Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 5, No. 5, October, 2021

Machine Learning Bias in Predicting High School Grades: A Knowledge Perspective

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

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

Abstract

This study focuses on the machine learning bias when predicting teacher grades. The experimental phase consists of predicting the student grades of 11th and 12th grade Portuguese high school grades and computing the bias and variance decomposition. In the base implementation, only the academic achievement critical factors are considered. In the second implementation, the preceding year's grade is appended as an input variable. The machine learning algorithms in use are random forest, support vector machine, and extreme boosting machine. The reasons behind the poor performance of the machine learning algorithms are either the input space poor preciseness or the lack of a sound record of student performance. We introduce the new concept of knowledge bias and a new predictive model classification. Precision education would reduce bias by providing low-bias intensive-knowledge models. To avoid bias, it is not necessary to add knowledge to the input space. Low-bias extensive-knowledge models are achievable simply by appending the student's earlier performance record to the model. The low-bias intensive-knowledge learning models promoted by precision education are suited to designing new policies and actions toward academic attainments. If the aim is solely prediction, deciding for a low bias knowledge-extensive model can be appropriate and correct.

Keywords:

Knowledge Bias;
Bias And Variance Decomposition;
Random Forest;
Support Vector Regression;
Precision Education;
Academic Achievement.

Article History:

Received:	07	June	2021
Revised:	05	September	2021
Accepted:	16	September	2021
Published:	01	October	2021

1- Introduction

Precision education stems directly from the concept of precision medicine [1]. Tuning medical treatment and health prevention based on detailed information about genetics, environment, and lifestyle constraints is the main motivation for developing the precision medicine conceptual framework [2]. Similarly, the definition of tailored educational practices hinges on detailed data on a student's genetic, neuronal, psychological, and environmental traits. As long as the learning process is perceived as the result of the joint dynamics amongst biological, genetic, and neuronal traits, and social and cultural pathways, it is appropriate to define precision education as an emergent interdisciplinary research field at the intersection of the social sciences and biology [3]. The constant collection and processing of sensitive personal data are paramount and go far beyond personalized learning, which tends to be restricted to the analysis of students' progress and results [4].

As it is with precision medicine, precision education raises important ethical questions. Data protection and personal privacy are obvious concerns in light of the emphasis on biological determinants [5]. It is therefore appropriate to take into account both the effectiveness and the risk associated with the most invasive precision technologies and tools, such as brain data collection neuro-technologies, genetic testing, and new personality and non-cognitive competencies tests.

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

DOI: http://dx.doi.org/10.28991/esj-2021-01298

Precision education presupposes an extensive database of the critical factors that influence the students' Academic Achievement (AA). In addition to the introduction of biological factors, it requires sharpening the metrics currently in use and improving their representative intake. Advanced data analytics will be needed to evaluate their importance and influence on AA, and machine learning algorithms will be used extensively for their greater predictive ability. The comprehensive and continuous data collection that is paramount in the precision education framework is an extension of the ongoing datafication process of the 21st century digital economy, perceived as a perpetual cycle of capital accumulation [6].

As precision education arose with the prospect of seriously augmenting the predictive ability of machine learning algorithms to anticipate teachers' grades and test scores, the present study focuses on ascertaining the specificities of the machine learning bias in the AA scientific domain. With no purpose of neglecting the profound ethical issues in letting an algorithm shape the future of human beings alone [7], the lack of success in using predictive models to assign grades also seems to corroborate the appropriateness of studying the machine learning bias. A remarkable example is the 2020 International Baccalaureate final exam [8]. Due to the SARS-Cov-2 pandemic crisis, the International Baccalaureate, an educational organization from Geneva that offers a worldwide high school program, has decided not to hold the final exam in 2020. Instead, the final scores were awarded by an algorithm that failed miserably, despite being allegedly based on the coursework and schools' predicted grades. Therefore, this study sheds light on both the structure of the machine learning bias that is bound to appear when predicting grades and the likely precision education effect on the performance of the algorithms.

We introduce the knowledge bias concept that fills an important gap in the predictive model classification. The knowledge space comprises every known and unknown critical factor that exerts some influence on the target concept [9]. The knowledge bias appears as the divergence between the input space composed by the actual critical factors in use and that theoretical optimal space. Depending on the low or high knowledge bias, a model is classified as an intensive-knowledge or extensive-knowledge model, respectively. The latter is suited only to evaluate the execution of policies and actions in a post-inception phase. When conceiving and planning, only intensive knowledge learning models are appropriate to assist the decision process of which critical factors should be swayed to produce the desired results. The knowledge bias is most important for classifying machine learning implementations in the social sciences, in which the longitudinal regularity of the target concept behaviour is stronger and the knowledge about the critical factors is weaker.

The conclusions are drawn from the simultaneous analysis of two different implementations, a base implementation relying on a feature space that includes only the variables related to the AA critical factors, and a second implementation in which the one-year lagged grade of the student is appended, emphasizing the influence of the student´s historical path. Bearing that in mind, we carry out various random forest, support vector machine, and extreme boosting machine regressors implementations not only to predict the grades (attributed by teachers) of 11th and 12th grade students in Portuguese public high schools but also to compute the bias and variance decomposition through a bootstrap procedure. In addition, we use the knowledge bias concept to feed the discussion and to build the conclusions. A Lasso procedure is used to select the input space variables along with a random forest feature importance structure analysis to operationalize the concept of knowledge bias. The research questions are the following:

- What are the factors that can explain the underperformance of machine learning algorithms when predicting student grades?
- Is precision education bound to improve machine learning bias when predicting grades?
- Is the machine learning bias an unbiased indicator of the model embedded knowledge?

The remainder of this paper is organized as follows: Section two proceeds with an AA critical factors literature review and presents the machine learning implementations that are appearing in the domain; Section three describes the methodology, the machine learning algorithms, and the research process in detail. Section four begins by presenting the data and how they were collected and organized. Then the results are shown and interpreted concerning the hyper-parameter optimization, the prediction, the bias and variance decomposition, and the knowledge intensity of the implementations. The duality between the implementations in terms of the generalization error and bias is demonstrated and compared with their incorporated knowledge. Section five discusses the results and introduces the knowledge bias concept as a means of differentiating effects and functions of the models. Finally, Section six presents the main conclusions and answers the research questions directly.

2- Literature Review

The literature has extensively confirmed the student's cognitive ability as the main determinant of AA [10, 11]. However, on average, it leaves unexplained 51-75% of the total variance [12]. Males more often develop a negative peer attitude toward school [13], corroborating the empirical evidence of a gender gap in favour of females that reaches high visibility in linguistics, although lower in mathematics [14–16]. Indeed, a personal attitude with an adequate level

of diligence, organization, focus, and resilience is conducive to overachievement [17]. The AA can vary according to ethnicity, as in the US, where white students seem to outperform consistently [18]. A similar gap can be found regarding immigrant groups [19]. Low Socioeconomic Status (SES) immigrant students from small communities whose parents have left their home countries due to political entanglements normally underperform [20].

Using the internet and personal computer to learning tasks easiness, attractiveness and diversification favours AA [21, 22]. However, if used excessively for leisure activities the use can be detrimental [23]. Parents' participation in the school activities motivates their children to outperform [24, 25] and is especially important amongst lower SES students [26]. The parental involvement forges a suitable and convenient attitude toward teachers and school tasks [27]. There is empirical evidence that supports a positive relationship between SES and AA [28, 29], magnifying the role played by a convenient endowment of social and cultural capital. Steinmayr et al. (2010) [30] show that parents' education is positively associated with AA even after controlling for student intelligence and personality. Using a concept of SES that includes parental education and occupation, household size, and possessions, Tesfagiorgis et al. (2020) [31] conclude that there is a positive association between SES and AA. Tomul and Savasci (2012) [32] found that parental educational status and the average income per capita were important positive factors related to AA.

The association between AA and class size is not straightforward. Hoxby (2000) [33] estimated that class size does not have a statistically significant effect on AA. Krueger (1999) [34] found otherwise – that the class size has a generally negative effect on AA and is stronger for minority students and those of lower SES. Wößmann and West (2006) [35] studied the effect of class size in 11 countries and concluded that its magnitude depends on the educational system itself and the teachers' lecturing abilities. In a less controversial stand, smaller schools seem to improve the academic outcomes of both lower SES students and those with greater learning needs [36, 37]. Schneider (2002) [38] highlighted the importance of schools' indoor environmental conditions such as noise, light, temperature, and comfort for teachers and students alike to be properly motivated. Furthermore, the architectural features of the school should embody the expectations of the school participants [39].

Lecturing ability inferred by panel data fixed effects emerges as a positive factor on AA [40, 41]. Rivkin et al. (2005) [42] concluded that the teachers' fixed effects on the 9th-grade math test score were substantial and educationally relevant. It is argued that the teacher's role in the AA is to a great extent related to unobservable personal characteristics and that the experience and education level of teachers have a minor role. In turn, Wayne & Youngs (2003) [43] add that teachers' college grades seem to be positively correlated to AA. Last, the teacher quality has not only a short term but also a long term positive effect on the student academic outcomes [44].

In the AA literature some studies have used machine learning algorithms to substantiate their conclusions (Table 1). However, there is a clear preference for solving classification instead of regression problems [45], and to the best of the authors' knowledge no published studies addressing bias and variance decomposition exist.

Table 1. Machine learning studies.

References	Data	Methods	Target
Costa-Mendes et al., (2020) [46]	362,127 High School Teacher Grades	Multilinear regression, random forest, support vector machine, artificial neural network, and extreme gradient boosting machine stacking ensemble.	High school end of the year teacher grades
Cruz-Jesus et al. (2020) [47]	110,267 High School Students	Artificial neural network, decision tree, extremely randomized trees, random forest, support vector machine, k-nearest neighbours, and logistic regression classifiers.	High school retentions
Miguéis et al. (2018) [48]	2,459 Higher Education Students	Naïve Bayes, support vector machine, decision tree, random forest, bagged trees, and adaptive boosting trees classifiers.	Five classes of achievement
Musso et al. (2020) [49]	655 University Students	Artificial neural network classifier	Low and high levels of three different measures of AA
Mengash (2020) [50]	2,039 Students	Artificial neural network, decision tree, support vector machine and naïve Bayes classifiers.	Evaluating the admission criteria of a Saudi University
Sorensen (2019) [51]	220,685 Students	Decision tree and support vector machine classifiers.	School dropout

3- Research Methodology

3-1- Supervised Learning Algorithms

Supervised learning consists of finding a mathematical function that efficiently maps the predictive variables input space into the target variables output space. In the learning phase a supervised learning algorithm uses the actual association between input and output variables to build a machine able to approximate the target outputs from the simple awareness of the input variables. Supervised learning is used for solving classification problems, in which the target variables are binary, and regression problems, in which the target variables are continuous [52].

