scientific reports

Check for updates

OPEN

A model for predicting academic performance on standardised tests for lagging regions based on machine learning and Shapley additive explanations

Mario Suaza-Medina[1,2], Rita Peñabaena-Niebles[2]✉ & Maria Jubiz-Diaz[2]

Data are becoming more important in education since they allow for the analysis and prediction of future behaviour to improve academic performance and quality at educational institutions. However, academic performance is affected by regions' conditions, such as demographic, psychographic, socioeconomic and behavioural variables, especially in lagging regions. This paper presents a methodology based on applying nine classification algorithms and Shapley values to identify the variables that influence the performance of the Colombian standardised test: the Saber 11 exam. This study is innovative because, unlike others, it applies to lagging regions and combines the use of EDM and Shapley values to predict students' academic performance and analyse the influence of each variable on academic performance. The results show that the algorithms with the best accuracy are Extreme Gradient Boosting Machine, Light Gradient Boosting Machine, and Gradient Boosting Machine. According to the Shapley values, the most influential variables are the socioeconomic level index, gender, region, location of the educational institution, and age. For Colombia, the results showed that male students from urban educational institutions over 18 years have the best academic performance. Moreover, there are differences in educational quality among the lagging regions. Students from Nariño have advantages over ones from other departments. The proposed methodology allows for generating public policies better aligned with the reality of lagging regions and achieving equity in access to education.

Keywords CRISP-DM, Educational data mining, Lagging region, Machine learning, Shapley additive explanations, Standardised test

The volume of available data during the last decade has grown exponentially. It has led companies, governments and individuals to exploit the knowledge provided by it[1]. Different economic sectors realised that they had accumulated data that could be useful with investment in technology. Therefore, researchers have turned data warehouses into sources of information where knowledge can be obtained from data to improve every organisation and make it more competitive[2]. To date, the tool used by institutions is descriptive statistics, i.e., tables, indicators and graphs, because they are easy to develop and interpret[3]. However, descriptive, predictive and prescriptive analytics can be performed with current technology to understand past information and predict future trends. Therefore, every organisation invests resources in robust tools for a better decision-making process[4].

The educational sector has been aware of the knowledge acquisition process based on data warehouses. In 2005, academics on data mining applied these techniques for solving problems in the education field to boost the growth and development of innovative technologies[5]. With this new trend, several studies have been focused on predicting academic performance, school dropout and student behaviour, among others[6–9]. Research has considered variables such as gender, family income, distance from educational institutions, type of school, absence, classroom environment demographic, psychographic and geographic variables supported by the

[1]Department of Informatics and Computer Science, Universidad de Zaragoza, Maria de Luna 1, Zaragoza 50018, Spain. [2]Department of Industrial Engineering, Universidad del Norte, Km 5 Via Puerto Colombia, Barranquilla 081007, Atlántico, Colombia. ✉email: rpena@uninorte.edu.co

different government and academic institutions databases[10]. Identifying these variables allows the analysis of the standardised tests that measure the added value given by each institution in the academic development of every student and their accessibility to higher education. Therefore, educational policies can be designed based on the results of descriptive and predictive analysis of those tests to impact students[11].

In Colombia, as in other countries, many institutions have stored data without exploiting knowledge. One of the government entities interested in the value of information is the Colombian Institute for the Evaluation of Education (ICFES). Its mission is to evaluate the training offered by educational institutions at different levels using standardised tests. These tests are applied in various stages of school. The results obtained in these exams determine the student's skills gained during their training process, and it is an instrument for improving the academic quality of institutions. It links the results with the synthetic index of educational quality (ISCE) to measure the quality improvement in the country's educational institutions over a determined period[12].

The results of standardised tests on developing countries (i.e., Colombia) have been heterogeneous, particularly in lagged regions. Economic factors, population growth and the aggregation in a geographical area should be analysed to identify them. Maldonado and Meza[13] defined lagged territories based on the economic differences among many geographic areas given by development, i.e., regions are identified, classified and organised according to the benefits each one has received from the country's development. Regions' growth also depends on geographic location, transportation network, natural resources and qualified human capital. According to several theories, economic development is related to physical capital, human capital and innovation[14]. Pugh[15] stated that education promotes regional economic growth since it is usually the centre of strategies and the core of economic emergence from knowledge.

According to the above, it is clear that the lagging is directly related to the quality of education. This implies that the opportunities for higher education are unequal for people in the same country due to differences in sociodemographic contexts. Therefore, this paper aims to identify the variables that affect academic performance in lagging regions using educational data mining (EDM) jointly with Shapley values (SHAP), as they are backwards in their education system. Moreover, the main contribution of this work is to identify the variables of lagging regions that may hinder the implementation of strategies designed by the government to improve the education system in the medium and long term. As they are lagging regions, the education solutions must be comprehensive, i.e., they must avoid solving isolated problems. Therefore, the factors that hamper access to higher education are addressed through school training. The proposed methodology provides a basis for designing governmental strategies that improve those aspects that influence the student's performance in their progress through elementary education. In this way, achieving equity in the quality of education in the different regions is possible.

The rest of the paper is organised as follows. Section 2 reviews recent related works on educational data mining. Section 3 presents the proposed methodology based on Cross Industry Standard Process for Data Mining (CRISP-DM). Section 4 details the results and discusses the obtained findings. Finally, 5 concludes this paper.

Related works

EDM is the field focused on applying statistical, machine learning, and data mining algorithms to different types of educational data. This concept appeared in 2005 and resulted from the interest of researchers in areas such as computer science, education, psychology, neuropsychiatry and human behaviour[16–19], agriculture[20], and statistics[21,22]. The main objectives of EDM are focused on behaviour detection, feedback for supporting teaching, performance prediction, recommendations to students, class planning, study groups, social network analysis, course construction, creation of early warnings and data visualisation[23].

Romero and Ventura[5] identified trends, the most used algorithms and listed contexts in which EDM could be useful. Baker and Yacef[24] described the evolution of this research field and highlighted recent contributions. Huebner[25] showed how EDM helps to improve students' success in the learning process. Peña-Ayala[26] identified the influence of education systems and how EDM is used for analysing them. Shahiri et al.[27] and Kumar and Singh[28] developed techniques for predicting students' performance, progress or regression, and outperformance in the classroom and analysed behavioural patterns and academic dropout.

Satyanarayana and Nuckowski[29] used three classification techniques: Decision Trees (DT), Random Forest (RF) and Naive Bayes (NB) to improve data quality and prediction metrics. The accuracy was enhanced by eliminating noise-generating variables and instances. Also, the results provided rules for understanding the factors that influence student outcomes. Venkatachalapathy et al.[30] analysed the most used tools in EDM that focus on visualisation, association, classification, text mining, pattern detection, clustering, web mining and logical rules. Widyahastuti and Tjhin[31] compared the results of linear regression and a multilayer perceptron (MLP) to predict the outcome of the last exam in a bachelor's degree, using a data set of 50 students. The MLP was the algorithm with the best results. Kumari et al.[32] studied how behavioural characteristics related to the learning experience during students' training affect academic performance. Four classification algorithms of the WEKA tool were used: ID3, K-nearest neighbour, SVM and NB. They concluded that an accurate prediction is achieved by performing a good analysis of behavioural characteristics. Adejo and Connolly[33] proposed a set of multiple models to predict the academic performance of undergraduate students. They used different data sources and three classification algorithms (DT, SVM and Artificial Neural Networks (ANN)). Results showed that implementing heterogeneous ensemble techniques helps identify students at dropout risk.

Ma and Zhou[34] used DTs and SVM to find relevant characteristics that affect students' approval rates based on two categories (pass and fail). The authors used demographic variables, analysed the dependence between them and included the orientation provided by experts. Optimised using grid search, SVM was the best performance algorithm. Kumar et al.[35] proposed data visualisation in EDM, focusing on graphic language. They implemented Python programming language, the Jupyter Notebooks development IDE, and the MatplotLib

and Seaborn libraries. The data set contained 22242 instances with categorical and quantitative variables. Based on the analysis of this set, correct predictions from the visualisation activities were achieved. Hellas et al.[36] and Muzzammel[37] conducted a literature review in data mining, analysing the variables and the different techniques used by researchers from 2009 to 2019. They provided an update on all the methodologies implemented so far. Khan et al.[38] used four algorithms of the WEKA tool: NB, DT, RF and MLP to analyse students' performance in educational institutions. DT outperformed the other algorithms with an accuracy of 0.88. The study helps students to predict their probably final grades and change their academic behaviour to achieve better results.

Livieris et al.[39] implemented semi-supervised learning techniques to predict the student's academic performance in high school. They used a data set with the grades of 3716 students who studied mathematics between 2007 and 2016. The use of these techniques improved the accuracy, even with unlabeled data. Prasanalakshmi and Farouk[40] also used classification algorithms to predict academic performance. They used the WEKA tool and found that DT and random forest algorithms gave the best results but with an unpromising accuracy of 0.56.[41] used NB, DT and ANN as classification algorithms with academic and demographic variables to develop a Kalboard platform. Results showed that the classifier with the highest value was the neural network, with an accuracy of 0.78.

