Main influential factors in academic performance. Analysis from data mining

Principales factores influyentes en el rendimiento académico. Análisis desde la minería de datos

Omar Danilo Castrillón[1*] https://orcid.org/0000-0003-3713-0696
Jaime Antero Arango[1] https://orcid.org/0000-0003-0626-6013
Luis Fernando Castillo-Ossa[1, 2] https://orcid.org/0000-0002-2878-8229

Recibido 23 de noviembre de 2023, aceptado 30 de diciembre de 2024
Received: November 23, 2023 Accepted: December 30, 2024

ABSTRACT

The objective of this article is to analyze the impact of a series of independent variables on the academic performance of an undergraduate student. A database is constructed with one dependent variable (average) and 23 independent variables: age, school, scholarship, occupation, income, transportation, housing, mother's education level, father's education level, number of siblings, parent's marital status, mother's occupation, father occupation, weekly study hours, non-scientific books read per year, scientific books read per year, class attendance, exam preparation, study method, class notes, attention in class, course participation, economic level. Subsequently, using machine learning and data mining software (WEKA) and applying intelligent techniques, a statistical selection and classification process is conducted to determine the most influential independent variables on the dependent variable. As a result of this process, it is found that, with effectiveness more significant than 81%, the most influential independent variables on the dependent variable, in order of importance, are the mother's education level, father's education level, and class notes. Other factors, such as base tuition scores and income (socioeconomic indicators), were less significant.

Keywords: Academic performance, data mining, weka, statistical selection and classification.

RESUMEN

El objetivo de este artículo es analizar la incidencia de una serie de variables independientes en el rendimiento académico de un estudiante de pregrado. Se construye una base de datos con una variable dependiente (Promedio) y 23 variables independientes: Edad, colegio, beca, trabaja_?, ingresos, transporte, vivienda, nivel educativo madre, nivel educativo padre, número hermanos, estado padres, ocupación madre, ocupación padre, horas estudio semanal, libros no científicos leídos por año, libros científicos leídos por año, asistencia clases, preparación exámenes, forma estudio, apuntes clase, atención clase, participación curso, nivel económico. Posteriormente, por medio del software para el aprendizaje automático y minería de datos (WEKA), y empleando técnicas inteligentes, se realiza un proceso de selección y clasificación estadística, con el fin de establecer las variables independientes más

[1] Universidad Nacional de Colombia. Facultad de Ingeniería y Arquitectura, Departamento de Ingeniería Industrial. Manizales, Colombia. Email: odcastrillong@unal.edu.co, jaarangom@unal.edu.co
[2] Universidad de Caldas. Facultad de Ingenierías, Departamento de Sistemas e Informática. Email: luis.castillo@ucaldas.edu.co, lfcastilloos@unal.edu.co
* Autor de correspondencia: odcastrillong@unal.edu.co

influyentes sobre la variable dependiente. Como resultado de este proceso, se encuentra que con una efectividad superior al 81%, las variables independientes más influyentes sobre la variable dependiente, en orden de importancia son: Nivel educativo madre, nivel educativo padre y apuntes de clase. Otros factores como puntaje básico de matrícula e ingresos, (indicadores socioeconómicos) no tienen una incidencia tan grande.

Palabras clave: *Rendimiento académico, minería de datos, weka, selección y clasificación estadística.*

INTRODUCTION

Academic performance is directly linked to all the elements involved in the learning process, acting as an indicator of these elements. It can be influenced by factors as diverse [1] as friendliness, perception of a healthy diet, and body mass index. Similarly, other authors [2] have shown that even environmental factors, such as ventilation and the concentration of CO and O_3 particles, can influence a student's academic performance. Various demographic, academic, cognitive, and personality factors have also been established as influential in student performance [3]. In some Latin American countries, [4] the incidence of overcrowding associated with poverty and socioeconomic status has been studied concerning academic performance, with findings that these factors negatively influence the learning of subjects such as mathematics and language. Additionally, social networks can significantly affect students' academic performance [5], as it has been observed to improve when these networks are used effectively.

Other studies [6] show that factors such as gender and city of origin do not influence academic performance. However, there is a direct relationship between academic performance and the development of social skills, study habits, and study methods. Similarly, studies based on categorical regression analysis methods [7] attempt to identify the relationship between academic performance and variables such as education, income, marital status, self-motivation, student-student relationships, student-teacher relationships, personal relationships, work experience, location, and environment. In these studies, academic performance is influenced by factors such as age, grades, attendance rate, and marital status. A direct relationship can also be observed between a student's satisfaction level and their academic performance [8]. In general, from a social perspective, a variety of factors have been studied that significantly influence academic performance.

From the perspective of artificial intelligence and mathematics, techniques such as decision trees, support vector machines, and nearest neighbor algorithms [9] have been used to establish the relationship between academic performance in elementary school students and their physical condition. These studies found an inverse relationship between the body mass index and academic performance and a direct relationship between limb strength, upper trunk strength, and aerobic endurance and academic performance. In a similar vein, there are studies [10] showing the relationship between severe nomophobia (fear of not being able to use a cell phone) and poor academic performance, although these results are not statistically significant. Additionally, models based on neural networks and structural equations have been developed to measure student satisfaction [11] linked to academic performance. According to some studies [8], there is also a direct relationship between academic performance and educational as well as socio-cultural services.

Similarly, research based on Machine Learning [12] has successfully predicted academic performance using radial basis function neural networks based on the student's academic history and cognitive abilities, with a success rate of over 93%. Additionally, structural equation models have studied the influence of parents and teachers on student academic performance [13], finding that in some subjects, previous achievements can predict current ones. Other academic performance prediction models based on intelligent techniques, such as neural networks, decision trees, machine learning, and genetic algorithms, can be found in various studies [14], and models based on collaborative learning [15], among others [16], [17].

