

A systematic review focused on the use of artificial intelligence techniques to analyze academic performance in higher education institutions. In their study, they compiled and examined several previous works, identifying the most commonly used methodologies and the most relevant findings in predicting and improving student performance using artificial intelligence tools. The research highpoints the potential of these technologies to optimize teaching and support decision making in higher education (Jimbo-Santana, Lanzarini, Jimbo-Santana & Morales-Morales, 2023).

Classification models categorize student performance based on various input variables. An investigation state that classification is an appropriate option when scores must be grouped into predefined categories without decimal values (Vega, Sanez, de la Cruz, Moquillaza & Pretell, 2022).

Some studies have applied KNN for educational purposes, demonstrating its potential for academic performance prediction. For instance, an investigation applied KNN to classify business-related data, achieving a classification accuracy of 70% (Nina-Asto & Vilca-Malpartida, 2018).

3. Methodology

For the implementation of this project, the CRISP-DM methodology was used, defined as a data mining process model that describes the way in which data mining experts approach the problem of adapting to a business objective such as that of an educational institution, through the completion of five phases: understanding the business, understanding the data, data preparation, modeling and evaluation (Galán-Cortina, 2015: page 21).

3.1. Participants

In coordination with the management of the Educational Institution, the grades of students in the 1st, 2nd and 3rd year of secondary education in 2019 were designated as a predefined sample, giving access to printed reports of the averages obtained in the three courses in the area of mathematics: algebra, mathematical aptitude and geometry. This sample was beneficial for the research because it covers the years 2019, 2020 and 2021, in such a way that the students who in 2019 were in 1st, 2nd and 3rd in 2021 would be in 3rd, 4th and 5th year of secondary school respectively, achieving the non-exclusion of information due to the completion of their secondary studies.

3.2. Instruments

Two instruments were used to carry out the research: a document review form and a questionnaire.

The document review form allowed for the collection of educational records of 1st, 2nd, and 3rd-year high school students for the years 2019, 2020, and 2021. This was done to obtain the academic performance in the subject of mathematics and related areas, which will be included in the dataset.

The questionnaire was administered to 711 students who make up the sample, to obtain predictive variables that will constitute the dataset. Before applying the questionnaire, it was previously subjected to an evaluation by five experts, yielding a Cronbach's alpha of 0.88, indicating that it is valid and shows good agreement.

3.3. Measures

To classify students' academic performance, the grades obtained in the mathematics area have been established as the output variable.

In the original dataset, there are three output variables corresponding to the weighted average in the mathematics courses: algebra, academic aptitude, and geometry. Then, during data preparation, these three variables were averaged to obtain the weighted average in the mathematics area on a scale from 0 to 20.

To discretize the values of the weighted average, a scale with three categories was established:

1: When the weighted average was in the range of 0 to 10.

2: When the weighted average was in the range of 11 to 15.

3: When the weighted average was in the range of 16 to 20.

3.3.1. Phase 1: Understanding of the Business

The educational institution of regular basic education is private, that is, in addition to having as its main objective to provide quality education, it has as its secondary objectives to generate profitability for its survival in such a competitive market and that faced great changes and challenges during the pandemic with the implementation of virtual education. It was thanks to this context that several processes were digitalized and automated, among them the process of grade management, which made it possible to demonstrate a deficit in academic performance in the area of mathematics. Based on the generation of the institution's own grade database, the objective was to carry out a study to establish the relationship between social, economic, cultural and technological factors, as well as grades in the area of mathematics, and the classification of academic performance.