Nahar et al.[42] analysed engineering students' behaviour and academic performance using DT, NB, PART, bagging, boosting and RF. The students' results were predicted based on the pre-requisite course performance and the grades until the midterm exam. The best models were DT and NB. Eleyan et al.[43] explored machine learning techniques such as classification trees, regression trees, logistic regression, and multiple regression to predict the final grades of secondary school students. The study concluded that classification trees and logistic regression outperformed the other methods. In addition, they help to identify students at risk and implement tailor-made interventions. Dinh-Thanh and Thi-Ngoc-Diem[44] used RF, XGBoost and Light GBM to predict and identify features that influence the academic performance of high school students. Results indicated that the cumulative GPA, age, class, father/mother occupation, and learning online contribute to the student's performance.

Ghosh[45] applied different classification techniques such as ANN, SVM, RF, and DT to predict academic performance. Findings suggested that these methods provide results that improve administrative and teaching staff in educational institutions. Alamgir et al.[46] implemented linear regression, RF regressor, and MLP to predict the final graduating cumulative GPA of undergraduate students based on the grades in pre-requisite courses and a relative grading scheme. Results showed that RF performed the best among all the classifiers. Alghamdi and Rahman[47] used NB and RF to predict academic achievement and support secondary school students. The models' performance was validated using several evaluation metrics, and results showed that NB had the best accuracy. Nayak et al.[48] analysed students' academic success, behaviour, and demographics using DT, NB, RF and MLP algorithms. Results showed that behavioural features influence the model's accuracy, indicating the importance of students' attitudes in achieving the desired academic outcomes.

Recently, Batool et al.[49] provided a systematic literature review of 260 research studies focused on data mining algorithms, tools, and students' attributes. Results concluded that DT is the most used algorithm, but ANN, SVM, and RF are trending. In addition, the authors indicated that academic records and demographic factors are the best predictors of students' future results. The literature reviewed shows a high application of machine learning tools to analyse academic performance in educational institutions like DT, RF, NB and ANN. However, these techniques have yet to be implemented to analyse academic performance in lagging regions where the socioeconomic context has a strong influence. Moreover, other algorithms have been successful in other research problems, such as Extreme Randomised Trees, Extreme Gradient Boosting Machine and Light Gradient Boosting Machine. This article proposes its implementation in the analysis of academic performance and compares it with those traditionally used to identify the best performer.

Proposed methodology

The proposed methodology was an adaptation of CRISP-DM[50], which has been applied by different authors in the education field[5,11,21,51–53]. This paper uses CRISP-DM to analyse the results of the standardised test taken by senior students in Colombia: the Saber 11 test. The exam is applied in all regions of Colombia to characterise academic performance, and its structure is similar to the final school year tests carried out in other countries. Its purpose is to measure the essential competencies that guarantee the student has the necessary knowledge to enter higher education. Also, it measures the added value the school has in students' training. The CRISP-DM methodology is widely used in academia and industry and has several modifications. Figure 1 presents the workflow of the CRISP-DM methodology.

Business understanding

This initial stage focuses on understanding the problem to determine its scope, objectives, and requirements for collecting the appropriate data and correctly interpreting the results. Based on the aim of this research, Figure 2 presents the designed plan for predicting the academic performance of the standardised test.

In detail, the plan results in the following steps:

1. Define what is a lagging area and which ones belong to this category.
2. Obtain the data set from the ICFES database.
3. Select variables for the study.
4. Clean the data obtained from the database.
5. Standardise and scale data.
6. Visualise and interpret data through an exploratory analysis.
7. Select and develop the most appropriate techniques of machine learning based on previous studies.

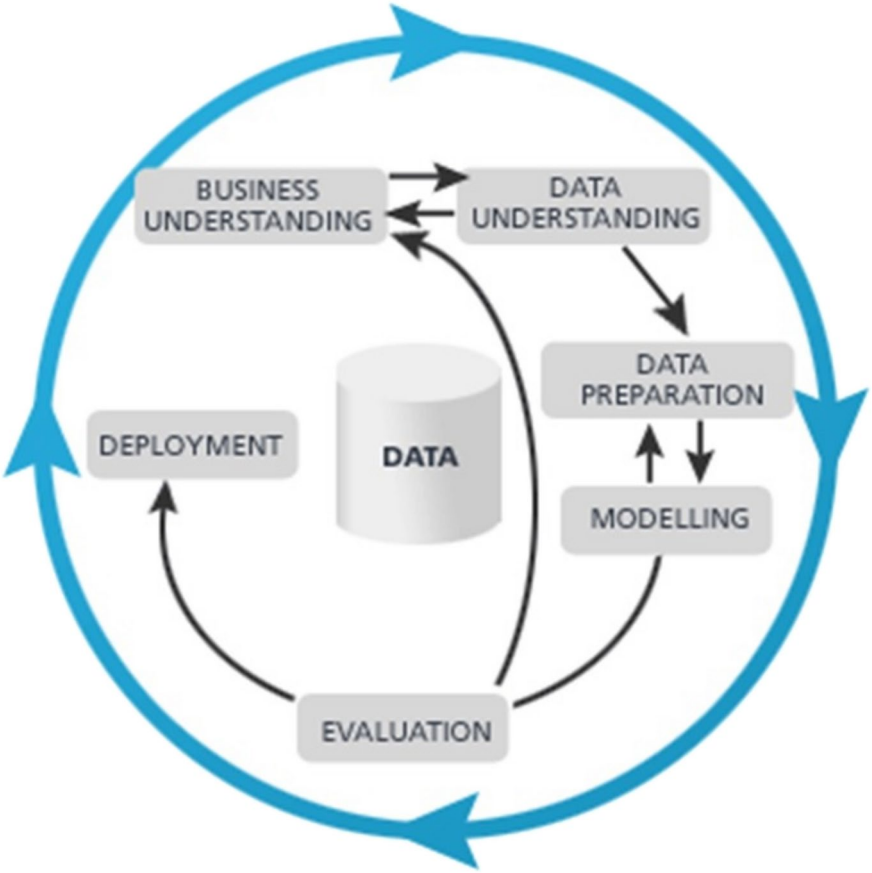

Fig. 1. CRISP-DM methodology: general outline.

8. Build a data set for training and validation.
9. Analyse and validate the proposed model.
10. Conclude based on the obtained results and propose recommendations.As mentioned in Section 1, the concept of lagging is based on the differences in economic development among geographical areas and how each has benefited from industrialisation and investment in infrastructure[13]. A region is considered a lagging one based on the economic factors, growth and the concentration of population in geographical locations[54]. Therefore, a territory with a higher population density, industrial progress, and infrastructure will have a higher level of economic development than other regions. On the other hand, a lagging region is one where the quality of life of its inhabitants remains the same or worsens due to factors such as poor transport links, low skilled human capital, and low educational quality[55].

Studies aim to identify a country's lagging regions based on poverty indices such as monetary poverty, the GINI coefficient and multidimensional poverty. The first one is calculated by dividing the current income of the unit of expenditure by the number of members of the same family. This result is compared with the cost of acquiring food or non-food goods with minimum subsistence capacity[56]. The second one measures the degree of inequality in income distribution. The third one is based on fifteen indicators grouped into five dimensions to determine whether a household can be considered poor (if deprivation on these indicators is at least 33%). According to a country's context, lagging regions are identified and analysed based on the proposed methodology. According to economic, demographic and social studies, the lagging areas in Colombia are shown in Table 1[57,58].

Then, academic and sociodemographic information of the students who took the test during the years considered in this study (2016, 2017 and 2018) was obtained from the ICFES-FTP server. This research was conducted for only three years because the test structure changed at the beginning of 2016, i.e., open questions were replaced by multiple-choice questions in which the type of rating was different. Table 2 presents data characteristics.

