Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 6, Special Issue, 2022

"Current Issues, Trends, and New Ideas in Education"

The Role of Technology in the Learning Process: A Decision Tree-Based Model Using Machine Learning

Yuri V. S. Mendonça [1], Paola G. Vinueza Naranjo [2], Diego Costa Pinto [1*]

[1] *NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal.*

[2] *Faculty of Engineering, Information Technology and Communication, National University of Chimborazo, Riobamba, Ecuador.*

Abstract

Machine learning approaches may establish a complex and non-linear relationship among input and response variables for the assessment of the Basic Education Development Index (IDEB) database and show indicators that may contribute to monitoring the quality of education. This paper uses extensive experimental databases from public schools, consisting of a case study in Brazil, to analyze data such as the physical and technological structure of schools and teacher profiles. The research proposes decision tree-based machine learning models for predictions of the best attributes to positively contribute to IDEB. It employs a newly developed SHapley Additive exPlanations (SHAP) approach to classify input variables, so to identify variables that impact the most the final model; a non-probabilistic sample was used, composed from three official databases of 450 schools, and 617 teachers. Results show that the number of computers per student, teachers' service time, broadband internet access, investments in technology training for teachers, and computer labs in schools are the variables that have the greatest effect on IDEB. The model applied shows high prediction accuracy for test data (MSE = 0.2094 and R^2 = 0.8991). This article contributes to improving efficiency when monitoring parameters used to measure the quality of a teaching-learning process.

Keywords:

Decision Tree; IDEB;
Machine Learning Approaches;
School Infrastructure;
Teacher Profile;
Learning Strategies.

Article History:

Received:	19	July	2022
Revised:	12	November	2022
Accepted:	04	December	2022
Published:	05	January	2023

1- Introduction

Advances in learning technologies and tools have grown significantly in recent years, showing the importance of using information technologies in teaching-learning processes, directly or indirectly contributing to students' and teachers' performance within basic education [1]. In addition, school infrastructure, scarcity of human and material resources, and teachers' qualifications and continuous training may also contribute to improving students' academic performance [2].

Nowadays, an increasing importance is given to predicting student performance due to how relevant this issue has been to the development of countries, as it depends entirely on the educational process leading to development in all aspects of life (scientific, social, economic, etc.). Also, the evaluation of teachers' and students' performance is a reflection of the efficiency of educational institutions. Therefore, focusing on the development of educational processes is one of the utmost necessities to push governments represented by educational institutions to make serious efforts regarding educational processes towards continuous and escalating development [3, 4].

There are different works in educational processes using decision tree, for example: creating adaptive dynamic tests for assessing student academic performance, while constantly comparing results of the assessment which exhibit the individual student profile with results of the decision tree's algorithm 2, which formulates a predictive model for

* **CONTACT**: dpinto@novaims.unl.pt

DOI: http://dx.doi.org/10.28991/ESJ-2022-SIED-020

students' knowledge level, according to the weights posed by the decision tree [5]. In 2016, the U.S. Department of Education launched the National Education Technology Initiative (Future Ready Learning: Reimagining the Role of Technology in Education), and defined how "personalized learning refers to teaching that is optimized for each learner's needs in terms of learning pace and teaching methods and requires that the learning objectives, learning content, and learning methods in the learning process should be different and adjustable according to the needs of learners". At the same time, both governments and educational authorities at all levels attach great importance to the development of personalized learning. In traditional classroom education, student learning activities are completely developed by teachers, and the practice of personalized learning can only rely on their teaching experience due to the constraints of time and space. Therefore, it is difficult to develop a personalized learning program for each student in such an educational learning model. In recent years, as the process of education informatization continues to deepen, more and more Internet information technologies have flooded into teaching processes within education, which has led to a significant innovation in the development of education and to an emerging online learning model that has gradually developed and become an integral part of educational learning [6].

In most countries, especially in South America, education requires a (re)construction considering education quality indicators of general aspect, that point to a need for re-evaluating pedagogical strategies and educational policies aimed at improving basic education quality. For example, the state of Par´a, in Northern Brazil, presented in 2019 the second worst basic education development index (IDEB) of the country for secondary education [7].

In this scenario, this research contributes to identifying parameters to better guide work plans and developed contents, aiming at better performances concerning the teaching-learning process directed at students, and, consequently, better external evaluation in IDEB results. Since most researchers present a generic point of view, employing only descriptive and/or inferential statistics to verify certain decisive characteristics of the educational development process [8, 9], there are few studies employing advanced machine learning techniques that can analyze many attributes to more accurately predict variables in order to positively contribute to students' academic performance [10].

Studies involving Data Mining (DM), Artificial Neural Networks (ANN), and Decision Tree (DT), with cross-validation and statistical analysis, have recently gained attention in the literature dedicated to monitoring school performance in large-scale high school evaluations [11, 12]. Some studies were able to observe, through Educational Data Mining, only a few variables; for example, if the level of access to computer and library resources offered by schools could contribute to the performance of high school students in large-scale evaluations [13].

This research contributes to the literature in a significant way, by examining a maximum number of variables, with the objective of improving the decision-making made by public policies oriented to the betterment of educational results. So, it covers both physical and technological school structures, and teacher profiles specifically for the first years of school by using decision tree (DT) approaches and official data from public schools situated in seven municipalities in the state of Pará, Brazil – all of which are part of the Pará State Department of Education (SEDUC) [*]–, and from Dashboard Google Platform for Education [14-16].

This work is structured as follows: Section 2 provides a literature review on how the investigated attributes impact the educational development index. Section 3 describes the methodology applied; in Section 4, the model proposed and developed is explained in detail in two phases; in Section 5, the mathematical approach of the SHAP method is shown; in Section 6, experimental results are presented using machine learning and SHAP values; and finally, some conclusions and recommendations are mentioned in Section 7.

2- Literature Review and the Basic Education Development Index (IDEB)

This section analyzes the relevant literature in which researchers observe one or more characteristics and/or teacher profiles, school infrastructure, and technological profiles and how these attributes interfere with student performance, as well as in the Basic Education Development Index (IDEB) (see Figure 1).

2-1- Teacher Profile Attributes

Different types of large-scale evaluations, applied both nationally and internationally, have contributed to the emergence of studies on teacher profiles, seeking to identify characteristics or attributes that may be associated with student academic success. In fact, some authors have recently dedicated themselves to basic education teacher profile analyses in Brazil [16], a very important step to better understand profile roles in student achievements. The Brazilian School Census has been of great aid in this regard, as sociodemographic information related to teacher profiles is also obtained. The Census, carried out annually in Brazil, characterizes teachers by gender, age, ethnicity, initial training, and teaching school locations, among other information, in terms of educational stages. The School Census is currently the largest available database in the country, comprising information on all operational schools, at all levels and types of education (municipal, state, or private). The Teaching and Learning International Survey (TALIS) consists in another means of obtaining teacher profile, teaching environment, and educational performance data. However, both the census and the TALIS do not, as in other studies, survey the characteristics, perceptions, and specific meanings that teachers attribute to different aspects and contexts of their work [17, 18].

