Education and Information Technologies (2021) 26:1527–1547
https://doi.org/10.1007/s10639-020-10316-y

A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach

Ricardo Costa-Mendes[1] · Tiago Oliveira[1] · Mauro Castelli[1] · Frederico Cruz-Jesus[1]

Received: 29 February 2020 / Accepted: 27 August 2020 / Published online: 5 September 2020

Abstract

This article uses an anonymous 2014–15 school year dataset from the Directorate-General for Statistics of Education and Science (DGEEC) of the Portuguese Ministry of Education as a means to carry out a predictive power comparison between the classic multilinear regression model and a chosen set of machine learning algorithms. A multilinear regression model is used in parallel with random forest, support vector machine, artificial neural network and extreme gradient boosting machine stacking ensemble implementations. Designing a hybrid analysis is intended where classical statistical analysis and artificial intelligence algorithms are blended to augment the ability to retain valuable conclusions and well-supported results. The machine learning algorithms attain a higher level of predictive ability. In addition, the stacking appropriateness increases as the base learner output correlation matrix determinant increases and the random forest feature importance empirical distributions are correlated with the structure of *p*-values and the statistical significance test ascertains of the multiple linear model. An information system that supports the nationwide education system should be designed and further structured to collect meaningful and precise data about the full range of academic achievement antecedents. The article concludes that no evidence is found in favour of smaller classes.

Keywords Machine learning · Stacking · Random forest · Support vector regression · Academic achievement · High school grades

✉ Ricardo Costa-Mendes
rmendes@novaims.unl.pt

Extended author information available on the last page of the article

1 Introduction

A nation's wealth, interconnected with the availability of human capital in its economy, hinges on its citizens' academic achievement and generally on the education system attainment level (Becker 1964; Hanushek and Wößmann 2010; Strenze 2007). Knowing the determinants of academic success in detail is an essential cornerstone in the pursuit of appropriate public policy designs. The ability to predict and anticipate student academic grades would enable policymakers, principals, and teachers to take timely action on preventing unfavourable results and provide a readily available solid conceptual framework, capable of feeding sound decision support systems (van der Scheer and Visscher 2018). The development and management of a nationwide schooling and education system database which brings together relevant information on the determinants of academic achievement is an investment that requires attention to the complexity of data collection and management at the school-teacher-student trinomial level. Still, it is an indispensable step in promoting conceptually well-designed policies which would confer an enhanced predictive ability to stakeholders.

Both in the scientific literature and practical institutional roles, academic performance prediction and the conceptual approach to its determinants have been massively based on the application of classical methods of statistical analysis, such as structural equation, multilinear, and panel data regression models. However, as machine learning and artificial intelligence algorithms show superior predictive ability, it is of great interest and pertinence to develop and deploy methodologies and methodological leads that intertwine approaches and simultaneously make it possible to seize the parametric nature of classical methods and the essentially empirical and predictive nature of machine learning algorithms.

Following this line of thought and research, the present study aims to predict high school academic scores by applying both frameworks simultaneously. A multilinear regression model is used to represent the classical approach. Random forest, support vector machine, artificial neural network, and extreme gradient boosting machine stacking ensemble models are implemented to constitute the machine learning outlook. Specifically, designing a hybrid analysis is intended where classical statistical analysis and artificial intelligence algorithms would blend and augment the ability to infer valuable conclusions. The answers to the following research questions, in particular, are sought:

1) Is machine learning algorithms' prediction accuracy superior to multilinear regression?
2) Are there any conclusions we can retain from the hybrid approach about the academic achievement conceptual framework?
3) Is there any empirical relationship between the classical conceptual analysis of the multilinear regression model and the feature importance empirical distributions of the random forest implementation?

In this sense, the study begins with an academic achievement literature review, followed by a presentation of the methodology and algorithms. The results and discussion sections come next, and finally, we present the conclusions.

2 Literature review

Research in academic achievement (AA) has not lost interest and relevance since the publication of “The Coleman Report” in 1966 (Coleman 1966). The report’s main statement was that the family background of students is a decisive AA determinant, conferring a secondary role to the resources and characteristics of schools and teachers.

Student cognitive ability has been undoubtedly identified as the major and fundamental determinant of AA as there is strong empirical evidence that confirms their generally accepted association (Jensen 1998). Indeed, 25 to 49% of AA variance is bound to be explained by differences in cognitive ability (Rohde and Thompson 2007). However, other student characteristics also exert their influence. Some personality traits, such as self-organization, attentiveness, perseverance, and focus on results, are associated with overachievement (Di Fabio and Busoni 2007). Females tend to have higher AA. Empirical evidence shows that the gender performance gap is wider in languages and literature and narrower in mathematics (Francis 2005; Lupart et al. 2004; Mensah and Kiernan 2010). Males seem to develop a more negative peer attitude relative to school (King 2016). There is also a non-neglectable performance gap between different ethnicities (Kuhfeld et al. 2018) and immigrant-origin countries (Levels et al. 2008; Perreira et al. 2006). Computer usage and internet access can have a positive effect on academic performance provided it is mindful and primarily directed towards learning support activities (Lei and Zhao 2007; Salomon and Ben-David Kolikant 2016). Nevertheless, in case its main target is leisure pursuits, it can be counterproductive (Kubey et al. 2001).

Some family background variables are also salient determinants of AA. Parental school involvement has a noticeable relationship with achievement (Fan and Chen 2001; Gilar-Corbi et al. 2019). Parental educational involvement enhances students’ ability to cope with schooling activities and promotes appropriate behavioural attitudes that lead to success (Hill and Taylor 2004). More importantly, participating in school activities seems to be more decisive for lower socioeconomic status parents (Benner et al. 2016), whom themselves are associated with lower scholarship outcomes (Sirin 2005). Indeed, academic performance is positively linked to parental income, education level, and type of occupation (Steinmayr et al. 2010; Tesfagiorgis et al. 2020; Tomul and Savasci 2012).