3.3.2. Phase 2: Understanding the Data

In order to build the dataset of the model, it was first necessary to determine the factors that are most directly related to academic performance (Gutiérrez-Monsalve, Garzón, Gonzalez-Gómez & Segura-Cardona, 2023; Jaimes-Medrano, Fossion, Flores-Lázaro & Caraveo-Anduaga, 2023; Osorio, Rodríguez & Zúñiga, 2023), by performing two activities:

The documentary review of the educational records of the students provided by the Institution, from which 30 factors were obtained: 6 personal factors such as Id_student, section, gender, age, debt/scholarship status, and belonging to an advanced mathematics group. We also obtained 24 academic factors such as: CTA final average, Attitudinal final average, Social science final average, Algebra 2019 final average, Algebra 2020 final average, Algebra first bimester 2021 average, Algebra second bimester 2021 average, Algebra average third bimester 2021, Algebra average fourth bimester 2021, Algebra performance indicator, Final math proficiency average 2019, Final math proficiency average 2020, Math proficiency average first bimester 2021, Mathematics proficiency average second bimester 2021, Mathematics proficiency average third bimester 2021, Mathematics proficiency average fourth bimester 2021, Mathematics proficiency performance indicator, Geometry final average 2019, Geometry final average 2020, Geometry average first bimester 2021, Geometry average second bimester 2021, Geometry average third bimester 2021, Geometry average fourth bimester 2021, Geometry performance indicator.

On the other hand, a questionnaire was applied to 711 students which was validated by the judgment of five experts with a content validity index of 0.88368 from which 7 factors were obtained: years in school, level of difficulty in the course, perception of performance in the area of mathematics, hours of study in mathematics courses, level of satisfaction with mathematics courses, amount of technological resources available and quality of internet service.

These two data sources were joined by digitizing all the information, obtaining 37 indicators related to academic performance and 711 rows corresponding to 711 students in the 1st, 2nd and 3rd grades of high school. The 37 variables (columns) in the dataset can be observed in Figure 1.

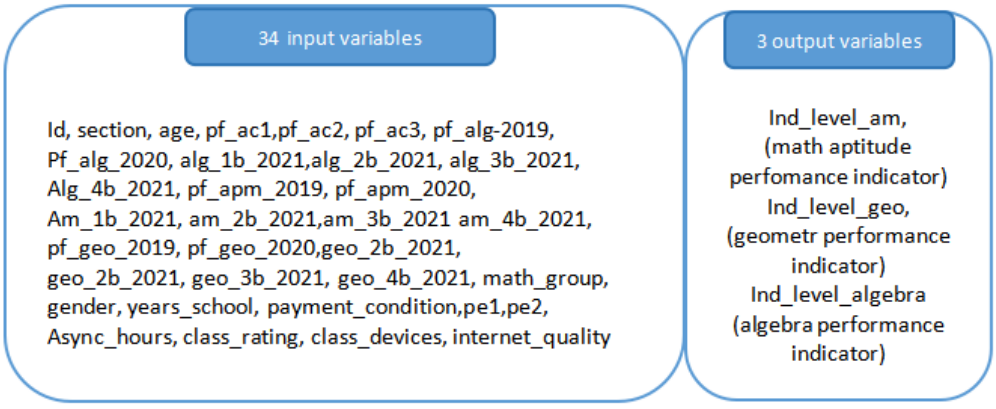


Figure 1. Variables of Dataset

3.3.3. Phase 3: Data Preparation

The first 5 records of each column of the dataset were visualized to verify that the data had been imported correctly, as shown in Figure 2. It can be seen that the qualitative variables are not quantified, i.e. no numerical values have been given to the categories of the qualitative variables.

In order to analyze how many variables are quantitative, i.e. with numerical values, we applied statistical measures such as counting, which additionally allows us to visualize if there are missing data, minimum value, maximum value, and median for each quantitative variable, as shown in Figure 3. It can be seen that

there are 29 numerical variables and that none have missing values, since in all the variables the count is 711 records.

	id	section	age	pf_ac1	pf_ac2	pf_ac3	pf_alg_2019	pf_alg_2020	alg_1b_2021	alg_2b_2021	...	school_years	payment_condition	pe1	pe2	async_hours	class_rating	class_devices	internet_quality
0	1	0	15	13	19	14	16	15	16	16	...	3	1	Easy	Good	3 hours	Satisfied	1	Good
1	2	0	14	16	19	17	16	17	17	16	...	3	1	Very easy	Excellent	From 4 hours to more	Very satisfied	4	Very good
2	3	0	14	11	18	11	11	13	12	12	...	3	1	Easy	Indiferent	1 hour	Indiferent	1	Medium
3	4	0	14	17	17	19	18	16	16	16	...	3	1	Very easy	Excellent	From 4 hours to more	Very satisfied	2	Very good
4	5	0	15	8	9	11	11	11	12	11	...	3	1	Neither easy nor difficult	Indiferent	1 hour	Indiferent	2	Medium