In line with the previous paragraphs, it is highlighted that many other studies related to the quality of education at various levels can be found in different investigations: the democratization of university management practices for high-quality education [18], violence and migration [19], the influence of demographic and environmental factors on academic performance [20], and the impact of Covid - 19 on university grades [21], among other aspects [22]-[24]. One of the challenges universities face (among many others) is to design strategies and policies that will enable them to address the factors affecting students' academic performance across various domains. These designs would enhance academic performance and improve the future of cities, regions, countries, and, consequently, the global population.

Although there is a wide variety of research related to student academic performance, it is important to note that there is no research applied to the central area of Colombia that attempts to predict academic performance through artificial intelligence techniques and that is also based on a dependent variable called "average" and 23 independent variables, which is the focus of this investigation. The independent variables are: age, school, scholarship, occupational status, income, transportation, housing, mother's education level, father's education level, number of siblings, parental marital status, mother's occupation, father's occupation, weekly study hours, non-scientific books read per year, scientific books read per year, class attendance, exam preparation, study method, class notes, attention in class, course participation, and economic level. As a result of this research, it was found that the most influential independent variables on a student's academic performance, in order of importance, are the mother's education level, the father's education level, and class notes. Other factors, such as basic tuition score, income, and socioeconomic indicators, do not have a significantly impact this process. For better comprehension, the document is structured as follows: 1) Materials and Methods, 2) Results, 3) Discussions, and 4) Conclusions. Finally, two sections are included, one for acknowledgments and another for references.

MATERIALS AND METHODS

The development of this methodology is based on the structure of an existing database referenced in [25] and employed in [26]. This database is restructured into 24 variables (23 independent and one dependent). A survey was prepared using this database as a reference and applied to a university in the central region of Colombia, aimed at a population of 140 students. For better clarity and analysis, the methodology developed in this research is based on the works of various authors [23], [24] and consists of the following steps: 1) Definition of the survey. 2) Definition of the sample size 3) Structure of the database. 4) Consistency of the survey 5) Definition of the WEKA file 6) Ranking and selection of variables 7) Prediction of the dependent variable's status 8) Decision tree 9) Comparison with other artificial intelligence techniques. 10) Identification by clusters.

Step 1: Definition of the survey. A survey was structured following the example in Table 1, using the database referred to in the previous paragraph as a source. This survey consists of 24 questions. The first 23 are associated with each independent variable, and the final question is related to the measurement of the dependent variable.

Step 2: Definition of the sample size. A simple random sampling method is used to determine the sample size, taking the total population of students in the undergraduate program as a reference. This analysis is performed uses the following equation (1) [27].

$$n = \frac{I^2 pqM}{ME^2 + I^2 pq} \tag{1}$$

Where: n: Sample size; I: Confidence interval for an inverted standard normal distribution; p and q: represent the success and failure variations, respectively; M: Student population; E: Sampling error.

Step 3. Structure of the database. The sample size obtained in Step 2 is used to select an equivalent random number of students for the survey. The application of this survey enables the measurement of each independent variable and allows for the prediction of the behavior of the dependent variable.

Step 4: Consistency of the survey. This step calculates Cronbach's Alpha index to validate the questionnaire and determine its internal consistency. This index is computed using the XRealStats Excel add-in. The calculation will generate a value between 0 and 1, with higher consistency as the value approaches 1. Suppose the Cronbach's Alpha index is very low

Table 1. Applied survey.

#	Question	R1	R2	R3	R4
1	Age: (R1: 15 > = and < 18; R2: > = 18 and < 21; R3 > =21 and < 24; R4 > = 24)				
2	School:(R1: private; R2: public; R3: Mixed; R4: other)				
3	Scholarship: (R1: > = 0% and < = 25%; R2: 25 > 0% and < = 50%; R3: >50% and < = 75%; R4 > 75% and < = 100)				
4	Employed: (R1: > = 0% and < = 25%; R2: 25 > 0% and < = 50%; R3: > 50% and < = 75%; R4 > 75% and < = 100)				
5	Income: (R1: very low; R2: low; R3: medium; R4: high)				
6	Transportation: (R1:Bus; R2: Taxi; R3: Car; R4: Other)				
7	Housing: (R1: individual apt; R2: shared apt; R3: Family; R4: Other)				
8	Mother's education level (R1: Elementary; R2: Secundary; R3: University; R4: Postgraduate)				
9	Father's education level (R1: Elementary; R2: Secundary; R3: University; R4: Postgraduate)				
10	Nº Siblings (R1: 0,1; R2: 2; R3: 3; R4: 4 or more)				
11	Parents' status (R1: Married; R2: Free Union; R3: Separated; R4: Either or both deceased)				
12	Mother's occupation (R1: retired; R2: home; R3: private job; R4: public job)				
13	Father's occupation (R1: retired; R2: home; R3: private job; R4: public job)				
14	Weekly study hours (R1: < = 5; R2: > 5 and < = 10; R3: > 10 and < = 15; R4 > 15 and < = 20)				
15	Nº of non-scientific books read per year (R1: 0; R2: 1 and 2 ; R3:3 and 4 ; R4: five or more)				
16	Nº of scientific books read per year (R1: 0; R2: 1 and 2 ; R3: 3 and 4 ; R4: five or more)				
17	Class attendance (R1: Almost never; R2: sometimes; R3: almost always; R4: always)				
18	Exam preparation (R1: A few hours prior; R2: The day prior; R3: A week prior; R4: constantly throughout the semester)				
19	Study method (R1: Alone; R2: With a classmate; R3: group; R4 Other)				
20	Class notes (R1: Doesn't take any; R2: sometimes; R3: Almost always; R4: Always)				
21	Attention in class (R1: Never; R2: sometimes; R3: Almost always; R4: Always)				
22	Course participation (R1: Never; R2: sometimes; R3: Almost always; R4: Always)				
23	Basic tuition score (R1: > = 0 and < 25; R2 > = 25 and < = 50; R3 > = 50 and < 75; R4: > = 75 and < = 100)				
24	Weighted Average on a scale of 1 to 5 (R1 > = 0 and < = 3.0; R2 > = 3.0 < 40; R3 > = 4.0 and < 4.5 and R4 > = 4.5 and < = 5.0) (socioeconomic indicator. R1 Terrible. R4 Excellent)				