Smaller schools seem to benefit both students with a lower economic and social status and students with a history of learning problems (Leithwood and Jantzi 2009). Class size seems to have a somehow blurred effect on achievement. The conclusions of Hoxby (2000) that class size does not influence AA contradict the findings of Krueger (1999) that performance increases in smaller classes, especially for students from ethnic minorities and lower social-economic statuses. In addition, the conclusions of Wößmann and West (2006) indicate that the effect of class size depends both on the educational system as a whole and on teachers’ general lecturing capabilities. Design quality and adequacy of school facilities in terms of environmental comfort can determine AA at least through students and teachers enhanced willingness to perform (Schneider 2002). In the same manner, physical design and spatial configuration matters and should meet users’ expectations (Woolner et al. 2007). Teacher quality, measured by teacher panel data fixed effects on student outcomes, has a positive relationship with AA (Aaronson et al. 2007; Rockoff 2004). Despite the low percentage of teacher quality variance explained by teacher education and experience (Rivkin et al. 2005), higher teacher college rankings and scores appear to be associated with higher academic outcomes (Wayne and Youngs 2003).

There are some studies in the literature that use machine learning in the context of AA. On a dataset of 110,267 high school students, Cruz-Jesus et al. (2020) applied artificial neural network, decision tree, extremely randomized trees, random forest, support vector machine, k-nearest neighbours and logistic regression classifiers to anticipate high school retentions. Miguéis et al. (2018) using a dataset of 2459 higher education students employed naïve Bayes, support vector machine, decision tree, random forest, bagged trees and adaptive boosting trees classifiers to address an academic achievement five classes' problem. Musso et al. (2020) called on a 655 university students' dataset and an artificial neural network to deal with a problem of classification between low and high levels of three different measures of AA. Mengash (2020) made use of artificial neural network, decision tree, support vector machine and a naïve Bayes classifiers to anticipate five classes of higher education AA from a sample of 2039 students in order to evaluate the admission criteria of a Saudi University. Sorensen (2019) collected a sample of 220,685 students from the North Carolina Department of Public Instruction and addressed a dropout classification problem with decision tree and support vector machine classifiers. Above all, it can be highlighted that the literature in the AA scientific domain tends to prefer to tackle classification instead of regression problems presumably as a result of the former having simpler probability functions and a much easier bias to handle.

3 Methodology

3.1 Learning algorithms

Supervised learning consists in estimating a mathematical function that maps a vector of predictor variables – feature space - to a vector of target or response variables through learning from a set of training data. The learning phase perdures until the function approximation is sufficiently accurate to produce sensible predictions. The target variables can either be binary - a classification or pattern recognition - or continuous - a regression (Murphy 2012).

In our case, the dataset was split into training and test datasets, 70% and 30% of the samples, respectively. The test dataset is a holdout dataset as it had no rule in the learning phase and was only used for testing and evaluating the generalisation performance of the applied algorithms. The training input variables vectors were standardised, and the result transformations were applied to the test dataset.

Before the learning phase and whenever applicable, a hyperparameter optimisation procedure was carried out. A hyperparameter subspace was built from several vectors of hyperparameters values. A five-fold cross-validation (Hastie et al. 2008; Mohri et al. 2018) grid search (Bergstra and Bengio 2012; Nievergelt 2000) was performed, and the best average cross-validation estimator score was elected. This blended method attempts to manage the bias and variance trade-off of the model generalisation performance (Briscoe and Feldman 2011; Hastie et al. 2008) and is a well-known method in the machine learning community. The randomness of the cross-validation procedure allows the choice of a hyperparameter combination that has a lower variance, decreasing the risk of overfitting to the training data. The grid search aims at ensuring a reduced bias through overcoming underfitting issues on the training data.

Follows the regression algorithms that were chosen to perform the prediction of our target variable - the student high school final grades.

3.1.1 Multiple linear regression

In a multiple linear regression (MLR), a linear mathematical function is assumed between the predictors and the target variable in an additive fashion (James et al. 2013). In the learning phase, the parameters are estimated using the ordinary least squares method (OLS). The multiple linear regression is a classical statistical model that has somehow strong assumptions. It imposes a linear relationship between response and predictors and assumes the sphericity of the error term when interpreting and statistically testing the regression coefficients β_i. This is a parametric approach with more emphasis on conceptual ascertainment as in augmenting the predictive accuracy.

3.1.2 Random forest

Random forest (Breiman 2001) is a machine learning ensemble method of randomised decision trees. Random forest is an ensemble method as the outcome is derived from the various decision tree scores produced by bagging or bootstrapping subsampling (Breiman 1996a). In the case of regression, the random forest outcome is the averages of the randomised decision trees scores. A decision tree is a machine learning algorithm that splits the predictor variables space sequentially in a set of partitions and sub-partitions to form homogenous classes in terms of target variables. In the case of regression, the split flow is interrupted when a further sub-partition is considered to non-significantly decrease the mean square error of the target variables. The final nodes of the tree are called leaves and consubstantiate the decision rules on which the target variable predictions are based. In a randomized decision tree, the search for the best split in each node is conducted through a random variable selection (Amit and Geman 1997).

The hyperparameter optimisation procedure included the number of trees in the forest, the minimum number of samples required to be at a leaf and the minimum number of samples required to split an internal node. For full reference, take notice of the scikit-learn python module that was used in the study (Pedregosa et al. 2011).

3.1.3 Support vector regression

The support vector regression algorithm (SVR) consists of trying to find the flattest mathematical function of the predictor variables whose deviation from the target is less than $\varepsilon \in \mathbb{R}^+$ for all the training data (Smola and Scholkopf 2004). That function is the backbone of a tube whose distance to both margins is $\varepsilon \in \mathbb{R}^+$. In contrast with the hard margin, the soft margin hyperplane SVR allows a deviation beyond $\varepsilon \in \mathbb{R}^+$ through the introduction of slack variables $\xi \geq 0$. For the primal form of the optimization problem for SVR see, e.g., Mohri et al. (2018).

The gaussian radial basis function (RBF) kernel was used, adding some nonlinearity and flexibility to the model. In this case, as the feature space dimension is infinite, it is not feasible to solve the optimisation problem through the primal form. Yet the dual form derived from the Lagrange multipliers method can be applied (Rivas-Perea et al. 2013).

The hyperparameter optimisation procedure included the hyperparameter C, a penalisation factor for the points placed outside the ε-tube, and γ that defines the radius of influence of the support vectors in the RBF kernel function.