For each dataset 70% of the examples were assigned to training and 30% to testing. A training set standardization procedure of the input variables was carried out and subsequently applied to the corresponding test set. In the learning phase the model is built upon the training dataset and is further evaluated in terms of generalization error on the holdout test set. In parallel, a four-fold cross-validation procedure on the training set was carried out to evaluate its consistency with the test dataset. Furthermore, as the 10th high school year's dataset was used specifically for both the Lasso feature selection procedure and the hyperparameter tuning, the cross-validation and the bias and variance decomposition bootstrap are virtually unbiased.

Before the training phase, the algorithms' hyperparameters were optimized through a four-fold cross-validation procedure [53, 54]. As soon as the hyperparameters to be optimized were selected, a search space was built, and a random grid search [55, 56] was carried out. The hyperparameters' combination that maximizes the algorithms' four-fold average performance was picked and further used in training, evaluation, and the bias and variance decomposition. The algorithms' implementations follow the scikit-learn python module documentation [57].

3-1-1- Random Forest

The Random Forest (RF) [58] is a randomized decision tree ensemble resulting from a bootstrap aggregating procedure. In the decision tree algorithm the input space is broken successively in a way that minimizes a cost function, normally purity-linked in case of a classification and pattern recognition, or the mean square error in case of regression. In each step, usually a pair of new nodes representing two different subsets of the input space is created. In a randomized decision tree the input variables that take part in the optimized split decision are selected randomly [59]. The partition process ends when the cost function gains are no longer perceived as significant. The final nodes are called leaves and deliver the decision rules guiding the target variable estimation and prediction. The random forest ensembles the randomized decision trees by majority vote in case of classification or by computing their scores' mean in case of regression.

3-1-2- Support Vector Regression

The Support Vector Regression (SVR) algorithm's main intention is to find a function that approximates a continuous target variable with a deviation not exceeding $\varepsilon \in \mathbb{R}^+$ [60]. In the soft margin SVR, some flexibility is added that augments the algorithm generalization ability by allowing a deviation beyond $\varepsilon \in \mathbb{R}^+$at a cost of C through the introduction of slack variables $\xi \geq 0$. For the primal form of the SVR optimization problem see, e.g., Mohri et al. (2018) [53].

The SVR Lagrange multiplier dual form of the mathematical optimization problem [61] highlights two fundamental characteristics of the algorithm. The approximated function depends solely on the inner products between the examples that lie outside the ε-tube – the support vectors – and every actual example, whichever the feature space used to represent them in. In our case and to add a nonlinear character to the approximation, the gaussian radial basis function (RBF) kernel was applied to compute the inner products of an extrapolated infinite-dimensional space.

3-1-3- Extreme Gradient Boosting Machine

Boosting is a machine learning ensemble method like bagging. Boosting consists of building a strong learner by training several weak learners in different training sets [62]. The main differences rely on both the training set resampling process, which is built specifically to generate complementary learning, and on the weak learner weights assignment, which is based on performance [63]. Essentially, and contrary to the case of bagging, the sample probability distribution is changed in each iteration to allow the next weak learners to focus on reducing the bias in the preceding worst-performing examples. The gradient boosting machine [54, 64] creates a chain in which each weak learner is moulded to minimize the generalization error of the previous iteration. In our case, the weak learners are regression decision trees, and the loss function is the square loss. To improve robustness, the extreme gradient boosting (XGB) machine [65] adds to the decision trees gradient boosting framework two regularization hyperparameters that control the size and the magnitude of the trees' scores.

3-2- Bias and Variance Decomposition for Regression

The following bias and variance decomposition is based on Mehta et al. (2019) [66]. Consider a target random variable y that can be approximated from a vector of independent variables X as follows:

$$y = F(X;\theta) + \varepsilon \tag{1}$$

where ε is an irreducible stochastic term, F is the unknown real function that maps X into y and θ is a vector of parameters.

Suppose that a dataset $D^N = (X, y)$ was randomly drawn from the population and a statistical learning procedure was carried out to estimate F. In regression, the square error is normally elected as the estimation cost function:

$$C\big(y, F(X;\theta)\big) = \sum_i (y_i - F(X_i;\theta))^2 \quad (2)$$

The optimization problem underlying the parameters' estimation can be formalized as follows:

$$\hat{\theta}_{D^N} = argmin\big[\theta, C\big(y, F(X;\theta)\big)\big] \quad (3)$$

Every dataset $D_j^N = (y_j, X_j)$ that can be randomly drawn from the population produces a different $\hat{\theta}_{D^N}$ and a specific value for the cost function. The cost function expected value for unseen data prediction, i.e., not belonging to the actual $D^N = (X, y)$ that was used to learn, comes as follows:

$$\begin{aligned} E_{D,\varepsilon}\left\{C\left(y, F\left(X;\hat{\theta}_{D^N}\right)\right)\right\} \\ = \sum_i \left(F(X_i;\theta) - E_D\left\{F\left(X_i;\hat{\theta}_{D^N}\right)\right\}\right)^2 + \sum_i E_D\left\{\left(F\left(X_i;\hat{\theta}_{D^N}\right) - E_D\left\{F\left(X_i;\hat{\theta}_{D^N}\right)\right\}\right)^2\right\} \\ + \sum_i E_\varepsilon\left\{\left(y_i - F(X_i;\theta)\right)^2\right\} \end{aligned} \quad (4)$$

$$E_{D,\varepsilon}\left\{C\left(y, F\left(X;\hat{\theta}_{D^N}\right)\right)\right\} = Bias^2 + Variance + N.\sigma^2{}_\varepsilon \quad (5)$$

where the bias measures the deviation of the model's expected value relative to the true value. In turn, the variance measures the model estimates sensitivity to sample variations. And finally, the irreducible variance refers to the structural noise that is inherent to the target variable. There is an empirical trade-off between bias and variance [54]. Although complex functions are being approximated, small training sets may require simple models that nonetheless asymptotically biased perform better in unseen data. A 200 samples train dataset bootstrap [67] was employed and the bias and variance decomposition upon the applicable test dataset was computed. The mean square error cost function was decomposed instead of the square error:

$$mse = mean\{Bias^2 + N.\sigma^2{}_\varepsilon\} + mean\{Variance\} \quad (6)$$

Note that as the true function $F(X_i;\theta)$ in Equation 5 is unknown, the bias and the irreducible variance cannot be empirically separated.

3-3- Feature Selection

Before the algorithms' hyperparameters tuning, an optimization procedure of the input space was undertaken, consisting of selecting the predictive variables according to the strength of their association with the target variable. The Lasso multilinear regression model [68] was used, comprising a classic multilinear regression and an L1 norm regularization term that exerts some pressure on the less important regression coefficients to converge to zero.

$$\left(\hat{\alpha}, \hat{\bar{\beta}}, \lambda\right) = arg\ min\left\{\sum_{i=1}^N \left(y_i - \alpha - \sum_{j=1}^p \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^p |\beta_j|\right\} \quad (7)$$

Through a four-fold cross-validation search grid procedure, the highest shrinkage pressure λ model, whose cost function was not higher than the optimum plus its cross-validation standard deviation, was picked and the null $\hat{\beta}_j$ variables were subsequently discarded. The model knowledge intensity can be inferred from the input space dimension, the number of critical factors in the model.

3-4- Methodology Steps

The order of the methodology's steps is the following (Figure 1):

- To select the variables of the input space, we used the Lasso multilinear regression model, the base implementation, and the 10th-grade dataset.
- To take into account any latent procedural bias, we used three different machine learning algorithms: the random forest, the support vector regression, and the extreme boosting machine. As the first is a bootstrapping ensemble, the second is a kernelized linear model, and the third is a boosting ensemble, we believe that together they constitute a comprehensive set of algorithms.
- To tune the hyperparameters, we performed the following sub-steps using the base implementation and the 10th-grade dataset:
  - We built a search space of hyperparameters to be optimized.
  - Then, we carried out a random grid search embedded in a four-fold cross-validation procedure.

- Finally, we selected the hyperparameters' combination that maximizes the algorithms' four-fold average mean absolute error (MAE).

- The training-test split was carried out at the grade level, assigning 70% of the examples to training and 30% to testing. The training dataset was standardized and the test dataset was transformed accordingly.
- The models were trained and their generalization error computed on the holdout test set. In addition, a four-fold cross-validation on the training set was used to evaluate its consistency with the test set.
- We made use of a bootstrap procedure to compute the bias and variance decomposition:
  - We generated 200 models from 200 subsamples of the training dataset [69].
  - With those models, we predicted the grades of the test dataset 200 times.
  - Then, we computed the mean square error (MSE) and the variance of those predictions.
  - Finally, we assigned to bias the difference between them.
- The knowledge intensity of a model was deduced from the number of relevant variables that are associated with the critical factors and from the structure of the random forest feature importance. We applied the Lasso multilinear regression model to the entire set of variables, using both base and second implementations and both 11th grade and 12th-grade datasets, aiming at finding the variables that are sufficiently important to participate in the learning model. Subsequently, we computed their random forest feature importance and aggregated them according to the related critical factor.

We specifically used the 10th-grade dataset to select the variables and to tune the hyperparameters to ensure the robustness of the bias and variance decompositions.

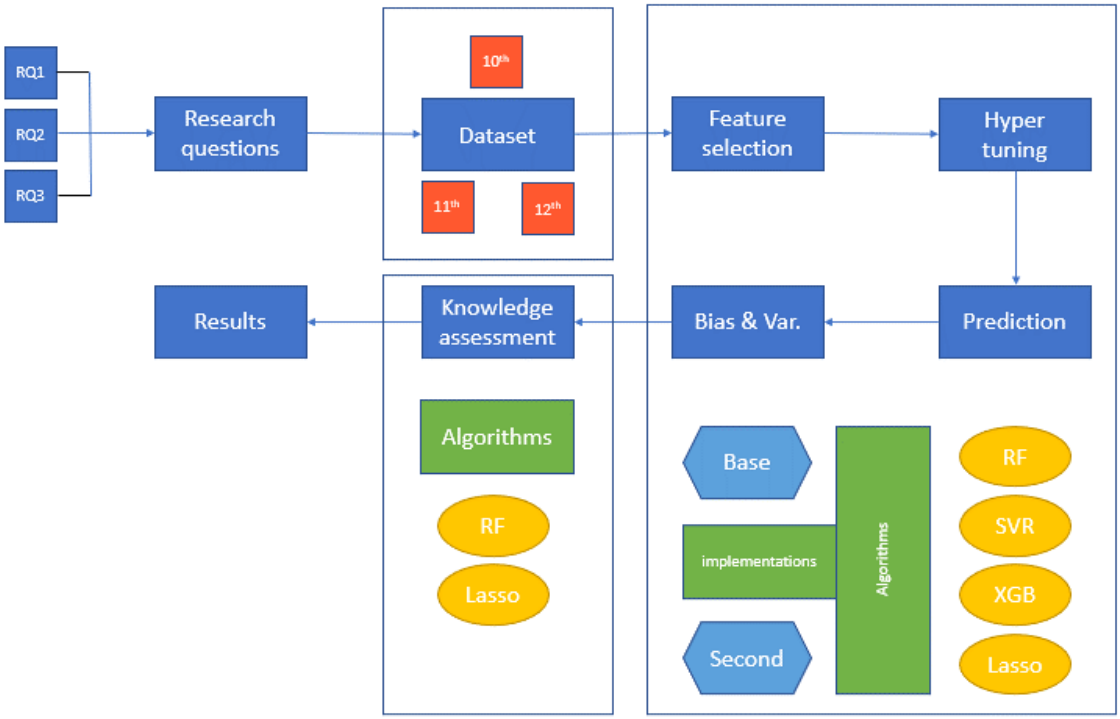


Figure 1. Methodological flowchart.