(less than 0.5); in this case, questions should be removed one at a time, and the index recalculated to eliminate those questions that generate low consistency in the survey. Only questions associated with the independent variables can be deleted.

Step 5. Definition of the WEKA file. Based on the results obtained in the previous step, the corresponding ARFF file (including header and detail) is structured to be interpreted by the J48 algorithm of the machine learning and data mining platform WEKA[28]. The interpretation of the data using this algorithm allows the construction of a decision tree, which illustrates how the states of the independent variables lead to a specific state of the dependent variable.

Step 6. Ranking and selection of variables. Using the Chi-Square test in the WEKA platform, a variables are ranked to identify the most influential independent variables on the dependent variable. Subsequently, through a Pareto analysis, the least influential independent variables are eliminated, ensuring that the percentage of accuracy in predicting the dependent variable is not significantly affected.

Step 7. Prediction of the dependent variable's status. Using variables selected in the previous step and through a validation and training process with 100% of the data, a prediction process for the dependent variable is performed using the J48 algorithm of the WEKA platform. Subsequently, the respective classification, precision, and confusion matrices are generated, and the relationships between the independent and dependent variables are defined.

Step 8. Decision tree. This step involves obtaining the decision tree, which facilitates predicting the behavior of the dependent variable and its various states. This tree will help identify the main factors (independent variables) influencing a student's academic performance (dependent variable).

Step 9. Comparison with other artificial intelligence techniques. The results obtained are compared with those of other artificial intelligence techniques, using the same calculation parameters to assess the effectiveness of the proposed methodology. The compared techniques are: Bayes, regression, lazy, rules and trees, among others.

Step 10. Identification by Clusters. Using the Expectation-Maximization (EM) algorithm implemented in the WEKA platform, the number of interest groups is determined based on a probability analysis. Once the number of interest groups has been established, these groups are formed using the SimpleKMeans algorithm from the same platform. This process aims to determine if any interest group can predict the behavior of the dependent variable.

RESULTS

As a result of applying each of the steps described in the proposed methodology, the following results are obtained:

Step 1: Definition of the survey. The survey, as defined in Table 1, is administered to a statistically valid sample from an undergraduate program with a population of 460 students at a university in Colombia's central region. In total, 140 surveys were conducted, which is sufficient based on the results obtained in Step 2, where it was established that the sample size should be 121. Table 2 and Table 3 present sections of this survey, though only a fragment is shown due to space limitations.

Step 2: Definition of the sample size. When equation (1) is applied for a population size m= 460, with the following parameters: I = 1.2815, representing the confidence interval obtained using the Excel function for inverted standard normal distribution (0.9); p = 0.5; q = 0.5; E = 0.05 (5% sampling error, for a 90% interval). The obtained sample size (n) is 121.

Step 3. Structure of the database. The database is prepared using the variables selected in the previous step as a reference. A fragment of this database is shown in Table 4.

Step 4: Consistency of the survey. When the Cronbach's Alpha index is calculated using the XRealStats Excel library, a value of 0.39 is generated. The index must be recalculated after each question is eliminated to exclude questions that have a negative effect on it. As a result of this process, questions P3, P4, P5, P8, P9, P12, P17, P18, P20, P21, P23, and P24, are retained (the rest are removed), resulting in a new Cronbach's Alpha index of 0.6, indicating a good consistency.

Table 2. Results of the applied survey. Variables 1-12.

Student	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
1	R2	R2	R4	R4	R2	R4	R3	R3	R3	R3	R3	R2
2	R2	R2	R1	R1	R1	R4	R3	R2	R2	R3	R1	R2
3	R2	R2	R4	R1	R1	R4	R3	R2	R2	R1	R2	R3
4	R4	R2	R1	R1	R1	R1	R3	R1	R1	R1	R1	R2
5	R2	R1	R2	R3	R3	R1	R2	R2	R2	R1	R3	R2
6	R1	R2	R1	R1	R3	R1	R2	R3	R1	R1	R1	R3
7	R2	R1	R1	R1	R3	R1	R3	R2	R2	R2	R3	R1
8	R2	R2	R1	R1	R2	R1	R2	R1	R2	R3	R2	R2
9	R1	R1	R1	R1	R2	R1	R2	R2	R1	R1	R2	R2
10	R2	R1	R1	R3	R3	R3	R4	R2	R2	R1	R1	R4
11	R2	R2	R1	R1	R3	R2	R3	R3	R3	R2	R1	R3
12	R3	R2	R1	R1	R3	R1	R3	R3	R3	R1	R4	R1

Table 3. Results of the Survey. Variables 13-24.