The scikit-learn module follows the libsvm implementation (Chang and Lin 2007). As the fit time complexity is more than quadratic with the number of samples, a subsampling without replacement procedure – pasting – was undertaken for the larger training datasets.

3.1.4 Multilayer feed-forward neural network

A multilayer feed-forward neural network is a multilayer perceptron. The perceptron architecture is the simplest form of a neural network. It is intended to solve linearly separable binary pattern classification problems through a nonlinear activate function – the threshold - which input is a linear combination of the predictor variables and a bias (Rosenblatt 1958). The perceptron just has an input layer and an output layer. The multilayer perceptron has hidden layers between them. The introduction of hidden layers in the perceptron architecture is bound to augment the ability to solve nonlinear problems. A fully connected multilayer feed-forward neural network, as is the case, is a multilayer perceptron, where each neuron in the input layer connects to every neuron in the first hidden layer, which in turn connects to every neuron of the next hidden or output layer. The input neurons collect the predictor variables values and send a signal to every neuron in the first hidden layer. The inputs of the first hidden layer neurons are linear combinations of the signals received plus a bias. The first hidden layer is the input layer of the second hidden layer, and the feed-forward process goes on up to the output layer. The nonlinear activation function of the hidden layer neurons, and of the output layer neurons in a classification problem, allow the neural network to learn and approximate complex functions between the predictor and the target variables. Indeed, neural networks are known as universal function approximators due to both the nonlinearity of the activation function and the existence of hidden layers (see,e.g., Basheer and Hajmeer 2000; Haykin 2009; Ramchoun et al. 2016).

The architecture has two hidden layers. The logistic activation function was used in every hidden neuron. However, the output neuron has no activation function. The feed-forward error-backpropagation algorithm was employed to search the connection weights and biases that minimise the square error loss function. It is a gradient descent optimisation algorithm that uses the derivative chain rule to compute the derivatives of the loss function with respect to each weight and bias in the network. During the learning phase, the weights and the biases change iteratively in the direction that minimises the cost function. As the cost function of a multilayer neural network with nonlinear activation functions is non-convex, the final solution is certain to be a local minimum with an expected good generalisation performance (Choromanska et al. 2015; Rumelhart et al. 1986).

The hyperparameter optimisation procedure included the hyperparameters hidden_layer_sizes, activation, alpha (L2 penalty) and learning_rate_init as defined by the scikit-learn documentation.

3.1.5 Extreme gradient boosting machine

Boosting is an ensemble method that improves the accuracy of any given machine learning algorithm. Its primary attention is to reduce bias in the learning phase. It consists in combining weak learners sequentially in order to build a strong ensemble learner (Bishop 2006). A weak learner is a function approximator machine learning algorithm whose architecture is simple and whose accuracy is slightly better than random guessing, in case of binary classification, or than a flat function, in case of regression. During the learning phase and in every boosting iteration, the distribution or the training data itself is changed in a manner that allows the next weak learner to primarily focus on the largest predictor error samples. The strong learner is built upon the weak learners by a majority voting scheme in case of classification or average and summation in regression. Normally the weak learners are weighted by accuracy (Schapire 2003).

Gradient boosting machine (Friedman 2001; Hastie et al. 2008) is an additive training ensemble algorithm that fits several weak learners sequentially on the last boosting iteration residual errors. In our case, the weak learner is a regression decision tree, and the objective function is the square loss.

The extreme gradient boosting machine (Chen and Guestrin 2016) is a regularised version of the gradient boosting machine. It introduces a penalisation to the complexity of the trees in the machine, defining it as the number of leaves and the squares of their scores.

In the stacking ensemble learning technique, the base level models are trained through a complete training set, and the meta-model is trained on the outputs of the base level models (Wolpert 1992). The aim is to improve generalisation performance (Breiman 1996b). In our case, the different regression algorithms were combined via an extreme gradient boosting machine. The input matrix of the regression tree extreme gradient boosting machine comprises the outputs of the other models except for the stand-alone SVR algorithm. Before learning, the input matrix underwent a principal components analysis (PCA) orthogonal transformation.

The hyperparameter optimisation procedure included the hyperparameters maximum tree depth for weak learners, boosting learning rate, number of trees to fit, L2 regularisation term on weights as defined by the scikit-learn documentation.

3.2 Feature selection

The feature selection consists in reducing the number of predictive variables that are used to learn a function that approximates a given response variable. The objective is to ease the model interpretability, reduce the complexity, enhance the computational efficiency and convergence of the algorithms, and at last to avoid overfitting.

In our case, the feature selection was undertaken by the lasso method (Tibshirani 1997). The method adds an L1 norm regularisation to the multilinear regression model objective function, forcing the parameters estimates of those predictive variables that are not sufficiently strong to converge to zero.

A search grid cross-validation procedure was carried out to find the shrinkage pressure that produces the best average score. The model whose score is immediately higher than the best mean score minus its standard deviation is chosen to lead the variable selection. Each variable with a null $\widehat{\beta}$ was dropped.

4 Data and results

4.1 Data

In the experimental phase, an anonymous 2014–15 school year dataset from the Information Systems Integration Mission Unit database, that supports the Directorate-General for Statistics of Education and Science (DGEEC) of the Portuguese Ministry of Education information system, was used. Its purpose is to centralize all educational data collection from pre-school, primary, and high school, as well as provide the respective institutes with the necessary information that will serve as the basis for the production of educational statistics to be used in decision-making processes. From the database, data from students and schools were collected through Microsoft ® SQL Server Management Studio queries. Students' residence area data such as urban, income, aging, employment, and cultural level indicators were added from Statistics Portugal to gather information on their socioeconomic background. Only broad attendance subjects were considered. The final teachers' high school grades per subject that were predicted are shown in bold in Table 1.

