Pesquisa Operacional (2024) 44: e275092 p.1-18
doi: 10.1590/0101-7438.2023.043.00275092

Printed version ISSN 0101-7438 / Online version ISSN 1678-5142
www.scielo.br/pope
APPLICATIONS TO THE PUBLIC SECTOR SPECIAL ISSUE ARTICLES

AN EARLY WARNING SYSTEM FOR SCHOOL DROPOUT IN THE STATE OF ESPÍRITO SANTO: A MACHINE LEARNING APPROACH WITH VARIABLE SELECTION METHODS

Guilherme Armando de A. Pereira[1*], Kiara de Deus Demura[2], Iago de Carvalho Nunes[3], Katia Cesconeto de Paula[4] and Pablo Silva Lira[5]

Received May 25, 2023 / Accepted October 19, 2023

ABSTRACT. School dropout has significant consequences for individuals and society, including increased crime, reduced productivity, and limited economic innovation. Identifying students at risk of dropping out is crucial. This paper aims to develop a logistic regression-based tool for predicting dropout in the state's public schools of Espírito Santo, Brazil. We utilized students' information, such as grades, school attendance, socioeconomic data, and others, provided by Espírito Santo State Education Secretariat and the National Institute of Educational Studies and Research Anísio Teixeira. Various regularization methods were employed. We compared three model specifications for students in the first year of high school in Espírito Santo using data from 2019 to 2022. The results indicated that the models could identify at-risk students satisfactorily, highlighting the effective use of available data by educational departments to identify potential dropouts. This tool can aid educators in creating targeted interventions to minimize dropout rates.

Keywords: education, machine learning methods, school dropout prediction.

1 INTRODUCTION

Education is one of the fundamental pillars for the social and economic development of a society. Over the years, researchers have studied the impacts on various dimensions of a nation (Psacharopoulos & Patrinos 2018).

*Corresponding author

[1] Federal University of Espírito Santo, Department of Economics, Vitória, ES, Brazil – E-mail: guilherme.aa.pereira@ufes.br – https://orcid.org/0000-0002-2833-1384

[2] Jones dos Santos Neves Institute, Vitória, ES, Brazil – E-mail: kiara.demura@ijsn.es.gov.br – https://orcid.org/0000-0001-9783-0892

[3] Jones dos Santos Neves Institute, Vitória, ES, Brazil – E-mail: iagodcn@gmail.com – https://orcid.org/0000-0001-6077-7135

[4] Jones dos Santos Neves Institute, Vitória, ES, Brazil – E-mail: katia.cesconeto@ijsn.es.gov.br – https://orcid.org/0009-0006-3747-4631

[5] Jones dos Santos Neves Institute, Vitória, ES, Brazil – E-mail: pablo.lira@ijsn.es.gov.br – https://orcid.org/0000-0002-2643-5219

Burgess (2016) argues that education brings numerous benefits, including improved quality of life and health, reduced teen pregnancy, lower crime rates, and decreased unemployment rate, among others. Furthermore, education is crucial in consolidating and strengthening democracy and citizenship.

High dropout rates, which occur when students discontinue their education during the school year, pose significant challenges for public administrators and educators. These high dropout rates are associated with lower wages, increased vulnerability during economic crises, higher crime rates, incarceration, and elevated mortality rates. From a social perspective, high dropout rates have several impacts, including reduced tax collection, increased spending on social programs, and difficulty in attracting businesses that require skilled workers (Parr & Bonitz 2015; Wood et al. 2017).

According to the Continuous National Household Sample Survey (PNAD) education module conducted by IBGE (2020), approximately 20.2% of individuals aged 14 to 29 years in Brazil did not complete high school. This means that 10.1 million people either dropped out of school before completing the academic cycle or never attended it. The dropout problem does not affect all students equally. It is more prevalent among men (58.3%) than women (41.7%). Furthermore, when considering race, the disparity between groups becomes even more significant, with non-whites accounting for 71.7% of total dropouts, while the dropout rate among whites is 27.3%.

Giuberti et al. (2018) assert that the reasons behind dropout are well-documented in the literature and are primarily associated with students' socioeconomic characteristics, family background, academic performance, school infrastructure, and quality of the teachers, among other factors.

Identifying students at risk of dropping out is crucial for developing strategies to mitigate this issue. Several models for predicting dropout have been proposed in the literature, including those by Bayer et al. (2012), Martinho et al. (2013), Wood et al. (2017), Giuberti et al. (2018), Adelman et al. (2018), and others. These models employ machine learning techniques such as logistic regression, support vector machine (SVM), random forest, adaptive boosting, classification, and regression tree (CART), naive Bayes, artificial neural networks, and decision trees, as discussed in studies (Sara et al. 2015; Costa et al. 2017; Bayer et al. 2012; Dekker et al. 2009; Sandoval-Palis et al. 2020; Jiménez-Gómez et al. 2015; Rovira et al. 2017), among others. Moreover, such tools have been employed by authorities around the world such as Norway (Sletten et al. 2022), in the state of Victoria in Australia (Lamb & Rice 2008), in the state of Wisconsin in the United States (Knowles 2015), Peru (Ministry of Education 2020) or in the Chile (Uldall & Rojas 2022).

In Espírito Santo, the Jones dos Santos Neves Institute (IJSN), in partnership with the Espírito Santo State Education Secretariat (SEGES), has been developing a logistic regression model to forecast student dropout, as described in detail by Giuberti et al. (2018). Ongoing studies are focused on enhancing this tool. One key question pertains to the selection of variables, given the increasing computational capabilities and the recognized significance of data collection, storage, and processing, resulting in a larger number of variables.