P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24
R3	R3	R2	R3	R4	R3	R1	R3	R4	R4	R1	R2
R4	R3	R2	R1	R4	R3	R1	R4	R4	R2	R2	R2
R3	R4	R4	R4	R4	R3	R2	R2	R3	R2	R1	R3
R3	R2	R1	R1	R3	R2	R1	R3	R3	R3	R1	R3
R2	R1	R1	R1	R4	R4	R2	R3	R4	R2	R1	R3
R3	R3	R3	R1	R4	R3	R2	R3	R4	R2	R1	R3
R4	R4	R1	R1	R4	R3	R1	R3	R3	R2	R2	R3
R3	R4	R2	R1	R3	R2	R1	R2	R3	R2	R1	R1
R4	R4	R2	R1	R4	R2	R1	R3	R3	R2	R2	R3
R3	R3	R3	R4	R1	R3	R1	R3	R4	R3	R3	R3
R1	R4	R2	R2	R4	R4	R3	R3	R3	R2	R3	R3
R4	R4	R1	R1	R1	R3	R4	R3	R4	R3	R2	R3

Table 4. Generated database.

Student	P3	P4	P5	P8	P9	P12	P17	P18	P20	P21	P23	P24
1	R4	R4	R2	R3	R3	R2	R4	R3	R3	R4	R10	R2
2	R1	R1	R1	R2	R2	R2	R4	R3	R4	R4	R20	R2
3	R4	R1	R1	R2	R2	R3	R4	R3	R2	R3	R10	R3
4	R1	R1	R1	R1	R1	R2	R3	R2	R3	R3	R10	R3
5	R2	R3	R3	R2	R2	R2	R4	R4	R3	R4	R10	R3
6	R1	R1	R3	R3	R1	R3	R4	R3	R3	R4	R10	R3
7	R1	R1	R3	R2	R2	R1	R4	R3	R3	R3	R20	R3
8	R1	R1	R2	R1	R2	R2	R3	R2	R2	R3	R10	R1
9	R1	R1	R2	R2	R1	R2	R4	R2	R3	R3	R20	R3
10	R1	R3	R3	R2	R2	R4	R1	R3	R3	R4	R30	R3
11	R1	R1	R3	R3	R3	R3	R4	R4	R3	R3	R30	R3

Step 5. Definition of the WEKA file. The header and details of the ARFF file are constructed using the database generated in step 4 for interpretation by the WEKA platform. Table 5 and Table 6 display each of these files.

Step 6. Ranking and selection of the variables. Using the Chi-Square test on the WEKA platform and based on the ARFF file created in the previous step, variables are ranked. This process yields the weight of each variable, as shown in Table 7.

A Pareto analysis of Table 7 enables the elimination of the last three variables. Since the first eight variables account for more than 80% of the information, they are sufficient to predict the behavior of the dependent variable. Consequently, the analysis proceeds with the first eight variables listed in Table 7.

Step 7. Prediction of the dependent variable's status. With the variables selected in the previous step, the ARFF file is restructured, and the prediction of

Table 5. File ARFF header.

@relation	factors
@attribute P3	{R1, R2, R3, R4}
@attribute P4	{R1, R2, R3, R4}
@attribute P5	{R1, R2, R3, R4}
@attribute P8	{R1 ,R2 ,R3 ,R4}
@attribute P9	{R1 ,R2 ,R3 ,R4}
@attribute P12	{R1 ,R2 ,R3 ,R4}
@attribute P17	{R1 ,R2 ,R3 ,R4}
@attribute P18	{R1 ,R2 ,R3 ,R4}
@attribute P20	{R1 ,R2 ,R3 ,R4}
@attribute P21	{R1 ,R2 ,R3 ,R4}
@attribute P23	{R1 ,R2 ,R3 ,R4}
@attribute P24	{R1,R2,R3,R4}

Table 6. File ARFF details.

P3	P4	P5	P8	P9	P12	P17	P18	P20	P21	P23	P24
R4	R4	R2	R3	R3	R2	R4	R3	R3	R4	R1	R2
R1	R1	R1	R2	R2	R2	R4	R3	R4	R4	R2	R2
R4	R1	R1	R2	R2	R3	R4	R3	R2	R3	R1	R3
R1	R1	R1	R1	R1	R2	R3	R2	R3	R3	R1	R3
R1	R1	R3	R3	R1	R3	R4	R3	R3	R4	R1	R3
R1	R1	R2	R1	R2	R2	R3	R2	R2	R3	R1	R1
R1	R1	R2	R2	R1	R2	R4	R2	R3	R3	R2	R3
R1	R3	R3	R2	R2	R4	R1	R3	R3	R4	R3	R3

Table 7. Ranking of the independent variables.

Variable	Ranking	Cumulative	% Cumulative
P8	35.997	35.997	0.2%
P9	28.85	64.847	36.07%
P20	20.582	85.429	47.52%
P4	14.872	100.301	55.79%
P17	13.651	113.952	63.38%
P21	13.56	127.512	70.93%
P23	13.444	140.9564	78.41%
P5	11.58	152.5364	84.85%
P3	11.12	163.6564	91.03%
P12	9.411	173.0674	96.27%
P18	6.699	179.7664	100%

the dependent variable is performed using the J48 algorithm from the WEKA platform. This process has a success rate of 81.42% when trained and validated with 100% of the data [29]. It is important to note that using all the variables from Table 7 results in only a 1.4% increase in classification accuracy, reaching 82.85%, indicating minimal improvement. The results of this classification process are shown in Table 8, Table 9, and Table 10.

Step 8. Decision tree. As a result of the previous step, the decision tree is obtained, which illustrates the

Table 8. Classification of the dependent variable. %. Success.

Classified Instances	114	81.4286%
Incorrectly Classified Instances	26	18.5714%
Kappa statistic	0.7008	
Mean absolute error	0.1322	
Root mean squared error	0.2571	
Relative absolute error	40.8447%	Correctly
Root relative squared error	64.0514%	
Total Number of Instances	140	

Table 9. Precision matrix.