Nine algorithms were selected based on previous studies for predicting academic performance using classification models in non-lagged regions. They are used in the modelling stage to design a model based on machine learning to identify the variables that influence academic performance in the Saber 11 test carried out in lagging regions. With the obtained results, decision-makers at academic institutions can focus on improving educational quality indices and increasing the number of students accessing higher education. It is suitable to mention that the proposed methodology has the following limitations:

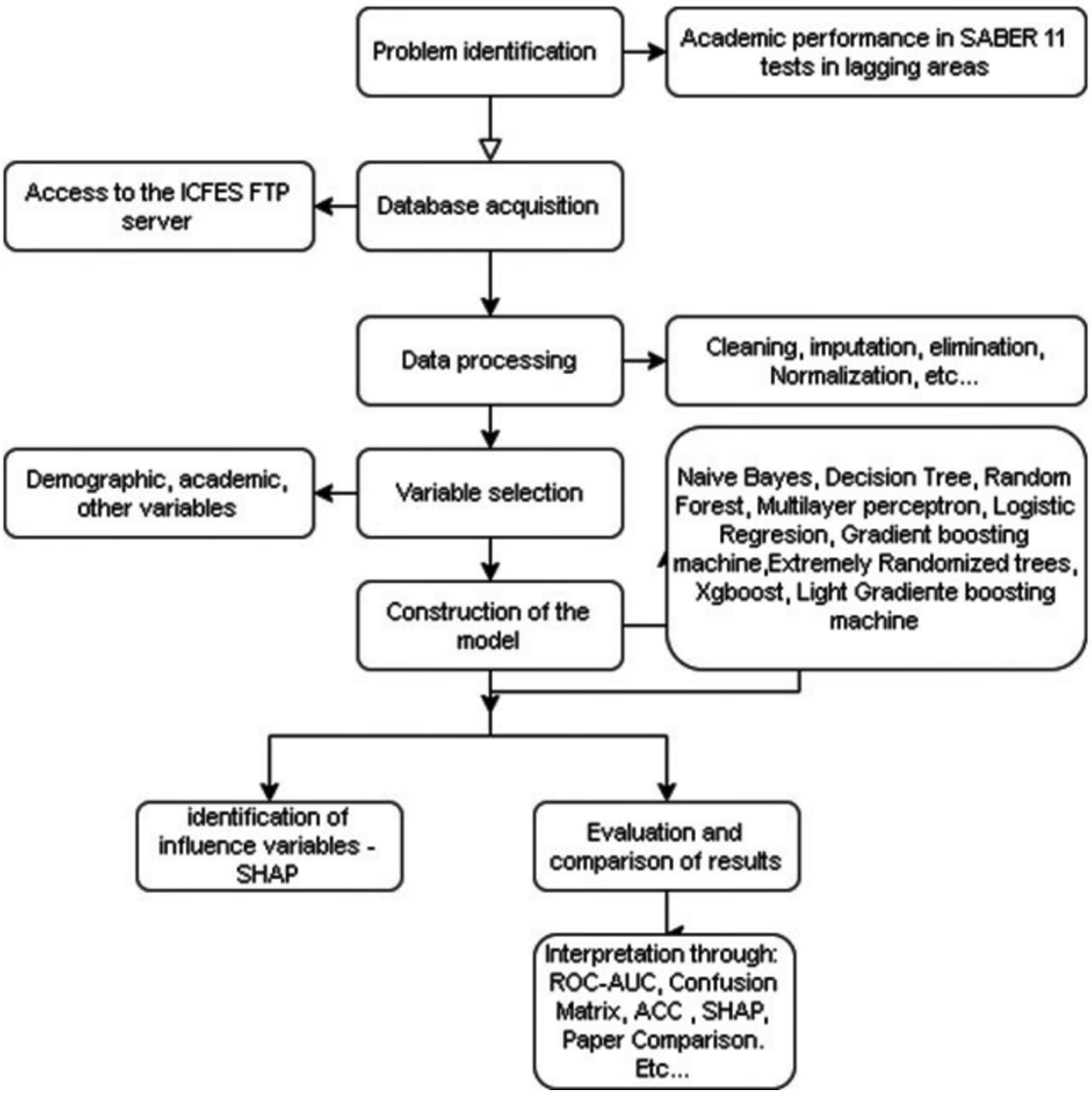

Fig. 2. Steps for designing the classification model based on CRISP-DM methodology.

Region	Department
Amazonía	Amazonas, Caquetá, Guainía, Guaviare, Putumayo, Vaupés
Andina	Norte de Santander
Caribe	Bolivar, Cesar, Cordoba, La Guajira, Magdalena, Sucre
Orinoquía	Arauca, Vichada
Pacífico	Cauca, Chocó, Nariño

Table 1. Lagged departments in Colombia.

Year	Imported format	Number of instances	Number of variables
2016-2	CSV	548206	83
2017-2	CSV	546162	86
2018-2	CSV	549934	83

Table 2. Available information from ICFES database.

- Information is limited, as access to the ICFES server requires membership in an educational institution.
- Data pre-processing must improve the quality of the information received by ICFES. Some instances were eliminated because they had incomplete information.

Data understanding

The FTP service provided the available data for 2016, 2017 and 2018. The three data sets were merged, and the dimensionality of the new database was solved by eliminating uncommon variables since each data set has a different number of variables. Fortunately, these variables were unnecessary as they are unrelated to students' academic performance. Examples are the value of school tuition fees, the career to be studied at the university and information about the university the student wishes to attend in the future. As a result, 67 variables (28 numerical and 39 categorical) were considered for the next stage. Both exploratory and correlation analyses were carried out to identify trends, data quality and highly related variables. The correlation analysis of the numerical and categorical variables (that provide valuable information to the study) was performed using the predictive power score proposed by Wetschoreck[59], as shown in Figure 3.

Results showed that most demographic variables correlate among them and with the response variables. Also, the scores and quartiles for analysing students' performance are related to the performance level per student. Therefore, they were excluded from the study. In addition, the results concluded that:

- A student's freedom status has a direct impact on all variables.
- Access to services and family conditions correlate with each other.
- Parental education is related to the results of standardised test scores.
- Ethnicity is related with other variables.

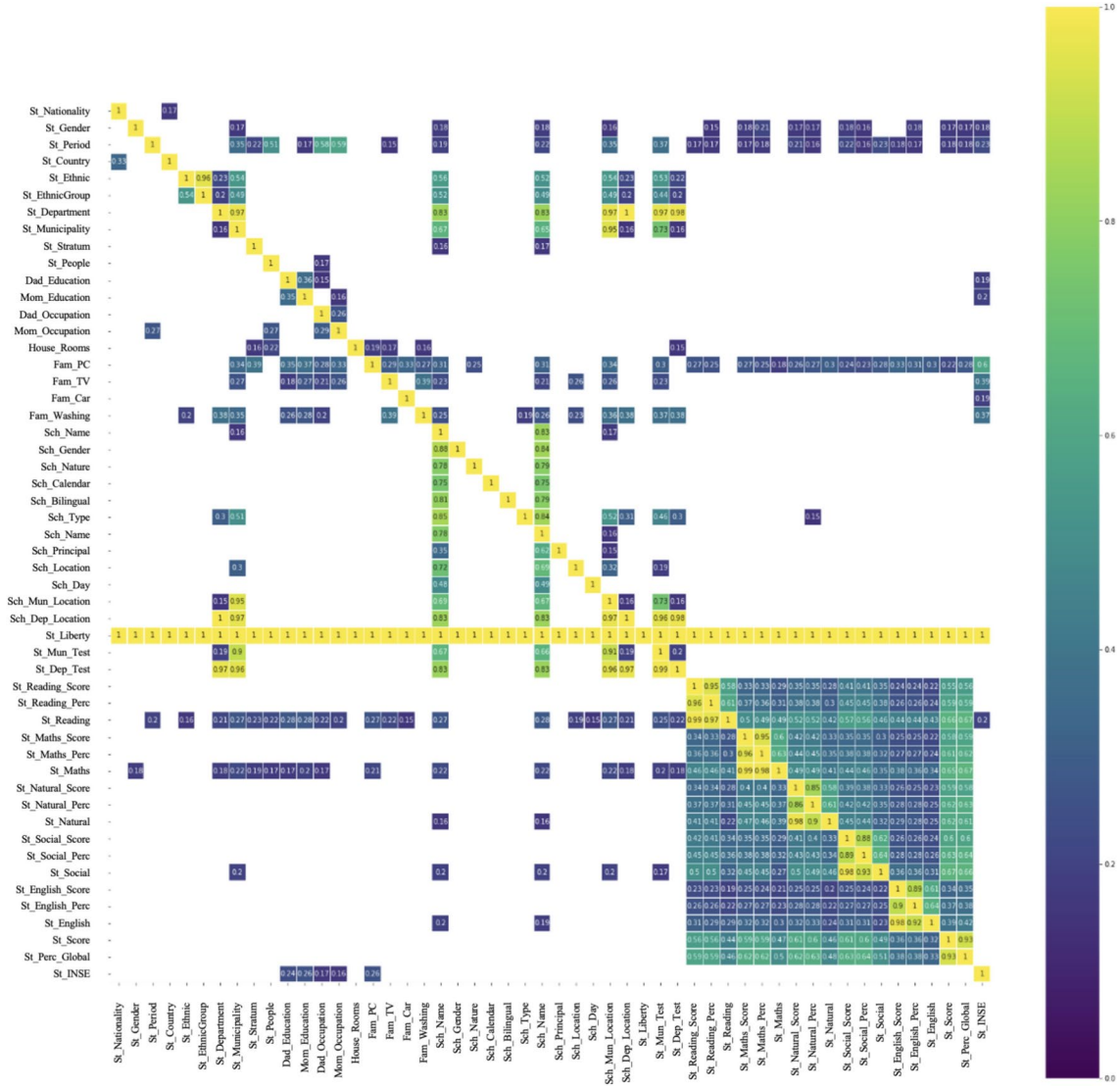

Fig. 3. Correlation analysis.

- Sociodemographic variables are correlated with each other.The analyses conducted to understand the data using exploration and correlation techniques were used as input for the next stage of the methodology.