[*] https://novo.qedu.org.br/

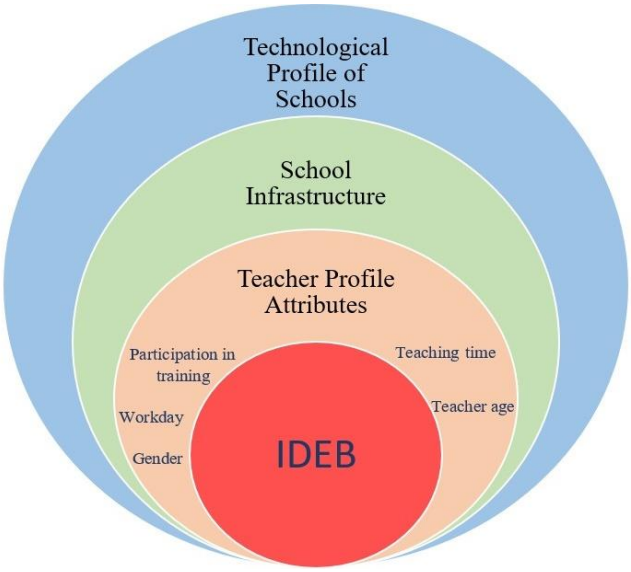

Figure 1. Brazilian Basic Education Development Index (IDEB) attributes

In this sense, teachers' role in the educational process can be contextualized through several characteristics and/or attributes such as good communication, creativity, organization, commitment, planning and content knowledge, as well as personal attributes like age, gender and beliefs. All these attributes interact with environmental factors [19], i.e., workplace infrastructure, number of students in each classroom, number of schools taught by each teacher, and teacher training quality, which influence their daily practice, manifested in student learning processes. In addition, with regards to teachers' knowledge in the field of information technology, the research [20] has shown that teachers' digital competence (TDC) is an important condition for an effective integration of technologies in education. Therefore, investing in IT training for teachers is of fundamental importance to improve the quality of teaching. Other characteristics displaying a direct relationship with teacher and, consequently, student performance, were evaluated herein as follows:

2-2- Teachers' Workload

Many teachers teach more than one class, school or discipline, demanding from these professionals work displacement and more willingness. In addition, to cover the need for schoolteachers, some educational networks increase the average teacher working day, making it difficult for those who work fragmented hours in different networks and schools [21]. These factors, in turn, directly affect teacher organization and work capacity and, consequently, their health and professional performance, as well as student learning quality [22].

In this scenario, teaching is noted as an exhausting profession, whose process is aggravated by poor working conditions, fatiguing hours, and tasks that go beyond the school environment, which further contribute to teacher exhaustion and, in some cases, illnesses [23]. Furthermore, teachers are not solely responsible for the student learning process, and their well-being and good performance are important factors in the teaching-learning process, as they result from adequate student-teacher interactions.

2-3- Gender

Gender is an important aspect that must be considered when reflecting upon and for understanding the characteristics a relationship formed between teachers and students in a school context has. It is also paramount in constructing teaching career policies. The 2017 Brazilian School Census indicated that 96% of teachers who work in the first years of primary school are women, as this profession allows women to reconcile family, domestic work, and the labor market [24].

3- Methods

3-1- Scenario and Participants

The datasets chosen to train the machine learning model include crossing information from three databases, as follows:

- ***Dataset 1 (DB1):*** The Pará State Department of Education (SEDUC), responsible for providing teachers' information. The data were provided as an Excel spreadsheet.
- ***Dataset 2 (DB2):*** The qedu.org.br website[*], which organizes data provided by the Ministry of Education, municipal and state secretariats of education, and non-governmental organizations linked to the educational sector in a

[*] *https://novo.qedu.org.br/*

didactic manner, providing data on the performance of both public and private schools, at all education levels, in addition to providing information on their physical and technological infrastructure.

- ***Dataset 3 (DB3):*** The Google for Education platform control panel, which allows for the identification of teacher accesses by their ID, characteristics and the school that uses or does not use the platform. This process was only possible due to the Google Platform being hired by the Pará State government, through the Inteceleri company, a partner in the Northern region of the country. The service providing contract indicates that the company would provide continuing education to teachers in state education network in 2017 and 2018, giving information on teacher profiles (i.e., gender and age), as well as a percentage of continuing attendance in technology education (use of Google tools) contracted by the government of Pará.

The attributes were then chosen to verify their target effects (school IDEB for 2019) (Figure 2). Data filter parameters comprised full or partial participation in continuing education promoted by the Google for Education platform, teaching in early basic education grades and, finally, working in schools in the metropolitan region of Belém, the state capital, amounting to total data from 450 schools in seven municipalities. After data refinement, the analysis presented a sample of 617 teachers.

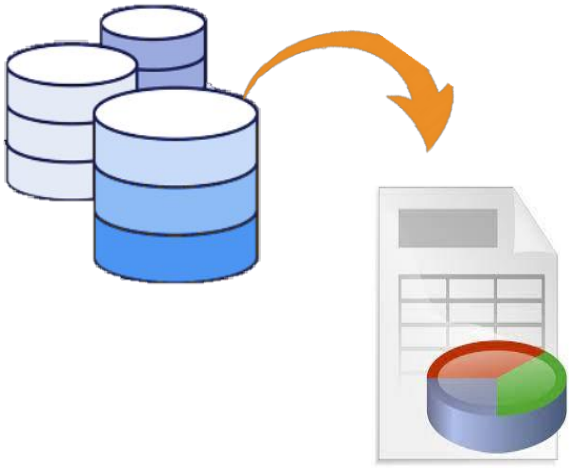

Figure 2. Attributes on the Development Index of Basic Education (IDEB)

3-2- Data Analysis

This study adopted a DT approach, a multivariate supervised learning algorithm consisting of a learning method used for objective classifications and predictions. DT-based models are, by design, easily interpretable in their predictions, but also a powerful non-parametric supervised learning method, which means that they do not assume any predefined type of probability distribution regarding the input data [25]. DT-based models can easily explain the reason for causal relationships, making this type of algorithm very useful when manipulated by decision-makers [26].

A variety of classic DT-based models can be found in the literature, such as the Chi-Square Automatic Interaction Detector [27] and Classification and Regression Trees [28]. The regression models used in this study are CART-type algorithms.

3-3- Variables

Building upon the literature review, the models proposed in this study were developed in four different instances (models).

In the first (M1), a model was trained to predict the Basic Education Development Index (IDEB) for primary schools, based on the following information detailed below, referring to teacher profile of each institution. In M1, five data points (X1, X2,..,X5) were extracted for each teacher. All levels and categories of variables have been renamed to be more understandable when they appear in a DT. The variables are described below:

- X1 = Gender: Gender of the teacher.
- X2 = Participation in ongoing training on technology: Frequency of teachers' participation in continuing technology education promoted by partner Google for Education.
- X3 = Employment period: Service time.
- X4 = Number of schools working: Number of schools in which teachers' working hours within the state public education system are distributed.
- X5 = Age: teachers' age.

Therefore, it was measured the impact variables exclusively related to teacher profile had on IDEB. In the second modeling step (M2), a model was trained to predict primary schools IDEB, based on the information listed below, referring to the schools physical infrastructure profile. In this model M2, six data points were extracted (X6, X7, X8, X9, X10 and X11).

- X6 = Reading room.
- X7 = Computer Lab.
- X8 = Science lab.
- X9 = Water supply or Filtered water.
- X10 = Kitchen.
- X11 = Restroom – WC.

As in the previous modeling, it was analyzed the impact that changes in physical infrastructure of schools had on IDEB. In the third stage of modeling (M3), a model was trained to predict elementary schools IDEB, based on the following information, referring to schools technological profile. In M3 model, four data points (X12, X13, X14, X15) were extracted.

- X12 = PCs by students.
- X13 = Broadband Internet.
- X14 = Projector.
- X15 = Printer.

$\chi_1 \epsilon \{0,1\},$

$If: \chi_1 = 1;\ female$

$= \chi_1 = 0;\ male$

$\chi_i \in Z^+, \quad i = 2,..,5$

$Where: \forall\ \chi_i > 0;$

$\chi_j \in \{0,1\}, \quad j = 6,..15$

$With: \chi_j = 1;\ presence$

$= \chi_j = 0;\ not\ presence$

For the first instance of modeling, a model of the random Forest (rf) type was used; for the second and third instance of modeling, models of the Extra Tree Regressor (ETR) type were used. These are ensemble type algorithms, that is, the final prediction of this type of model is based on the combination of N base algorithms, called Base Learners (BL), which are inside the ensemble model. In classification problems, the BL decide, by majority vote, which class/category will be the model output. In regression problems, the mean of the BL predictions is the final model output.