5 rows × 37 columns

Figure 2. First five records of the dataset without processing

	id	section	age	pf_ac1	pf_ac2	pf_ac3	pf_alg_2019	pf_alg_2020	alg_1b_2021	alg_2b_2021	...	school_years	payment_condition	pe1	pe2	async_hours	class_rating
count	711.000000	711.000000	711.000000	711.000000	711.000000	711.000000	711.000000	711.000000	711.000000	711.000000	...	711.000000	711.000000	711.000000	711.000000	711.000000	711.000000
mean	356.000000	3.518987	15.434599	14.115331	15.253165	17.571027	14.867792	14.807314	14.251758	14.510549	...	3.988748	0.933896	1.092827	0.998594	1.030942	1.011252
std	205.392308	2.468818	0.970486	2.643766	2.638403	2.018284	2.653080	2.670286	2.640405	2.670040	...	0.818716	0.248639	0.662766	0.601639	0.624972	0.618847
min	1.000000	0.000000	14.000000	6.000000	6.000000	9.000000	11.000000	6.000000	7.000000	5.000000	...	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	178.500000	1.000000	15.000000	12.000000	14.000000	16.000000	13.000000	13.000000	12.000000	13.000000	...	3.000000	1.000000	1.000000	1.000000	1.000000	1.000000
50%	356.000000	3.000000	15.000000	14.000000	16.000000	18.000000	15.000000	15.000000	14.000000	15.000000	...	4.000000	1.000000	1.000000	1.000000	1.000000	1.000000
75%	533.500000	5.000000	16.000000	16.000000	17.000000	19.000000	17.000000	17.000000	16.000000	16.000000	...	5.000000	1.000000	2.000000	1.000000	1.000000	1.000000
max	711.000000	9.000000	17.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	...	5.000000	1.000000	4.000000	2.000000	3.000000	3.000000

8 rows × 29 columns

Figure 3. Statistical measures of numerical variables of the dataset without processing

Concerning the qualitative variables, those containing alphanumeric values, such as: gender, payment condition, level of difficulty, perception of performance, hours of study, class evaluation, and Internet quality, their quantification was carried out, that is, converting the alphanumeric values into numerical values through categories as shown in Figure 4.

	id	section	age	pf_ac1	pf_ac2	pf_ac3	pf_alg_2019	pf_alg_2020	alg_1b_2021	alg_2b_2021	...	school_years	payment_condition	pe1	pe2	async_hours	class_rating	class_devices	internet_quality
0	1	0	15	13	19	14	16	15	16	16	...	3	1	1	1	1	1	1	1
1	2	0	14	16	19	17	16	17	17	16	...	3	1	1	0	0	0	4	0
2	3	0	14	11	18	11	11	13	12	12	...	3	1	2	2	3	2	1	2
3	4	0	14	17	17	19	18	16	16	16	...	3	1	1	0	0	0	2	0
4	5	0	15	8	9	11	11	11	12	11	...	3	1	2	2	3	2	2	2

5 rows × 37 columns

Figure 4. First five records of the categorized dataset

As the dataset has 3 output variables: ind_nivel_am, ind_nivel_geo, ind_nivel_algebra, and the KNN model needs only one output variable, a new variable called "Result" was created which will store the value of the average of the three initial variables, and subsequently these three variables will be eliminated. Therefore, the dataset will be left with 34 input variables and only 1 output variable ("Result").

Once the preparation was completed, the dataset was divided into training data (25%) and test data (75%), performing a data scaling with mixMaxScaler, which is a statistical procedure to avoid data dispersion.