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.556	0.008	0.833	0.556	0.667	0.971	R1
0.934	0.228	0.76	0.934	0.838	0.923	R2
0.796	0.07	0.878	0.796	0.835	0.942	R3
0.563	0.008	0.9	0.563	0.692	0.953	R4
0.814	0.128	0.826	0.814	0.809	0.937	Average

Table 10. Confusion matrix.

A	b	c	d	classified as
5	2	1	1	a = R1
0	57	4	0	b = R2
0	11	43	0	c = R3
1	5	1	9	d = R4

relationships between each independent variable and the specific states of the dependent variable. This tree is shown in Figure 1, Figure 2, Figure 3, and Figure 4:

Step 9. Comparison with other artificial intelligence techniques. When comparing the results obtained with other artificial intelligence techniques developed on the WEKA platform, the results illustrated in Table 11 are obtained.

Step 10. Identification by clusters. Four interest groups are identified using the WEKA platform's expectation maximization (EM) algorithm. Subsequently, the SimpleKMeans algorithm from the same platform is used to refine these groups, as shown in Table 12.

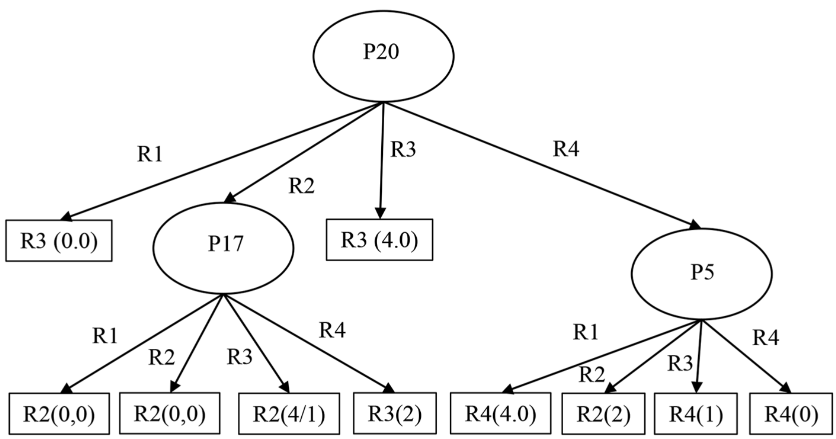

Figure 1. Decision tree for P8 = R1.

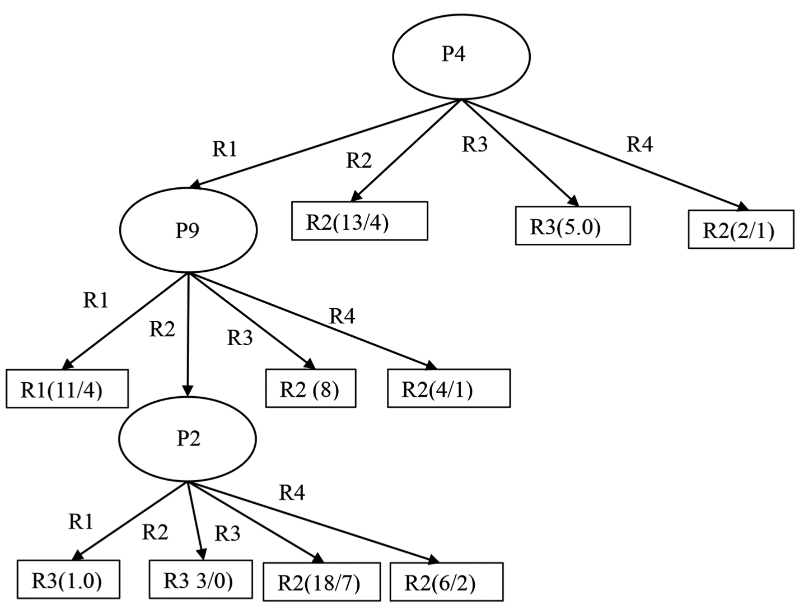

Figure 2. Decision tree for P8 = R2.

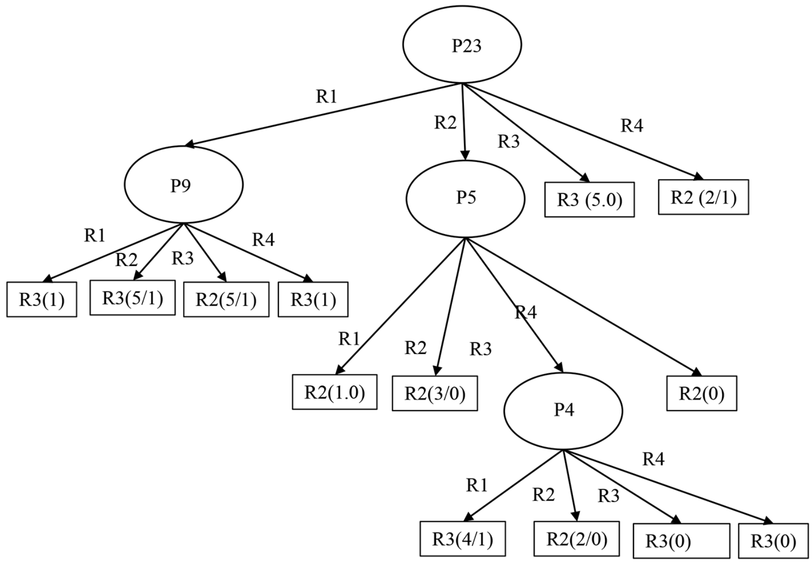

Figure 3. Decision tree for P8 = R3.