All modeling experiments in the data analysis stage were standardized, having as target variable the IDEB of lower elementary education, a training set corresponding to 80% of the available data, leaving 20% of the data for testing in cross-validation. All predictors were normalized using the Z-score method, which is also used to remove outliers in a standard way.

As performance metrics for LightGBM and ETR, it was used the R-squared, or coefficient of determination. It is a statistical measure that indicates how much of the variation in dependent variable is explained by independent variables in a regression model. Calculation of R-squared is shown in Equation 1.

$$R^2 = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \underline{Y})^2} \quad (1)$$

Where, Y is the actual value, while $\hat{Y}$ is the predicted value of Y, and $\underline{Y}$ is the mean of the Y value. The range of R-squared value is [0,1]. A higher score of R-squared means better modeling performance.

Finally, in the phase 2, Figure 3, the M4 model was developed. It was used the best performance predictor variables for each attribute, successfully repeating development of the predictive model to predict IDEB performance.

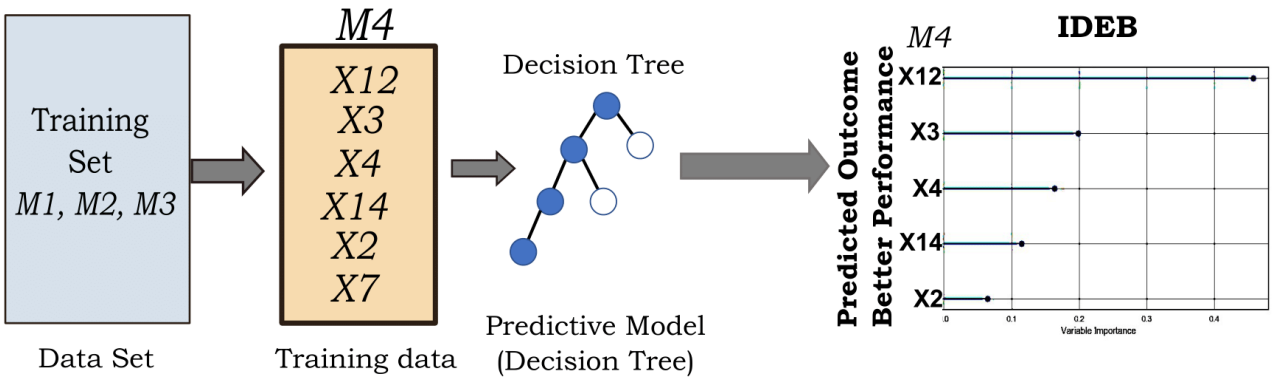

Figure 3. Phase 2, classification of predictors of better performance in the result

Then, after the three modeling steps, a final analysis was carried out with the best performance characteristics of each profile previously verified, continuing with a focus on IDEB.

4- Survey Methodology

The proposed models were developed in two phases. In the first phase, three different predictive models (M1, M2 and M3) were developed to predict the outcome of dimensions (attributes) in IDEB. These three predictive models are illustrated in Figure 4.

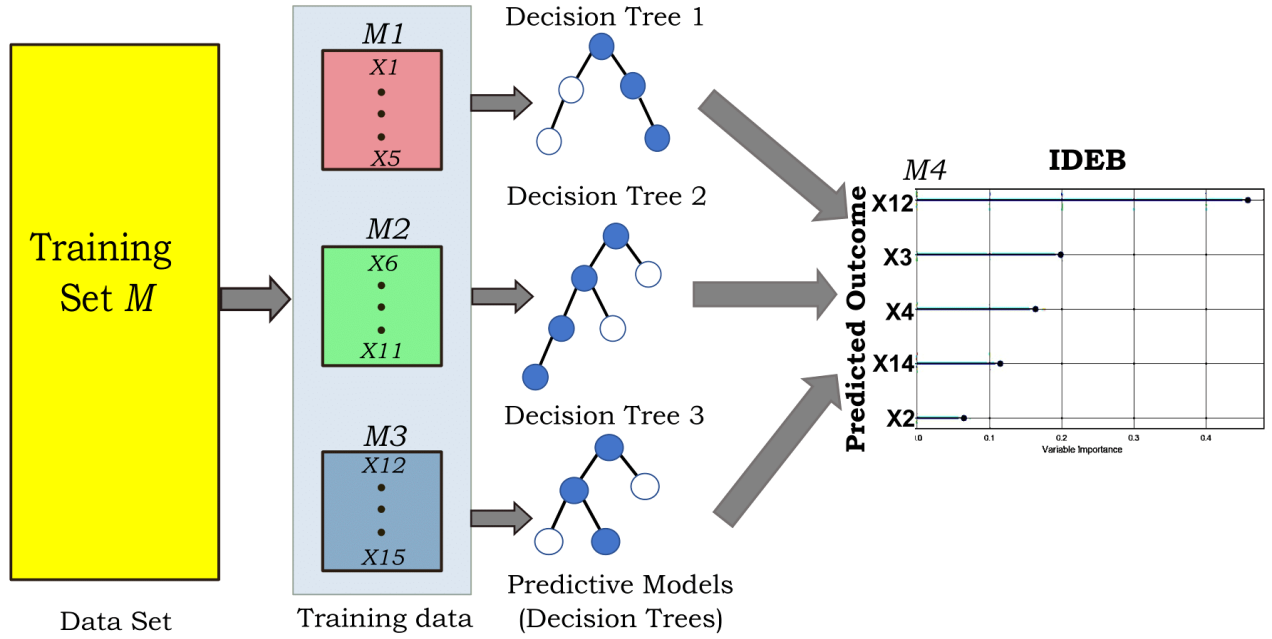

Figure 4. Attributes on the Development Index of Basic Education (IDEB)

The modeling in phase 2 is an attempt to jointly examine the attributes, and thus understand how the dimensions (physical, structural, and human) interfere in the result.

In all modeling instances, the following DT-based models were compared:

- Decision Tree Return;
- Random Forest Return;
- Extra Trees Return;
- AdaBoost Return;
- Light Gradient Boosting Machine (LightGBM).

Of all models considered in the modeling stage, only the one that stood out in the regression task was selected for optimization and, consequently, to generate the results presented in this work. For each experiment (or modeling instance) undertook, the same modeling parameters were replicated, so that a reliable result was achieved.

- Training set = 80%;
- Test set = 20%;
- Normalization via Z-score;
- Removal of outliers via Z-score;
- K-fold = 10.

The samples were randomly sorted by using random.sample function in python, and the first 80% of samples constituted the training sample set, while the remaining 20% samples constituted the testing sample set. A crossvalidation technique was adopted by using a suited early-stopping procedure in order to optimize the model complexity; this method was used during training so to avoid inconsistent variance data [29]. The approach to identifying importance variables that contribute to the final model is shown in Figure 5.