3.3.4. Phase 4: Modeling

One of the most important parameters in the KNN machine learning model (Begum & Padmannavar, 2023; Buslim, Zulfiandri & Lee, 2023; Zhang, Hu, Jing & Ren, 2023), is the number of neighbors "n_neighbors" that the model will contain. To determine the best value, tests were made when n_neighbors took values in the range from 1 to 10, finding the best value when it was 10 neighbors, obtaining an accuracy of 0.95 in the training and 0.91 in the test. Therefore, the KNN machine learning model will have as input 34 numerical variables and a single output variable that can contain three possible

values: 1 when the performance is poor, 2 when the performance is improvable and 3 when the performance is optimal, as shown in Figure 5.

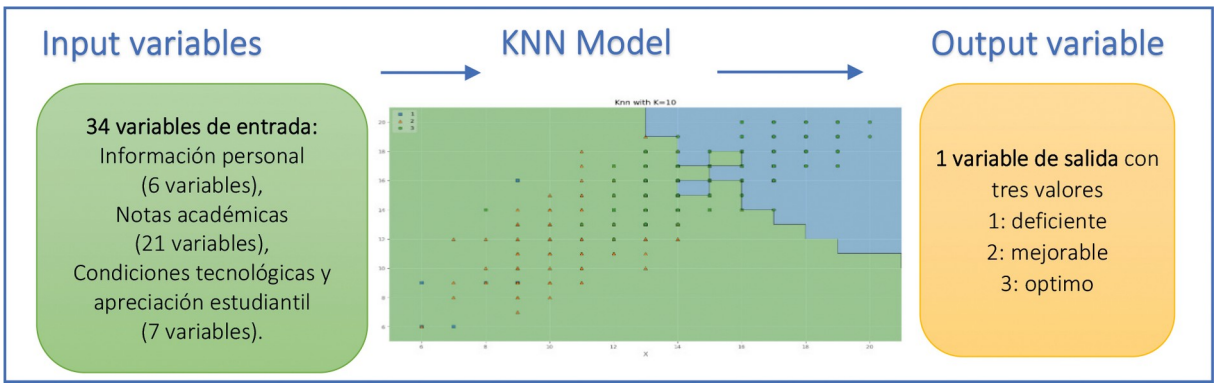


Figure 5. KNN Model Schematic

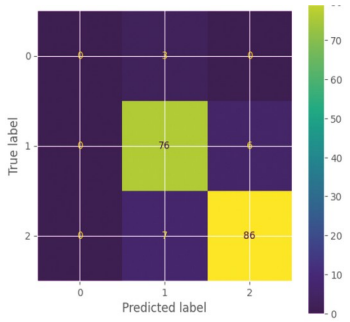
A correlation matrix was conducted to identify the variables most closely associated with the output variable "Result." These variables included grades from the last year 2021 across the second, third, and fourth bimesters, as well as the variables: student perceptions of their performance, assessments of subject difficulty, asynchronous study hours, and internet quality within a virtual education context, as shown in Figure 6.

id	0.342357
section	0.356556
age	-0.088166
pf_ac1	0.756321
pf_ac2	0.743423
pf_ac3	0.552957
pf_alg_2019	0.696653
pf_alg_2020	0.696472
alg_1b_2021	0.658051
alg_2b_2021	0.764317
alg_3b_2021	0.651344
alg_4b_2021	0.653561
pf_apm_2019	0.645596
pf_apm_2020	0.712433
am_1b_2021	0.693351
am_2b_2021	0.681732
am_3b_2021	0.611292
am_4b_2021	0.677561
pf_geo_2019	0.607094
pf_geo_2020	0.674344
geo_1b_2021	0.645569
geo_2b_2021	0.615263
geo_3b_2021	0.635486
geo_4b_2021	0.619384
math_group	0.255165
gender	-0.089887
school_years	-0.132732
payment_condition	-0.255165
pe1	-0.164108
pe2	-0.438995
async_hours	0.595086
class_rating	0.579625
class_devices	-0.022642
internet_quality	-0.172874
Result	1