Table 1 Dataset disaggregation

		Samples		
Subject	Quantitative & qualitative subjects	n	Outlier and incomplete records	Subtotal
English	Qual	**54,885**	**20**	**54,905**
Mathematics	Quant	**46,593**	**13**	**46,606**
Biology	Quant	16,451	2	16,453
Psychology	Qual	16,197	7	16,204
Portuguese	Qual	**75,035**	**26**	**75,061**
History	Qual	19,756	11	19,767
Philosophy	Qual	57,249	23	57,272
Geography	Quant	19,884	8	19,892
Physics and Chemistry	Quant	28,581	18	28,599
Biology and Geology	Quant	27,496	6	27,502
Quantitative Subjects	Quant	**139,005**	**47**	**139,052**
Qualitative Subjects	Qual	**223,122**	**87**	**223,209**
	Total	362,127	134	362,261

The samples that were removed from the dataset correspond either to extreme values of the predictive variables, e. g: outliers, or incomplete records. In accordance with Table 2, the dataset includes the high school final grades as the variable to be anticipated and a vector of 23 potential predictive variables that fed the algorithms input matrix through the Lasso feature selection method. The variables denoted in bold in Table 2 were dropped to avoid perfect collinearity problems.

4.2 Results

4.2.1 Training results

In the learning phase, the algorithms are fitted to the training dataset while the test dataset is kept aside. The stacking with the extreme gradient boosting algorithm (XGB) presents the best performances in any of the measures considered, mean absolute error

Table 2 Dataset variables

Variables	Description	Data Type
PermanentIncomeSupport_0	**No public support for family income**	**Binary**
PermanentIncomeSupport_1	High level of public support for family income	Binary
PermanentIncomeSupport_2	Medium level of public support for family income	Binary
PermanentIncomeSupport_3	Low level of public support for family income	Binary
Scholargrant_0	**No public support for education expenses**	**Binary**
Scholargrant_1	High level of public support for education expenses	Binary
Scholargrant_2	Low level of public support for education expenses	Binary
Gender_M	Male	Binary
AcYear_10	**Tenth academic year**	**Binary**
AcYear_11	Eleventh academic year	Binary
AcYear_12	Twelfth academic year	Binary
N_NoApprovals	Number of times the student fails to pass	Integer
Nationality	Foreign nationality	Binary
N_Enrollments	Number of times the student has been enrolled	Integer
Computer	The student has a computer	Binary
InternetAccess	The student has access to the internet	Binary
UrbanIndex	Population density measured by the school's county	Float
IncomeIndex	Income per capita measured by the school's county	Float
AgingIndex	Population ageing measured by the school's county	Float
UnemploymentIndex	Unemployment index by school's county	Float
CulturalLevelIndex	Cultural level index by school's county	Float
ClassSize	Number of class students	Integer
SchoolSize	Number of school students	Integer
N_SubjectsEnrolled	Number of subjects the student was enrolled	Integer
Gender_M__ClassSize	Class size if the student is male	Integer
FinalMark	Student's score in a 1–20 range	Integer

(MAE), mean square error (MSE) and determination coefficient (R^2) (see Table 3). In contrast, the classic multilinear regression model (OLS) has the worst performance, corroborating the well-known statement that machine learning algorithms have higher generalisation potential.

Just following the XGB algorithm, and focusing on MAE results, random forest (RF) comes in second concerning performance. Support vector regression (SV) ranks third in English and mathematics and fourth in Portuguese, surpassed by the artificial neural network (NN). The support vector regression bagging (BSV) was unable to exceed NN performance. Anyway, it did not drag as far behind as to consider it an unreasonable alternative to SV.

Another result worth mentioning refers to the gap between XGB and RF performance and its association with the stacking outputs correlation matrix determinant (Det [R]). For the same level of performance, the stacking efficiency is meant to be higher whenever the base algorithms outputs are uncorrelated. The correlation coefficient between the Det [R] and the performance gap between XGB and RF is 0.7911.

4.2.2 Test results

The main issue of our study is bias as made evident by the R^2 figures (see Table 4). Boosting is a class of algorithms whose primary focus is decreasing

Table 3 Learning phase results

		Train				
		English	Maths	Portuguese	Quantitative subjects	Qualitative subjects
MAE	OLS	2.38	2.68	1.85	2.44	2.26
	RF	2.24	2.54	1.72	2.34	2.17
	SV	2.34	2.63	1.79		
	BSV	2.35	2.65	1.81	2.40	2.22
	NN	2.34	2.64	1.79	2.40	2.22
	XGB	2.18	2.50	1.68	2.32	2.14
MSE	OLS	8.20	10.98	5.30	9.19	7.54
	RF	7.29	9.92	4.63	8.59	7.02
	SV	8.06	10.78	5.13		
	BSV	8.11	10.82	5.19	9.08	7.45
	NN	7.96	10.72	5.05	8.99	7.33
	XGB	7.04	9.71	4.47	8.50	6.90
R^2	OLS	12.88%	12.22%	15.27%	13.36%	11.19%
	RF	22.53%	20.64%	25.94%	18.96%	17.31%
	SV	14.30%	13.81%	17.96%		
	BSV	13.84%	13.49%	16.94%	14.32%	12.27%
	NN	15.37%	14.30%	19.13%	15.23%	13.71%
	XGB	25.18%	22.34%	28.43%	19.79%	18.73%
Det [R]	XGB	0.0011	0.0010	0.0013	0.0005	0.0010

Table 4 Generalization on the test dataset

		English	Maths	Portuguese	Quantitative subjects	Qualitative subjects
MAE	OLS	2.41	2.69	1.86	2.44	2.26
	RF	2.35	2.63	1.77	2.36	2.19
	SV	2.37	2.65	1.81		
	BSV	2.38	2.65	1.82	2.40	2.22
	NN	2.38	2.66	1.80	2.40	2.22
	XGB	2.36	2.66	1.77	2.36	2.18
MSE	OLS	8.31	10.97	5.34	9.16	7.51
	RF	7.94	10.67	4.93	8.75	7.15
	SV	8.23	10.81	5.20		
	BSV	8.24	10.82	5.26	9.05	7.42
	NN	8.16	10.79	5.12	8.96	7.32
	XGB	8.12	10.93	4.98	8.78	7.13
R^2	OLS	13.28%	11.66%	15.44%	13.77%	11.00%
	RF	17.18%	14.05%	21.92%	17.65%	15.29%
	SV	14.18%	12.95%	17.66%		
	BSV	14.10%	12.88%	16.78%	14.86%	12.06%
	NN	14.84%	13.07%	18.95%	15.63%	13.28%
	XGB	15.35%	11.97%	21.17%	17.38%	15.51%

the bias as is apparent in the training figures. However, the XGB performance drops substantially in unseen data, and RF presents better figures in English, mathematics and quantitative subjects. Mathematics is an extreme case as the XGB performance is the worst among the machine learning algorithms. XGB seems to overfit more on smaller datasets, exhibiting lower robustness in terms of a trade-off between bias and variance than RF. In this regard, OLS is the best method, only surpassed by BSV in quantitative subjects. The BSV algorithm has a narrower performance gap between training and testing than the SV parent. This enhanced robustness comes from the pasting randomisation.