Data preparation

This stage is focused on preparing the information for the final data set. As mentioned in Section 3.2, it is necessary to clean up the data to eliminate the noise variables identified in the previous stage. It is essential to mention that no algorithm was used for data pre-processing. The cleaning was carried out manually using the following steps:

- Variables must have the same font format.
- Delete spaces and accentuation.
- Each instance was checked to ensure that all students had complete information. If a student had missing information, that instance was completely deleted.
- Noise factors and correlated variables were deleted.The noise variables were codes that identified the student, city, department of residence, school information, and the place where the student took the exam. DANE uses this information to prepare the reports it publishes annually. These codes do not add value to the study because the information about the students, such as school characteristics and students' residence, are included in other variables of the final database. Therefore, they were eliminated to ensure the information is transparent to the classification algorithm. As a result, 30 variables were eliminated. Non-lagged regions were also deleted. The cleaning resulted in 29 categorical and 8 numerical variables with 365854 instances for training and validating the nine classification algorithms. Table 3 presents the selected variables.

Modelling

For the modelling stage, nine classification algorithms were selected. The best hyper-parameters for each algorithm were determined using a random search. The synthetic minority oversampling technique (SMOTE) improved the results. It created synthetic samples from the existing ones[60]. Data sets were divided into two groups: 70% for training and 30% for validation. The selected algorithms are the following:

- Extreme Gradient Boosting Machine (XGBM): This algorithm was proposed by Chen and Guestrin[61] as an extension of the Gradient Boosting Machine algorithm. It is characterised by its speed and prediction accuracy, and its modelling is composed of several weak classifiers, which improve based on the previous classifier at each iteration. Its main limitation is the treatment of categorical variables, as they require coding to be accepted by the model.
- Light Gradient Boosting Machine (LGBM): The algorithm was introduced by Microsoft developers. It is used in classification and regression problems and is characterised by its speed, low memory consumption and good modelling accuracy and precision[62]. Unlike other ensemble-type algorithms, it allows trees to grow vertically as leaves to obtain good accuracy.
- Multi-layer Perceptron (MLP): They are information processing systems that mimic the nature of biological neural networks[63]. They are composed of nodes (neurons) organised in layers and connected by communication links. The algorithm uses a MLP, where the output of each neuron depends on the output of all neurons in the previous layer. This technique is frequently applied in classification and regression problems.
- Decision Tree (DT): It is a technique that is widely used because of its easy visualisation. It has a hierarchical structure consisting of nodes and leaves, where each node is a particular attribute and each leaf is a class. Each node is responsible for classifying elements in the most accurate way[64].
- Gradient Boosting Machine (GBM): This technique involves many successive, shallow and weak trees based on sequential model building. A model is built based on the gradient of the loss function that evaluates the model performance, i.e., the smaller it is, the better the performance[65].
- Random Forest (RF): It is a machine learning algorithm where a set of trees is trained to create several decision tree models[66]. Each tree receives a data set, delivers a classification, and the forest takes the classifications to choose the most selected prediction[67].
- Naive Bayes (NB): It is one of the simplest algorithms for classification problems. It is characterised by its easy construction, application to large volumes of data and provides easy interpretation. The Bayesian classifier is based on the principle that one class is not related to another[68].
- Logistic Regression (LR): It can be used as a linear combination of two or more explanatory variables. The algorithm is characterised by predicting the probability of selecting a categorical value, i.e. it does not predict a numerical value[69].
- Extremely Randomised trees (EXRT): This technique, proposed by Geurts et al.[70], can create regression or classification models. It is characterised by being computationally efficient and working with multi-class problems. Its operation is slightly similar to random forests, but randomness within the training.A new technique called SHAP proposed by Lundberg and Lee[71] is used to determine the variables that influence academic performance in lagging areas. This approach explains the output of any machine-learning model and interprets the results of black-box models by calculating the contribution of each predictor. For this, the Shapley values are calculated based on the game theory in a coalition. These values indicate how to fairly distribute the prediction according to the marginal contribution[72,73]. Shapley values are defined by Equation 1.

$$\phi_i(f,x) = \sum_{S \subseteq al/\{i\}} \frac{|S|!(M-|S|-1)!}{M!}[f_x(S \cup \{i\} - f_x(S))] \tag{1}$$

Type	Variables	Description	Values
Categorical	Dad_Education	Father's maximum educational level	See Appendix A
	Dad_Occupation	Father's occupation	See Appendix A
	Fam_Car	Does the family have a car?	Yes, No
	Fam_PC	Does the family have a computer?	Yes, No
	Fam_TV	Does the family have a television?	Yes, No
	Fam_Washing	Does the family have a washing machine?	Yes, No
	House_Rooms	How many rooms does the student's	1,2,3,4,5,6,7,8,9,10
		house have?	or more
	Mom_Education	Mother's maximum educational level	See Appendix A
	Mom_Occupation	Mother's occupation	See Appendix A
	St_Age	Range of student's age when taking	Age < 15 years,
		the exam	15 $\leq$ Age $\leq$ 18 years,
			Age > 18 years
	St_Country	Country of residence	See Appendix A
	St_Department	Department of residence	See Table 1
	St_English	Student's performance level in English exam	1,2
	St_Ethnic	Does the student belong to an ethnic group?	Yes, No
	St_EthnicGroup	What ethnicity does the student belong to?	See Appendix A
	St_Gender	Student's gender	Male, Female
	St_Liberty	Is the student deprived of liberty?	Yes, No
	St_Nationality	Student's nationality	See Appendix A
	St_People	How many people are at student's home?	1-2,3-4,5-6,7-8,9 or more
	St_Stratum	Student's socioeconomic stratum	1,2,3,4,5,6,NA
	St_TypeDoc	Type of ID	TI,CR,CE,NES,PE,PEP
	Sch_Bilingual	Is the school a bilingual institution?	Yes, No
	Sch_Calendar	Type of academic calendar	A, B, Other
	Sch_Day	School day of the educational institution	Full day, Morning, Afternoon,
			Night, Saturday, Unique
	Sch_Gender	School's gender	Male, Female
	Sch_Location	Geographical region where the educational	Rural, Urban
		institution is located	
	Sch_Nature	Is the school a public or private institution?	Official, Not official
	Sch_Principal	Is the school the headquarters of	Yes, No
		the educational institution?	
	Sch_Type	Type of academic institution	Academic, Technician, NA
Numerical	St_INSE	Student's socioeconomic level index	From 1 to 100
	St_Maths	Student's performance level in	1,2
		mathematics exam	
	St_National	Student's performance level compared to	$\geq$ national average,
		the national average	$\leq$ national average
	St_Natural	Student's performance level in natural	1,2
		sciences exam	
	St_Period	Period in which the student takes the exam	20162, 20172, 20182
	St_Reading	Student's performance level in critical	1,2
		reading exam	
	St_Score	Student's overall score	From 1 to 500
	St_Social	Student's performance level in social	1,2
		sciences exam	

Table 3. Selected lagging regions variables for training.

Evaluation

As was mentioned before, this research aims to (1) design models with good predictive capabilities and (2) assess whether they provide answers to what is stated in the understanding phase of the problem. This paper considers the following metrics for testing the selected algorithms:

- Confusion matrix: It was introduced by Provost and Kohavi[74] to represent the performance of the algorithms in feature selection. This matrix compares the results between the predictive model classification and the actual classification values[75]. Table 4 shows its structure.

- In a confusion matrix, TP is a true positive, FP is false positive, FN is false negative and TN is true negative.
- Accuracy (*AC*): Evaluates the performance of a predictive model based on the proportion of correctly predicted instances. It can be calculated using the following expression:

$$AC = \frac{TP + TN}{TP + FN + FB + TN}$$

- Sensitivity (*TPR*): Represents the fraction of *TP* instances out of the total instances ($TP + FN$) to identify correctly classified true positives, as follows:

$$TPR = \frac{TP}{TP + FN}$$

- Specificity (*FPR*): Identify the correctly classified true negatives. It can be calculated as follows:

$$FRP = \frac{TN}{TN + FP}$$

- Precision (*PR*): Refers to the closeness of a measurement result to the true value. It cam be calculated with the following equation:

$$P = \frac{TN}{TN + FP}$$

- F-Score ($F - S$): It is an indicator that integrates the sensitivity and precision of the predictive model, as follows:

$$F - S = 2\left(\frac{P \times TPR}{P + TPR}\right)$$

Deployment

This stage depends on the conditions proposed for the research. It is optional because some studies require a detailed report, while others require implementing a whole process of data mining decision-making in institutions. This stage covers data acquisition to results presentation.

Results and discussion

This section describes the results obtained from the stages of the proposed methodology for designing a classification model to analyse the academic performance of lagged regions in Colombia. Based on the selected variables mentioned in Table 3, the complexity of the problem was decreased by reducing the number of classes. This problem can be considered a binary one due to the nature of the data and the fact that there is no qualification scale to assess the performance level of every student regarding the national average. Two classes were determined based on the general average of the three selected years. The prediction will be focused on classifying a student using a performance level, as follows: (1) calculate the overall average considering all the years as shown in Table 5 and (2) define two classes as "greater than or equal to the national average" and "lower than the national average.

	Target	
Model	**Positive**	**Negative**
Positive	TP	FP
Negative	FN	TN

Table 4. Confusion matrix.