4- Data and Results

4-1- Data

The experimental data come mainly from the Directorate-General for Statistics of Education and Science of the Portuguese Ministry of Education information system. The system was designed to assist the administrative management of the Portuguese public education system and to store information about students, schools, and teachers from pre-school and basic to high school. Through a series of Microsoft® SQL Server Management Studio queries it was possible to build a global dataset consisting of 96,346 grades from 10,364 high school historical student paths. It includes observations from 2014-2015 to 2017-2018 academic years. The subjects were aggregated into four classes, Portuguese language, foreign languages, quantitative and natural sciences, human and social sciences. A split into 10th, 11th, and 12th grades was also carried out to feed the intended implementations (see Table 2).

Table 2. Dataset.

Subject	Class	Grade	Samples n
Portuguese	Portuguese		45,043
English Spanish French German	Foreign Languages		7,465
Mathematics Physics Chemistry Biology and Geology Geography Descriptive Geometry Design	Quantitative and natural sciences		30,098
Philosophy History Economics	Human and social sciences		13,740
		10^{th}	32,706
		11^{th}	32,396
		12^{th}	31,244
Total			**96,346**

The dataset is composed of 40 features that are related to the AA critical factors identified in the literature review (see Table 3 and Annex for full feature description). The family non-classic dwellings, the collective dwellings, the literacy rate, the post-secondary schooling rate, the primary sector importance, the secondary sector importance, and the unemployment rate were retrieved from Statistics Portugal. Given the categorical features one-hot encoding procedure, the number of predictive variables available to be selected by the Lasso filter added up to 120.

Table 3. Features and variables.

Feature	Literature AA critical factor	Data Type	# variables
Subjects	N.A.	Categorical	3
Retentions	Cognitive ability	Integer	1
Enrolments	Cognitive ability	Integer	1
Gender	Gender	Categorical	1
Father nationality	Ethnicity	Categorical	6
Computer	Computer usage	Binary	1
Internet	Internet usage	Binary	1
Job situation	SES	Binary	1
Education guardian	SES	Categorical	4
Guardian job educational level	SES	Categorical	4
Father job educational level	SES	Categorical	4
Mother job educational level	SES	Categorical	4
Guardian job situation	SES	Categorical	8
Father job situation	SES	Categorical	8
Mother job situation	SES	Categorical	8
Guardian educational level	SES	Categorical	11
Father educational level	SES	Categorical	11
Mother educational level	SES	Categorical	11
Scholarship	SES	Categorical	2

Parish	SES	Binary	1
County	SES	Binary	1
Family non-classic dwellings	SES	Percentage	1
Collective dwellings	SES	Percentage	1
Illiteracy rate	SES	Percentage	1
Post-secondary schooling rate	SES	Percentage	1
Primary sector importance	SES	Percentage	1
Secondary sector importance	SES	Percentage	1
Unemployment rate	SES	Percentage	1
School size	School size	Integer	1
Class size	Class size	Integer	1
Teacher professional category	Lecturing quality	Categorical	6
Teacher educational level	Lecturing quality	Categorical	3
Teacher career step	Lecturing quality	Categorical	3
Teacher gender	Lecturing quality	Categorical	1
Temporary replacement	Lecturing quality	Binary	1
Educative support	Lecturing quality	Binary	1
Teacher age	Lecturing quality	Integer	1
Lecturing time	Lecturing quality	Integer	1
Non-lecturing time	Lecturing quality	Integer	1
Educative support time	Lecturing quality	Integer	1
Teacher grade	Target variable	Integer	
			120

4-2- Results

4-2-1- Feature Selection

A shrinkage pressure λ of 0.02 was used for the feature selection and 56 variables were subsequently dropped (Figure 2). The most important dropped variables were the internet usage, parish literacy rate, post-secondary schooling rate, and primary sector importance. The internet usage is strongly correlated with the computer usage and the shrinkage pressure tends to reject the weakest. The parish literacy rate, post-secondary schooling rate, and primary sector importance belong to a set of seven SES variables retrieved from Statistics Portugal. The dropping of the other variables corresponds to the clustering of homogeneous feature categories in terms of effect on the AA.

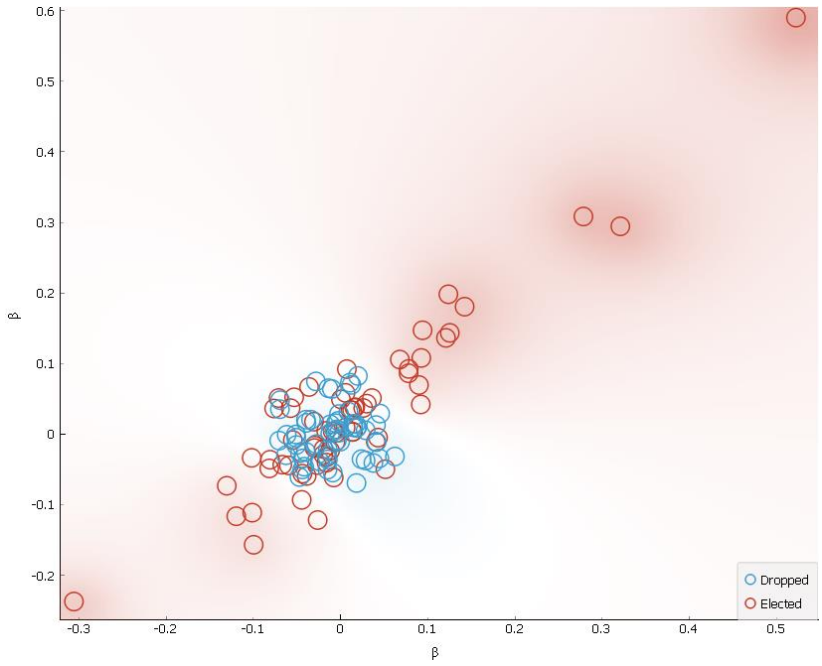


Figure 2. Feature selection.

4-2-2- Hyperparameter Optimization

The initial search space and the four-fold cross-validation random grid search results are shown in Table 4. The random grid search had 200 trials for each algorithm. The goal of the procedure is to minimize the cross-validation mean absolute error. According to the hyperparameter optimization procedure, the RFs were built from a 100% bootstrap of 420 trees. Two restrictions were imposed. First, the minimum number of examples required to be at a leaf could not be less than 0.009 of the dataset's length. Second, the minimum number of samples required to split an internal node could not be less than 0.001. The SVR hyperparameter optimization procedure set the penalty C to 9.541, and the RBF kernel γ to 0.004. Concerning the XGB, the procedure set the number of trees to 156, the subsample and column subsample to 1, the maximum tree depth to 20, the boosting learning rate to 0.42, the L2 regularization term on weights λ to 0.4, and the minimum number of instances in a child to 131.

Table 4. Hyperparameter search space

Algorithm	Hyperparameter	Minimum	Maximum	Cardinality	Best
RF	n_estimators	300	700	11	420
	min_samples_leaf	0.001	0.05	50	0.009
	min_samples_split	0.001	0.05	50	0.001
	Bootstrap	False	True	2	True
SVR	C	0.0001	100	50	9.541
	Γ	0.0001	100	50	0.004
XGB	max_depth	2	10	5	20
	min_child_weight	0	0.02	7	131
	Subsample	0.4	1	6	1
	colsample_bytree	0.4	1	6	1
	learning rate	0.01	1	198	0.42
	∧	0	20	17	0.4
	num_boost_round	1	999	999	155

The XGB performance was substantially improved by the hyper-optimization as shown by the large dispersion of the trial points on the scatter plot of Figure 3. The RF performance did not change greatly from trial to trial, inducing a concentrated cloud of points in the scatter plot. The SVR had a behaviour more in line with the RF than the XGB despite exhibiting a tendency to a higher overfitting.

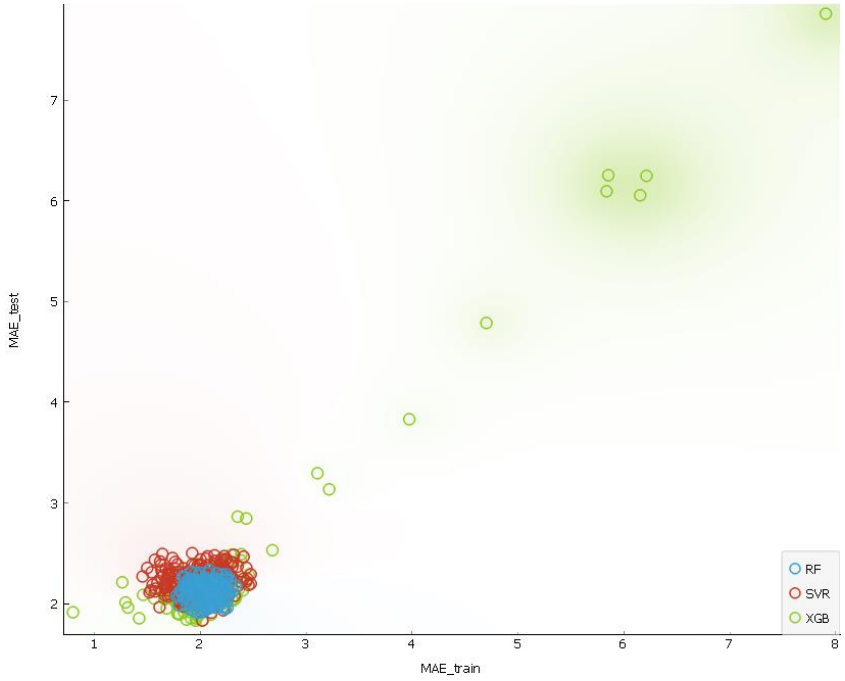


Figure 3. Random search trials

The average performances of the three algorithms were very similar (Figure 4). The RF had the smallest average MAE and the XGB the largest. In contrast, the MAE of the RF best trial, the elected hyperparameter combination, was 2.0377, while in the XGB was only 1.9073. The SVR fell into the middle with 2.0337. The flatness of the XGB empirical distribution curve in Figure 4 also highlights the bias focus of the algorithm.