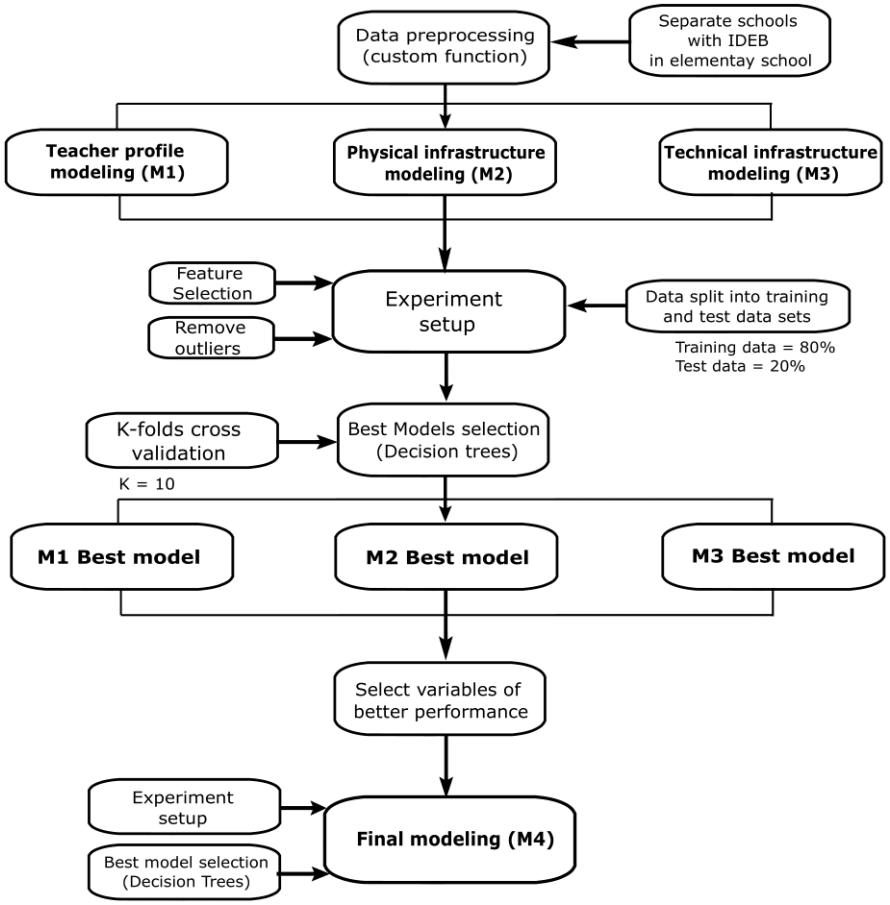

Figure 5. Flowchart of the research methodology

5- Shapley Additive Explanations (SHAP)

SHAP is a game-theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. To produce an interpretable model, SHAP uses an additive feature attribution method, i.e., an output model is defined as a linear addition of input variables. Assuming a model with input variables x = (x1, x2,.., xp), where p is the number of input variables, the explanation model g(x′) with simplified input x ′ for an original model f(x) is expressed as:

$$f(x) = g\,(x') = \phi_0 + \sum_{i=1}^{M} \phi_i\, x_i' \qquad (2)$$

In the SHAP Equation 2, M represents the number of input features, and ϕ0 represents the constant value when all inputs are missing. Inputs x′ and x are related through a mapping function x = hx(x′). Where ϕ0, ϕ1, ϕ2, and ϕ3 increase the predicted value of g(), while ϕ4 decreases the value of g(). As noted by Verwer & Zhang [30].

The SHAP graph can be interpreted as follows: the Y-axis indicates the variable name, in order of importance from top to bottom. The value next to them is the mean SHAP value. On the X-axis is the SHAP value, which indicates how much is the change in log-odds. From this number we can extract the probability of success. Gradient color indicates the original value for that variable. In booleans, it will take two colors, but in number it can contain the whole spectrum. Each point represents a row from the original dataset.

Based on SHAP game theory, information gain rate graph and regression performance indicators were used to interpret the models. The SHAP method helps us to interpret machine learning models more easily, through direct and indirect relationships between predictors. SHAP values are measures of contribution that each predictor has in a given model. However, they not only quantify the importance of each predictor in the regression task, but also the direction in the causal relationship [31]. In the case of this work, the importance lies in an increase or decrease in the value of primary schools IDEB. Difficulties in interpreting machine learning (ML) models and their predictions limit the practical applicability of and confidence in ML in education research.

To this end, the SHAP methodology has recently been introduced. SHAP approach enables the identification and prioritization of features that determine compound classification and activity prediction using any ML model. Herein, we further extend the evaluation of the SHAP methodology by investigating the most important variables for exact calculation of Shapley values for DT methods, and systematically compare these variables in IDEB predictions with the model independent SHAP method. Moreover, new applications of the SHAP analysis approach are presented, including interpretation of ensemble regression models for IDEB prediction.

6- Experimental Results: Machine Learning Using SHAP Values

Machine learning models have been widely used to accelerate the interpretation and highlight hidden patterns in the data. However, as the complexity of the model increases, interpretating the results can become quite challenging. The SHAP technique developed here in this research provides a measurement on the importance of each input attribute on the model's output. We illustrate the value of the SHAP technique using a decision tree machine learning implementation to classify the educational performance of different school cycles using data science methods.

This section presents the modeling steps results based on the training phases of machine learning models, using the SHAP graphs. Initially the results will be demonstrated by predictive models. Decision tree relevance provides a score that indicates how useful or valuable each variable was in building the DTs driven within the model presented here. The more an attribute is used to make important decisions with DTs, the greater its relative importance. This importance is calculated for each attribute in the dataset, allowing the attributes to be ranked and compared against each other.

Importance is calculated for a single DT by the amount each attribute split point improves the performance measure, weighted by the number of observations for which the node is responsible. The performance measure can be purity (Gini index) used to select split points or another more specific error function. Feature importance is then calculated across all DTs within the model.

The results of the present study use large databases, being able to analyze the profile of teachers (M1), physical infrastructure (M2), technological infrastructure of schools (M3). Thus, this approach helps to fill in gaps from the previously 17 published works.

In the following section, it was analyzed the importance of databases variables by using SHAP approach.

6-1- Predictive Model M1: Profile of Teachers

Five predictors were used to model teacher profiles, with two of them being the most prominent for the regression task. The Figure 6 indicates the performance rate of the selected variables with greater effect on IDEB. The variables considered are: X1 = Gender; X2 = Participation in ongoing training on technology X3 = Employment period; X4 = Number of schools working; and X5 = Age.

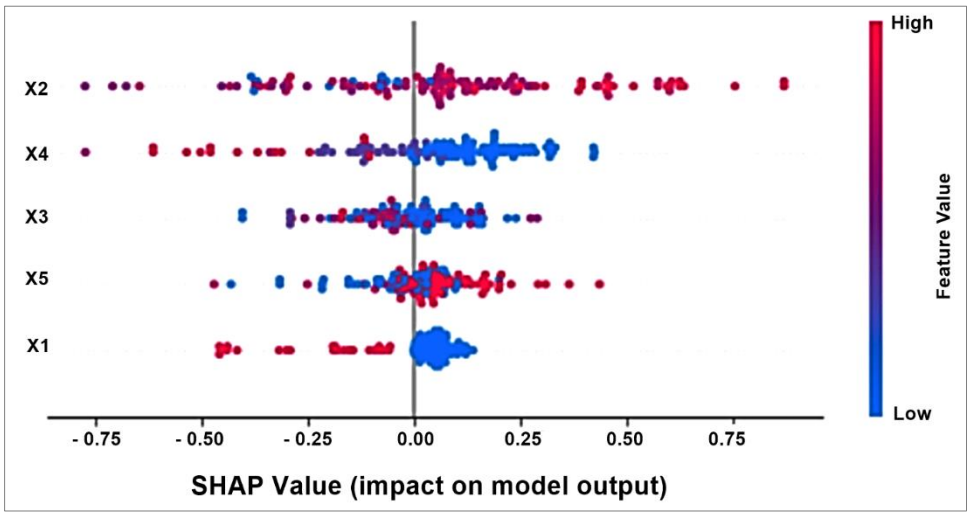

Figure 6. Test - SHAP method applied to predictive model M1

Figure 6 indicates that input variables in order of importance from top to bottom, X2 and X4, provide the greatest contributions to the M1 model. Thus, we can evaluate an average contribution of the variables in model responses. Considering, for instance, X2 = Participation in ongoing training on technology, we see that its average contribution is around 100% for the positive category.