Therefore, the primary objective of this study is to construct a dropout identification model. To accomplish this, we utilized confidential students' information, such as grades, school attendance, socioeconomic data, and others, provided by Espírito Santo State Education Secretariat and the National Institute of Educational Studies and Research Anísio Teixeira. These methods allow for tracing the profiles of students likely to drop out. Then, the gathered information can be used for generating a report on potential dropouts, enabling educators to intervene proactively. Moreover, establishing and understanding the profiles of students in dropout situations provide valuable insights for planning long-run strategies to mitigate this problem.

As a secondary goal, we analyze whether the students' information used in this research provides relevant insights for identification. This is particularly important as similar data are typically available in other educational departments across Brazilian states, indicating that this type of modeling can be implemented elsewhere.

Our case study focuses on the first year of the state's high school in Espírito Santo. We constructed three different models using distinct tools for variable selection.

The results validate the importance of the information available in the SEGES and INEP datasets for predicting dropout, and the applied methodologies successfully identified students at risk of dropping out. Lastly, it was observed that students who are behind their grade level, possess low grades, and have a high number of absences are more susceptible to school dropout.

This work is organized as follows. Section 2 presents related works. Section 3 gives the mathematical tools employed. Section 4 presents the results, estimated models, and performances in predictive terms. Conclusions are in Section 5.

2 RELATED WORKS

Education data mining (EDM) is the process of applying data mining techniques to extract knowledge from educational datasets in order to improve teaching-learning process. For instance, EDM studies can evaluate universities' technical efficiency, as demonstrated by Visbal-Cadavid et al. (2019), analyse the determinants of school performance, such as Soares & Mendonça (2003), and predict school dropout, which is the focus of this paper. Dol & Jawandhiya (2023) conducts an extensive review of 142 articles published between 2010 and 2020 on EDM where the importance of EDM is presented via distinct applications and mathematical tools.

Although the reasons that lead students to leave school before completing the academic cycle are multifactorial, often stemming from socioeconomic vulnerability, school infrastructure, historical performance, and individual psychological characteristics, the use of computational tools for dropout prediction has proven effective worldwide. There are various studies for predicting dropout using machine learning methods. Nevertheless, it is worth noting that each individual study or model possesses its own distinctiveness, as a consequence of the idiosyncratic nature of data and student attributes within each specific country, city, school, and educational tier. For these reasons there is a lack of a global model the brings all the possible variables that concerns the dropout (Oqaidi, Aouhassi & Mansouri 2022).

In terms of educational level, there is a considerable focus on higher education. Dekker et al. (2009) employed machine learning methods[1] to predict student dropouts in the electrical engineering department of Eindhoven University of Technology. The results demonstrate the effectiveness of several classification techniques, achieving accuracies ranging from 75% to 80%.

At Masaryk University, Bayer et al. (2012) utilized predictive models[2] to identify applied computing students prone to course evasion. In addition to personal and teachers' information, the authors incorporated data on student access to the computer platforms, emails exchanged, shared files, and forums. Based on this behavioral information, they created new variables through network analysis, which were subsequently used in the prediction models. The accuracy of the models varied approximately from 50% to 92%.

Márquez-Vera et al. (2016) introduced a methodology based on genetic programming. The model was tested on 419 students from the Autonomous University of Zacatecas. The results indicate that the proposed method effectively predicts high school dropout.

Rovira et al. (2017) analyzed machine learning methods[3] for early dropout detection and course grade prediction. The study involved a dataset of 4,434 students enrolled in three different courses (Law, Mathematics, and Computer Science) at the University of Barcelona. The results revealed that it is not possible to establish a universal model due to variations in performance across different courses. For the Law course, random forest and adaptive boosting were the most suitable methods. In contrast, logistic regression and naïve Bayes demonstrated better performance in the Mathematics and Computer Science courses.

Sandoval-Palis et al. (2020) evaluated the performance of the logistic regression and the artificial neural networks to predict dropout in the leveling courses of the Escuela Politécnica Nacional. The dataset used for analysis included academic and socioeconomic information of 2,097 students. The results revealed that the neural network model outperformed the logistic regression model, achieving a predictive accuracy of 76%.

Niyogisubizo et al. (2022) proposed an ensemble model based on random forest, gradient boosting, extreme gradient boosting, and neural networks at the Constantine the Philosopher University in Nitra using data gathered about 261 student's on-line learning environment actives and partial achievements. The results indicated high performance of the ensemble model, with accuracies varies from 76,67% to 92,18%.

Kim et al. (2023) developed a hybrid student dropout prediction system, based on boosting and clustering algorithms, capable of identifying students at risk and classifying the reasons for dropping out. The study considered a dataset composed of academic data, academic records, personal information, facility use history, and website use history from 67,060 students from Gyeongsang National University, South Korea. The model's accuracy was approximately 0,98.

[1] CART, C4.5, a Bayesian classifier, a logistic model, a rule-based learner, and random forest.

[2] J48 decision tree learner, IB1 lazy learner, PART rule leaner, support vector machine and naïve Bayes classifier.

[3] Logistic regression, naïve Bayes, SVM, random forest and adaptive boosting.