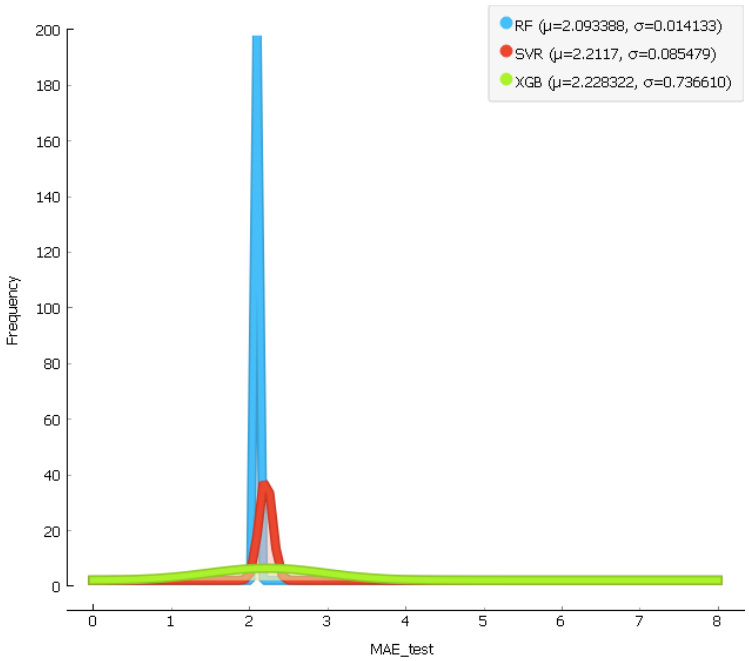


Figure 4. Search trials distributions.

The elected hyperparameter combinations are within the surface of the search spaces far from the edges, ensuring that at least a local optimum was reached.

4-2-3- Prediction Training Phase

To evaluate the algorithms' performance, the MSE, the MAE, and the coefficient of determination (R2) are shown in Table 5. It is apparent that the second implementation, which includes the lagged student grade as an input variable, has overwhelming results when compared to the base implementation, which considers only the critical factors. The base implementation led us to poor fits to the training data. On the other hand, the second implementation reaches a good accuracy level. This is true regardless of which algorithm is considered. The XGB has the best results overall, in which the edge is much more pronounced in the base implementation. Boosting is a machine learning method the principal objective of which is to reduce bias even if it is more prone to incurring overfitting. The RF comes next, being surpassed by SVR only in the 12th-year base implementation.

The duality between base and second implementations in favour of the latter is well represented in Figure 5. Only the XGB shortens the distance between both implementations. However, it is accomplished by overfitting the training data and does not revert to its generalization ability.

Table 5. Training results.

		Train			
		Implementation			
		11th year		12th year	
		Base	Second	Base	Second
MSE	RF	6.051	1.992	5.467	1.515
	SVR	6.124	2.087	5.411	1.523
	XGB	2.609	1.265	2.322	0.817
MAE	RF	2.002	1.073	1.900	0.937
	SVR	1.964	1.094	1.817	0.929
	XGB	1.279	0.863	1.196	0.693
R2	RF	0.227	0.746	0.294	0.804
	SVR	0.218	0.733	0.301	0.803
	XGB	0.667	0.838	0.700	0.894

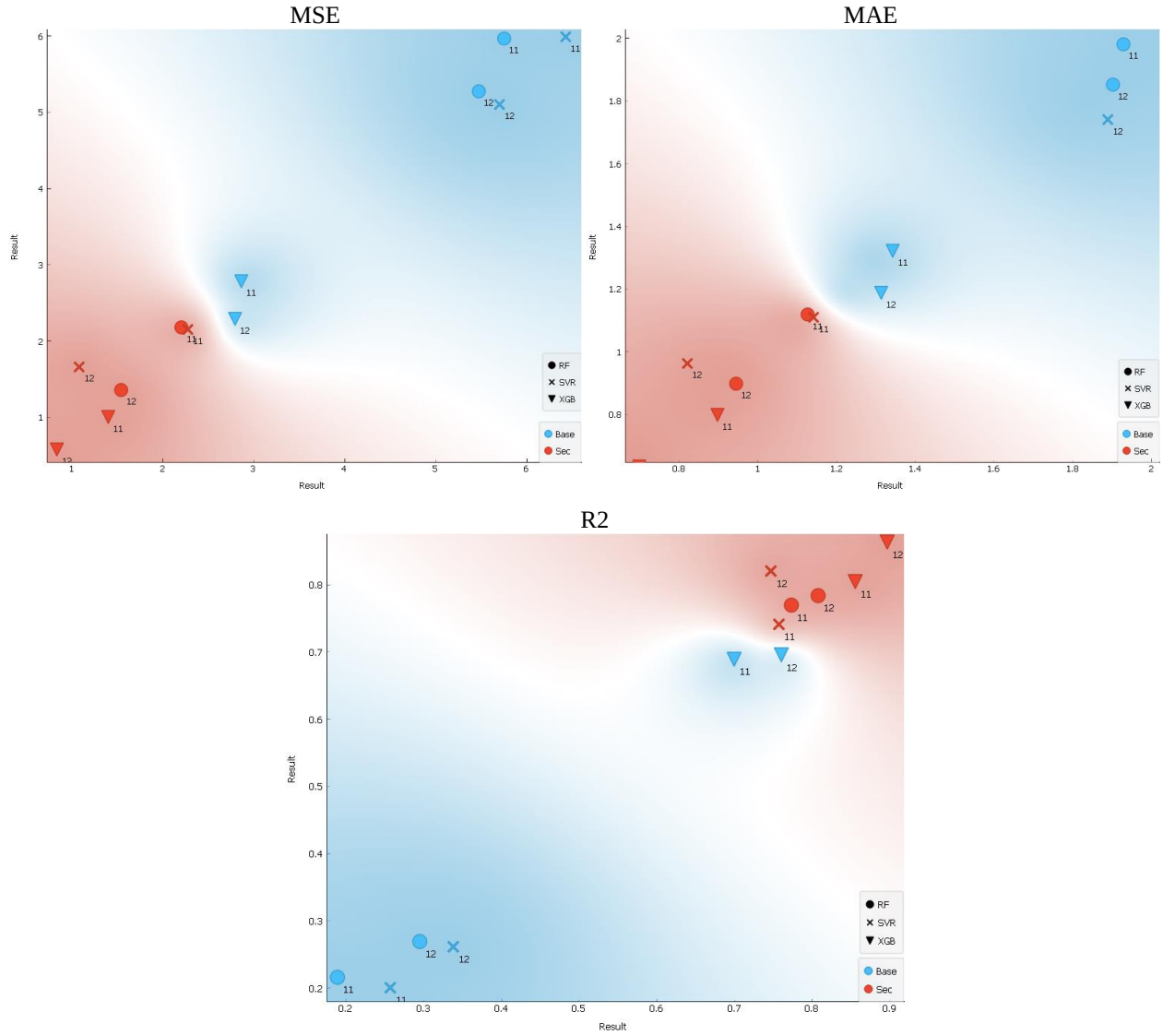


Figure 5. Base and second implementation duality in training.

4-2-4- Prediction Test Phase

The test results are shown in Table 6. They are poorer than the training results, highlighting the existence of overfitting. Figure 6 illustrates the difference between train and test phases. The deterioration is generally more acute in the base implementation. Every algorithm exhibits at least some overfitting, but it is intense in the XGB case, especially in the base implementation, which is invariably located on the graphs upper right corner. The second implementation still presents an appropriate accuracy and seems to yield a good level of robustness.

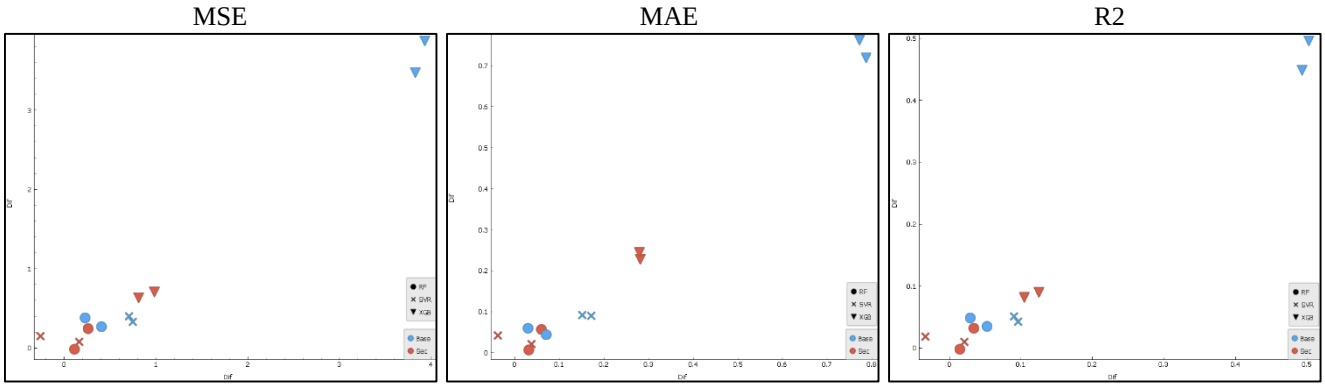


Figure 6. Overfitting and train-test gap.

In the base implementation the XGB training edge is significantly shortened and in the second implementation virtually disappears. Indeed, the SVR even takes the lead in the 12th-year second implementation. The training four-fold cross-validation results converge with the test results, as both the features selection and hyperparameter optimization were undertaken on the 10th-year base implementation dataset. The XGB cross-validation standard error is in line with RF and SVR, indicating that the strong overperformance of the XGB base implementation in training and its further fall in the test evaluation are almost certainly due to noise retention.

Table 6. Generalization on test dataset.

		Cross validation Implementation				Test Implementation			
		11th year		12th year		11th year		12th year	
		Base	Second	Base	Second	Base	Second	Base	Second
MSE	RF	6.489	2.104	5.869	1.608	6.440	2.094	5.834	1.553
	SVR	6.617	2.118	5.956	1.575	6.576	2.109	6.000	1.528
	XGB	6.360	2.153	5.817	1.615	6.069	2.087	5.699	1.545
δMSE	RF	0.076	0.045	0.054	0.066				
	SVR	0.090	0.045	0.006	0.045				
	XGB	0.072	0.050	0.056	0.055				
MAE	RF	2.073	1.104	1.970	0.965	2.068	1.095	1.966	0.949
	SVR	2.074	1.106	1.949	0.952	2.074	1.098	1.966	0.934
	XGB	2.017	1.126	1.919	0.970	1.971	1.105	1.895	0.948
δMAE	RF	0.012	0.009	0.015	0.010				
	SVR	0.016	0.010	0.007	0.007				
	XGB	0.020	0.010	0.016	0.009				
R2	RF	17.08%	73.12%	24.21%	79.23%	17.15%	73.06%	23.80%	79.71%
	SVR	15.45%	72.93%	23.08%	79.65%	15.40%	72.87%	21.62%	80.04%
	XGB	18.74%	72.49%	24.87%	79.15%	21.92%	73.15%	25.56%	79.82%
δR2	RF	0.008	0.006	0.004	0.006				
	SVR	0.012	0.006	0.006	0.006				
	XGB	0.006	0.007	0.004	0.007				

When predicting student grades, the second implementation is better than the base implementation regardless of which algorithm is taken. The duality between the implementations deepens in the test phase as we evaluate the generalization ability of the algorithms in unseen data. The XGB test results are not blurred with overfitting issues and end by converging to the other algorithms' performances. In Figure 7 the dual zones of the base and second implementations are much clearer.