Figure 6. Correlation matrix of the variables in the dataset

3.3.5. Phase 5: Evaluation

To evaluate the model, we used the confusion matrix shown in Figure 7, observing that in the main diagonal of the matrix we obtained 162 hits in total, 0 from category 1, 76 from category 2 and 86 from category 3, and on the sides we can observe the degrees of error that are 16 in total. He failed 3 times when he said it was 2 and it was really 1, he failed 6 times when he said it was 3 and it was really 2 and he failed 7 times when he said it was 2 and it was really 3, being much higher the number of successes than errors.




```
Accuracy of K-NN classifier on training set: 0.95
Accuracy of K-NN classifier on test set: 0.91
[[ 0  3  0]
 [ 0 76  6]
 [ 0  7 86]]
```

Figure 7. Model Confusion Matrix

4. Results

The performance evaluation of the KNN model was performed using key classification metrics such as accuracy, precision, recall and F1 score. The detailed results are summarized in Figure 8.

	precision	recall	f1-score	support
1	0.00	0.00	0.00	3
2	0.88	0.93	0.90	82
3	0.93	0.92	0.93	93
accuracy			0.91	178
macro avg	0.61	0.62	0.61	178
weighted avg	0.90	0.91	0.90	178

```
KNN Accuracy: 91.01%
KNN F1-score: 90.26%
KNN precision: 89.55%
KNN recall: 91.01%
```

Figure 8. KNN Model Evaluation Metrics

The model achieved an overall accuracy of 91.01%, indicating good overall performance. The weighted average F1 score was 90.26%, reflecting an adequate balance between accuracy and recall in all classes. In addition, the model obtained an accuracy of 89.55% and recall of 91.01%, demonstrating its ability to correctly classify instances while maintaining a good detection rate.

However, a closer examination of the classification report reveals variations in performance among the different classes, which can be attributed to imbalance in the dataset. Class 1, corresponding to students with superior academic performance in mathematics (mean scores between 16 and 20), had only 7 of 711 students in the data set. This limited representation affected the model's ability to correctly predict this category, resulting in an accuracy, recall, and F1 score of 0.00 in category 1.

In contrast, class 2 (medium-performing students) had 349 students, and class 3 (low-performing students) had 355 students. These two classes showed significantly superior prediction performance: class 2 achieved 88% accuracy and 93% recall, and class 3 achieved 93% accuracy and 92% recall.

The mean F1 macroscore was 61%, further highlighting the disparity in class performance. The poor classification of class 1 suggests that the model was unable to learn meaningful patterns due to the small sample size.

5. Discussion

The results obtained in this research reaffirm the potential of machine learning models in the prediction of academic performance, in line with the findings of previous studies. Álvarez-Rodriguez (2020), demonstrated that neural networks could achieve accuracies above 90%, which is consistent with the 91% accuracy obtained in our study. This reinforces the applicability of machine learning techniques to identify patterns of academic performance.

Similarly, Orihuela-Maita (2019), reported that the Logistic Regression and Random Forest models reached accuracies above 75%. The superior accuracy obtained in our study suggests that KNN can be a competitive alternative for the prediction of academic performance when properly adjusted and adapted to the characteristics of the data set.

In addition, Kaur et al. (2023) developed, a model to detect and classify Alzheimer's disease, achieving an accuracy of 93.68% with standard KNN classifiers. In our study, the model achieved a 91% recall, which, although slightly lower, is still very competitive given the complexity of predicting academic performance compared to medical image classification.

Supervised learning models have proven to be effective tools for predicting academic performance. In this context, the study by Vargas-Quispe and Prieto-Luna (2024) evaluated the effectiveness of three algorithms: K-Nearest Neighbors (KNN), Naive Bayes (NB) and Decision Tree (DP). Their findings indicated that the KNN model achieved an accuracy of 81.97%, ranking it as the most accurate among those evaluated. This result highlights the potential of KNN as a suitable method for identifying at-risk students and optimizing early intervention strategies. The relevance of this study lies in the applicability of machine learning in educational settings, providing evidence on the feasibility of these algorithms to improve decision making in academic institutions.