Year	Annual national average	General average
2016	258.2	254.7
2017	255.2	
2018	250.8	

Table 5. Overall average from the annual averages of years 2016, 2017 and 2018.

Algorithm	AC	AUC	CV AC	Std deviation AC	CV AUC	Std deviation AUC
EXRT	0.77	0.85	0.76	0.00	0.86	0.00
LGBM	0.77	0.85	0.76	0.00	0.86	0.00
GBM	0.77	0.83	0.75	0.00	0.83	0.00
RF	0.76	0.84	0.76	0.00	0.84	0.00
XGBM	0.76	0.85	0.76	0.00	0.85	0.00
MLP	0.74	0.83	0.74	0.00	0.82	0.00
LR	0.71	0.79	0.71	0.00	0.79	0.00
DT	0.69	0.75	0.69	0.00	0.75	0.00
NB	0.53	0.54	0.53	0.00	0.54	0.01

Table 6. Algorithm validation using accuracy and ROC curve cross-validation.

Author	Algorithm	AC
Algorithms used in this paper	EXRT	0.77
	LGBM	0.77
	GBM	0.77
	RF	0.76
	XGBM	0.76
	MLP	0.74
	LR	0.71
	DT	0.69
	NB	0.53
Timaran-Pereira et al.[11]	DT	0.67
Pereira et al.[76]	DT	0.66
Solano et al.[77]	LTM	0.73

Table 7. Algorithm comparison.

The number of instances per class was obtained to determine whether the data were unbalanced. Results showed that 148471 students reached a score greater or equal to the national average, while 218383 scored lower than the national average. Since data is unbalanced, the SMOTE oversampling technique[60] was applied before executing the algorithms to get more synthetically created instances that will have characteristics of the main classes. Data were divided into 70% and 30% for training and validation, respectively. Each algorithm was executed, and the results were analysed based on the selected performance metrics. With this, it is possible to verify the capabilities of each algorithm to perform a correct classification on a given data set. Table 6 shows the accuracy of each algorithm. Appendices Appendix B and Appendix C present the rest of the performance metrics and the confusion matrix per algorithm. According to Table 7, the algorithms EXRT, LGBM and GBM have the highest accuracy of 0.77. They are followed by XGBM and RF with an accuracy of 0.76, while MLP with 0.74. The other algorithms did not perform acceptably. Also, EXRT, RF, XGBM, GBM and LGBM had good accuracy and area under the curve (AUC), as shown in Table 6 and Figure 4.

EXRT, LGBM and GBM are the ones with the best accuracy. However, the two first mentioned have an AUC of 0.85, while the third one has an AUC of 0.83. Checking the number of instances correctly classified by LGBM and EXRT based on the confusion matrix (see Appendix C), the results of true positives and true negatives are very similar. Since both have equal performance in the mentioned metrics, the one with the shortest computational time is the best: LGBM. This algorithm also has a precision-recall curve of 0.84, as shown in Figure 5, and the cross-validation of 10 folds yielded an average accuracy of 0.75 and AUC of 0.86 (see Table 6). Therefore, the algorithm achieves a well-performed classification of instances.

Most of the algorithms used in this research are black-box types. Their primary objective is to provide results but no detail on how each variable influences decision-making, leaving aside the interpretability given to academic stakeholders. As a solution, Lundberg and Lee[71] proposed SHAP, a new technique for interpreting

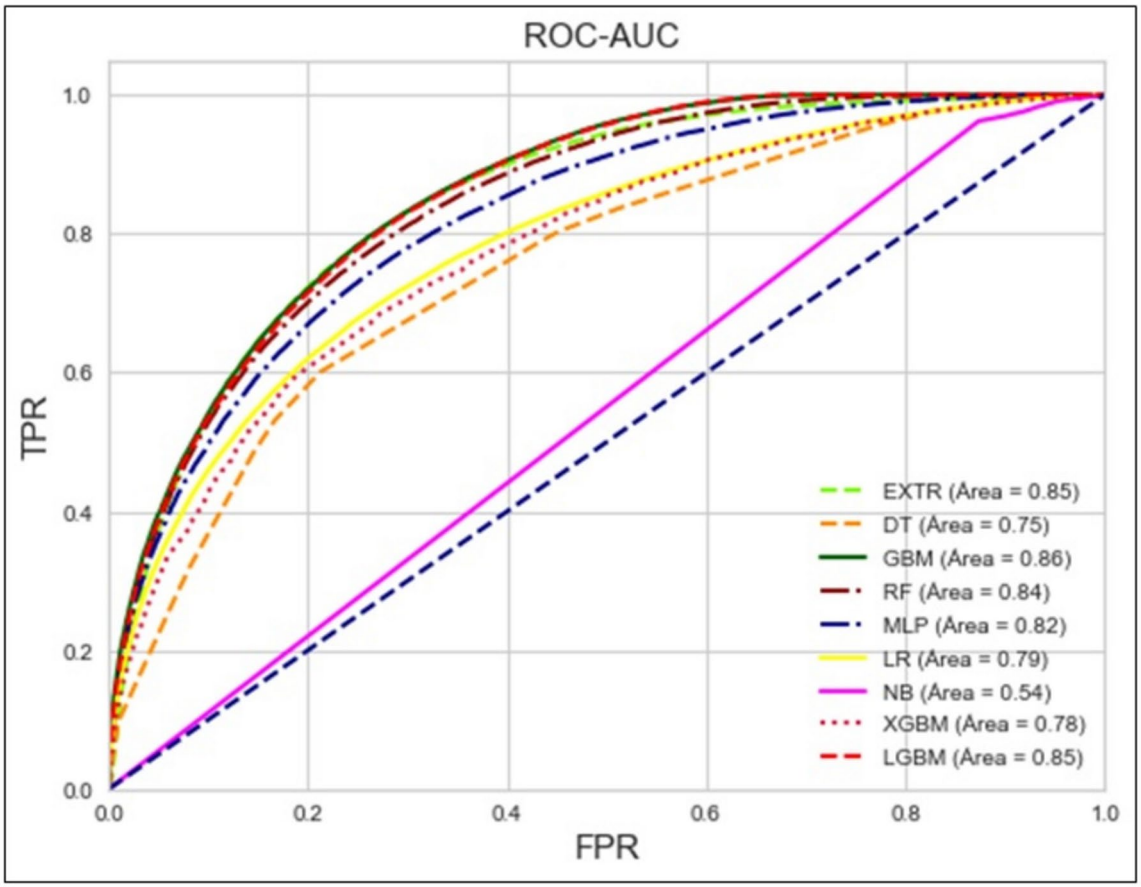

Fig. 4. AUC per classification algorithm.

black-box models. This method calculates the Shapley value from a game theory coalition. Based on this, the Shapley values show how the prediction is equally distributed among the characteristics according to their marginal contribution[72,73]. Figures 6 and 7 show the variables influencing academic performance in Colombian lagging areas and their impacts.

From these results, the following features can be highlighted:

- The most relevant variable is the socioeconomic level index. It has a high negative impact on the performance level, i.e., the higher the index, the higher the performance will be.
- Belonging to an official school has a high and positive effect on performance.
- The different school hours have a high and positive impact. Studying at night, on Saturday and in the morning influence the student's academic performance.
- The mother's education has both positive and negative effects on performance.
- Male gender has a high and negative impact on academic performance, i.e., female students may have a better performance on the exam.
- Access to goods and services at home has a positive and high impact.
- Be over 18 years old effects in a high and positive way. This can be related to the abilities the students gain in the passing of years.
- Belonging to an afro-descendant community positively affects performance.
- The school location in urban areas has both positive and negative impact.
- Belonging to an ethnic group has a high and positive effect. However, these communities have very limited access to services and cultural progress.
- The number of rooms affects positively as this quantity increases.
- Being located in a rural area impacts negatively academic performance since these people do not have the adequate conditions for effective learning processes.Based on these findings, the following actions could be suggested to address educational performance disparities:
-
- Enforce financial assistance and social support programs to encourage students from lower socioeconomic backgrounds.
- Increase funding and resources for public schools to maintain and enhance their positive impact on student performance.