For fully on-line courses, Bañeres, et al. (2023) developed a system considering learner's profile information, performance data within the course, and daily clickstream data about the course on-line platform to reduce dropout at the assessable activity level. The model was evaluated on 581 students from the Faculty of Economics and Business at the Universitat Oberta de Catalunya. The results confirms that early intervention can reduce the dropout rates and increase the engagement in the course. Considerable studies have also focused on primary and/or secondary school.

In Denmark, Sara et al. (2015) conducted a study to predict high school dropouts using various data mining techniques[4]. The authors considered 72,598 students and aimed to identify those who were likely to abandon school in the next three months. The random forest model demonstrated the highest performance, achieving an accuracy of 93.5%.

Adelman et al. (2018) employed linear probability models in Guatemala and Honduras. The model is based on socio-demographic information, school characteristics, and municipality factors. The findings suggest that these simple models yield satisfactory results, accurately identifying 80% of sixth-grade students at risk of dropping out.

Uldall & Rojas (2022) compared traditional machine learnings methods[5] to the four-year period of all schools from Chile, excluding the private ones. The data employed consists of socioeconomic data, registration information, attendance records, school grades and teacher data. The models reached an accuracy between 93% to 98%.

Mnyawami, Maziku & Mushi (2022) employed an Automated Machine Algorithm in a dataset containing features related to students, family, and school of 206,855 students of secondary schools in Tanzania. The results indicated an accuracy between 97% and 99%, approximately.

In possession of longitudinal data, in Chile, Rodríguez et al. (2023) developed a model that considers the individual trajectories of the students. Thus, the proposed model is capable of identify trends and patterns that lead to abandon. The model considered administrative data and the results indicated that the proposed model performs satisfactorily. The proposed approach is quite interested, as it is known that the dropout is the final stage of a process that evolves over the time. On the other hand, the main drawback is the limitation of the dataset, for this reason most dropout models consider cross-section data.

In Brazil, some related studies were developed. Different from the previously mentioned works, where the goal was to identify students likely to evade school early, Nascimento et al. (2018) and Nascimento et al. (2022) employed statistical learning methods to predict dropout rates in schools using public information provided by the Anísio Teixeira National Institute of Educational Studies and Research (INEP).

Cunha et al. (2016) used cluster analysis to detect the reasons for dropout and failure at the Federal Institute of Education, Science, and Technology of Rio Grande do Norte to define the student's profile prone to drop out. The results indicate that students from public schools, with a

[4] SVM, random forest, CART, and naïve Bayes.

[5] Logit, decision trees, random forests, and neural networks.

family income less than one minimum wage, with low school performance, who live with their parents, who are unemployed or are minors constitute the profile of students who do not complete the academic cycle.

Martinho et al. (2013) developed a *Fuzzy-ARTMAP Neural Network* model based on academic and socioeconomic information of students of the technology courses of the Federal Institute of Mato Grosso. The results indicate an accuracy of approximately 85%.

Barros, et al. (2019) compared different ML algorithms[6] trained with original data and with balanced data[7] composed of 7,718 students of Integrated Education (secondary education with training in professional education through technical courses) of the Federal Institute of Rio Grande do Norte, Brazil. The variables considered are related to academic performance, demographic and socio-economic characteristics. The results indicated a superiority in favour of balanced bagging.

Krüger et al. (2023) used ensemble classifier based on decision tree, logistic regression, random forest, adaboost and xgboost to predict the dropout of a group of Brazilian private schools. Considering socio-economic information related to the school's local region and school's educational system the model could predict satisfactorily students at dropout risk.

The review of related works found many studies that employ ML algorithms to school dropout. One potential limitation of these approaches is their lack of interpretability, as most of these models are black-box systems. This limitation hinders educators from interpreting the results and understanding the potential reasons behind student dropout. To address this concern, our research utilized a logistic regression model, which offers more interpretability.

Furthermore, some of the cited studies had a limited number of features, leading them to include all available in their models. In contrast, other researches opted to pre-select certain variables before the training process. In our present paper, we considered regularization models, which not only select the most relevant variables but also estimate the parameters.

Finally, it is worth noting that in general, these models are often tested within specific schools or universities. However, our present study is conducted at a state-wide level in Brazil, encompassing all public schools in the State of Espírito Santo. This means that our work operates within a high-dimensional data environment, which presents unique challenges and opportunities.

3 MATHEMATICAL TOOLS

3.1 Logistic regression

The logistic regression, also known as the logit model, when dealing with a binary dependent variable, has been widely employed for classification problems in various areas. These include, for instance, credit scoring modeling, economic recessions, dropout prediction, football games,

[6] Decision trees, neural networks, and balanced bagging

[7] Imbalaced data problem occurs when classes are considerable disproportional. This might lead to a poor performance in ML models. To solve this issue, we can employ pre-processing algorithms in order to balance the data.

image classification (Alves et al. 2010, Ng 2012, Selau & Ribeiro 2013, Ferreira et al. 2013, Nazish et al. 2021, Rovira et al. 2017, Hu et al. 2021).