4-2-5- Bias and Variance Decomposition

As the irreducible variance and the target variable stochastic process are not supposed to vary with the implementations, the bias and irreducible variance aggregation are further referred to as bias.

As in the prediction, the second implementation provides a pronounced improvement over the base implementation with an MSE maximum decline of 71.10% in the 12th year SVR and a minimum of 63.18% in the 11th year XGB (see Table 7). The decrease in the bias explains a major percentage of the MSE improvement, reaching a maximum of 98.61% in the RF and a minimum of 82.25% in the XGB, both for the 11th year. Though far from being decisive, the variance also decreases, contributing to the MSE improvement (see Figure 8).

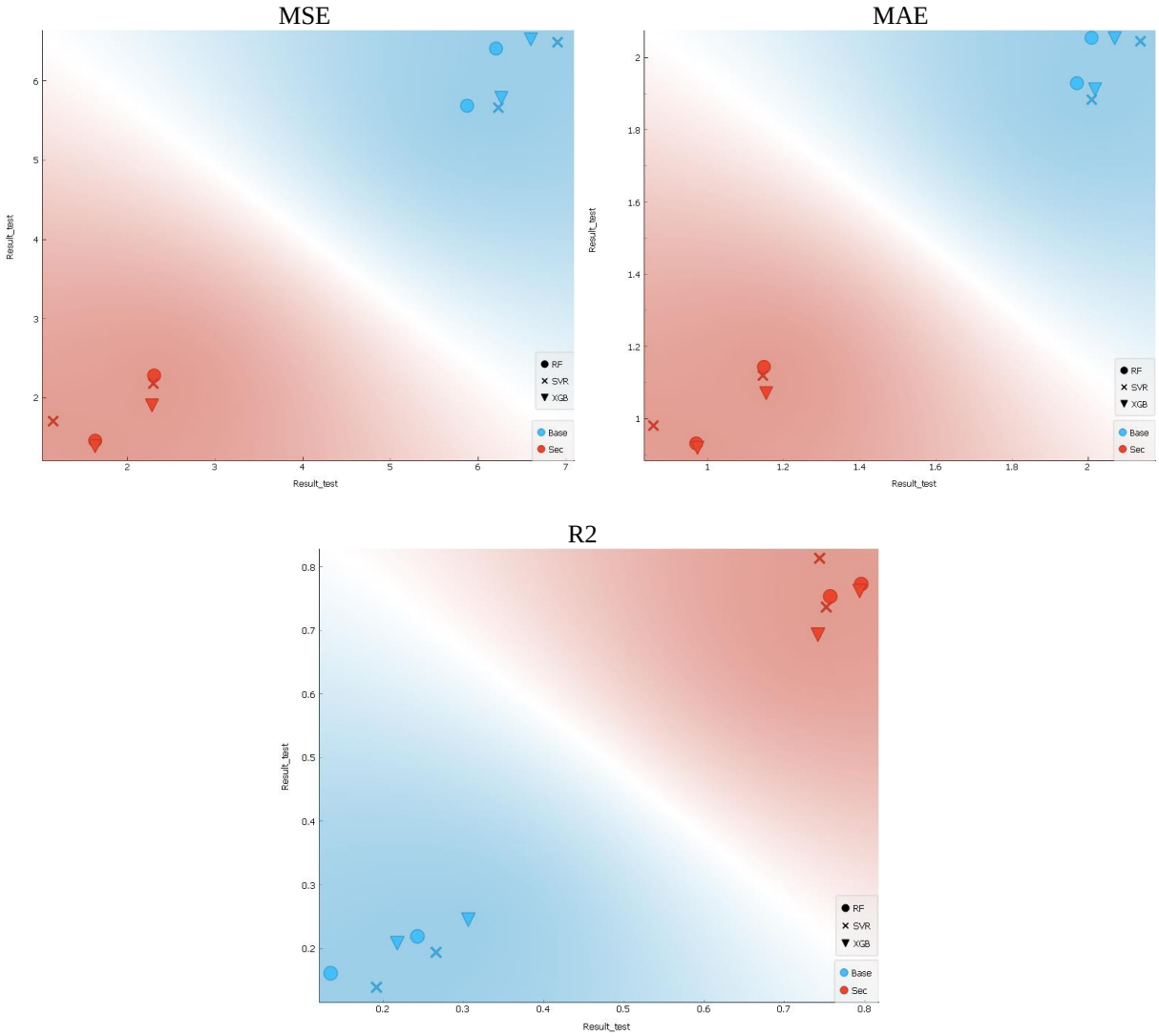


Figure 7. Base and second implementation duality in test.

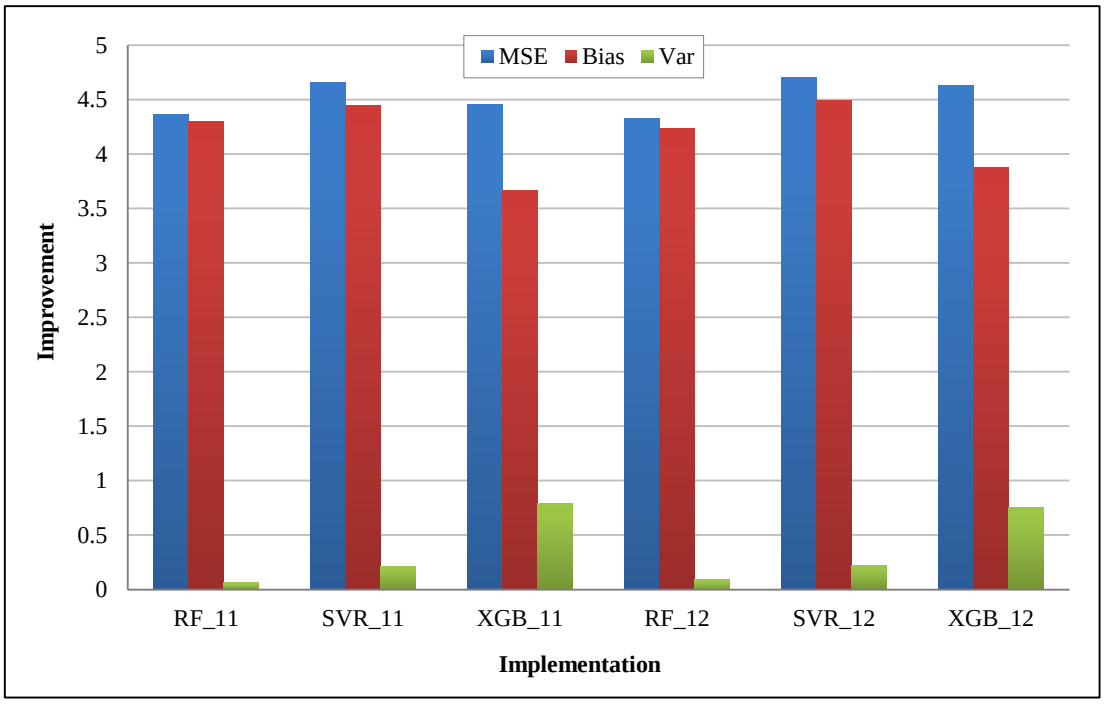


Figure 8. Decomposition of the MSE improvement.

The best bias results correspond to XGB implementations, which are consistent with the machine learning boosting technique's main purpose. In turn, the RF presents the MSE best results, which were essentially built upon the variance performance the inherent RF bootstrap is meant to provide. The SVR improves performance in the second implementation and is quite effective in adapting to the lagged teacher grade strong signal. The described duality between the base and second implementation generalization ability in the prediction sections corresponds to a bias duality in the bias and variance decomposition (see Figure 9). The duality in terms of variance does not appear perfect because of the XGB variance comparing poorly with any other algorithm implementation, a classic example of the well-known bias and variance trade-off [70].

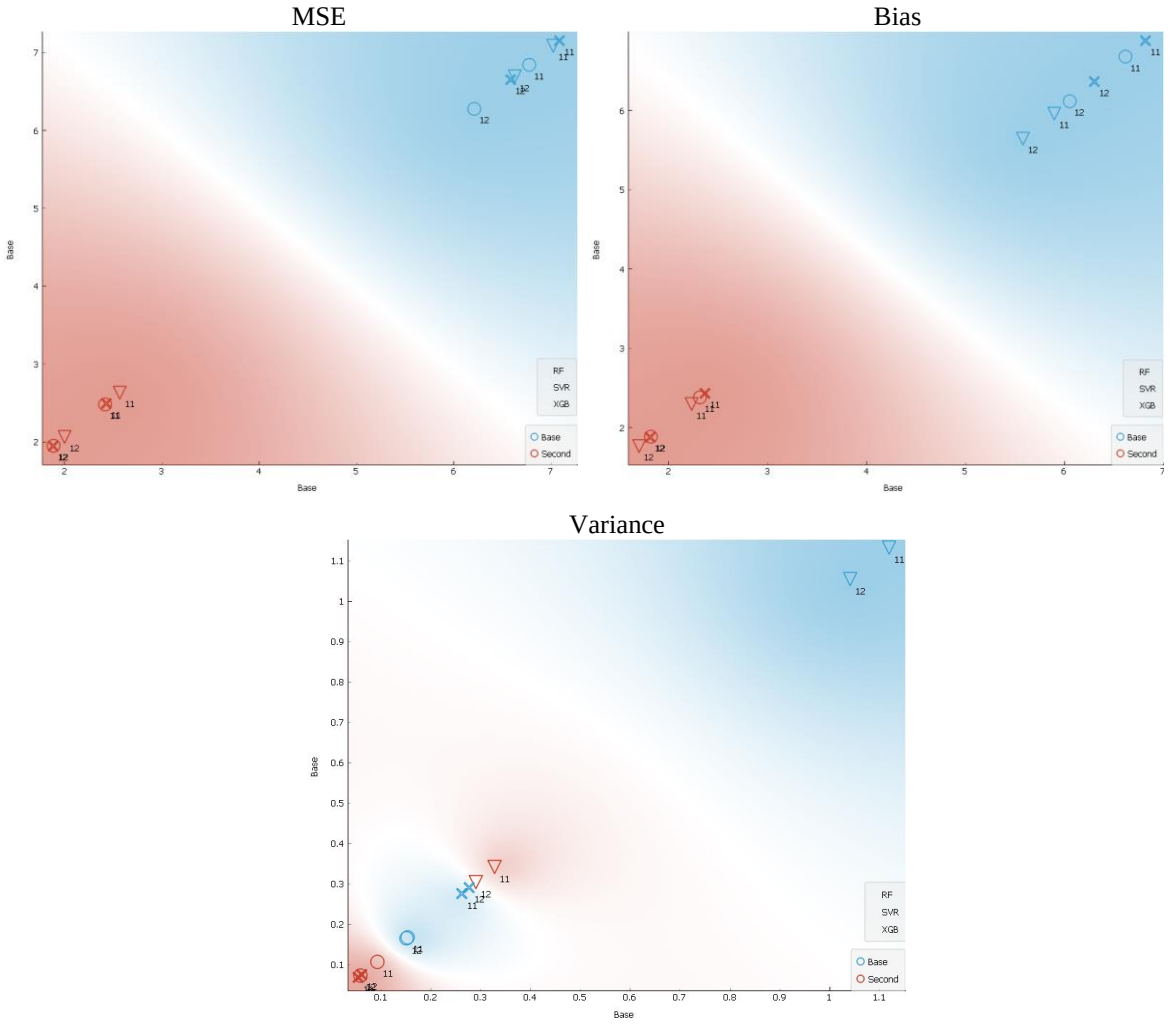


Figure 9. Base and second implementation duality.