To improve the interpretability of the model, a strategy adopted was to use the python SHAP library, as it is based on SHAP, which is very useful to explain different types of models such as Kernel, Tree, DL and others. One of the reasons why the SHAP chart has been so used is the quality of its result interpretation. As in Fig.6, showing SHAP analysis method, which must be employed considering the dimensions [intensity of the variable effect on the target (high or low) vs. effect direction (positive or negative)]. Evaluation of the SHAP chart can be done as follows:

- The Y-axis is the variables of our model in order of importance.
- The X-axis is the SHAP values. As our reference is the positive category, positive values indicate support for the reference category (it contributes to the model responding positive category at the end) and negative values indicate support for the opposite category.
- Each point on the graph represents a sample, so each sample has a value for that variable. Note that these point clouds at some point expand vertically. This occurs given the density of values of that variable in relation to SHAP values.
- Finally, the colors represent the increase/decrease in the value of the variable. Redder tones are high values, and bluish tones are lower values.

In general, we will look for variables that have the following characteristics:

- Ones to have a very clear division of colors, that is, red and blue in opposite places. This information shows that they are good predictors; after all, only by changing their value can the model verify in a simpler way its contribution to a class.
- Associated with this, the greater the range of SHAP values, the better that variable will be for the model.

Thus, with regards to the side color bar (Y-axis), Figure 6 indicates when a predictor value is high or low, with the blue dots indicating a low predictor value compared to other values for that same predictor, whereas the orange/reddish dots indicate a high value compared to the rest of the values of that same predictor. The X-axis (horizontal axis), on the other hand, indicate the direction (positive or negative) of the independent variable effect on the target (IDEB).

The Y-axis indicates the input variables in order of importance from top to bottom. Each dot is colored by the value of a input variable, from low (blue) to high (red). Density represents the distribution of points in the data set, i.e., whether it contains a range of values or selected ranges. The Figure 6 represents the input X2, X4, X3, X5 and X1 respectively for M1 model.

The variable X2 = participation in ongoing training on technology is the most important in Model 1. Figure 6 shows that the greater the value of X2, the greater its SHAP value, and the greater its impact on IDEB. So, X2 in some situations presents SHAP values around 0.9, that is, a 90% contribution to the model result (due to 100% being the maximum any variable can reach).

The variables X4 and X3 are the next to be evaluated, as it can be seen that low values for X4 and X3 contribute positively to IDEB. The X5 variable, although less significant, tends to increase the IDEB score with the increase in its value. On the other hand, low values of X1, meaning gender = F, in blue color, increase the SHAP value and the associated probability of increasing IDEB.

The variables X2 and X4 present these two mentioned characteristics. Now, on the X5 variable, note that: overall it is a confusing variable, as its SHAP values are around 0 (weak contributions) and with a clear mix of colors. Also, you cannot see an increase/decrease trend of this variable in the final answer. It is also worth mentioning the X1 variable, which does not have such a wide range as X2, but demonstrates a clear division of colors.

In this sense, regarding teacher profiles for variable X1, which represents gender, we assumed 0 for females and 1 for males, with a difference concerning gender effect on the target.

Variable X2, which represents teacher age, exhibited the greatest target effect, although Fig.6 demonstrates that this characteristic does not necessarily have a positive or negative effect on IDEB. Still, variable X3, which represents teacher service periods in the state education system, indicates a clear trend towards the positive effect of a low length of service in the target.

This same trend is followed by variable X4, which reflects the fact that teachers with more concentrated activities have a positive effect on IDEB.

Finally, X2 which stands for teacher attendance in technological training promoted by the state government in partnership with Google, proved to be the most important variable for the M1 model, as to potential target effects for IDEB.

6-2- Performance Metrics (Error Measures) in Machine Learning Regression - M1 Model

Results were evaluated by the coefficient of determination (R^2). The coefficient of determination is the proportion of the variance in response variable that is explained by the model. R^2 is an accuracy statistic which allows to assess a regression model.

It is said to be an accuracy of the regression model; in summary, this coefficient of determination R-squared is more informative than MAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation, as shown in Soper [29].

The models were trained with k-Fold Cross-Validation = 10, a procedure that is a standard method for estimating the performance of a machine learning algorithm on a dataset. For MAE, MSE, RMSE, R^2, RMSLE, and MAPE, we adopted the default options provided by the software platform for training and model regularization in the python, respectively. Performances of the regression models to be considered were investigated by means of several simulation tests, carried out using data from 450 schools. All data processing and analysis were performed using Python v3.10. The prediction performances were measured by metrics commonly adopted for this type of problem, which were independent of the said procedure for data normalization: NMSE, MARE, MSE and MAPE.

- Normalized Mean Square Error (NMSE);

$$NMSE = \frac{\sum_{t=1}^{N_T}(Y_t - \tilde{Y}_t)^2}{\sum_{t=1}^{N_T}(Y_t - \bar{Y}_t)^2} \quad (3)$$

where $\tilde{y}$ is the average value of samples y_t in the test set

- Mean Absolute Range Error (MARE);

$$MARE = \frac{1}{N_T}\sum_{t=1}^{N_T}\frac{|y_t - \tilde{y}_t|}{y_{max} - y_{min}} \quad (4)$$

- Mean Square Error (MSE);

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i^{estimated} - X_i^{measured})^2 \quad (5)$$

- Mean Absolute Percentage Error (MAPE);

$$MAPE = \left\{Mean\left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i^{estimated} - X_i^{measured}}{X}\right|\right)\right\} \quad (6)$$

The results of the modeling steps based on the training phases, i.e., the results will be demonstrated by predictive models:

Since a lot of models and methods are presented in this work, a remark must be done on how these solutions are evaluated in terms of forecasting performance. In general, the interest is to know how much accurate a prediction is, considering the best forecasting method as the most accurate one, in terms of some error metric. This section discusses ways of measuring prediction accuracy.

Table 1 presents a performance comparison between models based on DTs considered in this study.

Table 1. Finding the best model for M1 by analyzing: Random Forest Regressor (RFR); Light Gradient Boosting Machine (LGBM); AdaBoost Regressor (AR); Extra Trees Regressor (ETR); Decision Tree Regressor (DTR)

Results								
Models		**MAE**	**MSE**	**RMSE**	**$\mathbb{R}^2$**	**RMSLE**	**MAPE**	**TT (Sec)**
rf	**RFR**	0.3818	0.3066	0.5494	0.8139	0.0921	0.0782	0.2180
lightgbm	**LGBM**	0.4145	0.3110	0.5535	0.8034	0.0930	0.0852	0.2030
ada	**AR**	0.4726	0.3527	0.5910	0.7939	0.1006	0.0989	0.0290
et	**ETR**	0.3964	0.3712	0.6011	0.7498	0.1009	0.0814	0.8620
dt	**DTR**	0.4155	0.4216	0.6423	0.7737	0.1073	0.0854	0.0220

Table 1 indicates that the Random Forest Regressor model presents better results for the M1 model performance, R^2 = 0.8139. The model opti mization was performed through the Bayesian Optimization method (see Table 2), a model optimization method that takes a long time to converge, recommended for models with less than 20 predictors [30].

Table 2. Applying Bayesian optimization to the featured for M1 model

Hyperparameter optimization						
Metric	**MAE**	**MSE**	**RMSE**	**$\mathbb{R}^2$**	**RMSLE**	**MAPE**
Mean	0.3941	0.2910	0.5355	0.8550	0.0899	0.0809

The training and validation process of the DT was conducted with ”tpe” algorithm from optuna library. This library is particularly designed for machine learning, where:

search algorithm=“tpe” - means Bayesian optimization algorithm called Tree-structured Parzen Estimator;

choose better=True - when set to True, the returned object is always better performing; the metric used for comparison is defined by the optimize parameter.

6-3- Predictive Model M2: - Physical Infrastructure

Considering the analyzed variables in the M2 model:

X6 = Reading room;

X7 = Computer Lab;

X8 = Science lab;

X9 = Water supply or Filtered water; X10 = kitchen;

X11 = Restroom – WC;

The SHAP graph in Figure 7 shows that the variable X6 is the one that can most positively impact the IDEB final result. As well as the x7 and x8 variables, which showed high impact.