Formally, the model can be defined as

$$Prob\left(Y=1/\mathbf{X}\right)=F\left(\beta_0+\beta_1X_1+\ldots+\beta_kX_k\right), \tag{1}$$

where $F(\cdot)$ is a logistic cumulative distribution function given by so that $F(u)=\frac{\exp(u)}{1+\exp(u)}$, $0< \ F(u)<1,\ u\in\mathscr{R}$. The response variable Y represents the student's dropout situation and is defined as follows:

$$y_i=\begin{cases}0\ (negative), & \textit{if the student i does not drop out.}\\ 1\ (positive), & \textit{if the student i drop out.}\end{cases} \tag{2}$$

The estimation is done via maximization of the log-likelihood function with the support of non-linear optimization algorithms. For more details, see Hosmer & Lemeshow (2000). The estimation is based on a set $\mathbf{X}=(X_1,\ldots,X_k)$ that only contains variables relevant to the problem that experts previously selected. It is common to encounter problems where the number of variables available for selection is much greater than the number of variables required for the problem. Thus, correctly defining which variables should compose the model is not trivial. This selection can be conducted via variable selection methods, such as shrinkage methods. More details can be found in James et al. (2021) and Hastie et al. (2017).

3.1.1 Variable selection

This work employed regularization methods such as LASSO (*least absolute shrinkage and selection operator*), Ridge, and Elastic-Net. These allow the estimation of the model and the selection of variables simultaneously by inserting a penalty term in the likelihood function. A description of these methods can be found below.

Shrinkage or regularization methods impose penalties on the model's parameters to make the coefficients of the non-relevant explanatory variables close to or equal to zero. Several regularization methods exist, such as LASSO, ridge regression, and elastic net.

The likelihood function of logistic regression with LASSO is given by

$$Max\left\{\sum_{i=1}^{N}\left[y_i\left(\beta_0+\beta'\mathbf{X}_i\right)-log\left(1+e^{\beta_0+\beta'\mathbf{X}_i}\right)\right]-\lambda\sum_{j=1}^{k}|\beta_j|\right\}, \tag{3}$$

where $\beta=[\beta_1,\ldots,\beta_k]$, represents the parameter vector and k indicates the number of available exploratory variables. In this equation, λ $(\lambda\geq 0)$ is the tuning parameter that controls the size of the shrinkage, that is, the strength at which the parameters $\beta_j's$ are forced to zero. High lambda values will shrink to zero more elements of β (those associated with variables that have a small contribution to the objective function). On the other hand, if $\lambda=0$, none of the variables are shrunk to zero, so Equation (3) becomes a simple logit model that uses all exploratory variables.

When $\lambda \neq 0$ the final model is composed of a subset of the most relevant explanatory variables in terms of the likelihood function.

Other regularization methods originate from different penalty terms. For example, ridge regression is obtained using $\lambda \sum_{j=1}^{k} \beta_j^2$. In this case, the final model will comprise all available variables, but the less important ones will have coefficients close to zero. On the other hand, elastic-net can be understood as a convex linear combination between ridge regularization and lasso. Thus, some coefficients will be close to zero, while others will equal zero. Mathematically, the elastic-net regularization is given by $\lambda \sum_{j=1}^{k} \left(\alpha\beta_j^2 + (1-\alpha)|\beta_j|\right)$. For more details, James et al. (2021) and Hastie et al. (2017).

It is worth mentioning that in the regularization methods presented, the value of λ is fundamental and defined beforehand. In practical terms, different values are tested, and the one that presents the best fit is selected.

3.2 Evaluation Metrics

This section introduces the metrics employed to evaluate the classifications. These are based on the confusion matrix, which compares the actual (real classes) and the predicted condition. We may define the number of instances correctly classified with status 0 (true negative) and the number of instances correctly classified with status 1. Two mistakes can be made. The false negative error occurs when the actual status is 1, and the model predicts the instance as 0. The *false positive* error occurs when the model classifies the instance as 1 while the actual condition is 0. Table 1 depicts how a confusion matrix can be displayed.

Table 1 – Confusion matrix.

		Predicted condition (status)	
		0	1
Actual condition (status)	0	*True Negative (TN)*	*False Positive (FP)*
	1	*False Negative (FN)*	*True Positive (TP)*

Some metrics are defined following a confusion matrix. The most common is accuracy, given by the percentage of correct predictions. Mathematically, the accuracy is estimated as (TN + TP)/(TN + TP + FN + TP).

The sensitivity is obtained by TP/(FN+TP). It indicates how many students the model correctly predicts, considering only those who dropped out.

On the other hand, specificity represents how many students the model correctly predicts, considering only those who did not drop out. Formally it is given by TN/(TN + FP).

Precision is the ratio of correct predictions labeled as 1 and the total of predictions labeled as 1. Formally, it is estimated by TP/(TP+FP).

Finally, the F1 metric can be understood as an average between precision and sensitivity. Mathematically it is given by 2TP/(2TP + FP + FN).

We should consider different metrics to evaluate the model since misclassifications may have distinct impacts. For instance, the consequences of a false negative, that is, classifying a student as not likely to abandon when she is a student who will leave the school, have a substantial negative consequence since no intervention would be carried out to reverse this scenario. On the other hand, in a false positive, some intervention would be performed to avoid the abandonment of a student who would not abandon the studies. There may have some costs associated with it, but the educator still would achieve the goal of keeping the student in school.

It is worth mentioning that the results of a logistic regression model are provided in terms of probability instead of a dichotomous variable, $y = \{0,\ 1\}$. For this reason, it is necessary to specify some rule for the classification. Usually, it can be defined as $\widehat{y_i} = 1$, if $Prob\,(y_i = 1|\mathbf{X}) \geq \tau$ and $\widehat{y_i} = 0$, if $Prob\,(y_i = 1|\mathbf{X}) < \tau$. In this case, τ is known as cutoff point. It is not difficult to see that the value of the cutoff impacts the predictive performance. Thus, one can select, for example, the τ that maximizes a predefined metric. For more details, see Hosmer & Lemeshow (2000).