4.2.3 Feature space analysis

The feature space analysis aims to assess the importance of each input variable in the predicting process of the target variable. The quantitative subjects and the qualitative subjects' results are based on agglomerated datasets that condense information and generate a more general outlook.

In our case, the feature space is the input subspace that is formed by the independent variables that have overcome the lasso feature selection filter (non-zero lasso βs).

The random forest feature importance used follows the scikit-learn implementation. In a regression tree the node importance NI_j comes as follows:

$$NI_j = W_j mse_j - \left(W_j^{left} mse_j^{left} + W_j^{right} mse_j^{right} \right) \tag{1}$$

Where W_j is the proportion of the samples that reach the node j, *left* and *right* refer to the classes after a split and *mse* is the mean square error. The node importance increases with the proportion of samples and the decrease of the mean square error due to the split. The feature importance FI_i is the proportion of the importance of those nodes whose splits are based on the feature i:

$$FI_i = \frac{\sum_j NI_{ij}}{\sum_j NI_j} \tag{2}$$

The feature importance is normalised to a value between 0 and 1 as follows:

$$\overline{FI_i} = \frac{FI_i}{\sum_i FI_i} \tag{3}$$

In a random forest, the feature importance is the feature importance mean of all trees (T):

$$\overline{FI_{i,rf}} = \frac{\sum_k^T \overline{FI_{ik}}}{T} \tag{4}$$

The dummy variables AcYear_11 and AcYear_12 just capture the gap between the academic years' average scores. As they are related to a natural partition of the target variable that is ever-present, the random forest feature importance was adjusted to cancel out their influence in the relative feature importance (%*).

Quantitative subjects In the study of the quantitative subjects (Table 5), the eleventh-year average score was almost coincident with the tenth year score, but the twelfth-year average score was about 1.99 points higher as shown by MLR βs in Table 5 for AcYear_11 and AcYear_12 variables.

The study corroborates the assertion that the permanent income support and scholar grant variables, both capturing the neediest students, should have a negative relationship with achievement. Furthermore, the second level of the scholar grant variable that includes the most impoverished students has the most significant negative effect. However, the second level of permanent income is not statistically significant for a confidence level of 5%. The RF feature importance empirical distribution seems to corroborate it. In fact, the correlation coefficient between p values and the RF feature importance coefficient of variation is a striking 0.98. In addition, first and third levels of the permanent income support were dropped in the lasso feature selection filter and the adjusted random forest feature importance, when considering them all together, reached

Table 5 Feature space analysis: quantitative subjects

Variables	Quantitative subjects						
	β				RF feature importance		
	Lasso	MLR	Literature expected sign	p-value	%	δ	%*
PermanentIncomeSupport_1	0.00	n.a.	–	n.a.	n.a.	n.a.	n.a.
PermanentIncomeSupport_2	−0.01	−0.13	–	0.09	0.16%	0.20%	0.21%
PermanentIncomeSupport_3	0.00	n.a.	–	n.a.	n.a.	n.a.	n.a.
Scholargrant_1	−0.07	−0.37	–	0.00	1.35%	0.24%	1.78%
Scholargrant_2	−0.12	−0.64	–	0.00	2.10%	0.20%	2.77%
Gender_M	−0.18	−0.49	–	0.00	3.64%	0.27%	4.81%
AcYear_11	0.00	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
AcYear_12	0.64	1.99	n.a.	0.00	24.27%	0.58%	0.00%
N_NoApprovals	−0.76	−1.24	–	0.00	41.76%	0.79%	55.14%
Nationality	0.00	n.a.	–	n.a.	n.a.	n.a.	n.a.
N_Enrollments	−0.09	−0.34	–	0.00	0.99%	0.22%	1.31%
Computer	−0.02	−0.19	- +	0.00	1.26%	0.23%	1.66%
InternetAccess	0.00	n.a.	- +	n.a.	n.a.	n.a.	n.a.
UrbanIndex	0.00	n.a.	+	n.a.	n.a.	n.a.	n.a.
IncomeIndex	0.00	n.a.	+	n.a.	n.a.	n.a.	n.a.
AgingIndex	0.00	n.a.	+	n.a.	n.a.	n.a.	n.a.
UnemploymentIndex	−0.01	−0.03	–	0.00	7.62%	0.65%	10.06%
CulturalLevelIndex	0.00	n.a.	+	n.a.	n.a.	n.a.	n.a.
ClassSize	0.00	n.a.	–	n.a.	n.a.	n.a.	n.a.
SchoolSize	0.07	0.0004	+	0.00	11.83%	0.73%	15.62%
N_SubjectsEnrolled	0.07	0.25	+	0.00	5.01%	0.45%	6.62%
Gender_M__ClassSize	0.00	n.a.	–	n.a.	n.a.	n.a.	n.a.

only a modest 4.77%. Take note that none of the socio-economic indexes could pass the lasso step except the unemployment index with an importance of 10.06%. As the indexes are highly interrelated, the lasso feature selection tends to select just one or at most just a few of them.

Male students have, on average, 0.49 points lower scores than female students. It is a somehow narrow gender gap with a feature importance of only 4.81%.

As they are highly correlated, either the computer usage or internet access was dismissed by the lasso feature selection step. As a result, they should be interpreted jointly. Thus, their negative effect on achievement was not unexpected as high school students tend to use computers and the internet for recreational purposes (Salomon and Ben-David Kolikant 2016). Anyway, their negative effect is only 0.19 points on final scores. Nationality does not play a relevant role in the models. The students were split between nationals and non-nationals, but no significant pattern was inferred. The school size variable appears positively related to achievement with a feature importance of 15.62%. In turn, class size was not selected by lasso, being considered dispensable for predictive purposes.