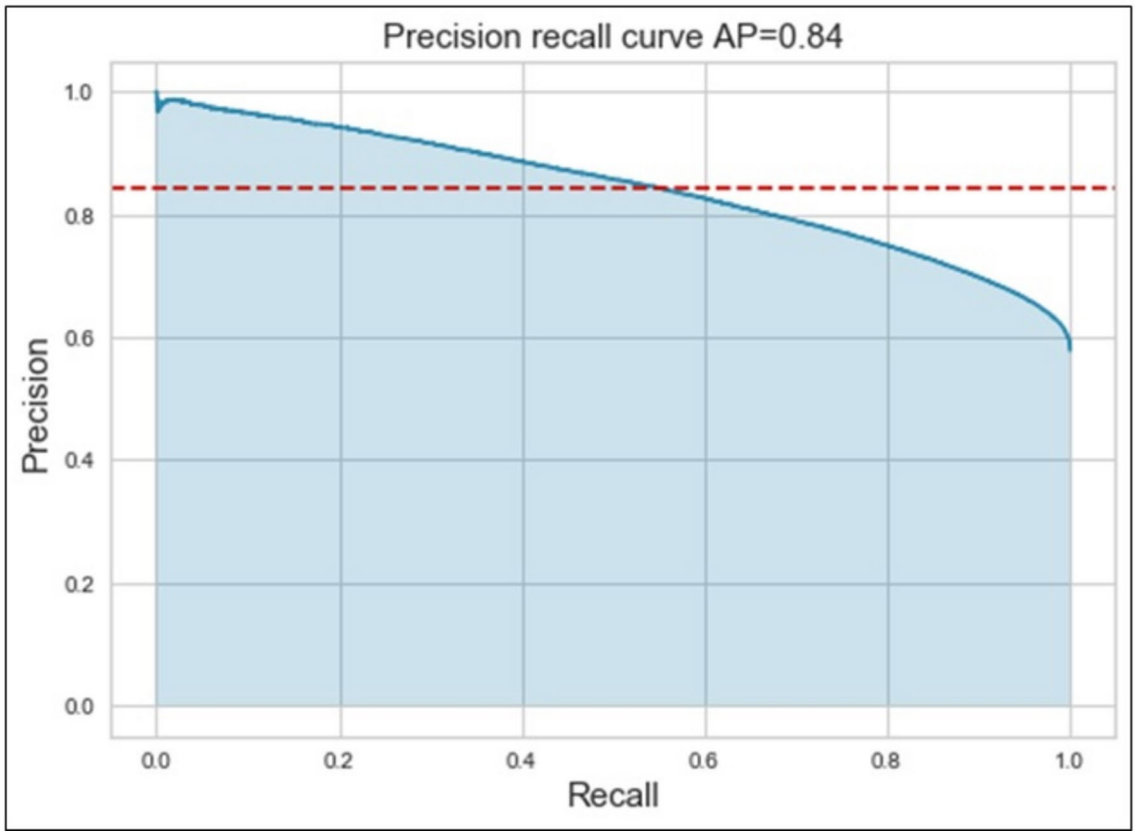

Fig. 5. Precision-recall curve for LGBM algorithm.

- Increase funding and resources for public schools to maintain and enhance their positive impact on student performance.
- Develop educational programs and support for parents, especially mothers.
- Enhance access to educational resources and services for Afro-descendant and other ethnic communities.
- Invest in infrastructure improvements in rural areas, such as better school buildings, reliable internet access, and transportation services, to overcome the geographic and resource-based challenges.
- Provide targeted training for teachers in rural schools to enhance their skills and adaptability and create equitable and effective learning environments for all students.

Comparison with other studies

Numerous studies have examined the Saber 11 exam to understand the factors influencing student performance and the existing educational gaps. In recent years, the application of machine learning models has emerged as a promising tool for analysing test data and extracting valuable knowledge to improve academic quality. Therefore, the obtained results were compared with the several studies. Timaran-Pereira et al.[11] applied DTs for academic performance classification of all Colombian regions with an accuracy of 0.67. Fortunately, the LGBM algorithm outperforms those results, and most of the selected algorithms reached an accuracy greater than Timaran-Pereira et al.[11] proposal. It is suitable to mention that Timaran-Pereira et al.[11] did not use other performance metrics such as AUC, F-Score and Precision to test the performance of his model, a gap that this paper intends to fill.

According to Timaran-Pereira et al.[11], the socioeconomic stratum and the family income are related to academic performance on the Saber 11 test. They stated that students of low strata tend to score under the national average while students of high strata perform above the national average. They also found that academic schedule influences academic performance on tests. Students with a full-time schedule will be above the national average. In addition, the use of technology for school learning enhances academic performance.

These results are similar to the ones obtained for lagging regions. Relevant variables for lagging regions include other socioeconomic factors due to the context of these geographical areas. The location of the school in urban areas has an impact on performance. These schools generally have more resources to invest in tools that facilitate students' learning. However, it does not guarantee good performance in the exam as it can have both a positive and negative impact. Family size in terms of the number of rooms also affects academic performance. In these areas, large families usually have low purchasing power, i.e. the greater the number of members, the poorer the family tends to be. The above implies that the student needs to have the necessary conditions at home to appropriate the knowledge received at school through independent study. In addition, the department where the student resides influences academic performance. Notably, these areas (Nariño, Norte Santander, Bolivar

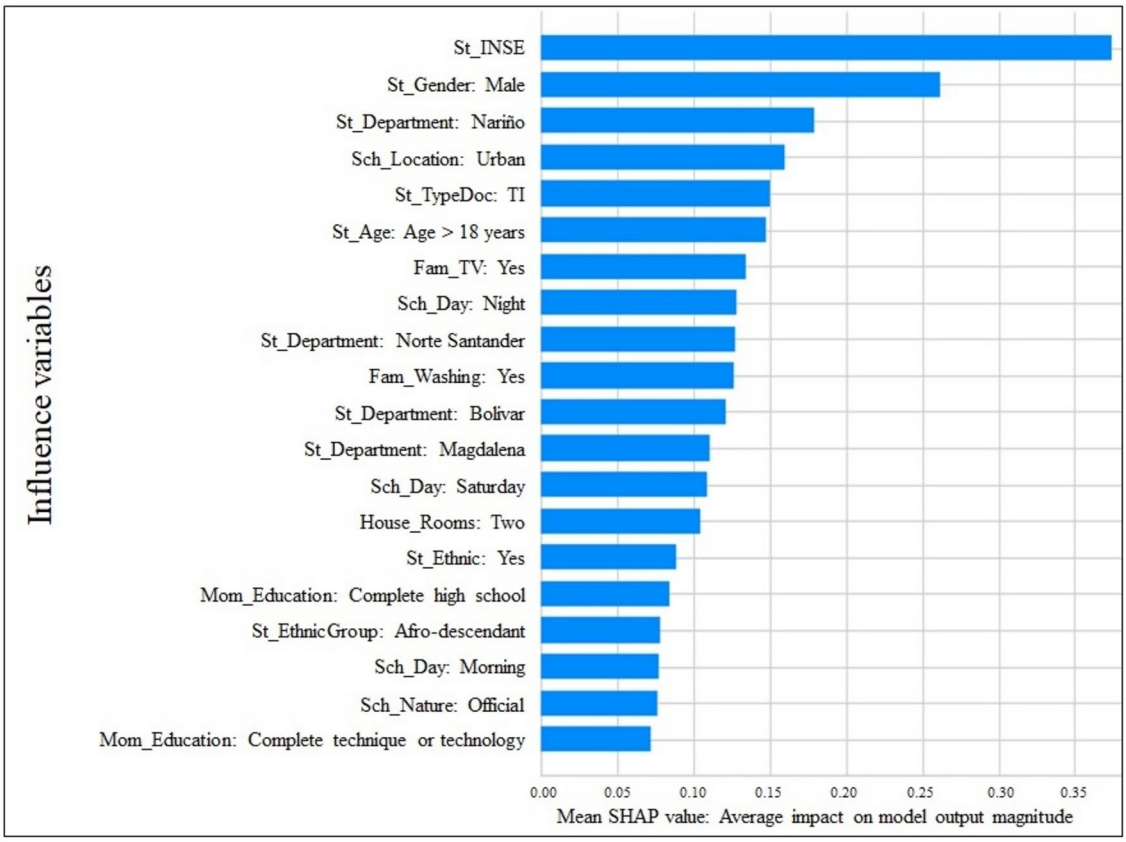

Fig. 6. Mean SHAP values for variables that influences national performance.

and Magdalena) are in those departments with the highest monetary poverty index. The latter confirms one of the main results of this research: the socioeconomic level index is the most influential variable on academic performance.

The comparison of the results of this study and the one developed by Timaran-Pereira et al.[11] indicates that government policies and strategies should indeed improve the living conditions of the inhabitants in the lagging regions. This will directly affect the students' performance, increase the opportunity to access higher education and achieve equality.

Other studies have used machine learning models to identify factors associated with academic performance in the Saber 11 and predict future results. Pereiera et al.[76] used decision trees to discover performance-related patterns in Critical Reading, finding that socioeconomic factors, such as stratum and technological conditions, such as internet access, are determinants. Similarly, Solano et al.[77] evaluated machine learning models to predict performance on the test, concluding that the Logistic Model Tree (LMT) algorithm offered the best accuracy. The study also explored the impact of computer and internet access at home on test results, finding a positive correlation, especially in English.

Rodríguez-Pose and Ketterer[14] analysed showing significant differences in variables such as internet access and the number of books in the home. They also addressed the prediction of academic performance using machine learning models, finding that variables such as gender, geographic location and family support are relevant. Alcázar et al.[78] confirmed the existence of a correlation between socioeconomic status and test results, highlighting the need to strengthen the education system in lower strata. Cardozo Anaya[79] built a regression model to predict the expected Saber 11 score for students in a specific school, using data from simulations and the actual test. Table 7 shows the AC of the algorithms developed in this paper compared to similar studies.

These studies show the potential of machine learning models to analyse Saber 11 data, identify key factors that influence academic performance, and contribute to the understanding of the education gap in Colombia. Applying these techniques can provide new information for designing public policies that promote equity and improve the quality of education in the country.