4 CASE STUDIES

4.1 Data

The dataset employed in this research comes from two different sources. The first one is provided by the Espírito Santo State Education Secretariat (SEDU) via the State's system of Educational Management of Espírito Santo (SEGES system) and is composed of individual information of the students enrolled at the state's public schools. The second data source originates from the National Institute of Educational Studies and Research Anísio Teixeira (INEP).

By merging these two datasets, we can have interesting data on the student's level. The final dataset consists of age; sex; class shift (morning, afternoon, night, full-time); average school grade in mathematics; average school grade in Portuguese; average grade of the class in mathematics; average grade of the class in Portuguese; student grades in Portuguese, mathematics, history, geography, physics, chemistry, and biology; the proportion of absences in Portuguese, mathematics, history, geography, physics, chemistry and biology; color/race; *dummies* that indicate whether the student has changed the class shift, the class, the school or the city; how many changes of class shift, classes, schools and cities the student made and; the number of schools, cities, shift class and class the student attended in the year.

In this paper, we analyzed the first high school grade in the state's public schools of Espírito Santo. The time coverage of the data is 2019-2022. The codes were implemented in R using the package caret, Kuhn (2022).

4.2 Results

We carried out three case studies in a cross-section framework. First, we consider only the year 2019, both for the estimation and for the model validation. Thus, we split the data into two sub-datasets, one for estimation and another for validation. We estimate the model exclusively with data available in the year's first quarter and try to identify dropouts throughout 2019.

The second case study is more realistic, as we considered the model estimated in 2019 and looked at its performance to predict dropout in 2020. This is how this tool can be employed in real situations. In this case, new data containing personal information regarding student performance are available immediately after the end of the year's first quarter. These data feed the previously estimated model to identify those students likely to drop out the school by the end of the year.

In the third case study, the focus is forecasting dropout for 2022. For this task, we calibrated the model with all data available at the end of 2021, that is, 2019, 2020, and 2021. Our aim is to demonstrate that information gathered in previous years can be effectively utilized for predicting attrition in future years, as we exclusively employ variables that are not time-dependent.

Three distinct specifications were taken in consideration. All models were constructed via regularization methods (lasso, ridge, and elastic-net), while dummies for schools were available for selection. The objective is to control for unobserved heterogeneity among schools, including characteristics of the schools or their neighborhoods, such as infrastructure and violence, which could impact students' decisions to discontinue their studies. Subsequently, we present the results for the three case studies.

4.2.1 Case 1 – First grade of high school in 2019

In the first case study, we considered only the information related to the first quarter of 2019 to verify the dropout throughout the corresponding year. We separated the database into two sets: the training set has approximately 85%, and the validation set has 15% of the sample. The training sample is used for estimation, whereas the validation set is employed to verify the model's performance. It is important to note that the historical dropout rate is preserved in these two samples.

The performance metrics and the τ (cut-off) were defined via *10-fold cross-validation.* Moreover, this procedure is also crucial for determining the regularization constants α and λ.

Regarding the regularization methods, the estimated parameters were $\lambda = 0.001$ for Lasso, $\lambda = 0.0081$ for ridge and $\alpha = 0.1$, and $\lambda = 0.00005$ for elastic-net. As already mentioned, such hyper-parameters were obtained through cross-validation by maximizing the area under the receiver operating characteristic curve.

Table 2 summarizes the explanatory variables. The ”+” sign indicates that an increase in the variable increases the dropout probability. On the other hand, the symbol ”-” means the opposite. Besides that, blank spaces indicate that the model does not employ the variable.

Table 2 – Relationship between variables and dropout probability.

	Model 1	**Model 2**	**Model 3**
Age	+	+	+
Male		+	+
School average score (Portuguese)		-	-
School average score (Mathematics)		-	-
Afternoon school shift		+	+
Evening school shift	+	+	+
Full-time school shift	-	-	-
Class average score (Portuguese)		+	+
Class average score (Mathematics)		-	-
Student’s score (Portuguese)	-	-	-
Student's score (Mathematics)	-	-	-
Student’s score (History)		+	+
Student’s score (Geography)	-	-	-
Student’s score (Biology)	-	-	-
Student's score (Chemistry)	-	-	-
Student’s score (Physics)	-	-	-
Proportion of absences (Portuguese)	+	+	+
Proportion of absences (Mathematics)	+	+	+
Proportion of absences (History)	+	+	+
Proportion of absences (Geography)	+	+	+
Proportion of absences (Biology)	+	+	+
Proportion of absences (Chemistry)	+	+	+
Proportion of absences (Physical)		-	-
White race	-	-	-
Indigenous race	+	+	+
Undeclared race		+	+
Multiracial race		+	+
Black race		+	+
Did the student change schools?		-	-
How many times did the student change schools?		-	-
How many schools did the student attend?		-	-
Did the student change cities?		+	+
How many times did the student change cities?	-	-	-
In how many cities did the student live?			
Did the student change classes?		+	+
How many times did the student change classes?		+	+
How many classes did the student attend?		+	+
Did the student change shifts?		-	+
How many times did the student change shifts?		+	-
How many shifts did the student attend?		+	-
School Fixed Effects	Partial	Yes	Yes

Source: Own Elaboration.

Analyzing the selected variables, we can see that age appears in all models with a positive sign. The older the student, the greater the probability of dropping out. The students' average grades were also selected in all models. The higher the student's grade, the lower the chances of dropping out.