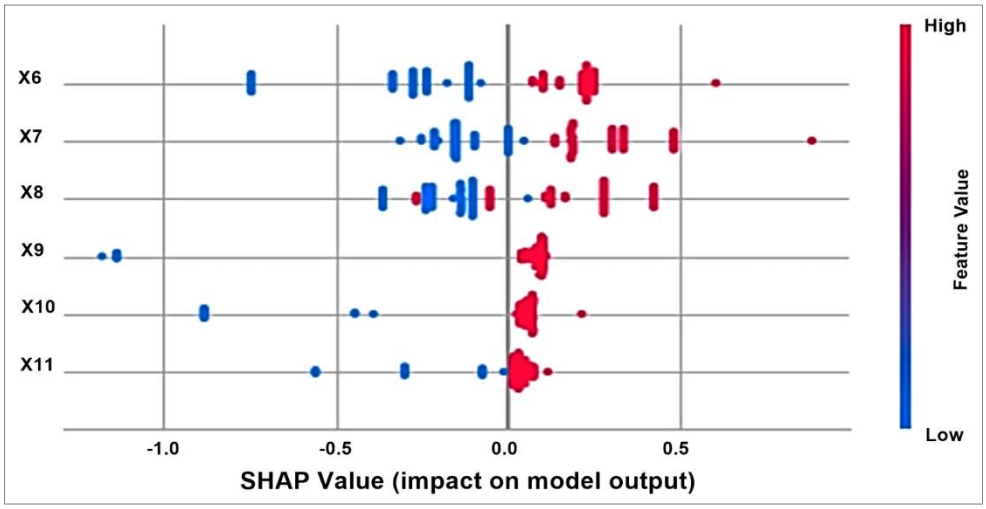

Figure 7. Test - SHAP method applied to predictive model M2

The performance analysis indicates the effects of each variable in the model (Figure 7). Where:

The SHAP graph in Figure 7 exhibits the direction and intensity of the variable effects, specifically high values of X6, X7, X8, X9, X10 and X11 variables (in pink) impacting positively in the final prediction result for IDEB, while low values (in blue) would have a negative impact. It is worth remembering that, in our database, the presence of resource is represented by “1” and the non-presence by “0”, so when we say that it has a positive impact, we mean that it increases the probability of being “1”.

Table 3 demonstrates DT model performances for the predictive M2 regression.

Table 3. Finding the best model for M2 by analyzing: Random Forest Regressor (RFR); Light Gradient Boosting Machine (LGBM); AdaBoost Regressor (AR); Extra Trees Regressor (ETR); Decision Tree Regressor (DTR)

		Results						
Models		**MAE**	**MSE**	**RMSE**	$\mathbb{R}^2$	**RMSLE**	**MAPE**	**TT (Sec)**
et	**ETR**	0.3537	0.2316	0.4783	0.8444	0.0808	0.0722	0.1700
dt	**DTR**	0.3537	0.2316	0.4783	0.8444	0.0808	0.0722	0.0170
rf	**RFR**	0.3587	0.2336	0.4803	0.8398	0.0811	0.0732	0.2040
lightgbm	**LGBM**	0.3884	0.2460	0.4926	0.8127	0.0833	0.0793	0.0220
ada	**AR**	0.4226	0.2992	0.5432	0.7834	0.0902	0.0849	0.0180

All the adjustment and performance indicators were better predicted by the Extra Trees Regressor - ETRs model, with an initial value of $R^2 = 0.8444$. An ETR optimization procedure was also performed (see Table 4), obtaining $R^2 = 0.8488$.

Table 4. Applying Bayesian optimization to the featured for M2 model

Hyperparameter optimization						
Metric	**MAE**	**MSE**	**RMSE**	**$\mathbb{R}^2$**	**RMSLE**	**MAPE**
Mean	0.3563	0.2300	0.4766	0.8488	0.0805	0.0726

6-4- Predictive Model M3: - Technological Infrastructure

Variables analyzed in the M3 model were: X12=Computers by students; X13 = Broadband Internet; X14 = Projector; X15 = Printer.

The SHAP graph, Figure 8, shows that variable X12, which represents the number of computers per student, is the most important for the model, and high values of this variable represent a positive impact on the IDEB final result. As well as high values of X13 and X14 variables, which showed high impact. Meanwhile, the X15 variable proved to have few contributions to the M3 model.

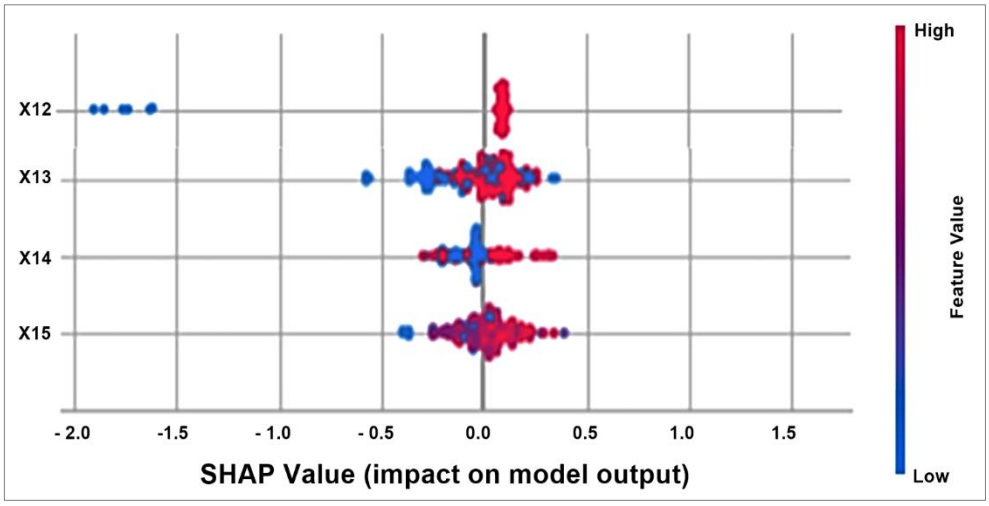

Figure 8. Test - SHAP method applied to predictive model M3

The performance analysis indicates the effects of each variable in the model (see Table 5).

Table 5. Finding the best model for M3 by analyzing: Random Forest Regressor (RFR); Light Gradient Boosting Machine (LGBM); AdaBoost Regressor (AR); Extra Trees Regressor (ETR); Decision Tree Regressor (DTR)

Results								
Models		**MAE**	**MSE**	**RMSE**	**$\mathbb{R}^2$**	**RMSLE**	**MAPE**	**TT (Sec)**
et	**ETR**	0.3829	0.2698	0.5149	0.8698	0.0883	0.0802	1.0120
dt	**DTR**	0.3836	0.2700	0.5151	0.8691	0.0884	0.0803	0.0210
rf	**RFR**	0.3890	0.2710	0.5159	0.8660	0.0886	0.0814	0.2020
lightgbm	**LGBM**	0.4529	0.3145	0.5573	0.8454	0.0956	0.0948	0.2350
ada	**AR**	0.4494	0.3288	0.5694	0.8154	0.0974	0.0937	0.0170

All the adjustment and performance indicators were better predicted by the Extra Trees Regressor - ETRs model, with an initial value of $R^2 = 0.8698$. An ETR optimization procedure was also performed (see Table 6), obtaining $R^2 = 0.8752$.

Table 6. Applying Bayesian optimization to the featured for M4

Hyperparameter optimization						
Metric	**MAE**	**MSE**	**RMSE**	**$\mathbb{R}^2$**	**RMSLE**	**MAPE**
Mean	0.3899	0.2692	0.5145	0.8752	0.0882	0.0818

6-5- Predictive Model M4: Outcome

This study uses an experimental database composed of 450 schools, and 617 teachers, in order to suggest a methodology for creating a classification label for variables that most influence the teaching-learning process, being a case study of school performance.