Conclusions

This paper aims to design a model for predicting academic performance in standardised tests based on machine learning to detect the influencing variables and decide how to improve the quality of education in lagging regions. The literature revealed that many studies have been conducted in different educational institutions, but only some have been executed on standardised tests. Also, research focused on lagging areas was not found. According to the gaps in this research, the proposed methodology is based on CRISP-DM integrated

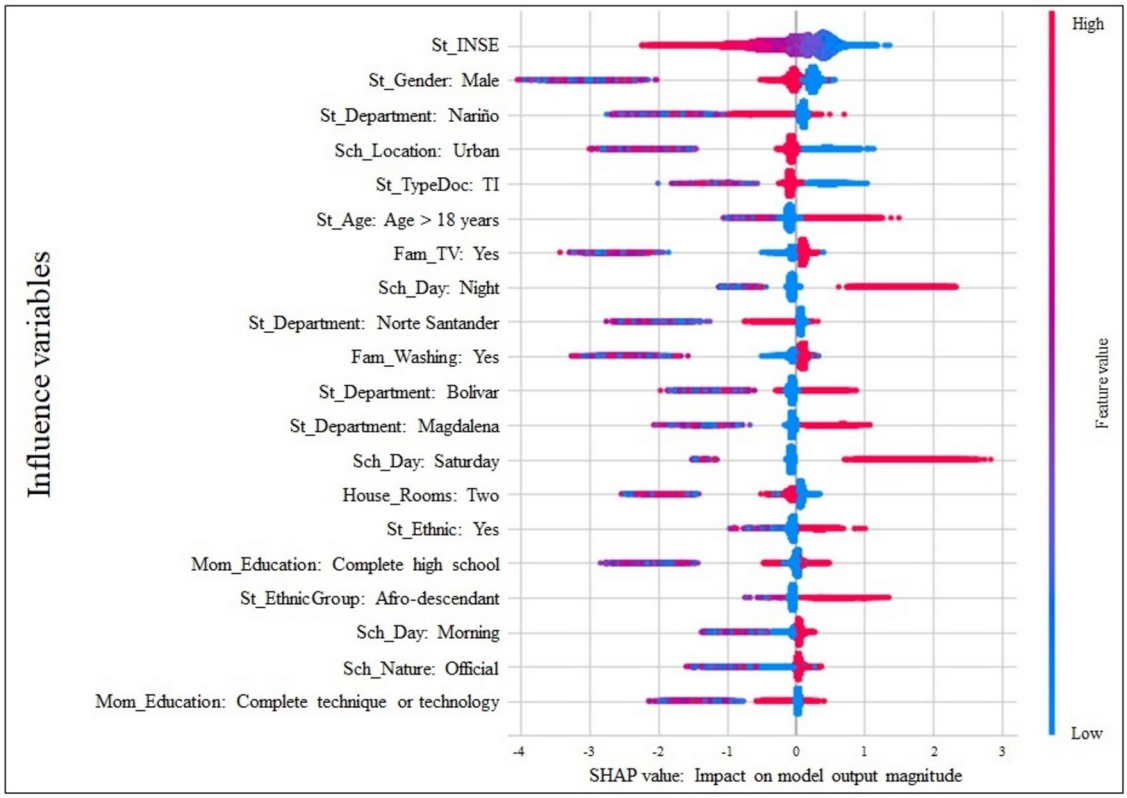

Fig. 7. SHAP values for variables that influences national performance.

with machine learning models and the SHAP values technique to determine the influencing variables and their impact on academic performance.

EGBM and LGBM were the algorithms with the best results in student classification performance metrics (accuracy of 0.77 and AUC of 0.85) and computational cost. The presented work outperformed Timaran-Pereira et al.[11] algorithm with higher accuracy. Although the algorithm has a high predictive capacity, it is a black-box algorithm, i.e. it is not explanatory. The problem was solved using the SHAP technique, which has shown effective results. SHAP values showed that socioeconomic level index, gender, school location, age, access to services such as TV and washing machine, and school day are the variables most influential on academic performance in the Saber 11 test.

Moreover, there are differences in educational quality among the lagging regions. The proposed methodology applies to the educational system of any country as long as it considers the structure of the standardised test and the database to determine the variables that influence academic performance. They provide a clear vision of how public policies should be oriented and strengthened to allow better opportunities for access to higher education in the lagging regions. Therefore, the expected impact of this study is to lead the development of comprehensive educational policies that help students from lagging regions access the same academic quality as those in other socio-demographic conditions.

Future works should apply deep learning techniques using Tensorflow, PyTorch, or Keras libraries since they have shown excellent results in other areas. Artificial intelligence must also be merged with EDM and big data methodologies such as Spark, Hive, and Hadoop. In addition, machine learning models can support other elements of educational management and academic outcomes. They can identify at-risk students for early intervention, personalise learning experiences, and optimise resource administration. These models also help analyse student engagement, automate assessments and feedback, and enhance student experience. Additionally, they detect anomalies and fraud, forecast course demand, evaluate the effectiveness of educational programs, and optimise student recruitment and admissions processes. Implementing machine learning in these areas leads to a more efficient, student-centred educational management system that improves administrative processes and academic results. Moreover, educational policies should be developed based on machine learning models.

Data availibility

The data used for the study can be found at https://github.com/jubizm/Academic_performance.git. The database used for this study was obtained from the public repository of ICFES, available at https://www.icfes.gov.co/data-icfes.

Received: 4 April 2024; Accepted: 15 October 2024

Published online: 25 October 2024

References

1. Barbu, M., Vilanova, R., Lopez Vicario, J., Pereira, M.J., Alves, P., Podpora, M., Ángel Prada, M., Morán, A., Torreburno, A., Marin, S., et al.: Data mining tool for academic data exploitation: literature review and first architecture proposal. Projecto SPEET-Student Profile for Enhancing Engineering Tutoring (2017)
2. Blytt, M. Big challenges for visual analytics: Assisting sensemaking of big data with visual analytics. Norwegian University of Science and Technology (2013)
3. Fisher, M. J. & Marshall, A. P. Understanding descriptive statistics. *Australian Critical Care***22**(2), 93–97 (2009).
4. Gagliardi, J., Parnell, A., & Carpenter-Hubin, J. The analytics revolution in higher education. Change: The Magazine of Higher Learning 50(2), 22–29 (2018)
5. Romero, C. & Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications***33**(1), 135–146 (2007).
6. Athani, S.S., Kodli, S.A., Banavasi, M.N., & Hiremath, P.S. Student academic performance and social behavior predictor using data mining techniques. In: 2017 International Conference on Computing, Communication and Automation (ICCCA), pp. 170–174 (2017). IEEE
7. Costa, E. B., Fonseca, B., Santana, M. A., Araújo, F. F. & Rego, J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior***73**, 247–256 (2017).
8. Devi, A., & Kaur, J. A survey on data mining and its current research directions. International Journal of Advanced Research in Computer Science and Software Engineering **8**(4) (2017)
9. Burgos, C. et al. Data mining for modeling studentsâ€™ performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering***66**, 541–556 (2018).
10. Fernandes, E. et al. Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil. *Journal of Business Research***94**, 335–343 (2019).
11. Timaran-Pereira, R., Hidalgo, A., Caicedo, J., & Benavides, J. Discovering factors associated with academic performance of high school students in saber 11th test using educational data mining techniques. In: E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, pp. 188–197 (2018). Association for the Advancement of Computing in Education (AACE)
12. Ministerio de Educación Nacional: Colombia mejora progresivamente en calidad educativa, así lo evidencian los resultados del ISCE. urlhttp://mineducacion.gov.co (2018)
13. Maldonado, F.J.B., & Meza, M.V.G. El rezago social en áreas metropolitanas de méxico. Estudios económicos, 265–297 (2013)
14. Rodríguez-Pose, A. & Ketterer, T. Institutional change and the development of lagging regions in europe. *Regional Studies***54**(7), 974–986 (2020).
15. Pugh, R. Universities and economic development in lagging regions:â€˜triple helixâ€™policy in wales. *Regional Studies***51**(7), 982–993 (2017).
16. Sharma, M. Research and google trend for human neuropsychiatric disorders and machine learning: a brief report. Psychiatria Danubina 33(br 3), 354–357 (2021)
17. Monga, P., Sharma, M., & Sharma, S.K. Performance analysis of machine learning and soft computing techniques in diagnosis of behavioral disorders. In: Electronic Systems and Intelligent Computing, pp. 85–99. Springer, ??? (2022)
18. Wang, C. & Du, C. Optimization of physical education and training system based on machine learning and internet of things. *Neural Computing and Applications***34**(12), 9273–9288 (2022).
19. Xie, C. et al. Influence of artificial intelligence in education on adolescentsâ€™ social adaptability: A machine learning study. *International Journal of Environmental Research and Public Health***19**(13), 7890 (2022).
20. Sharma, A.K., Ghodke, P.K., Goyal, N., Nethaji, S., & Chen, W.-H. Machine learning technology in biohydrogen production from agriculture waste: Recent advances and future perspectives. Bioresource Technology, 128076 (2022)
21. Romero, C., & Ventura, S. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40(6), 601–618 (2010)
22. Sharma, M., Sharma, S. & Singh, G. Performance analysis of statistical and supervised learning techniques in stock data mining. *Data***3**(4), 54 (2018).
23. Bakhshinategh, B., Zaiane, O. R., ElAtia, S. & Ipperciel, D. Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies***23**(1), 537–553 (2018).
24. Baker, R.S., & Yacef, K. The state of educational data mining in 2009: A review and future visions. JEDM| Journal of Educational Data Mining 1(1), 3–17 (2009)
25. Huebner, R.A. A survey of educational data-mining research. Research in Higher Education Journal **19** (2013)
26. Peña-Ayala, A. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications***41**(4), 1432–1462 (2014).
27. Shahiri, A. M. et al. A review on predicting student's performance using data mining techniques. *Procedia Computer Science***72**, 414–422 (2015).
28. Kumar, M., & Singh, A. Evaluation of data mining techniques for predicting student's performance. International Journal of Modern Education & Computer Science **9**(8) (2017)
29. Satyanarayana, A., & Nuckowski, M. Data mining using ensemble classifiers for improved prediction of student academic performance. In: ASEE Mid-Atlantic Section Spring 2016 Conference, pp. 1–7 (2016). George Washington University
30. Venkatachalapathy, K., Vijayalakshmi, V., & Ohmprakash, V. Educational data mining tools: a survey from 2001 to 2016. In: 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), pp. 67–72 (2017). IEEE
31. Widyahastuti, F., & Tjhin, V.U. Predicting students performance in final examination using linear regression and multilayer perceptron. In: 2017 10th International Conference on Human System Interactions (HSI), pp. 188–192 (2017). IEEE
32. Kumari, P., Jain, P.K., & Pamula, R. An efficient use of ensemble methods to predict students academic performance. In: 2018 4th International Conference on Recent Advances in Information Technology (RAIT), pp. 1–6 (2018). IEEE
33. Adejo, O.W., & Connolly, T. Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education (2018)
34. Ma, X., & Zhou, Z. Student pass rates prediction using optimized support vector machine and decision tree. In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp. 209–215 (2018). IEEE
35. Kumar, J.D., Shankar, K., & Saravanaguru, R. An investigation on educational data mining to analyze and predict the studentâ€™s academic performance using visualization. In: Information Systems Design and Intelligent Applications, pp. 179–188. Springer, ??? (2019)
36. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S.N. Predicting academic performance: a systematic literature review. In: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, pp. 175–199 (2018)
37. Muzzammel, R. Intelligent Technologies and Applications, vol. 932. Springer (2019)

