Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 6, Special Issue, 2022

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

Deep Learning in Predicting High School Grades: A Quantum Space of Representation

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 paper applies deep learning to the prediction of Portuguese high school grades. A deep multilayer perceptron and a multiple linear regression implementation are undertaken. The objective is to demonstrate the adequacy of deep learning as a quantitative explanatory paradigm when compared with the classical econometrics approach. The results encompass point predictions, prediction intervals, variable gradients, and the impact of an increase in the class size on grades. Deep learning's generalization error is lower in the student grade prediction, and its prediction intervals are more accurate. The deep multilayer perceptron gradient empirical distributions largely align with the regression coefficient estimates, indicating a satisfactory regression fit. Based on gradient discrepancies, a student's mother being an employer does not seem to be a positive factor. A benign paradigm shift concerning the balance between home and career affairs for both genders should be reinforced. The deep multilayer perceptron broadens the spectrum of possibilities, providing a quantum solution hinged on a universal approximator. In the case of an academic achievement-critical factor such as class size, where the literature is neither unanimous on its importance nor its direction, the multilayer perceptron formed three distinct clusters per the individual gradient signals.

Keywords:

Data Science Applications in Education;
Secondary Education;
Social Sciences; Deep Learning;
Artificial Neural Networks;
Academic Achievement.

Article History:

Received:	15	June	2022
Revised:	17	August	2022
Accepted:	02	September	2022
Published:	10	September	2022

1- Introduction

An artificial neural network (ANN) [1] is a machine learning algorithm built upon simple interconnected processing units, known as artificial neurons or nodes. The nodes are displayed in layers, allowing efficient parallel and distributed processing of knowledge and information [1]. Deep learning has been one of the most important developments in computer science in the last decade. It consists of using neural networks with at least two hidden layers to address various problems in different domains. The widespread use of powerful hardware and graphics processing units has allowed the construction of learning systems with numerous parameters trained on large datasets. Although affecting a plethora of domains, the most relevant contributions of deep learning have appeared in the computer vision, speech recognition, natural language processing, and robot control domains [2, 3].

Most recently, the widespread use of the internet, e-learning platforms, educational software, and the establishment of public education systems' databases have generated a substantial increase in the availability of educational data. In fact, Internet-based education systems have allowed the accumulation of enormous amounts of digital data from different sources, formats, and granularities, inducing the emergence of the Learning Analytics (LA) and Educational Data Mining (EDM) fields in the mid-2000s [4]. Deep learning is already being applied in education to predict and explain students'

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

DOI: http://dx.doi.org/10.28991/ESJ-2022-SIED-012

academic achievement (AA). EDM and LA often use neural networks to study educational realities and extract valuable knowledge from digital platform data. Learning systems with the ability to anticipate students at risk of failing are a promising development for improving learning contexts and academic attainment. However, there still seems to be an ongoing preference for traditional methods such as multiple linear regression [5]. On the other hand, EDM and LA develop extensive knowledge models suitable for predictive analysis alone. These models do not have the traditional explanatory nature built upon the measurement of the literature-based AA determinants [6]. To the best of the authors' knowledge, there is no educational econometrics study that has considered deep learning as the explanatory quantitative method. This article aims to fill this important scientific gap.

The study of the determinants of AA is crucial to promoting accurate educational policies. Moreover, the success of a country's education system can leverage the entire nation's wealth [7]. Promoting an improvement in the conceptual framework or in the quantitative approach that supports it is a meaningful and necessary breakthrough. Applying deep learning to infer relationships between concepts is not the same as using it for purely predictive purposes. Since deep learning is based on a universal approximator, the vast underlying parameters make interpretation and knowledge retention more challenging. It is necessary to ensure that developments in scientific experimentation do not bring any spurious complications. Any added complexity will lead to a better approximation of the reality under scrutiny. Thus, deciphering the deep learning black box is a valuable scientific undertaking [8]. In this study we address this challenging task by computing the deep gradients for each variable-observation pair and comparing their distributions with the traditional βs of the multilinear regression.