In this analysis, the best predictors of each of the attributes of previous models were detected, analyzing the following databases: teacher profile (M1), physical infrastructure (M2), technological infrastructure of schools (M3). That is, the best predictors of each of the attributes of previous models (M1, M2, M3) were used in this last step of the search. Using these databases as input to the model, it was possible to explore attribute selection techniques and predictive algorithms, aiming to develop a model of a predictor to identify which factors positively impact schools IDEB (Basic Education Development Index), in regards to a case study in Brazil.

The following variables were used in this last step to develop the predictive model, with the output node M4:

The bests predictors of each of the attributes of the previous models (M1, M2, M3) were used:

X12 = PCs per students, X3 = Employment period; X4 = Number of schools working; X14 = Broadband Internet; X2 = Participation in ongoing training on technology: (Frequency of participation of teachers in continuing education in technology promoted by partner Google for Education).; X7 = Computer Lab;

Figure 9 demonstrates the performance of the variables trained for the target (IDEB). In this case, the SHAP graph shows that variable X12 is the most important for the final model. The graph shows that high values of X12, X14 and X7 have positive effects on IDEB, while low values of X3 and X4 contribute positively to IDEB. The X2 variable proved to have few contributions to the final model.

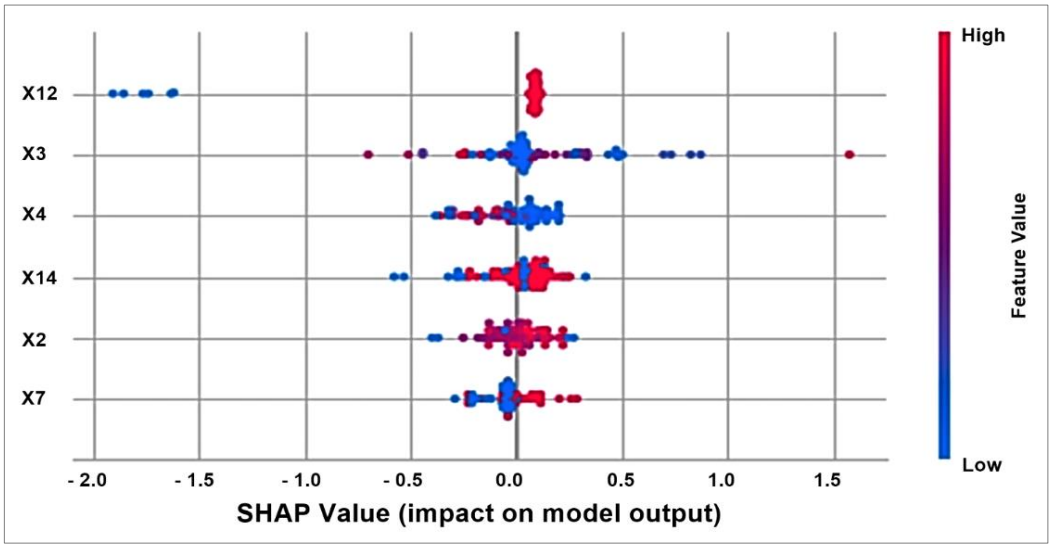

Figure 9. Test - SHAP method applied to predictive model M4

This output of the final model, M4, shows relevant results in the task of identifying attributes that have great importance in the evaluation of educational performance in different school cycles. This is an important result, which contributes to the decision-making process, showing public managers how greater investments in technologies, for example, can positively contribute to student performance, variable X12, X14 and X7. The work also shows that teachers need time to prepare their classes more efficiently and with more quality, and, consequently, can improve the quality of teaching, variable X3 and X4.

All these educational indicators (X12, X3, X4, X14, X2, X7) used here are metrics that help in assessing the education system quality. They are often associated with economic and social factors suggested to contribute to good school performance. The main objective of this work was to evaluate factors related to school performance. Using a dataset composed of Brazilian school performance variables (IDEB), socioeconomic and school structure variables, we generated different models.

The techniques proposed in the present study are noteworthy since they can aid in evaluating important parameters in the field of education, being relevant for identifying variables that should receive, for example, more investments. The work contributes to the area of school management and the decision-making process. Future research should

consider different case studies and different databases, other prediction models, and possible improvements that can be obtained using input resource selection methods, as well as the distributed processing of multiple data sources and multiple schools, even in other countries, showing its international relevance.

The performance analysis indicates the effects of each variable in the model (see Table 7).

Table 7. Finding the best model for M4 by analyzing: Random Forest Regressor (RFR); Light Gradient Boosting Machine (LGBM); AdaBoost Regressor (AR); Extra Trees Regressor (ETR); Decision Tree Regressor (DTR)

Models		MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	**ETR**	0.3064	0.2408	0.4796	0.8905	0.0799	0.0615	0.8220
dt	**DTR**	0.3597	0.3291	0.5627	0.8749	0.0946	0.0725	0.0160
rf	**RFR**	0.3465	0.2519	0.4924	0.8722	0.0824	0.0700	0.2070
lightgbm	**LGBM**	0.4460	0.3128	0.5552	0.7357	0.0933	0.0912	0.2150
ada	**AR**	0.3976	0.2613	0.5098	0.7244	0.0865	0.0819	0.0230

All the adjustment and performance indicators were better predicted by the Extra Trees Regressor - ETRs model, with an initial value of $R^2 = 0.8905$. An ETR optimization procedure was also performed (see Table 8), obtaining $R^2 = 0.8991$.

Table 8. Applying Bayesian optimization to the featured for M4 model

Hyperparameter optimization						
Metric	**MAE**	**MSE**	**RMSE**	$\mathbb{R}^2$	**RMSLE**	**MAPE**
Mean	0.3154	0.2694	0.4514	0.8991	0.0757	0.0642

7- Conclusions

Machine learning approaches are being used increasingly in the analysis of parameters that influence the teaching-learning process. It is crucial to understand why a data-driven model makes any prediction based on particular input data to: 1) understand model decisions, 2) understand complex under- lying non-linear relationships, and 3) assess the applicability of the model for further analysis and evaluation.

This study analyzes the importance of financial investments in key areas in order to improve measures such as the Basic Education Development Index (IDEB). Extensive experimental data are used in this study; the data are randomly split into a training set and a test set to construct the machine learning model. The ensemble methods were established based on the training set, with the goal to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability/robustness over a single estimator. In this research, the two best-known families of ensemble methods in machine learning, known as averaging and boosting methods, were implemented.

In averaging methods, the principle implemented here was to build multiple estimators independently and then average their predictions. On the other hand, in the boosting methods, the base estimators were built sequentially, which reduced the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble. In the final model, the extra-tree regressor (ETR) had shown satisfactory results, with MSE = 0.2094 and $R^2 = 0.8991$ for the test data. In machine learning approaches, this class implements a meta estimator that fits a number of randomized DTs on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

The model is then explored with SHAP to identify the feature importance and decode the complex underlying relationships between IDEB (target) and input variables. The X12, computers per student, has the greatest influence on failure mode. However, other variables also have significant influence, like X13, Broadband Internet, and X14, Projector. An increase in X13 and X14 also increases the value of SHAP and the associated probability of predicting improvements in the final grade of IDEB, while variable X15, Printer, is the least significant to the model.

Although the current study is limited to predicting variables that can improve IDEB, the SHAP approach to interpreting machine learning models is relevant and applicable to other problems in the field of education. More associated studies are needed to reach firm conclusions. Interpretable machine learning models in education can contribute to decision-making processes that improve teaching quality in many critical cases.

Therefore, it is necessary that public managers, as well as managers of educational entities, seek to carefully evaluate all the variables pertaining to performance in the predictive models trained with a target in IDEB so that decisions regarding educational strategies can be made to improve the quality of teaching.