These comparisons highlight the relevance of our study in the broader context of machine learning applications. While our results demonstrate high accuracy, they also reveal key challenges, particularly in predicting underrepresented categories. The classification of students with superior academic performance (class 1) suggests that data imbalance significantly affects the model's ability to generalize across categories.

Addressing this limitation through data balancing techniques, selection of more representative features, hybrid models or alternative classification techniques could improve the fairness of classification and enhance the predictive capacity of the system and ensure more equitable results across all groups of students. These findings highlight the need for more robust approaches in the prediction of academic performance, which opens new lines of research to optimize the application of machine learning models in educational contexts.

Other limitation of this study is that it was conducted at a single institution; however, the specificity of our sample provides a controlled environment for in-depth analysis, allowing us to identify key patterns of academic performance that may be present at other institutions. Thus, this research not only contributes to the field of academic performance prediction, but also highlights the need to refine machine learning models to handle imbalanced datasets more effectively. Future work should explore hybrid model approaches and the incorporation of additional features that could improve classification results, particularly for students in high-achieving categories.

6. Conclusion

According to the results obtained from the Machine Learning KNN model, it can be concluded that it is possible to classify the academic performance of 1st, 2nd and 3rd-year high school students of a regular basic education private school in three possible values: poor, improvable and optimal, with an accuracy level of 91%. The algorithm chosen for the present research was the K-Nearest Neighbors KNN algorithm, due to its easy adaptability and the types of data it handles, the qualitative variables having been

converted to categorical, so that the model only works with numbers, which are easier to handle and help improve the model's performance.

The dataset that fed the model was built in the research, through a documentary review of the records provided by the educational institution regarding personal information and academic data of the students, and through a survey employing a questionnaire on technological conditions and student appreciation of mathematics subjects, which was previously validated by expert judgment, obtaining a Cronbach's alpha of 0. 88368, which after unifying the variables with the data preprocessing, a final dataset of 35 variables was obtained, 34 input variables and 1 output variable. Factors such as student self-perception of achievement and grades in the last year of mathematics subjects were the most related to the classification of academic achievement, as analyzed in the correlation matrix.

7. Limitations and Future Research

One limitation of the model could be the number of predictor variables. In future research, it is possible to increase the number of predictors in the model to enhance accuracy. Another improvement that can be made in future research is to implement additional classification models to obtain a variety of results and a broader perspective, based on factors different from those employed in this study.

8. Data Availability Statement

The student data was acquired from CIMA college, and they have not given their permission for researchers to share their data because there are grades within the dataset. Data requests can be made in Spanish to CIMA college via this email: cima@speedy.com.pe, or in the web page: https://admision.colegiocima.edu.pe/contact-us.

Acknowledgment

Special thanks to the directors and teaching staff of the CIMA school of Chiclayo for allowing the application of the documentary review instruments and the questionnaires applied to the members of the educational community.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Abualhaj, M.M., Abu-Shareha, A.A., Shambour, Q.Y., Alsaaidah, A., Al-Khatib, S.N., & Anbar, M. (2024). Customized K-nearest neighbors' algorithm for malware detection. *International Journal of Data and Network Science*, 8(1), 431-438. https://doi.org/10.5267/j.ijdns.2023.9.012

Álvarez-Rodriguez, R. (2020). *Predicción del rendimiento académico en las Matemáticas de la educación Secundaria mediante Redes Neuronales*. Available at: https://oai.e-spacio.uned.es/server/api/core/bitstreams/23f9fde1-9476-4e1c-ad8b-2b7f8eb2a1ed/content

Begum, S., & Padmannavar, S.S. (2023). Student Performance Analysis using Bayesian Optimized Random Forest Classifier and KNN. *International Journal of Engineering Trends and Technology*, 71(5), 132-140. https://doi.org/10.14445/22315381/IJETT-V71I5P213

Buslim, N., Zulfiandri, Z., & Lee, K. (2023). Ensemble learning techniques to improve the accuracy of predictive model performance in the scholarship selection process. *Journal of Applied Data Sciences*, 4(3), 264–275. https://doi.org/10.47738/jads.v4i3.112

Candia-Oviedo, D.I. (2019). *Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático.* Available at: https://repositorio.unsaac.edu.pe/bitstream/handle/20.500.12918/4120/253T20191024_TC.pdf?sequence=1&isAllowed=y

Contreras-Bravo, L., Fuentes, H., & Rivas, E. (2021). Análisis del rendimiento académico mediante técnicas de aprendizaje automático con métodos de ensamble. *Revista Boletín Redipe*, 10(13), 171-190. https://doi.org/10.36260/rbr.v10i13.1737

Cubas, J.V., & Niño, G.M. (2022). Machine learning model in the detection of phishing websites. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, 2022(E52), 161-173.