38. Khan, I., Al Sadiri, A., Ahmad, A.R., & Jabeur, N. Tracking student performance in introductory programming by means of machine learning. In: 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), pp. 1–6 (2019). IEEE
39. Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A. & Pintelas, P. Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of Educational Computing Research***57**(2), 448–470 (2019).
40. Prasanalakshmi, B. & Farouk, A. Classification and prediction of student academic performance in king khalid university-a machine learning approach. *Indian Journal of Science and Technology***12**, 14 (2019).
41. Sana, B., Siddiqui, I.F., & Arain, Q.A. Analyzing students' academic performance through educational data mining. 3C Tecnología, 402–421 (2019)
42. Nahar, K., Shova, B. I., Ria, T., Rashid, H. B. & Islam, A. S. Mining educational data to predict students performance: A comparative study of data mining techniques. *Education and Information Technologies***26**(5), 6051–6067 (2021).
43. Eleyan, N., Al Akasheh, M., Malik, E.F., & Hujran, O. Predicting student performance using educational data mining. In: 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 1–7 (2022). IEEE
44. Dinh-Thanh, N., & Thi-Ngoc-Diem, P. Predicting academic performance of high school students. In: International Conference on Nature of Computation and Communication, pp. 123–135 (2022). Springer
45. Ghosh, P. Data mining approach to predict academic performance of students. *BOHR International Journal of Computer Science***2**(1), 21–31 (2023).
46. Alamgir, Z., Akram, H., Karim, S., & Wali, A. Enhancing student performance prediction via educational data mining on academic data. Informatics in Education (2023)
47. Alghamdi, A. S. & Rahman, A. Data mining approach to predict success of secondary school students: A saudi arabian case study. *Education Sciences***13**(3), 293 (2023).
48. Nayak, P., Vaheed, S., Gupta, S., & Mohan, N. Predicting students' academic performance by mining the educational data through machine learning-based classification model. Education and Information Technologies, 1–27 (2023)
49. Batool, S. et al. Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies***28**(1), 905–971 (2023).
50. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., et al. Crisp-dm 1.0: Step-by-step data mining guide. SPSS Inc. 9, 13 (2000)
51. Ashraf, M., Zaman, M., Ahmed, M., & Sidiq, S.J. Knowledge discovery in academia: a survey on related literature. International Journal of Advanced Research in Computer Science and Software Engineering **8**(1) (2017)
52. Zaffar, M., Savita, K., Hashmani, M. A. & Rizvi, S. S. H. A study of feature selection algorithms for predicting students academic performance. *International Journal of Advanced Computer Science and Applications***9**(5), 541–549 (2018).
53. Carrascal, A. I. O. & Giraldo, J. J. Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas saber-pro. *Revista Politécnica***15**(29), 128–140 (2019).
54. Fernández, M., Guerra-Curvelo, W., & Meisel-Roca, A. Políticas para reducir las desigualdades regionales en colombia. Technical report, Banco de la Republica de Colombia (2007)
55. DNP: Índice Departamental de Innovación para Colombia (IDIC) 2015. urlhttp://dnp.gov.co (2015)
56. DANE: Pobreza monetaria por departamentos en Colombia. https://www.dane.gov.co (2019)
57. DNP: Convergencia y desarrollo regional–Caracterización, dinámica y desafíos. urlhttps://colaboracion.dnp.gov.co (2010)
58. DANE: Pobreza Multidimensional por departamentos 2018. urlhttps://www.dane.gov.co (2019)
59. Wetschoreck, F. Rip correlation. introducing the predictive power score. Towards Data Science (2020)
60. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering***21**(9), 1263–1284 (2009).
61. Chen, T., & Guestrin, C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
62. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems***30**, 3146–3154 (2017).
63. Hilera González, J.R., & Martínez Hernando, V.J., et al. Redes Neuronales Artificiales: Fundamentos, Modelos Y Aplicaciones. Alfaomega, ??? (2000)
64. Superby, J.-F., Vandamme, J., & Meskens, N. Determination of factors influencing the achievement of the first-year university students using data mining methods. In: Workshop on Educational Data Mining, vol. 32, p. 234 (2006). Citeseer
65. Tian, W. Predicting and interpreting students performance using supervised learning and shapley additive explanations. PhD thesis, Arizona State University (2019)
66. Diedrichs, A. L., Bromberg, F., Dujovne, D., Brun-Laguna, K. & Watteyne, T. Prediction of frost events using machine learning and iot sensing devices. *IEEE Internet of Things Journal***5**(6), 4589–4597 (2018).
67. Breiman, L. Random forests. *Machine Learning***45**(1), 5–32 (2001).
68. Wu, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems***14**(1), 1–37 (2008).
69. Seufert, E.B. Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue. Elsevier, ??? (2013)
70. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Machine Learning***63**(1), 3–42 (2006).
71. Lundberg, S., & Lee, S.-I. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874 (2017)
72. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering***2**(10), 749–760 (2018).
73. Movahedi, A., & Derrible, S. Interrelated patterns of electricity, gas, and water consumption in large-scale buildings. engrXiv (2020)
74. Provost, F. & Kohavi, R. Glossary of terms. *Journal of Machine Learning***30**(2–3), 271–274 (1998).
75. Gutiérrez, G., Canul-Reich, J., Zezzatti, A. O., Margain, L. & Ponce, J. Mining: Students comments about teacher performance assessment using machine learning algorithms. *International Journal of Combinatorial Optimization Problems and Informatics***9**(3), 26 (2018).
76. Pereira, R. T., Troya, A. H. & Zambrano, J. C. Factores asociados al desempeño académico en lectura crítica en las pruebas saber 11 con árboles de decisión. *Investigación e Innovación en Ingenierías***8**(3), 29–37 (2020).
77. Solano, J. A., Cuesta, D. J. L., Ibáñez, S. F. U. & Coronado-Hernández, J. R. Predictive models assessment based on crisp-dm methodology for students performance in colombia-saber 11 test. *Procedia Computer Science***198**, 512–517 (2022).
78. Alcázar, J.G., Ruiz, I.R.H., Paternina, W.A., Cuesta, L.F.R.T., & Ortega, A.V.T. Análisis de influencia de estrato socio–económico en resultados de pruebas saber 11 (2022)
79. Cardozo Anaya, J.M. Modelo de regresión para predecir el puntaje esperado en la saber 11 para los estudiantes del colegio cajasai (2023)

Declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-76596-3.

Correspondence and requests for materials should be addressed to R.P.-N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.