8- Declarations

8-1- Author Contributions

Conceptualization, Y.V.S.M. and P.G.V.N.; methodology, Y.V.S.M. and Y.V.S.M; software, Y.V.S.M.; validation, Y.V.S.M. and P.G.V.N.; formal analysis, Y.V.S.M.; investigation, Y.V.S.M.; resources, Y.V.S.M; data curation, Y.V.S.M and P.G.V.N.; writing—original draft preparation, Y.V.S.M., D.P., and P.G.V.N.; writing—review and editing, Y.V.S.M. and P.G.V.N.; supervision, P.G.V.N. and D.P. All authors have read and agreed to the published version of the manuscript.

8-2- Data Availability Statement

The data presented in this study are available in the article.

8-3- Funding

This work received partial support from national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

8-4- Institutional Review Board Statement

Not applicable.

8-5- Informed Consent Statement

Not applicable.

8-6- Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

9- References

[1] Raffaghelli, J. E., Rodríguez, M. E., Guerrero-Roldán, A. E., & Bañeres, D. (2022). Applying the UTAUT model to explain the students' acceptance of an early warning system in Higher Education. Computers and Education, 182, 104468. doi:10.1016/j.compedu.2022.104468.

[2] Barrett, P., Treves, A., Shmis, T., Ambasz, D., & Ustinova, M. (2019). The Impact of School Infrastructure on Learning: A Synthesis of the Evidence. In The Impact of School Infrastructure on Learning: A Synthesis of the Evidence. The World Bank, Washington, United States. doi:10.1596/978-1-4648-1378-8.

[3] Li, S., & Liu, T. (2021). Performance Prediction for Higher Education Students Using Deep Learning. Complexity, 2021. doi:10.1155/2021/9958203.

[4] Fernandez Rivas, D., Boffito, D. C., Faria-Albanese, J., Glassey, J., Cantin, J., Afraz, N., Akse, H., Boodhoo, K. V. K., Bos, R., Chiang, Y. W., Commenge, J. M., Dubois, J. L., Galli, F., Harmsen, J., Kalra, S., Keil, F., Morales-Menendez, R., Navarro-Brull, F. J., Noël, T., … Weber, R. S. (2020). Process intensification education contributes to sustainable development goals. Part 2. Education for Chemical Engineers, 32, 15–24. doi:10.1016/j.ece.2020.05.001.

[5] Matzavela, V., & Alepis, E. (2021). Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. Computers and Education: Artificial Intelligence, 2, 100035. doi:10.1016/j.caeai.2021.100035.

[6] Jiang, X. (2021). Online English Teaching Course Score Analysis Based on Decision Tree Mining Algorithm. Complexity, 2021, 1–10. doi:10.1155/2021/5577167.

[7] SEDUC (2020). Secretary of States for Education. State office of education in Belém, Brazil. Available online: http://www.seduc.pa.gov.br/portal/escola/consulta_matricula/RelatorioMatriculas.php (accessed on May 2022). (In Portuguese).

[8] Ajayi, I. A., & Ekundayo, H. T. (2011). Factors determining the effectiveness of secondary schools in nigeria. Anthropologist, 13(1), 33–38. doi:10.1080/09720073.2011.11891174.

[9] Pangeni, K. P. (2014). Factors determining educational quality: Student mathematics achievement in Nepal. International Journal of Educational Development, 34(1), 30–41. doi:10.1016/j.ijedudev.2013.03.001.

[10] Khanna, L., Singh, S. N., & Alam, M. (2016). Educational data mining and its role in determining factors affecting students academic performance: A systematic review. 2016 1st India International Conference on Information Processing (IICIP). doi:10.1109/iicip.2016.7975354.

[11] Byeon, H. (2017). Chi-Square Automatic Interaction Detection Modeling for Predicting Depression in Multicultural Female Students. International Journal of Advanced Computer Science and Applications, 8(12), 179–183. doi:10.14569/ijacsa.2017.081222.

[12] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, 1–24. doi:10.7717/PEERJ-CS.623.

[13] Parreñas, R. (2020). The International Division of Reproductive Labor. In Servants of Globalization. Servants of Globalization, Stanford University Press, Redwood City, United States. doi:10.1515/9780804796187-003.

[14] Shamim, A., Hussain, H., & Maqbool Uddin Shaikh. (2010). A framework for generation of rules from decision tree and decision table. 2010 International Conference on Information and Emerging Technologies. doi:10.1109/iciet.2010.5625700.

[15] Priyanka, & Kumar, D. (2020). Decision tree classifier: A detailed survey. International Journal of Information and Decision Sciences, 12(3), 246–269. doi:10.1504/ijids.2020.108141.

[16] Liang, J., Qin, Z., Xiao, S., Ou, L., & Lin, X. (2021). Efficient and Secure Decision Tree Classification for Cloud-Assisted Online Diagnosis Services. IEEE Transactions on Dependable and Secure Computing, 18(4), 1632–1644. doi:10.1109/TDSC.2019.2922958.

[17] Alves, M. da C. P., Barros, R. C. B. de, & Carrozza, G. (2018). O Ideb e seus efeitos de sentido na Educação Básica do Brasil. Interfaces, 9(2), 29–40. doi:10.5935/2179-0027.20180020. (In Portuguese).

[18] de Carvalho, M. R. V. (2018). Basic education teacher profile. Relatos de Pesquisa, (41), 68-68. (In Portuguese).

[19] Batunacun, Wieland, R., Lakes, T., & Nendel, C. (2021). Using Shapley additive explanations to interpret extreme gradient boosting predictions of grassland degradation in Xilingol, China. Geoscientific Model Development, 14(3), 1493–1510. doi:10.5194/gmd-14-1493-2021.

[20] Chávez Hernani, M., & Vieira, S. D. R. (2020). Reflexões sobre os cursos de Formação de Professores no Peru e no Brasil. Educación, 29(57). doi:10.18800/educacion.202002.001. (In Portuguese).

[21] Klassen, R. M., Durksen, T. L., Al Hashmi, W., Kim, L. E., Longden, K., Metsäpelto, R. L., Poikkeus, A. M., & Györi, J. G. (2018). National context and teacher characteristics: Exploring the critical non-cognitive attributes of novice teachers in four countries. Teaching and Teacher Education, 72, 64–74. doi:10.1016/j.tate.2018.03.001.

[22] Cattaneo, A. A. P., Antonietti, C., & Rauseo, M. (2022). How digitalised are vocational teachers? Assessing digital competence in vocational education and looking at its underlying factors. Computers and Education, 176, 104358. doi:10.1016/j.compedu.2021.104358.

[23] Elacqua, G., & Marotta, L. (2020). Is working one job better than many? Assessing the impact of multiple school jobs on teacher performance in Rio de Janeiro. Economics of Education Review, 78, 102015. doi:10.1016/j.econedurev.2020.102015.

[24] Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How Teacher Turnover Harms Student Achievement. American Educational Research Journal, 50(1), 4–36. doi:10.3102/0002831212463813.

[25] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. CRC Press, Boca Raton, United States.

[26] Loh, W. (2011). Classification and regression trees. WIREs Data Mining and Knowledge Discovery, 1(1), 14–23. Portico. doi:10.1002/widm.8.

[27] Timofeev, R. (2004). Classification and regression trees (CART) theory and applications. Humboldt University, Berlin, Germany.

[28] Codo, W. (2006). For a Psychology of Work: Essays. Casa do Psicólogo, Brazil. (In Portuguese).

[29] Soper, D. S. (2021). Greed is good: Rapid hyperparameter optimization and model selection using greedy k-fold cross validation. Electronics (Switzerland), 10(16). doi:10.3390/electronics10161973.

[30] Verwer, S., Zhang, Y. (2017). Learning Decision Trees with Flexible Constraints and Objectives Using Integer Optimization. Integration of AI and OR Techniques in Constraint Programming. CPAIOR 2017. Lecture Notes in Computer Science, 10335, Springer, Cham, Switzerland. doi:10.1007/978-3-319-59776-8_8.

[31] Frazier, P. I. (2018). A tutorial on Bayesian optimization. arXiv Preprint, arXiv:1807.02811. doi:10.48550/arXiv.1807.02811.