Table 7. Bias and variance decomposition.

	Implementation	11th year			12th year		
		RF	SVR	XGB	RF	SVR	XGB
MSE	Base	6.81	7.12	7.05	6.24	6.62	6.66
	Second	2.45	2.46	2.60	1.92	1.91	2.03
	Difference	4.36	4.66	4.46	4.33	4.70	4.63
Bias	Base	6.65	6.85	5.93	6.08	6.33	5.61
	Second	2.35	2.40	2.26	1.85	1.84	1.73
	Difference	4.30	4.45	3.66	4.23	4.49	3.88
Variance	Base	0.16	0.27	1.13	0.16	0.28	1.05
	Second	0.10	0.06	0.33	0.07	0.07	0.30
	Difference	0.06	0.21	0.79	0.09	0.22	0.75

4-2-6- Knowledge Intensity

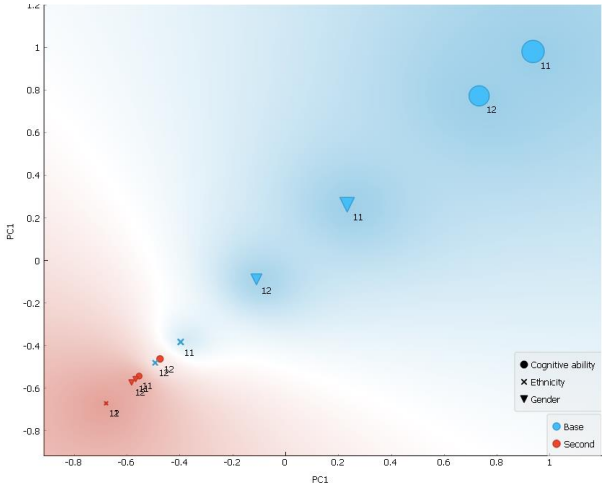
Table 8 shows the knowledge incorporated in the different implementations per AA critical factor. (Subjects) refers to the classes presented in Table 2 and it is not related to any critical factor. The base implementations of the 11th and 12th grades have 49 and 52 input variables respectively, contrasting with the 16 and 25 input variables of the second implementations (Table 8). Due to the introduction of the lagged teacher grade as a predictive variable, the Lasso method of selecting relevant input variables discards a much larger number of predictive variables associated with AA critical factors. Through the analysis of the RF feature importance structure, it is concluded that the critical factors that most contribute to the final solution in the base implementations are the cognitive ability and the SES. However, the importance of the lagged teacher grade of 96.7% for the 11th year and 96.2% for the 12th year overpowers any contribution of the critical factors to the final solution in the second implementations. Thus, the base implementations are considered knowledge-intensive when compared to the second implementations.

Table 8. Knowledge.

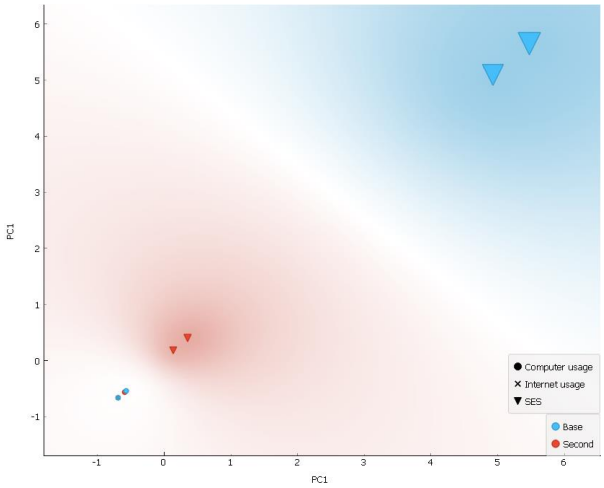
Literature AA critical factors	# variables after Lasso procedure				RF feature importance			
	Implementation				Implementation			
	Base		Second		Base		Second	
	11th	12th	11th	12th	11th	12th	11th	12th
(Subjects)	3	3	2	3	0.286	0.436	0.008	0.018
Cognitive ability	2	2	1	2	0.158	0.136	0.004	0.002
Gender	1	1	1	1	0.091	0.053	0.002	0.000
Ethnicity	3	2	0	0	0.001	0.000	0.000	0.000
Computer usage	0	1	0	1	0.000	0.003	0.000	0.001
Internet usage	0	0	0	0	0.000	0.000	0.000	0.000
SES	32	34	8	10	0.356	0.275	0.010	0.013
School size	1	1	1	0	0.060	0.038	0.008	0.000
Class size	1	1	0	0	0.025	0.024	0.000	0.000
Lecturing quality	6	7	3	8	0.024	0.035	0.002	0.003
(lagged teacher grade)							0.967	0.962
Total	**49**	**52**	**16**	**25**	**1**	**1**	**1**	**1**

The graphs in Figure 10 were built upon the first component of a Lasso variables and RF feature importance principal components analysis. It is strong and positively correlated with both variables and explains 91.3% of total variance. Concerning knowledge (Figure 10) there is also a duality between the base and second implementation. However, in this case, the base implementation takes the lead and incorporates more knowledge than the second implementation.

Cognition, gender, and ethnics



Computer, internet, and SES



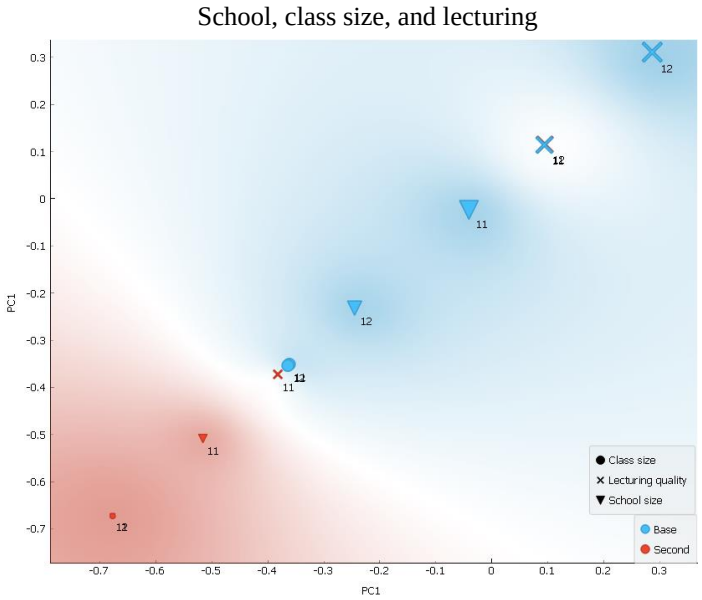


Figure 10. Base and second implementation knowledge duality.

5- Discussion

5-1- Discussion

In machine learning, bias can refer to any factor, embedded either in the algorithm architecture or in the concept representation form, which leads to a decision of preferring one learning generalization to another that is inconsistent with the ground knowledge of the experimental examples [71]. The procedural bias or algorithm bias focuses on the appropriateness of the search heuristics preferences on paths and approaches that assist the learning process. One example is the problem of structural bias that consists of the inability of the evolutionary algorithms to carry out an impartial search that includes every part of the search space [72]. A set of well-known state of the art machine learning algorithms – RF, SVR, and XGB – was purposefully called on for factoring in the procedural bias. Its influence can be regarded negligible, as the algorithms' performances are quite similar throughout the implementations.

The representational bias focuses on the adequateness of the search space to define, explain, and predict the target concept [73]. The dataset bias problem of the image object detector domain that limits the generalization ability to test datasets within the learning source is an example of representational bias [74]. Another example can be found in the size of Big Data datasets extracted from the digital platforms, which often leads researchers to generalize the conclusions to the entire population when in fact they represent only individuals with a special propensity to use them [75]. The second implementation presents an adequate performance in terms of generalization error and bias, despite the AA critical factors' small role in the definition of its input space. Its accuracy is built upon the student's historical path. The base implementation shows poor performance in terms of generalization error and bias due to the lack of precision in the critical factors' measurement. In the current study the poor performance of machine learning algorithms when predicting student grades is related to the input space's poor precision and the lack of a sound student historical path. Indeed, the representational bias is set to a minimum when the search space imprints every tone of the target concept. However, it is not decisive about whether it is established upon a differed measurement of the same target concept or upon a comprehensive knowledge and precise measurement of its determinants.

The concept of knowledge bias refers to the gap between the target concept knowledge space and the input search space. The former can have unknown dimensions and includes every element that affects the target concept. In turn, the input search space normally has only a subset of those elements, adding knowledge bias to the learning model. The concept of knowledge bias is pivotal to frame the precision education effect on machine learning bias. The base implementations have poorer performance and wider machine learning bias relative to the second implementations. However, the knowledge bias is weaker in the former as the base search space invariably has more critical factor components with greater RF feature importance. Therefore, it is possible to avoid machine learning bias and augment the generalization ability of a model without adding knowledge. Precision education would improve the machine learning bias through a knowledge bias decrease. More precisely, precision education would mostly improve high bias knowledge-intensive machine learning models and the effect in low bias knowledge-extensive models as the second implementation would be marginal. By no means is its role diminished. First, it is worth mentioning that the expansion of knowledge about the AA critical factors is important in the design of novel conceptualizations in the AA domain [9]. Second, low bias intensive-knowledge learning models are crucial to design new policies and actions, as the goal is to mould the critical factors in such a way that is conducive to AA attainments. Last, low bias extensive-knowledge learning

models are suitable to evaluate the same policies and actions but only in a post-design phase. Indeed, they do not assist the education stakeholders in the design of policies, as do low bias intensive-knowledge learning models. On the other hand, as long as there is irreducible variance, the grades predicted by any algorithm have an ever-present quantum bias. The individual essential foundation resides in the quantum, and student evaluation through real life assessments is a way to ensure the freedom of being.

5-2- Limitations

This study has several limitations. The cognitive ability is not directly represented by student intelligence quotient data and there is no measure pointing to the student attitude toward school activities and the corresponding parental involvement. The set of SES variables does not include income data and family size. Furthermore, the comfort of the school infrastructure, its adequateness, and the teachers' lecturing abilities are also omitted. The lack of depth and scope in the dataset can explain a non-significant part of the performance differences reported in the results. The adoption of a precise and data-driven approach in the management and storage of education data is a pivotal cornerstone in the implementation of a precision education framework.