Galán-Cortina, V. (2015). *Aplicación de la Metodologia CRISP-DM a un Proyecto de Mineria de Datos en el Entorno Universitario.* Available at: https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

Gupta, T., Arora, P., Rani, R., Jaiswal, G., Bansal, P., & Dev, A. (2022). Classification of Flower Dataset using Machine Learning Models. *AIST 2022 - 4th International Conference on Artificial Intelligence and Speech Technology*. https://doi.org/10.1109/AIST55798.2022.10065178

Gutiérrez-Monsalve, J.A., Garzón, J., Gonzalez-Gómez, D., & Segura-Cardona, A.M. (2023). Factors related to academic performance among engineering students: a descriptive correlational research study [Factores asociados al rendimiento académico en ingenieria: un estudio correlacional descriptivo]. *DYNA (Colombia)*, 90(227), 35-44. https://doi.org/10.15446/dyna.v90n227.107150

Jaimes-Medrano, A.L., Fossion, R., Flores-Lázaro, J., & Caraveo-Anduaga, J.J. (2023). Cognitive flexibility and academic performance in first-year medical students [Flexibilidad cognitiva y rendimiento académico en estudiantes de primer año de medicina]. *Investigacion En Educacion Medica*, 12(47). https://doi.org/10.22201/fm.20075057e.2023.48.23523

Jimbo-Santana, P., Lanzarini, L.C., Jimbo-Santana, M., & Morales-Morales, M. (2023). Inteligencia artificial para analizar el rendimiento académico en instituciones de educación superior: Una revisión sistemática de la literatura. *Cátedra*, 6(2), 30-50. https://doi.org/10.29166/catedra.v6i2.4408

Kaur, M., Thacker, C., Goswami, L., Thamizhvani, T.R., Abdulrahman, I.S., & Raj, A.S. (2023). Alzheimer's Disease Detection using Weighted KNN Classifier in Comparison with Medium KNN Classifier with Improved Accuracy. *Institute of Electrical and Electronics Engineers (IEEE)* (715-718). https://doi.org/10.1109/ICACITE57410.2023.10183208

Maquen-Niño, G.L.E., Ayelen, A., Carrión-barco, G., & Adrianzén-Olano, I. (2023). Brain Tumor Classification Deep Learning Model Using Neural Networks. *International Journal of Online and Biomedical Engineering (IJOE)*, 19(9), 81-92. https://doi.org/10.3991/ijoe.v19i09.38819

Maquen-Niño, G.L.E., Nuñez-Fernandez, J.G., Taquila-Calderon, F.Y., Adrianzén-Olano, I., de la Cruz-VdV, P., & Carrión-Barco, G. (2024). Classification Model Using Transfer Learning for the Detection of Pneumonia in Chest X-Ray Images. *International Journal of Online and Biomedical Engineering (IJOE)*, 20(05), 150-161. https://doi.org/10.3991/IJOE.V20I05.45277

Morales-Hernández, M.Á., González-Camacho, J.M., Robles-Vásquez, H., del Valle-Paniagua, D.H., & Durán-Moreno, J.R. (2022). Algoritmos de aprendizaje automático para la predicción del logro académico. *Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, 12(24). https://doi.org/10.23913/ride.v12i24.1180

Moujahid, A., Inza, I., & Larrañaga, P. (2008). *Métodos Matemáticos en Ciencias de la Computación- Clasificadores K-NN.* Available at: http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t9knn.pdf