The proportion of absences was widely used for the models. As it can be seen, an increase in absences increases the chance of dropout. Variables related to race/color were also relevant to predictions. Overall, the probability of abandonment is higher in non-white races.

In general, these results establish the profile of the dropout. In other words, students older than expected for their grades, low grades on exams, high absences are susceptible to dropping out.

It is important to emphasize that we cannot infer a cause-and-effect relationship from Table 2. The proposed model is not a causal inference model. For example, the fact that we find a negative sign for the student's score variable does not mean that we can reduce dropout simply by artificially increasing the student's grade. The lower score could reflect family and personal issues that demotivate the student from studying, negatively impacting the grade and ultimately resulting in dropout.

It is worth noting that models 2 and 3 should be analyzed carefully. Model 2, for example, does not eliminate any variables, even if they are irrelevant to the problem. In this case, the regularization method approximates the values of the irrelevant coefficients to zero. As mentioned, Model 3 (elastic-net) is a balance between lasso and ridge regression. So, it eliminates some irrelevant variables and sets other irrelevant coefficients close to zero. Model 1 (lasso) is the only one that strictly selects variables, discarding all the irrelevant ones.

In terms of predictive performance, Tables 3 and 4 present the metrics for the training and validation sets. The cut-offs varied from 0.0251 to 0.0357 and were obtained through 10-fold cross-validation.

Table 3 – Performance in the training set - 2019.

Specification	Accuracy	Sensitivity	Specificity	Precision	F1
Model 1	0.866	0.865	0.866	0.154	0.261
Model 2	0.882	0.882	0.882	0.172	0.288
Model 3	0.890	0.890	0.890	0.186	0.307

Source: Own Elaboration.

Table 4 – Performance in the test set - 2019.

Specification	Accuracy	Sensitivity	Specificity	Precision	F1
Model 1	0.866	0.800	0.867	0.142	0.241
Model 2	0.876	0.807	0.878	0.154	0.259
Model 3	0.886	0.885	0.886	0.176	0.293

Source: Own Elaboration.

In both the training and validation sets, the three evaluated models perform quite similarly, with Model 1 being slightly inferior. Overall, the metrics obtained by the models are in line with the accuracies observed in the literature.

The accuracy found in both datasets varies approximately from 0.86 to 0.89. This means that more than 86% of the students are correctly identified. The rate of correct predictions of students who drop out (*sensitivity*) ranges from approximately 0.80 to 0.89. In contrast, the proportion of correct classifications of the students who do not leave the school varies from 0.86 to 0.89. The results indicate many corrected identification cases; however, the precision found indicates a considerable level of false positive cases.

4.2.2 Case 2 – First grade of high school in 2020

The second case study uses the model calibrated with the data from the first quarter of 2019 to predict the situation of school abandonment throughout 2020. We considered 273 public schools with 34.954 students. Nine hundred thirty-seven (937) students dropped out throughout 2019, while 1207 abandoned the school during 2020.

This is how this kind of tool can be applied to real problems for educators. The model must have been calibrated beforehand at the beginning of the year. For this task, we may employ the data from the previous years. Thus, as soon as the year's first quarter ends, we can feed the calibrated model with new data. In other words, in this simulation exercise, we have estimated the model with data from 2019. At the end of the first quarter of 2020 (when new data is available), we predict the students likely to drop out during the whole year of 2020.

Table 5 display the results. Overall, all models exhibit similar performance, and the choice of the best model depends on the metric selected by the user. Sensitivity is particularly crucial in this context, and according to this metric, Model 1 emerges as the best. However, in terms of F1 score, a commonly metric that balances precision and sensitivity, Model 2 achieves the highest performance.

Table 5 – Performance of the models for the year 2020.

Specification	Accuracy	Sensitivity	Specificity	Precision	F1
Model 1	0.846	0.843	0.846	0.163	0.274
Model 2	0.879	0.778	0.883	0.191	0.307
Model 3	0.879	0.621	0.889	0.166	0.262

Source: Own Elaboration.

4.2.3 Case 3 – First grade of high school in 2022

To forecast student dropout for 2022, we employed data spanning from 2019 to 2021 for model estimation, and subsequently, we projected the abandon for the entire year of 2022. This exercise illustrates that previously collected data from preceding years can be effectively utilized for predicting future trends. In this context, it is even possible to enhance the training dataset by incorporating additional years, contingent upon data availability.

The training set comprises 249 schools with total of 90,099 students, of which 2,584 left the school. For the year 2022, there are 22,014 students, with 517 dropouts.

Table 6 presents the metrics for the third case study. As evident, all three methods exhibit remarkably similar performance, with no significant differences in any of the considered metrics. Once again, these findings align with other dropout models documented in the literature and correspond with the results obtained in the case studies conducted in this paper.

Table 6 – Performance of the models for the year 2022.

Specification	Accuracy	Sensitivity	Specificity	Precision	F1
Model 1	0,826	0,822	0,826	0,102	0,182
Model 2	0,827	0,818	0,827	0,102	0,181
Model 3	0,826	0,816	0,826	0,101	0,180

Source: Own Elaboration.

5 CONCLUSIONS

Education plays a crucial role in our society, but unfortunately, it remains inaccessible to everyone in developing countries. A prevalent issue in this context is the dropout rate, which refers to students leaving school before completing the academic cycle.