6- Conclusion

As for the first research question, we conclude that the poor performance of machine learning algorithms when predicting student grades is related to the input space's poor precision and the lack of a sound student historical path. To anticipate student's grades through a machine learning implementation, we must collect either a comprehensive dataset that includes the entire range of the critical factors or the most recent preceding grades. On the other hand, the information systems that support the national education cluster should be designed in such a way as to allow every important piece of information about the AA critical factors to be collected. This is a most needed background if the aim is to implement machine learning models that would be decisive both in educational policy planning and in the decision-making process of the educational stakeholders. Regarding the second research question, precision education would mostly improve high bias knowledge-intensive machine learning models and the effect in low bias knowledge-extensive models as the second implementation would be marginal. If the education stakeholders' objective is to design policies and define new actions, a low bias knowledge-intensive model, the search space of which is formed by every critical factor, is almost mandatory, as it produces less biased estimates of the effects of the critical factors. The precision education framework adoption can provide them. If the aim is to anticipate student's grades, a knowledge-extensive model can be sufficient and appropriate, depending solely on the generalization error it conveys.

Concerning the third research question, the second implementation has a greater knowledge bias when compared to the base implementation even though it has a lower machine learning bias. Therefore, it is possible to reduce the machine learning bias without adding knowledge to the learning model. It can be accomplished by simple deferred observation of the target concept.

7- Declarations

7-1- Author Contributions

Conceptualization, R.C.M., F.C.J., T.O., and M.C.; methodology, R.C.M., F.C.J., T.O., and M.C.; software, R.C.M. and M.C.; validation, F.C.J., T.O., and M.C.; formal analysis, R.C.M., F.C.J., T.O., and M.C.; investigation, R.C.M.; resources, R.C.M., F.C.J., T.O., and M.C.; data curation, R.C.M. and F.C.J.; writing—original draft preparation, R.C.M.; writing—review and editing, R.C.M.; visualization, R.C.M.; supervision, F.C.J., T.O., and M.C.; project administration, F.C.J., T.O., and M.C.; funding acquisition, F.C.J., T.O., and M.C. All authors have read and agreed to the published version of the manuscript.

7-2- Data Availability Statement

3rd Party Data: Restrictions apply to the availability of these data. Data were obtained from DGEEC- Direção Geral de Estatísticas da Educação e da Ciência and are available from the authors upon reasonable request with the permission of DGEEC- Direção Geral de Estatísticas da Educação e da Ciência.

7-3- Funding

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

7-4- Conflicts of Interest

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

8- References

[1] Lu, Owen HT, Anna YQ Huang, Jeff CH Huang, Albert JQ Lin, Hiroaki Ogata, and Stephen JH Yang. "Applying learning analytics for the early prediction of Students' academic performance in blended learning." Journal of Educational Technology & Society 21, no. 2 (2018): 220-232. Available online: http://www.jstor.org/stable/26388400. (accessed on May 2021).

[2] Faulkner, Eric, Anke-Peggy Holtorf, Surrey Walton, Christine Y. Liu, Hwee Lin, Eman Biltaj, Diana Brixner, et al. "Being Precise About Precision Medicine: What Should Value Frameworks Incorporate to Address Precision Medicine? A Report of the Personalized Precision Medicine Special Interest Group." Value in Health 23, no. 5 (May 2020): 529–539. doi:10.1016/j.jval.2019.11.010.

[3] Youdell, Deborah. "Bioscience and the Sociology of Education: The Case for Biosocial Education." British Journal of Sociology of Education 38, no. 8 (January 26, 2017): 1273–1287. doi:10.1080/01425692.2016.1272406.

[4] Selzam, S, E Krapohl, S von Stumm, P F O'Reilly, K Rimfeld, Y Kovas, P S Dale, J J Lee, and R Plomin. "Predicting Educational Achievement from DNA." Molecular Psychiatry 22, no. 2 (July 19, 2016): 267–272. doi:10.1038/mp.2016.107.

[5] Lupton, Deborah, and Ben Williamson. "The Datafied Child: The Dataveillance of Children and Implications for Their Rights." New Media & Society 19, no. 5 (January 23, 2017): 780–794. doi:10.1177/1461444816686328.

[6] Sadowski, Jathan. "When Data Is Capital: Datafication, Accumulation, and Extraction." Big Data & Society 6, no. 1 (January 2019): 205395171882054. doi:10.1177/2053951718820549..

[7] Rességuier, Anaïs, and Rowena Rodrigues. "AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics." Big Data & Society 7, no. 2 (July 2020): 205395172094254. doi:10.1177/2053951720942541.

[8] Broussard, Meredith. "When algorithms give real students imaginary grades." The New York Times (2020).

[9] Hatchuel, Armand, and Benoit Weil. "A new approach of innovative Design: an introduction to CK theory." In DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm. 2003.

[10] A. R. Jensen, "The g Factor: The Science of Mental Ability." Choice Reviews Online 36, no. 04 (December 1, 1998): 36–2443–36–2443. doi:10.5860/choice.36-2443.

[11] Georgiou, George K., Kan Guo, Nithya Naveenkumar, Ana Paula Alves Vieira, and J.P. Das. "PASS Theory of Intelligence and Academic Achievement: A Meta-Analytic Review." Intelligence 79 (March 2020): 101431. doi:10.1016/j.intell.2020.101431.

[12] Rohde, Treena Eileen, and Lee Anne Thompson. "Predicting Academic Achievement with Cognitive Ability." Intelligence 35, no. 1 (January 2007): 83–92. doi:10.1016/j.intell.2006.05.004.

[13] King, Ronnel B. "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 (2016): 60-71. doi:10.1016/j.lindif.2016.10.006.

[14] Francis, Becky, and Christine Skelton. "Reassessing Gender and Achievement" (November 22, 2005). doi:10.4324/9780203412923.

[15] Lupart, Judy L., Elizabeth Cannon, and Jo Ann Telfer. "Gender Differences in Adolescent Academic Achievement, Interests, Values and Life‐role Expectations." High Ability Studies 15, no. 1 (September 2004): 25‐42. doi:10.1080/1359813042000225320..

[16] Mensah, Fiona K., and Kathleen E. Kiernan. "Gender Differences in Educational Attainment: Influences of the Family Environment." British Educational Research Journal 36, no. 2 (April 2010): 239–260. doi:10.1080/01411920902802198.

[17] Di Fabio, Annamaria, and Lara Busoni. "Fluid Intelligence, Personality Traits and Scholastic Success: Empirical Evidence in a Sample of Italian High School Students." Personality and Individual Differences 43, no. 8 (December 2007): 2095–2104. doi:10.1016/j.paid.2007.06.025.

[18] Kuhfeld, Megan, Elizabeth Gershoff, and Katherine Paschall. "The Development of Racial/ethnic and Socioeconomic Achievement Gaps During the School Years." Journal of Applied Developmental Psychology 57 (July 2018): 62–73. doi:10.1016/j.appdev.2018.07.001.

[19] Perreira, Krista M., Kathleen Mullan Harris, and Dohoon Lee. "Making It in America: High School Completion by Immigrant and Native Youth." Demography 43, no. 3 (August 1, 2006): 511–536. doi:10.1353/dem.2006.0026.

[20] Qin, Desiree Baolian. "The Role of Gender in Immigrant Children's Educational Adaptation." Current Issues in Comparative Education 9, no. 1 (2006): 8-19. doi:10.1177/000312240807300507.

[21] Lei, Jing, and Yong Zhao. "Technology Uses and Student Achievement: A Longitudinal Study." Computers & Education 49, no. 2 (September 2007): 284–296. doi:10.1016/j.compedu.2005.06.013.

[22] Salomon, Adi, and Yifat Ben-David Kolikant. “High-School Students’ Perceptions of the Effects of Non-Academic Usage of ICT on Their Academic Achievements.” Computers in Human Behavior 64 (November 2016): 143–151. doi:10.1016/j.chb.2016.06.024.

[23] Kubey, Robert W., Michael J. Lavin, and John R. Barrows. “Internet Use and Collegiate Academic Performance Decrements: Early Findings.” Journal of Communication 51, no. 2 (June 1, 2001): 366–382. doi:10.1111/j.1460-2466.2001.tb02885.x.

[24] Fan, Xitao, and Michael Chen. "Parental involvement and students' academic achievement: A meta-analysis." Educational psychology review 13, no. 1 (2001): 1-22. doi:10.1023/A:1009048817385.

[25] Gilar-Corbi, Raquel, Pablo Miñano, Alejandro Veas, and Juan-Luis Castejón. “Testing for Invariance in a Structural Model of Academic Achievement Across Underachieving and Non-Underachieving Students.” Contemporary Educational Psychology 59 (October 2019): 101780. doi:10.1016/j.cedpsych.2019.101780.

[26] Benner, Aprile D., Alaina E. Boyle, and Sydney Sadler. “Parental Involvement and Adolescents’ Educational Success: The Roles of Prior Achievement and Socioeconomic Status.” Journal of Youth and Adolescence 45, no. 6 (February 5, 2016): 1053–1064. doi:10.1007/s10964-016-0431-4.

[27] Hill, Nancy E., and Stracie A. Craft. "Parent-school involvement and school performance: Mediated pathways among socioeconomically comparable African American and Euro-American families." Journal of educational psychology 95, no. 1 (2003): 74. doi:10.1111/j.0963-7214.2004.00298.x.

[28] Sirin, Selcuk R. “Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research.” Review of Educational Research 75, no. 3 (September 2005): 417–453. doi:10.3102/00346543075003417.

[29] Papay, John P., Richard J. Murnane, and John B. Willett. “Income-Based Inequality in Educational Outcomes.” Educational Evaluation and Policy Analysis 37, no. 1_suppl (May 2015): 29S–52S. doi:10.3102/0162373715576364.

[30] Steinmayr, Ricarda, Felix C. Dinger, and Birgit Spinath. "Parents’ education and children's achievement: The role of personality." European Journal of Personality 24, no. 6 (2010): 535-550. doi:10.1002/per.755.

[31] Tesfagiorgis, Mussie, Samuel Tsegai, Tedros Mengesha, Jana Craft, and Mussie Tessema. “RETRACTED: The Correlation Between Parental Socioeconomic Status (SES) and Children’s Academic Achievement: The Case of Eritrea.” Children and Youth Services Review 116 (September 2020): 105242. doi:10.1016/j.childyouth.2020.105242.

[32] Tomul, Ekber, and Havva Sebile Savasci. “Socioeconomic Determinants of Academic Achievement.” Educational Assessment, Evaluation and Accountability 24, no. 3 (May 2, 2012): 175–187. doi:10.1007/s11092-012-9149-3.

[33] Hoxby, C. M. “The Effects of Class Size on Student Achievement: New Evidence from Population Variation.” The Quarterly Journal of Economics 115, no. 4 (November 1, 2000): 1239–1285. doi:10.1162/003355300555060.

[34] Krueger, A. B. “Experimental Estimates of Education Production Functions.” The Quarterly Journal of Economics 114, no. 2 (May 1, 1999): 497–532. doi:10.1162/003355399556052.