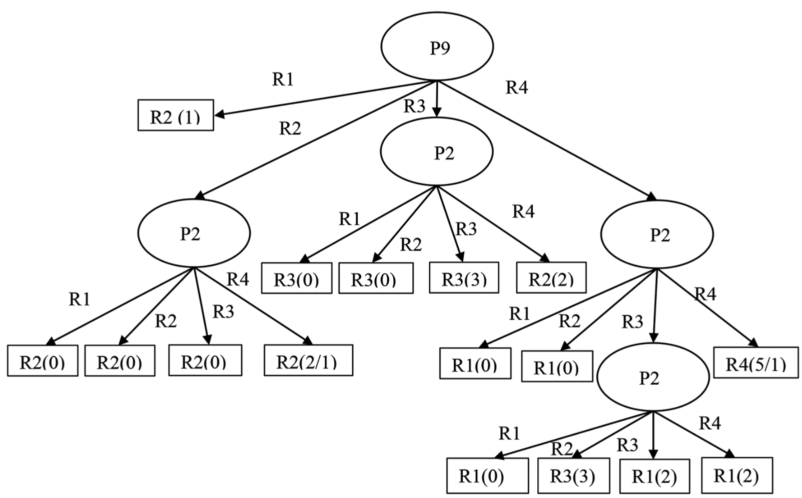

Figure 4. Decision tree for P8 = R4.

Table 11. Comparison with other artificial intelligence techniques.

Technique	% effectiveness
Bayes	67.14
MultiLayer Percerptron	96.42%
SMO	72.85%
IB1	96.42%
IBK	97.14
BFTree	80.71%
J48GRAF	81.42
RandomForest	94.28%

DISCUSSIONS

In the literature reviewed for this research, no similar work is found that predicts student academic performance based on the independent factors outlined in this article. Some authors [6], [7], [8], [11] utilize intelligent techniques to identify relationships between factors such as age, gender, marital status, income level, satisfaction, and academic performance. Other authors [8]-[10] employ artificial intelligence to explore the relationship between academic performance and factors related to physical conditions such as

Table 12. Clusters.

Attribute	Full Data	Grupo 0 (71)	Grupo 1 (37)	Grupo 2 (25)	Grupo 3 (7)
P4	R1	R1	R1	R1	R1
P5	R3	R2	R3	R3	R3
P8	R2	R2	R2	R3	R4
P9	R2	R2	R2	R4	R4
P17	R4	R4	R4	R4	R3
P20	R3	R3	R3	R4	R3
P21	R4	R4	R3	R4	R3
P23	R1	R1	R2	R3	R2
P24	R2	R2	R2	R3	R1

nomophobia and academic services. Although, in general, many related investigations can be found [12], [13], [16], [17], none predict a student's academic average as this research does, nor do they achieve a reduction of more than 62% of the independent variables. It is important to note that this research is contextualized within the central region of Colombia, and its results establish the relationships between independent variables that lead to a specific average, as shown in Table 13.

Table 13. Relationships between the independent variables and the dependent variable.

Relationships	Average
P8 = R1→P20 = R4→P5 = (R1 o R3 o R4)	R4
P8 = R4→P9 = R4→P21 = R4	R4
P8 = R1→P20 = (R1 o R3)	R3
P8 = R1→P20 = R2→P17 = R4	R3
P8 = R2→P4 = R3	R3
P8 = R2→P4 = R1→P9 = R1	R3
P8 = R2→P4 = R1→P9 = R2→P20 = (R1 o R2)	R3
P8 = R3→P23 = R1→P9 = (R1 o R2 o R4)	R3
P8 = R3→P23 = R3	R3
P8 = R3→P23 = R2→P5 = R3→P4 = (R1 o R3 o R4)	R3
P8 = R4→P9 = R3→P21 = (R1 o R2 o R3)	R3
P8 = R4→P9 = R4→P21 = R3→P20 = R2	R3
P8 = R1→P20 = R2→P17 = (R1 o R2 o R3)	R2
P8 = R1→P20 = R4→P5 = R2	R2
P8 = R2 →P4 = (R2 o R4) →	R2
P8 = R2→ P4 = R1→P9 = (R3 o R4)	R2
P8 = R2→ P4 = R1→P9 = R2→P20 = (R3 o R4)	R2
P8 = R3→P23 = R1→P9 = R3	R2
P8 = R3→P23 = R2→P5 = (R1 o R2 o R4)	R2
P8 = R3→P23 = R2→P5 = R3→P4 = R2	R2
P8 = R3→P23 = R4	R2
P8 = R4→P9 = R1	R2
P8 = R4→P9 = R2→P20 = (R1 o R2 o R3)	R2
P8 = R4→P9 = R3→P21 = R4	R2
P8 = R4→P9 = R2→P20 = R4	R1
P8 = R4→P9 = R4→P21 = (R1 o R2)	R1
P8 = R4→P9 = R4→P21 = R3→P20 = (R1 o R3 o R4)	R1

Table 13 illustrates the conditions under which students can achieve a specific average. Students obtain an excellent average (Higher than 90%) when the mother's educational level is low. However, class notes are outstanding, or when the mother and father's educational level are high and class attention is excellent. Students can also achieve a good average (between 80% and 90%) with good class attendance. It is also observed that aspects such as work can influence the academic average, with a lower amount of work associated with a better average.

Furthermore, Table 13 indicates that a good average is obtained when the mother's educational level is good, combined with a good or regular economic level. In contrast, regular averages (Between 60% and 80%) are observed in the following cases: If the mother's educational level is regular or dire, with poor class notes and regular or bad class attendance. Even if the class notes are good, a low-income level can negatively affect academic performance. Similarly, a low basic tuition score can result in a poor average, even if the parents' educational level is good. Part-time employment also influences academic performance. Finally, Table 13 shows that even with excellent educational levels for both parents, the average is bad if attention in class is insufficient (less than 60%).