In this context, we utilized private student information provided by SEGES and INEP and applied statistical learning methods to develop an early warning system. It is important to note that this data is confidential, although educational authorities in Brazilian states may have access to it.

Our approach involved employing logistic regression with different regularization methods. Inspired by Adelman et al. (2018), we also explored the potential improvement of predictive performance by incorporating school-fixed effects.

Three case studies were conducted. The first one uses only data regarding 2019, so we trained and validated the model with two subsamples of this year. The second case study used the model calibrated with the data from the first quarter of 2019 (case study one) to predict the situation of abandonment throughout 2020. Finally, in the third case study, we assessed the model calibrated with data from 2019 to 2021 to predict the student's abandon for the year 2022. These last two case studies illustrate how this tool can be applied in real-world situations.

Comparing the performance of our models with those found in the literature, we confirmed the relevance of the information contained in the SEGES and INPE datasets for predicting dropout. Additionally, our methodologies effectively identified students at risk of dropping out. It is of utmost importance to highlight that our approach allows us to calibrate the model using data from different previous years. Furthermore, there is the potential to enlarge the training dataset by integrating data from earlier years.

Regarding variable selection methods, the results indicated that all tested methods perform similarly, making it impossible to consistently establish the best one. Notably, lasso regularization produced models with fewer parameters, allowing for easier interpretation of the results. The findings also underscore the advantages of sophisticated selection methods. This becomes even more attractive if more information from other databases is added to the tool. Moreover, using automatic selection methods, the computational tool can easily estimate different models for different grades, expanding the tool's use in all grades. For future studies, exploring alternative machine learning models and addressing the challenge of imbalanced data is recommended, as the proportion of non-dropout students is significantly higher. This imbalance may introduce biases that could affect the overall performance of the models.

Acknowledgements

The authors would like to thank the Department of Education of the State of Espírito Santo (SEDU-ES) and the Espírito Santo Research and Innovation Support Foundation (FAPES) for the financial support. The authors would also like to thank SEDU-ES for providing the data.

References

ADELMAN M, HAIMOVICH F, HAM A & VAZQUEZ E. 2018. Predicting school dropout with administrative data: New evidence from Guatemala and Honduras. *Education Economics*, **26**(4): 356–372.

ALVES A, MELLO J, RAMOS T & SANT'ANNA A. 2011. Logit models for the probability of winning football games. *Pesquisa Operacional*, **31**(3).

BARROS T, NETO P, SILVA I & GUEDES L. 2019. Predictive models for imbalanced data: a school dropout perspective. *Education Sciences*, **9**(4).

BAYER J, BYDZOVSKÁ H, GÉRYK J, OBSÍVAC T & POPELÍNSKÝ L. 2012. Predicting dropout from social behaviour of students. In: *International Conference on Educational Data Mining (EDM*. p. 7. Chania, Greece: EDM.

BAÑERES D, RODRÍGUEZ GONZÁLEZ M, GUERRERO ROLDÁN A & CORTADAS P. 2023. An early warning system to identify and intervene online dropout learners. *International Journal of Educational Technology in Higher Education*, **20**: 3.

BURGESS S. 2016. Human capital and education: The state of arrt in the economics of education.

COSTA E, FONSECA B, SANTANA M, ARAÚJO F & REGO J. 2017. 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.

CUNHA J, MOURA E & ANALIDE C. 2016. Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. In: CORREIRA A, ADELI H, REIS L, TEIXEIRA A & ROCHA M (Eds.), *New Advances in Information Systems and Technologies*. p. 773. Switzerland: Springer International Publishing.

DEKKER G, PECHENIZKIY M & VLEESHOUWERS J. 2009. Predicting Students Drop Out: A Case Study. In: *International Conference on Educational Data Mining (EDM*. p. 10. Cordoba, Espanha: EDM.

DOL S & JAWANDHIYA P. 2023. Classification Technique and its combination with clustering and associatino rule mining in educational data mining - A survey. *Engineering applications of artificial intelligence*, **122**.

FERREIRA P, LOUZADA F & DINIZ C. 2013. Credit scoring modeling with state-dependent sample selection: a comparision study with usual logistic modeling. *Pesquisa Operacional*, **35**(1).

GIUBERTI A, RANGEL L, CASTRO M, SANTOS M, VAZZOLER M, FRANCO S, FERREIRA T, OLIVEIRA T & GOMES V. 2018. *Modelo de predição do abandono escolar*. vol. 58. Instituto Santos dos Jones Neves, Vitória, Espírito Santo, Brazil.

HASTIE T, TIBSHIRANI R & FRIEDMAN J. 2017. The elements of statistical learning: data mining, inference, and prediction.

HOSMER D & LEMESHOW S. 2000. *Applied logistic regression*. John Wiley & Sons.

HU Y, ZHONG Z, WANG R, LIU H, TAN Z & ZHENG W. 2021. Data Augmentation in Logit Space for Medical Image Classification with Limited Training Data. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2021*. p. 469–479.

IBGE (BRAZILIAN INSTITUTE OF GEOGRAPHY AND STATISTICS). 2020. Cartilha Educação. Accessed 05/10/2022. Available at: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101736_informativo.pdf.

JAMES G, WITTEN D, HASTIE T & TIBSHIRANI R. 2021. *An introduction to statistical learning with applications in R*. Second ed. Springer texts in statistics. Springer.

JIMÉNEZ-GÓMEZ M, LUNA J, ROMERO C & VENTURA S. 2015. Discovering clues to avoid middle school failure at early stages. In: *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*. p. 300–304.