The primary variable that is related to a student's cognitive ability is the number of non-approvals that has a feature importance of 55.14%. The other cognitive ability proxy variables are the number of enrollments and the number of subjects enrolled that have a feature importance of 1.31% and 6.62% respectively.

Qualitative subjects In the qualitative subjects' case (Table 6), the eleventh and the twelfth year average scores were higher than the tenth year scores by 0.94 and 1.47 points respectively, as shown by MLR βs in Table 6 for AcYear_11 and AcYear_12 variables.

As in the quantitative subjects' case, the permanent income support and the scholar grant variables appear to have a negative relationship with achievement. As expected, the second level of the scholar grant had the most significant effect. However, the two

Table 6 Feature space analysis: qualitative subjects

Variables	Qualitative subjects						
	B				RF feature importance		
	Lasso	MLR	Literature expected sign	p-value	%	δ	%*
PermanentIncomeSupport_1	−0.02	−0.14	–	0.02	0.65%	0.79%	0.72%
PermanentIncomeSupport_2	0.00	n.a.	–	n.a.	n.a.	n.a.	n.a.
PermanentIncomeSupport_3	0.00	n.a.	–	n.a.	n.a.	n.a.	n.a.
Scholargrant_1	−0.12	−0.46	–	0.00	1.89%	0.19%	2.08%
Scholargrant_2	−0.16	−0.48	–	0.00	2.28%	0.85%	2.51%
Gender_M	−0.17	0.02	–	0.81	1.19%	1.71%	1.31%
AcYear_11	0.34	0.94	n.a.	0.00	6.04%	0.37%	0.00%
AcYear_12	0.38	1.47	n.a.	0.00	3.20%	0.55%	0.00%
N_NoApprovals	−0.70	−1.12	–	0.00	48.27%	0.62%	53.18%
Nationality	0.00	n.a.	–	n.a.	n.a.	n.a.	n.a.
N_Enrollments	−0.14	−0.47	–	0.00	2.94%	0.25%	3.24%
Computer	0.00	n.a.	- +	n.a.	n.a.	n.a.	n.a.
InternetAccess	0.05	0.15	- +	0.00	0.76%	0.18%	0.84%
UrbanIndex	−0.03	0.00	+	0.00	4.74%	0.51%	5.22%
IncomeIndex	0.00	n.a.	+	n.a.	n.a.	n.a.	n.a.
AgingIndex	0.02	0.00	+	0.00	5.72%	0.53%	6.30%
UnemploymentIndex	0.00	n.a.	–	n.a.	n.a.	n.a.	n.a.
CulturalLevelIndex	0.00	n.a.	+	n.a.	n.a.	n.a.	n.a.
ClassSize	0.05	0.02	–	0.00	4.54%	0.52%	5.00%
SchoolSize	0.07	0.00	+	0.00	8.18%	0.62%	9.01%
N_SubjectsEnrolled	0.11	0.26	+	0.00	5.13%	0.59%	5.65%
Gender_M__ClassSize	−0.06	−0.02	–	0.00	4.48%	1.77%	4.94%

last levels of the permanent income support were dropped in the lasso feature selection and the joint adjusted random forest feature importance is only 5.31%. Among the socioeconomic indexes, just urban and ageing indexes were strong enough to overcome the lasso step with a joint importance of 11.52%.

Male students have an average score comparable with female students. Notice that the gender MLR β is not statistically significant, and again, the RF importance standard deviation is greater than the RF feature importance itself. In fact, the correlation coefficient between p values and RF feature importance coefficient of variation is 0.74.

Computer usage did not pass the lasso feature selection. Contrary to the case of quantitative subjects, internet access and computer usage have a modest but positive joint effect on AA. Nationality has the same result as quantitative subjects, not playing a relevant role in the models. The school size variable appears to be positively related to achievement, and its feature importance reaches 9.01%. Male students scores appear to be insensitive to class size, but females tend to thrive in larger classes, as can be inferred from the ClassSize and Gender_M_ClassSize joint results.

Concerning the influence of student's cognitive ability on AA, the figures are as bold as the quantitative subjects' case. The number of non-approvals, number of enrollments and number of subjects enrolled have feature importance of 53.18%, 3.24% and 5.13% respectively.

5 Discussion

5.1 Academic achievement

Despite the indecisiveness that pervades the literature about the school size effect on AA (Schwartz et al. 2013), in this case, they appear positively associated. It should be interpreted with care because it can be related to omitted variables such as teaching quality, school design, management soundness, and student social economic status (Opdenakker and Van Damme 2007). The Portuguese public education system, the ambit of this study, is much centralized: Schools are quite homogeneous in the physical conditions they provide; Teachers are hired and allocated nationwide through a centralized selection process; Parents cannot choose freely, being restricted to enrol their children in neighbourhood schools. Teachers' evaluation cornerstones are uniform across the country. Furthermore, the target variable is teachers' grades, instead of grades of a unified test that would shed light on their lecturing quality. Thus, the effects of the variables' omission are considered minor despite the toll they inflict on the final robustness of the model. However, and most importantly, as the nation observes a significant income gap between coastal areas, where most of the large schools are located, and the countryside, along with the socioeconomic variables' lack of detail and scope, it is appropriate to accept that the school size importance is being distorted by the school socioeconomic status average.

The importance of the socioeconomic status variables in AA prediction was not as strong as might be expected. First, the socioeconomic indexes have an inherent limited predictive power due to their municipal level scope much larger than the student or family scope optimal. Secondly, the relationship between family income and AA is moderated by resources accessibility and the education system in question belongs to a medium-high-income country where poor people can enjoy a reasonable rapport of crucial social capital. Indeed, the differences between publicly funded schools in terms of physical conditions and teacher capabilities are almost inexistent, decreasing the effect of school choice on AA that is typically driven by socio-economical differences (Sirin 2005). The public high school education system being free and widely accessible contributes to narrow the socioeconomic gap in AA. However, if we consider that the importance of school size reflects differences in school socioeconomic status average, the results shall otherwise be read as inconclusive.

The AA gender gap, given by the difference between mean scores, is narrow. This result is in accordance with the country's PISA results (OEDC 2016) where fifteen-year-old male students outperformed female students in science and mathematics and behaved above the OEDC average when lagging in reading. Clearly, the gender gap does not seem to be as profound as it is in other countries, even though male students' dropout and retention rates are significantly higher than female ones.