Nina-Asto, C.M., & Vilca-Malpartida, A.J. (2018). *Búsqueda de patrones de comportamiento usando machine learning, para la toma decisiones gerenciales en la empresa Chuchuhuasi.* Universidad Andina del Cusco. Available at: http://repositorio.uandina.edu.pe/bitstream/UAC/2795/1/Christian%7B%5C_%7DJean%7B%5C_%7DTesis%7B%5C_%7Dbachiller%7B%5C_%7D2019.pdf

OCDE (2023). Pisa 2022. In *Perfiles Educativos,* 46(183). https://doi.org/10.22201/iisue.24486167e.2024.183.61714

Orihuela-Maita, G.Y. (2019). *Aplicación de Data Science para la Predicción del Rendimiento Académico de los Estudiantes de la Facultad de Ingenieria de Sistemas de la Universidad Nacional del Centro del Perú.* Universidad Nacional del Centro del Perú. Available at: http://repositorio.uncp.edu.pe/bitstream/handle/20.500.12894/5837/TESIS.pdf?sequence=1%7B%5C&%7DisAllowed=y

Osorio, S., Rodríguez, A., & Zúñiga, J. (2023). Sociodemographic Factors and Academic Performance of Medicine and Surgery Students in a Course of Gross Human Anatomy[Factores Sociodemográficos y Rendimiento Académico de Estudiantes de Medicina y Cirugía en un Curso de Anatomía Humana Macroscópica]. *International Journal of Morphology*, 41(5), 1372-1381. https://doi.org/10.4067/S0717-95022023000501372

Pérez-Mármol, M., Castro-Sánchez, M., Chacón-Cuberos, R., & Gamarra-Vengoechea, M.A. (2023). Relationship between academic performance, psychosocial factors and healthy habits in secondary school students. *Aula Abierta*, 52(2), 281-288. https://doi.org/10.17811/rifie.52.3.2023.281-288

Quezada-Mora, P.A., & Pardo-Frias, V.F. (2018). *El entorno social y el aprendizaje*. Available at: https://www.researchgate.net/publication/327403136%7B%5C_%7DEl%7B%5C_%7Dentorno%7B%5C_%7Dsocial%7B%5C_%7Dy%7B%5C_%7Del%7B%5C_%7Daprendizaje

Vargas-Quispe, A.A., & Prieto-Luna, J.C. (2024). Predicción del rendimiento académico estudiantil usando algoritmos de aprendizaje supervisado en una universidad de la selva peruana. *Revista Amazonía Digital*, 3(1), e292. https://doi.org/10.55873/rad.v3i1.292

Vega-Huerta, H., Pantoja-Pimentel, K.R., Quintanilla-Jaimes, S.Y., Maquen-Niño, G.L.E., de la Cruz-VdV, P., & Guerra-Grados, L. (2024). *Classification of Alzheimer's Disease Based on Deep Learning Using Medical Images*. 20(10), 101-114.

Vega, H., Sanez, E., de la Cruz, P., Moquillaza, S., & Pretell, J. (2022). Intelligent System to Predict University Students Dropout. *International Journal of Online and Biomedical Engineering (IJOE)*, 18(07), 27-43. https://doi.org/10.3991/IJOE.V18I07.30195

Yauri, J., Lagos, M., Vega-huerta, H., de la Cruz-VdV, P., Maquen-Niño, G.L.E., & Condor-Tinoco, E. (2023). Detection of Epileptic Seizures Based-on Channel Fusion and Transformer Network in EEG Recordings. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 14(5). https://doi.org/10.14569/IJACSA.2023.01405110

Zegarra-Vargas, E. (2020). *Estrategias lúdicas en el aprendizaje de la matemática del cuarto grado de Educación Secundaria Institución Educativa San José, Chiclayo.* Universidad César Vallejo. Available at: https://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/52830/Zegarra_VE-SD.pdf?sequence=8

Zhang, Y., Hu, N., Jing, R., & Ren, L. (2023). Research on Predictive Model Technology for Student Academic Development Based on Machine Learning. *ACM International Conference Proceeding Series* (793-797). https://doi.org/10.1145/3603781.3603922