The adoption of deep learning as an experimental approach in educational and social sciences alike has remarkable advantages beyond its predictive capacity. The paradigm does not depend on a specific mathematical form to express relationships between concepts and has a particular aptitude to represent social phenomena whose heterogeneity is paramount [9]. The treatment of conceptual heterogeneity is undertaken naturally and spontaneously. By widening the spectrum of possibilities, deep learning introduces a capacity to anticipate nonconformities, which induces the search for fairer and more equitable policies. Any policy measure that brings about changes in the critical factors of AA is evaluated within the heterogeneous spectrum of both the possible outcomes and the underlying gradient structure. For example, there is room for a critical factor with an average positive impact on the student's grades to have a detrimental effect in a hypothetical individual example. This study undertakes this comprehensive analysis for the critical AA factor of class size, for which the literature is unanimous on neither its importance nor its direction.

This paper aims therefore to apply deep learning to predict upper secondary students' AA, highlighting the revolutionary character of its widespread adoption. It seeks to reflect on the repercussions for the AA domain (and for the social sciences in general) resulting from the use of a paradigm that has the intrinsic ability to create a quantic space of representation of social phenomena. For this purpose, we implement deep learning and multilinear regression simultaneously to predict the upper secondary grades assigned by a Portuguese education system teacher at the end of the 2018-19 school year. The discussion that follows stems from the interpretation and comparison of the results regarding point and interval predictions, independent variables gradients, and the likely effect of a generalized increase in class size.

The remainder of the document is organized as follows: first, a review of AA literature is presented, followed by a detailed description of the methodology and the underlying algorithms. Then, the empirical results are shown and interpreted, followed by the discussion and conclusions.

2- Literature Review

In the scientific literature, AA determinants are commonly classified into student, parents, and school critical factors [10]. A thorough assessment of the conditional background induced by those three analytical axes is of utmost importance when explaining students' AA. Cognitive ability has long been considered the most essential determinant of AA [11, 12]. Not surprisingly, students' scores can be anticipated accurately from their Intelligence Quotient [13], despite the significant role that is left for other important factors [14]. When it comes to gender, females generally attain better scores in school, especially in languages, and less so in Math [15–17]. The tendency to create a negative peer view of the school activities undermines males' levels of engagement, motivation, and achievement [18]. There is a relationship between certain personality traits, such as organization and steadiness of effort, and overachievement [19].

There is an AA gap between different ethnic groups. Black students in the US are invariably bound to underperform [20]. Even though not extendable to the following generations in the US, first-generation children of African, Asian, and Hispanic origins achieve higher education levels than did their parents [21]. The AA tends to be poorer if the origin country has a low economic development level and better if the origin country is politically stable [22]. Using personal computers at school can improve AA. However, students tend to use them primarily for unhelpful leisure activities such as emailing friends and navigating the Internet [23]. There is a negative relationship between the non-academic use of information and communication technologies and student grades [24]. Greater use of internet applications is also associated with sleeping late, fatigue, class absence, and AA underperformance [25].

Parents' expectations about their children's education attainment positively affect their AA, which is more significant than a proper home structure and supervision [26]. Underachieving students are bound to benefit from good relationships between parents and school [27]. Furthermore, parental involvement seems to especially help low socioeconomic status (SES) students [28]. There is a strong positive relationship between the SES of the student's family and AA, highlighting education inequalities and the importance of resources and cultural capital [29]. The association between parents' education and AA remains even after controlling for variables associated with intelligence and personality [30], underlining the prominent role of schools in providing cultural experiences and additional stimuli that are lacking at home. In addition, having private lessons, which is associated with parental education and family income, can be decisive for students' AA [31].

There is some controversy surrounding the relationship between class size and AA in the literature. Hoxby (2000) [32] concluded that the class size effect is insignificant even for minor effects. By contrast, Krueger (1999) [33] concluded that smaller classes improve AA. The most benefitted are the minority and impoverished students. Smaller classes appear to have a favourable effect on AA in education systems, where the lecturing quality seems to be lower [34]. In a more convergent tone, smaller schools promote AA, providing the greatest benefit to students with learning difficulties and lower SES [35]. An adequate school environment and design are conducive to overachievement. Students and school stakeholders should be provided with a peaceful and comfortable learning environment with clean air and good light [36]. When introducing changes in the school environment, an inclusive design process is recommended that welcomes genuine inputs of teachers and students [37].