Concerning public policies, as female students thrive in larger classes in qualitative subjects and male students seem to be insensitive to class size, no evidence is shown that smaller classes improve AA.

The importance of cognitive ability variables corroborates the central assertion of the literature. However, they do have limited predictive power as they fundamentally rely on the number of times a student has failed to ascertain overall student cognitive ability.

5.2 Machine learning

Even though the XGB stacking input matrix determinant is ever close to zero in our case and the potential for stacking enhancing is limited, the XGB went further on average relatively to RF whenever the Det [R] was higher.

The correlation between the *p*-values and RF feature importance coefficient of variation is a significant empirical result. As the tree feature space was not randomised, when the RF importance standard deviation is much larger than the RF feature importance itself, it is because the variable was unable to sustain a proper level of relevance in each tree that belongs to the forest. The variable importance varies bluntly from tree to tree because it is not decisive in the process of decreasing the mean square error of the outcomes.

5.3 Limitations

The study underfitting is expected. The number of times a student has not passed and the number of times he is enrolled in a particular subject are only weak proxy variables for cognitive ability. Analogously, the permanent income support, the scholarship grant and the county socioeconomic indexes cannot replace socioeconomic variables at the student or family level without a significant loss of information and model ableness of pattern retention. Moreover, teaching quality is an omitted variable, and the school facilities are reduced to size and location.

6 Conclusions

Concerning the first research question, the results show that all machine learning algorithms have attained a higher level of predictive ability when compared with the classical multiple linear regression model. However, data appropriateness in detail and scope is nevertheless of utmost importance.

Pasting the support vector machine regressor had appreciable results, especially for the larger datasets where the similarity matrix computation cost is much higher.

Stacking efficiency is enhanced if the outputs of the base learners are uncorrelated. Empirically, it was shown that the stacking appropriateness increases as those outputs' correlation matrix determinant also increases. In this regard, the objective function and the performance measure that supports the tuning of the base learners can be adapted to take-into-account the efficiency of the stacking step.

Concerning the second research question and the hypothetical implications for public policies, and given the purpose of deploying accurate and robust predictive models, an information system that supports the nationwide education system should be designed and further structured as to collect meaningful and precise data about the full range of academic achievement antecedents.

As female students thrive in larger classes in qualitative subjects and male students seem to be insensitive to size, no evidence is shown that smaller classes would improve AA.

Concerning the last research question, the random forest feature importance empirical distributions are correlated with the structure of p values and statistical significance test ascertains of the multiple linear model. This conclusion can lead to a new line of research in terms of using the random forest algorithm to develop conceptual specification tests. It is not cumbersome to put forward the hypothesis of a mismatching specification relationship between the target and predictor variables when the p values structure is not in accordance with the inherent RF feature importance empirical distributions.

Availability of data and material The data that support the findings of this study are available from DGEEC- Direção Geral de Estatísticas da Educação e da Ciência but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of DGEEC- Direção Geral de Estatísticas da Educação e da Ciência.

Authors' contributions All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript.

Authors' contributions as follows:

Conception and design of the study: Ricardo Costa-Mendes, Mauro Castelli, Tiago Oliveira and Frederico Cruz-Jesus.

Collection of data: Frederico Cruz-Jesus.

Analysis and interpretation of data: Ricardo Costa-Mendes, Frederico Cruz-Jesus.

Drafting the manuscript: Ricardo Costa-Mendes.

Revising the manuscript critically for important intellectual content: Mauro Castelli, Tiago.

Oliveira, Frederico Cruz-Jesus.

Approval of the version of the manuscript to be published: Tiago Oliveira, Ricardo Costa-Mendes, Mauro Castelli, Frederico Cruz-Jesus.

Funding This study was funded by FCT – Fundação para a Ciência e Tecnologia (DSAIPA/DS/0032/2018).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Code availability The custom code is available from the authors upon request.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*, 95–135.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation, 9*, 1545–1588.

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods, 43*, 3–31.

Becker, G. S. (1964). Human capital, a theoretical and empirical analysis with special reference to education. In *General series (National Bureau of economic research) (vol. 80)*. New York: National Bureau of Economic Research : Distributed by Columbia University Press.

Benner, A. D., Boyle, A. E., & Sadler, S. (2016). Parental involvement and adolescents' educational success: The roles of prior achievement and socioeconomic status. *Journal of Youth and Adolescence, 45*, 1053–1064.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305.

Bishop, C. M. (2006). *Pattern recognition and machine learning, Information Science and Statistics*. Berlin: Springer.

Breiman, L. (1996a). Bagging predictors. *Machine Learning, 24*, 123–140.

Breiman, L. (1996b). Stacked regressions. *Machine Learning, 24*, 49–64.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition, 118*, 2–16.

Chang, C.-C., & Lin, C.-J. (2007). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, Article No: 27.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining August 2016 (pp. 785–794).

Choromanska, A., Henaff, M., & Mathieu, M. (2015). The loss surfaces of multilayer networks. *Journal of Machine Learning Research, 38*, 192–204.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington.

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon, 6*, e04081.

Di Fabio, A., & Busoni, L. (2007). Fluid intelligence, personality traits and scholastic success: Empirical evidence in a sample of Italian high school students. *Personality and Individual Differences, 43*, 2095–2104.

Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational Psychology Review, 13*, 1–22.

Francis, B. (2005). *Reassessing gender and achievement, Questioning contemporary key debates*. New York: Routledge.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*, 1189–1232.

Gilar-Corbi, R., Miñano, P., Veas, A., & Castejón, J. L. (2019). Testing for invariance in a structural model of academic achievement across underachieving and non-underachieving students. *Contemporary Educational Psychology, 59*, 101780.

Hanushek, E. A., & Wößmann, L. (2010). *Education and economic growth, international encyclopedia of education*. Oxford: Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference, and prediction second editon*. Springer.

Haykin, S. (2009). *Neural networks and learning machines third edition*. Pearson.

Hill, N. E., & Taylor, L. C. (2004). Parental school involvement and children's academic achievement pragmatics and issues. *Current Directions in Psychological Science, 13*, 161–164.

Hoxby, C. M. (2000). The effects of class size on student achievement : New evidence from population variation. *Quarterly Journal of Economics, 115*, 1239–1285.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer Texts in Statistics: Springer.