In general, Table 13 indicates that, although many factors affect the academic average, the mother's educational level (P8) is the most influential variable, followed by other variables such as the father's educational level (P9) and class notes (P20). However, factors such as Basic Tuition Score (P23) and income level (P5) (socioeconomic indicators of a person) do not significantly impact the academic average. These aspects are confirmed in Table 7.

When comparing the results obtained in this methodology to those from other artificial intelligence techniques (Table 11), it is found that although many other artificial intelligence techniques yield better results, only the J48 algorithm (which behaves similarly to Bayesian algorithms) produces a decision tree (Figure 1-4) that illustrates the relationships between the independent variables and the dependent variable. Other intelligent techniques encapsulate the results, and while there are methods to extract them, they are complex and imprecise. Similarly, the clustering analysis of the problem reveals that, on average, specific values for the variables P8, P9, P20, P4, P17, P21, P23, and P5 lead to an average R2 (> = 3.0 and < 4.0), as illustrated in groups 0 and 1 of Table 12. Additionally, it is found that another series of values for the independent variables leads to an average R3 (> = 4.0 and < 4.5) and R1 (> = 0 and < 3.0), as shown in groups 2 and 3 of Table 12; this provides an alternative method for predicting the behavior of the dependent variable, which is not found in the existing literature. However, it is important to note that there is no average value of these independent variables (P8, P9, P20, P4, P17, P21, P23, P5) that results in an average R4 (> = 4.5 and < = 5.0)

Finally, it is noted that in the existing literature, no study establishes the direct relationship between each of these independent variables and a student's specific average nor achieves a reduction of more than 62% of the independent variables; this is of great importance for universities, as it provides a basis for identifying the specific aspects they should focus on if they wish to improve the academic performance of their students. Such improvements would, therefore, contribute to the development of the university, region, and country.

CONCLUSIONS

This research identifies the main independent variables that influence student's academic averages, which can be quantified within a series of ranges. The most influential variables in this process, listed in order of importance and with an effectiveness rating of over 81%, are the mother's educational level, the father's educational level, and class grades. Socioeconomic indicators such as basic enrollment score and income have less impact on this process. Furthermore, the methodology presented in this research is developed under a free machine learning and data mining platform, WEKA. This aspect facilitates its recreation and adaptation to other contexts and regions, and by following the proposed steps, the methodology allows for a reduction of more than 62% of the independent variables, simplifying their management for universities.

REFERENCES

[1] A. Beltrán-Velasco, M. Donoso-González, and V. Clemente-Suarez, "Analysis of Perceptual,

Psychological, and Behavioral Factors that Affect the Academic Performance of Education University Students," *Physiology & Behavior*, vol. 238, Art. no. 113497, pp. 1-4, 2021, doi: 10.1016/j.physbeh.2021.113497.

[2] A. Kabirikopaei, J. Lau, J. Nord, and J. Bovaird, "Identifying the K-12 Classrooms' Indoor Air Quality Factors that Affect Student Academic Performance," *Science of the Total Environment*, vol. 786, Art. no. 147498, pp. 1-7, 2021, doi: 10.1016/j.scitotenv.2021.147498.

[3] P. Victoria, D. Powis, T. Levett-Jones, and S. Hunter, "Factors influencing Nursing Students' Academic and Clinical Performance and Attrition: An Integrative literature Review," *Nurse Education Today*, vol. 32, pp. 903-913, 2012, doi:10.1016/j.nedt.2012.04.011.

[4] D. Contreras, J. Delgadillo, and J. Riveros, "Is home overcrowding a significant factor in children's academic performance? Evidence from Latin America," *International Journal of Educational Development*, vol. 67, pp. 1-17, 2019, doi: 10.1016/j.ijedudev.2019.01.006.

[5] M. Shafiq and K. Parveen, "Social media usage: Analyzing its effect on academic performance and engagement of higher education students," *International Journal of Educational Development*, vol. 98, Art. no. 102738, pp. 2-9, 2023, doi: 10.1016/j.ijedudev.2023.102738.

[6] H. Erdem, "A Cross-Sectional Survey in Progress on Factors Affecting Students' Academic Performance at a Turkish University," *Procedia - Social and Behavioral Sciences*, vol. 70, pp. 691-695, 2013, doi:10.1016/j.sbspro.2013.01.110.

[7] C. Çilan and M. Can, "Measuring factors effecting MBA students' Academic Performance by Using Categorical Regression Analysis: A Case Study of Institution of Business Economics, Istanbul University," *Procedia - Social and Behavioral Sciences*, vol. 122, pp. 405-409, 2014, doi: 10.1016/j.sbspro.2014.01.1362.

[8] T. Van, K. Thi, and H. Thi, "Data survey on the factors affecting students' satisfaction and academic performance among private universities in Vietnam," *Data in Brief,* vol. 33, Art. no. 106357, pp. 1-7, 2020, doi: 10.1016/j.dib.2020.106357.

[9] K. Xu and Z. Sun, "Predicting academic performance associated with physical fitness of primary school students using machine learning methods," *Complementary Therapies in Clinical Practice*, vol. 51, Art. no. 101736, pp. 1-7, 2023, doi: 10.1016/j.ctcp.2023.101736.

[10] M. Qutishata, E. Lazarusa, A. Razmyb, and S. Packianathand, "University students' nomophobia prevalence, sociodemographic factors and relationship with academic performance at a university in Oman," *International Journal of Africa Nursing Sciences*, vol. 13, Art. no. 100206, pp. 2-6, 2020, doi: 10.1016/j.ijans.2020.100206.