[35] Wößmann, Ludger, and Martin West. “Class-Size Effects in School Systems Around the World: Evidence from Between-Grade Variation in TIMSS.” European Economic Review 50, no. 3 (April 2006): 695–736. doi:10.1016/j.euroecorev.2004.11.005.

[36] Leithwood, Kenneth, and Doris Jantzi. “A Review of Empirical Evidence About School Size Effects: A Policy Perspective.” Review of Educational Research 79, no. 1 (March 2009): 464–490. doi:10.3102/0034654308326158.

[37] Gershenson, Seth, and Laura Langbein. “The Effect of Primary School Size on Academic Achievement.” Educational Evaluation and Policy Analysis 37, no. 1_suppl (May 2015): 135S–155S. doi:10.3102/0162373715576075.

[38] Schneider, Mark. “Do School Facilities Affect Academic Outcomes?,” National Clearinghouse for Educational Facilities and Educational Resources Information Center, Washington DC, 2002.

[39] Woolner, Pamela, Elaine Hall, Steve Higgins, Caroline McCaughey, and Kate Wall. “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, no. 1 (February 2007): 47–70. doi:10.1080/03054980601094693.

[40] Aaronson, Daniel, Lisa Barrow, and William Sander. “Teachers and Student Achievement in the Chicago Public High Schools.” Journal of Labor Economics 25, no. 1 (January 2007): 95–135. doi:10.1086/508733.

[41] Rockoff, Jonah E. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” American Economic Review 94, no. 2 (April 1, 2004): 247–252. doi:10.1257/0002828041302244..

[42] Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. “Teachers, Schools, and Academic Achievement.” Econometrica 73, no. 2 (March 2005): 417–458. doi:10.1111/j.1468-0262.2005.00584.x..

[43] Wayne, Andrew J., and Peter Youngs. “Teacher Characteristics and Student Achievement Gains: A Review.” Review of Educational Research 73, no. 1 (March 2003): 89–122. doi:10.3102/00346543073001089.

[44] Lee, Se Woong. "Pulling Back the Curtain: Revealing the Cumulative Importance of High-Performing, Highly Qualified Teachers on Students' Educational Outcome." Educational Evaluation and Policy Analysis 40, no. 3 (April 20, 2018): 359–381. doi:10.3102/0162373718769379.

[45] Papamitsiou, Zacharoula K., and Anastasios A. Economides. "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence." J. Educ. Technol. Soc. 17, no. 4 (2014): 49-64. Available online: http://www.jstor.org/stable/jeductechsoci.17.4.49. (accessed on May 2021).

[46] Costa-Mendes, Ricardo, Tiago Oliveira, Mauro Castelli, and Frederico Cruz-Jesus. "A Machine Learning Approximation of the 2015 Portuguese High School Student Grades: A Hybrid Approach." Education and Information Technologies 26, no. 2 (September 5, 2020): 1527–1547. doi:10.1007/s10639-020-10316-y.

[47] Cruz-Jesus, Frederico, Mauro Castelli, Tiago Oliveira, Ricardo Mendes, Catarina Nunes, Mafalda Sa-Velho, and Ana Rosa-Louro. "Using Artificial Intelligence Methods to Assess Academic Achievement in Public High Schools of a European Union Country." Heliyon 6, no. 6 (June 2020): e04081. doi:10.1016/j.heliyon.2020.e04081.

[48] Miguéis, V.L., Ana Freitas, Paulo J.V. Garcia, and André Silva. "Early Segmentation of Students According to Their Academic Performance: A Predictive Modelling Approach." Decision Support Systems 115 (November 2018): 36–51. doi:10.1016/j.dss.2018.09.001.

[49] Musso, Mariel Fernanda, Carlos Felipe Rodríguez Hernández, and Eduardo C. Cascallar. "Predicting key educational outcomes in academic trajectories: a machine-learning approach." (2020). doi:10.1007/s10734-020-00520-7.

[50] Mengash, Hanan Abdullah. "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems." IEEE Access 8 (2020): 55462–55470. doi:10.1109/access.2020.2981905..

[51] Sorensen, Lucy C. "'Big Data' in Educational Administration: An Application for Predicting School Dropout Risk." Educational Administration Quarterly 55, no. 3 (September 27, 2018): 404–446. doi:10.1177/0013161x18799439.

[52] Murphy, K. P., Machine Learning: A probabilistic perspective. MIT Press, (2012).

[53] Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. MIT press, 2018.

[54] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" Second Edition. Springer, (2008).

[55] Bergstra, James, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. "Algorithms for hyper-parameter optimization." Advances in neural information processing systems 24 (2011): 2546–2554.

[56] Bergstra, James, and Yoshua Bengio. "Random search for hyper-parameter optimization." Journal of machine learning research 13, no. 2 (2012).

[57] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[58] Breiman, Leo. "Bagging Predictors." Machine Learning 24, no. 2 (August 1996): 123–140. doi:10.1007/bf00058655..

[59] Amit, Yali, and Donald Geman. "Shape Quantization and Recognition with Randomized Trees." Neural Computation 9, no. 7 (October 1, 1997): 1545–1588. doi:10.1162/neco.1997.9.7.1545.

[60] Smola, A.J., Schölkopf, B., "A tutorial on support vector regression," Statistics and Computing, vol. 14, no. 3, (August 2004): 199–222. doi:10.1023/B:STCO.0000035301.49549.88.

[61] Rivas-Perea, Pablo, Juan Cota-Ruiz, David Garcia Chaparro, Jorge Arturo Perez Venzor, Abel Quezada Carreón, and Jose Gerardo Rosiles. "Support Vector Machines for Regression: A Succinct Review of Large-Scale and Linear Programming Formulations." International Journal of Intelligence Science 03, no. 01 (2013): 5–14. doi:10.4236/ijis.2013.31002..

[62] C. Bishop, Christopher M. "Pattern recognition." Machine learning 128, no. 9 (2006).

[63] R. E. Schapire, "The Boosting Approach to Machine Learning: An Overview," in Nonlinear Estimation and Classification. Lecture Notes in Statistics, vol 171, D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, Eds. New York: Springer, (2003): 149–171. doi:10.1007/978-0-387-21579-2_9.

[64] Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." The Annals of Statistics 29, no. 5 (October 1, 2001). doi:10.1214/aos/1013203451.

[65] Chen, Tianqi, and Carlos Guestrin. "XGBoost." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (August 13, 2016). doi:10.1145/2939672.2939785.

[66] Mehta, Pankaj, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. "A High-Bias, Low-Variance Introduction to Machine Learning for Physicists." Physics Reports 810 (May 2019): 1–124. doi:10.1016/j.physrep.2019.03.001.

[67] Efron, Bradley, and Trevor Hastie. Computer age statistical inference. Vol. 5. Cambridge University Press, 2016.

[68] Tibshirani, Robert. "The lasso method for variable selection in the Cox model." Statistics in medicine 16, no. 4 (1997): 385-395. doi:10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

[69] Domingos, Pedro. "A unified bias-variance decomposition." In Proceedings of 17th International Conference on Machine Learning, (2000): 231-238.

[70] Geman, Stuart, Elie Bienenstock, and René Doursat. “Neural Networks and the Bias/Variance Dilemma.” Neural Computation 4, no. 1 (January 1992): 1–58. doi:10.1162/neco.1992.4.1.1.

[71] Gordon, Diana F., and Marie Desjardins. “Evaluation and Selection of Biases in Machine Learning.” Machine Learning 20, no. 1–2 (1995): 5–22. doi:10.1007/bf00993472.

[72] Caraffini, Fabio, and Anna V. Kononova. “Structural Bias in Differential Evolution: A Preliminary Study” (2019). doi:10.1063/1.5089972.

[73] Li, Yi, and Nuno Vasconcelos. “REPAIR: Removing Representation Bias by Dataset Resampling.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019). doi:10.1109/cvpr.2019.00980.

[74] Tommasi, Tatiana, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. “A Deeper Look at Dataset Bias.” Advances in Computer Vision and Pattern Recognition (2017): 37–55. doi:10.1007/978-3-319-58347-1_2.

[75] McFarland, Daniel A, and H Richard McFarland. “Big Data and the Danger of Being Precisely Inaccurate.” Big Data & Society 2, no. 2 (December 1, 2015): 205395171560249. doi:10.1177/2053951715602495.

Appendix I: Features Description

Feature	Description
Subjects	Portuguese, foreign languages, quantitative and natural sciences, human and social sciences
Retentions	Number of student retentions
Enrolments	Number of student enrolments
Gender	Feminine and masculine gender
Father nationality	Portugal, Africa, Brazil, China, East Europe, developed countries and others
Computer	The student owns a personal computer
Internet	The student has access to the internet
Job situation	The student works
Education guardian	Mother, father, himself, close relative, and guardian
Guardian job educational level	Unknown, basic, secondary, college graduation, and post-graduate
Father job educational level	Unknown, basic, secondary, college graduation, and post-graduate
Mother job educational level	Unknown, basic, secondary, college graduation, and post-graduate
Guardian job situation	Unknown, employee, unemployed, self-employed, employer, home affairs, retired, student, and other
Father job situation	Unknown, employee, unemployed, self-employed, employer, home affairs, retired, student, and other
Mother job situation	Unknown, employee, unemployed, self-employed, employer, home affairs, retired, student, and other
Guardian educational level	Unknown, no formal education, basic I, basic II, basic III, secondary, bachelor, university degree, post-graduate, master, PhD, and other
Father educational level	Unknown, no formal education, basic I, basic II, basic III, secondary, bachelor, university degree, post-graduate, master, PhD, and other
Mother educational level	Unknown, no formal education, basic I, basic II, basic III, secondary, bachelor, university degree, post-graduate, master, PhD, and other
Scholarship	No support, half support, and full support
Parish	Student's home is located in the school parish
County	Student's home is located in the school county
Family non-classic dwellings	Percentage of family non-classic dwellings that exist in the student's home parish
Collective dwellings	Percentage of collective dwellings that exist in the student's home parish
Illiteracy rate	Student home parish illiteracy rate
Post-secondary schooling rate	Student home parish post-secondary schooling rate
Primary sector importance	Student home parish primary sector activities importance
Secondary sector importance	Student home parish secondary sector activities importance
Unemployment rate	Student home parish unemployment rate
School size	Number of school students
Class size	Number of class students
Teacher professional category	School definitive permanent staff, school cluster definitive permanent staff, pedagogical zone definitive permanent staff, school non-definitive permanent staff, school cluster non-definitive permanent staff, pedagogical zone non-definitive permanent staff, and fixed-term staff
Teacher educational level	Bachelor degree, master and PhD, and other
Teacher career step	Non-existent, low, medium, and high
Teacher gender	Feminine and masculine gender
Temporary replacement	The teacher is replacing a temporarily unavailable colleague
Educative support	The teacher delivers further support to the students that are at risk of failing
Teacher age	the age of the teacher
Lecturing time	Teacher time dedicated to lecturing in hours
Non-lecturing time	Teacher time not dedicated to lecturing in hours
Educative support time	Teacher time dedicated to educative support in hours
Teacher grade	End of the year teacher grade (0-20)