KIM S, CHOI E, JUN Y & LEE S. 2023. Student Dropout Prediction for University with High Precision. *Applied Sciences*, **13**(10).

KNOWLES J. 2015. Of needles and haystacks: Building and accurate statewide dropout early warnings system in Wisconsin. 7(3),18-67. *Journal of Educational Data Mining*, p. 18–67.

KRÜGER J, BRITTO JR A & BARDDAL J. 2023. An explainable machine learning approach for student dropout prediction. *Expert Systems With Applications*, **233**.

KUHN M. 2022. Caret: Classification and regression training, R package version 6.0-93. Available at: https://CRAN.R-project.org/package=caret.

LAMB S & RICE S. 2008. *Effective strategies to increase school completion report: Repost to the Victorian department of educatino and early childhood development*. Communications Division, Department of Education and Early Childhood Development.

MARTINHO V, NUNES C & MINUSSI C. 2013. An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. In: *2013 IEEE 25th International Conference on Tools with Artificial Intelligence. IEEE*. p. 159–166.

MINISTRY OF EDUCATION - PERU. 2020. Accessed 11/8/2023. Available at: https://www.gob.pe/institucion/minedu/noticias/306531-minedu-implementa-alerta-escuela-un-sistema-de-alerta-temprana-para-identificar-estudiantes-con-riesgo-de-abandonar-el-sistema-educativo.

MNYAWAMI Y, MAZIKU H & MUSHI J. 2022. Enhanced Model for Predicting Student Dropouts in Developing Countries Using AutomatedMachine Learning Approach: A Case of Tanzanian's Secondary Schools. *Applied Artificial Intelligence*, **36**.

MÁRQUEZ-VERA C, CANO A, ROMERO C, MOHAMMAD A, FARDOUN H & VENTURA S. 2016. Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, p. 107–124.

NASCIMENTO R, FAGUNDES R & SOUZA R. 2022. Statistical Learning for Predicting School Dropout in Elementary Education: A Comparative Study. *Annals of Data Science*, **9**: 801–828.

NASCIMENTO R, JUNIOR R & NETO MAA ND FAGUNDES R. 2018. Educational data mining: An application of regressors in predicting school dropout. In: PERNER P (Ed.), *Lecture Notes in Computer Science*. p. 246–257. Springer.

NAZISH S, SALAM A, ULLAH W & IMAD M. 2021. COVID-19 Lung Image Classification Based on Logistic Regression and Support Vector Machine. In: *Artificial Intelligence Systems and the Internet of Things in the Digital Era*.

NG E. 2012. Forecasting US recessions with various risk factors and dynamic probit models. *Journal of Macroeconomics*, **34**(1): 112–125.

NIYOGISUBIZO J, LIAO L, NZIYUMVA E, MURWANASHYAKA E & NSHIMYUMUKIZA E. 2022. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, **3**.

OQAIDI K, AOUHASSI S & MANSOURI K. 2022. Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms. *International Journal of Emerging Technologies in Learning*, **18**.

PARR A & BONITZ V. 2015. Role of family background, student behaviors, and school-related beliefs in predicting high school dropout. *The Journal of Educational Research*, **108**(6): 504–514.

PSACHAROPOULOS G & PATRINOS H. 2018. Returns to investment in education. A decennial review of the global literature.

RODRÍGUEZ P, VILLANUEVA A, DOMBROVSKAIA L & VALENZUELA J. 2023. A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. *Education and Information Technologies*, **28**: 10103–10149.

ROVIRA S, PUERTAS E & IGUAL L. 2017. Data-driven system to predict academic grades and dropout. *PLoS one*, **12**(2).

SANDOVAL-PALIS I, NARANJO D, VIDAL J & GILAR-CORBI R. 2020. Early Dropout Prediction Model: A Case Study of University Leveling Course Students. *Sustainability*, **12**(22).

SARA N, HALLAND R, IGEL C & ALSTRUP S. 2015. High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. In: *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence*. Bruges, Belgium.

SELAU L & RIBEIRO J. 2012. A systematic approach to construct credit risk forecast models. *Pesquisa Operacional*, **31**(1).

SLETTEN M, TOGE A & MALMBERG-HEIMONEM I. 2022. Effects of an early warning system on student absence and completion in Norwegian upper secondary schools: a cluster-randomised study. *Scandinavian Journal of Educational Research*, .

SOARES T & MENDONÇA M. 2003. Construção de um modelo de regressão hierárquico. *Pesquisa Operacional*, **23**(3).

ULDALL J & ROJAS C. 2022. An application of machine learning in public policy early warning prediction of school dropout in the chilean public education system. *Multidisciplinary Business Review*, **15**(1): 20–35.

VISBAL-CADAVID D, MENDOZA A & HOYOS I. 2019. Prediction of efficiency in colombian higher educatino institutions with data envelopment analysis and neural networks. *Pesquisa Operacional*, **39**(2).

WOOD L, KIPERMAN S, ESCH R, LEROUX A & TRUSCOTT S. 2017. Predicting dropout using student-and school-level factors: An ecological perspective. *School Psychology Quarterly*, **32**(1): 35–49.

How to cite

PEREIRA GAA, DEMURA KD, NUNES I, PAULA KC & LIRA PS. 2024. An early warning system for school dropout in the state of Espírito Santo: a machine learning approach with variable selection methods. *Pesquisa Operacional*, **44**: e275092. doi: 10.1590/0101-7438.2023.043.00275092.