[11] S. Aghaei, Y. Shahbazi, M. Pirbabaei, and H. Beyti, "A hybrid SEM-neural network method for modeling the academic satisfaction factors of architecture students," *Computers and Education Artificial Intelligence*, vol. 4, Art. no. 100122, pp. 1- 41, 2023, doi: 10.1016/j.caeai.2023.100122.

[12] O. Olabanjo, A. Wusu, and M. Manuel, "A machine learning prediction of academic performance of secondary school students using radial basis function neural network," *Trends in Neuroscience and Education*, vol. 29, Art. no. 100190, pp. 1-18, 2022, doi: 10.1016/j.tine.2022.100190.

[13] C. Nunes, T. Oliveira, M. Castelli, and F. Cruz-Jesus, "Determinants of academic achievement: how parents and teachers influence high school students' performance," *Heliyon*, vol. 9, no. 2, Art. no. e13335, pp. 1-32, 2023, doi: 10.1016/j.heliyon.2023.e13335.

[14] N. Beckham, L. Akeh, G. Pratama, and J. Moniaga, "Determining factors that affect student performance using various machine learning methods," *Procedia Computer Science*, vol. 216, pp. 597-603, 2023, doi: 10.1016/j.procs.2022.12.174.

[15] J. Muñoz, C. Simón de Blas, F. Rodríguez, and A. García, "Collaborative Learning in Management Subjects to University Students a Multilevel Research to Identify Group Profile, Engagement and Academic Performance," *The International Journal of management Education*, vol. 21, no. 1. Art. no. 100762, pp. 1-25, 2023, doi: 10.1016/j.ijme.2022.100762.

[16] S. Rajendran, S. Chamundeswari, and A. Sinha, "Predicting *the Academic Performance*

of Middle- and High-School Students Using Machine Learning Algorithms," *Social Sciences & Humanities Open*, vol. 6, no. 1, Art. no. 100357, pp. 1-33, 2002, doi: 10.1016/j.ssaho.2022.100357.

[17] X. Wang, Y. Zhao, C. Li, and P. Ren, "ProbSAP: A comprehensive and high-performance system for student academic performance prediction," *Pattern Recognition*, vol. 137, Art. no. 109309, 2023, doi: 10.1016/j.patcog.2023.109309.

[18] A. Sen, K. Kabak, F. Tuysus, and D. Kuzalic, "Democratization of university management for quality higher education," *Procedia - Social and Behavioral Sciences*, vol. 58, pp. 1491-1504, 2012, doi: 10.1016/j.sbspro.2012.09.1136.

[19] M. Padilla-Romo and C. Pelufo, "Violence-induced migration and peer effects in academic performance," *Journal of Public Economics*, vol. 217, Art. no. 104778, 2023, doi: 10.1016/j.jpubeco.2022.104778.

[20] B. Minaravesh and O. Aydin, "Environmental and demographic factors on childhood academic performance in Los Angeles county: A generalized linear elastic net regression model," *Remote Sensing Applications: Society and Environment*, vol. 30, Art. no. 100942, 2023, doi: 10.1016/j.rsase.2023.100942.

[21] J. Ferrer, E. Iglesias, I. Blanco-Gutiérrez, and J. Estavillo, "Analyzing the impact of COVID-19 on the grades of university education: A case study with economics students," *Social Sciences & Humanities Open*, vol. 7, no. 1, Art. no. 100428, pp. 2-8, 2023, doi: 10.1016/j.ssaho.2023.100428.

[22] M. Yukhymenko-Lescroart, "The role of achievement goals and identity in academic performance and misconduct of college athletes: Considering sport-to-school spillover," *Heliyon*, vol. 9, no. 2, Art. no. e13435, pp. 1-15, doi: 10.1016/j.heliyon.2023.e13435.

[23] O. Castrillón, W. Sarache y S. Ruiz-Herrera, "Predicción del rendimiento académico por medio de técnicas de inteligencia artificial", *Formación Universitaria*, vol. 13, no. 1, pp. 93-102, 2020, doi: 10.4067/S0718-50062020000100093.

[24] O. Castrillón, W. Sarache y S. Ruiz-Herrera, "Predicción de las principales variables que con llevan al abandono estudiantil por medio de técnicas de minería de datos", *Formación Universitaria*, vol. 13, no. 6, pp. 217-228, 2020, doi: 10.4067/S0718-50062020000600217.

[25] D. Dua and C. Graff, "UCI Machine Learning Repository. Irvine, CA: University of California," archive.ics.uci.edu. [Online]. Available: http://archive.ics.uci.edu/ml

[26] N. Yilmaz and B. Sekeroglu, "Student performance classification using artificial intelligence techniques," in *10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019. Advances in Intelligent Systems and Computing,* vol. 1095, R. Aliev, J. Kacprzyk, W. Pedrycz, M. Jamshidi, M. Babanli, and F. Sadikoglu, Eds. 2020, pp. 596-603.

[27] C. Valdivieso, R. Valdivieso y O. Valdivieso, "Determinación del tamaño muestral mediante el uso de árboles de decisión", *Investigación & Desarrollo*, vol. 11, pp. 148-176, 2011, doi:10.23881/idupbo.011.1-4e.

[28] I. Witten, M. Hall, E. Frank, and J. Christopher, *Data Mining. Practical Machine Learning Tools and Techniques*, Cambrige, USA: Morgan and Kaufman, 2017, doi: 10.1016/C2015-0-02071-8.

[29] L. Castillo "Influentilfactors," github.com [Online]. Available: https://github.com/luisfercastilloCOL/Influentilfactors