Jensen, A. R. (1998). *The G factor: The science of mental ability*. London: Praeger.

King, R. B. (2016). Gender differences in motivation, engagement and achievement are related to students' perceptions of peer—But not of parent or teacher—Attitudes toward school. *Learning and Individual Differences, 52*, 60–71.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114*, 497–532.

Kubey, R. W., Lavin, M. J., & Barrows, J. R. (2001). Internet use and collegiate academic performance decrements: Early findings. *The Journal of Communication, 51*, 366–382.

Kuhfeld, M., Gershoff, E., & Paschall, K. (2018). The development of racial/ethnic and socioeconomic achievement gaps during the school years. *Journal of Applied Developmental Psychology, 57*, 62–73.

Lei, J., & Zhao, Y. (2007). Technology uses and student achievement: A longitudinal study. *Computers in Education, 49*, 284–296.

Leithwood, K., & Jantzi, D. (2009). A review of empirical evidence about school size effects : A policy perspective. *Review of Educational Research, 79*, 464–490.

Levels, M., Kraaykamp, G., & Dronkers, J. (2008). Immigrant children's educational achievement in western countries: Origin, destination, and community effects on mathematical performance. *American Sociological Review, 73*, 835–853.

Lupart, J. L., Cannon, E., & Telfer, J. A. (2004). Gender differences in adolescent academic achievement, interests, values and life-role expectations. *High Ability Studies, 15*, 25–42.

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access, 8*, 55462–55470.

Mensah, F. K., & Kiernan, K. E. (2010). Gender differences in educational attainment: Influences of the family environment. *British Educational Research Journal, 36*, 239–260.

Miguéis, V. L., Freitas, A., Garcia, P. J. V., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36–51.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning second edition. In F. Bach (Ed.), *Adaptive computation and machine learning series*. MIT Press.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge: MIT Press.

Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. *Higher Education*.

Nievergelt, J. (2000). Exhaustive search, combinatorial optimization and enumeration : Exploring the potential of raw computing power. In V. Hlaváč, K. G. Jeffery, & J. Wiedermann (Eds.), *Lecture notes in computer science vol. 1963, SOFSEM 2000: Theory and practice of informatics - 27th Conference on Current Trends in Theory and Practice of Informatics Milovy, Czech Republic, November 25 – December 2, 2000 proceedings* (pp. 18–35). Springer.

OEDC. (2016). *PISA 2015 volume I: Results excellence and equity in education*. Paris: OECD Publishing.

Opdenakker, M. C., & Van Damme, J. (2007). Do school context, student composition and school leadership affect school practice and outcomes in secondary education? *British Educational Research Journal, 33*, 179–206.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhoffer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Perreira, K. M., Harris, K. M., & Lee, D. (2006). Making it in America: High school completion by immigrant and native youth. *Demography, 43*, 511–536.

Ramchoun, H., Idrissi, M. A. J., Ghanou, Y., & Ettaouil, M. (2016). Multilayer perceptron: Architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Intelligence, 4*, 26.

Rivas-Perea, P., Cota-Ruiz, J., Chaparro, D. G., Venzor, J. A. P., Carreón, A. Q., & Rosiles, J. G. (2013). Support vector Machines for Regression: A succinct review of large-scale and linear programming formulations. *International Journal of Intelligent Science, 03*, 5–14.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement : Evidence from panel data. *The American Economic Review, 94*, 247–252.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*, 83–92.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the BRAIN. *Psychological Review, 65*, 386–408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning internal representations by error propagation, in: Parallel distributed processing: Explorations in the microstructure of cognition volume I: Foundations Institute for Cognitive Science University of California, San Diego*. London: MIT Press.

Salomon, A., & Ben-David Kolikant, Y. (2016). High-school students' perceptions of the effects of non-academic usage of ICT on their academic achievements. *Computers in Human Behavior, 64*, 143–151.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear estimation and classification. Lecture notes in statistics, vol 171* (pp. 149–171). New York: Springer.

Schneider, M. (2002). *Do school facilities affect academic outcomes?* Washington DC: National Clearinghouse for Educational Facilities and Educational Resources Information Center.

Schwartz, A. E., Stiefel, L., & Wiswall, M. (2013). Do small schools improve performance in large, urban districts? Causal evidence from New York City. *Journal of Urban Economics, 77*, 27–40.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*, 417–453.

Smola, A. J., & Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 199–222*.

Sorensen, L. C. (2019). "Big data" in educational administration: An application for predicting school dropout risk. *Educational Administration Quarterly, 55*, 404–446.

Steinmayr, R., Dinger, F. C., & Spinath, B. (2010). Parents' education and Children's achievement: The role of personality. *European Journal of Personality, 24*, 535–550.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*, 401–426.

Tesfagiorgis, M., Tsegai, S., Mengesha, T., Craft, J., & Tessema, M. (2020). The correlation between parental socioeconomic status (SES) and children's academic achievement: The case of Eritrea. *Children and Youth Services Review, 116*, 105242.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine, 16*, 385–395.

Tomul, E., & Savasci, H. S. (2012). Socioeconomic determinants of academic achievement. *Educational Assessment, Evaluation and Accountability, 24*, 175–187.

van der Scheer, E. A., & Visscher, A. J. (2018). Effects of a data-based decision-making intervention for teachers on students' mathematical achievement. *Journal of Teacher Education, 69*, 307–320.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*, 89–122.

Wolpert, D. H. (1992). Stacked generalization. *Elsevier Neural Networks, 5*, 241–259.

Woolner, P., Hall, E., Higgins, S., McCaughey, C., & Wall, K. (2007). A sound foundation? What we know about the impact of environments on learning and the implications for Building Schools for the Future. *Oxford Review of Education, 33*, 47–70.

Wößmann, L., & West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review, 50*, 695–736.

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

Affiliations

Ricardo Costa-Mendes[1] · Tiago Oliveira[1] · Mauro Castelli[1] · Frederico Cruz-Jesus[1]

Tiago Oliveira
toliveira@novaims.unl.pt

Mauro Castelli
mcastelli@novaims.unl.pt

Frederico Cruz-Jesus
fjesus@novaims.unl.pt

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