Lecturing ability and teacher quality are important for AA in general and influence underperforming students in particular [38]. There is a positive relationship between teachers' ability and college grades [39]. However, many measurable teacher characteristics seem to be unrelated to teacher quality, which is intrinsically linked to unobservable factors. This finding points to policies favouring teaching evaluation based on students' performances [40]. In the same line, Rivkin et al. (2005) [41] corroborated that lecturing effectiveness is undoubtedly a significant AA determinant. However, in the same study the education and teacher experience revealed only a weak effect.

The LA/EDM field is a predictive branch of the AA domain that uses machine learning to disclose relevant behaviour patterns embedded in the educational databases. The increase of LA/EDM research is an ongoing process. However, there are only a few regression studies, as most of them are designed to solve classification and clustering problems [42, 43]. Normally, the LA/EDM learning systems resort to socio-demographic variables, digital log data, and course assignment scores to anticipate the students' AA. They are extensive knowledge models appropriate for predictive but not explanatory analysis [6]. It has also been proved that ANN performs among the best when predicting grades [44]. Table 1 shows a representative set of the studies that use ANN in the experimental phase. It is worth mentioning that our research goes far beyond their scope and depth. For instance, none of those in Table 1 involves estimating prediction intervals, the computation of the deep learning gradients, and further analysis of the results of political measures.

Table 1. Artificial neural network studies

Authors	Dataset	Machine learning algorithms	Output variable
Feng et al. (2022) [45]	Three datasets of 46, 61, and 51 university student records	K-means clustering, discriminant analysis, and convolutional neural network	Four categories of AA
Nabil et al. (2021) [46]	4,266 university students and 12 features	Deep neural net, decision tree, logistic regression, support vector classifier, k-nearest neighbour	Fail the course
Al-Tameem et al. (2021) [47]	2013-2014 data from two virtual social sciences modules	Spearman's correlation and deep neural net	Fail the module
Costa-Mendes et al. (2020) [48]	362,127 high school grades	Multilinear regression, random forest, support vector machine, artificial neural network, and extreme gradient boosting machine stacking ensemble	High school grades
Cruz-Jesus et al. (2020) [10]	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
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
Aydoğdu (2020) [51]	3,518 students and 22,979 grades	Artificial neural network classifier	Successful or unsuccessful
Li et al. (2019) [52]	480 students of an online course	Support vector machine, artificial neural network, naïve Bayesian and decision tree classifiers	Low, middle, high grades classes
Altaf et al. (2019) [53]	900 students	Artificial neural network classifier	Needs assistance and does not need assistance
Lau et al. (2019) [54]	1,085 university students	Artificial neural network regressor	Cumulative grade point average
Arunachalam & Velmurugan, (2018) [55]	1,300 undergraduate students	Artificial neural network, probabilistic neural network, and evolutionary neural network classifiers	First, second, third, and fail
Mondal & Mukherjee (2018) [56]	480 students	Artificial neural network, deep neural network, and recurrent neural network classifiers	High, medium, and low classes

3- Methodology

Supervised learning involves learning a function that maps an input to an output based on a set of input-output pairs. The function is inferred from labelled data consisting of training examples. Each example is a pair of an input vector and an output value in supervised learning, also called a supervision signal. Each component of the input vector corresponds to a feature or attribute. A supervised learning algorithm analyses the training data and infers a function to be used to map new examples. The learned function should accurately anticipate the class labels in case of classification or the numeric target variable in case of regression. The learning algorithm should have the statistical quality of properly generalizing from training to unseen data [57].

The dataset was split into 60% for training, 20% for validation, and 20% for testing. All the variables were standardized. The deep multilayer perceptron (MLP) implementation includes training and test performance statistics, test prediction intervals, training, test gradients, and the analysis of the effects on grades of a class size increase. The multilinear regression (MLR) implementation does not include the computation of gradients because they coincide with the regression coefficients. The core of the experimental phase involved eight main steps. The first consisted of a feature selection procedure based on the Lasso regression algorithm. The multilinear regression results were computed in both the training and test sets in the second step. In the third step a thorough architecture-topology and hyperparameter optimization procedure of the deep MLP was undertaken. Next, the deep MLP was trained. Then the deep MLP test prediction intervals were calculated. In the sixth step, the training and test gradients were determined. Finally, the seventh and eighth step comprised predicting class size effects on grades for both MLR and deep MLP. Figure 1 displays the research methodology followed in this study.

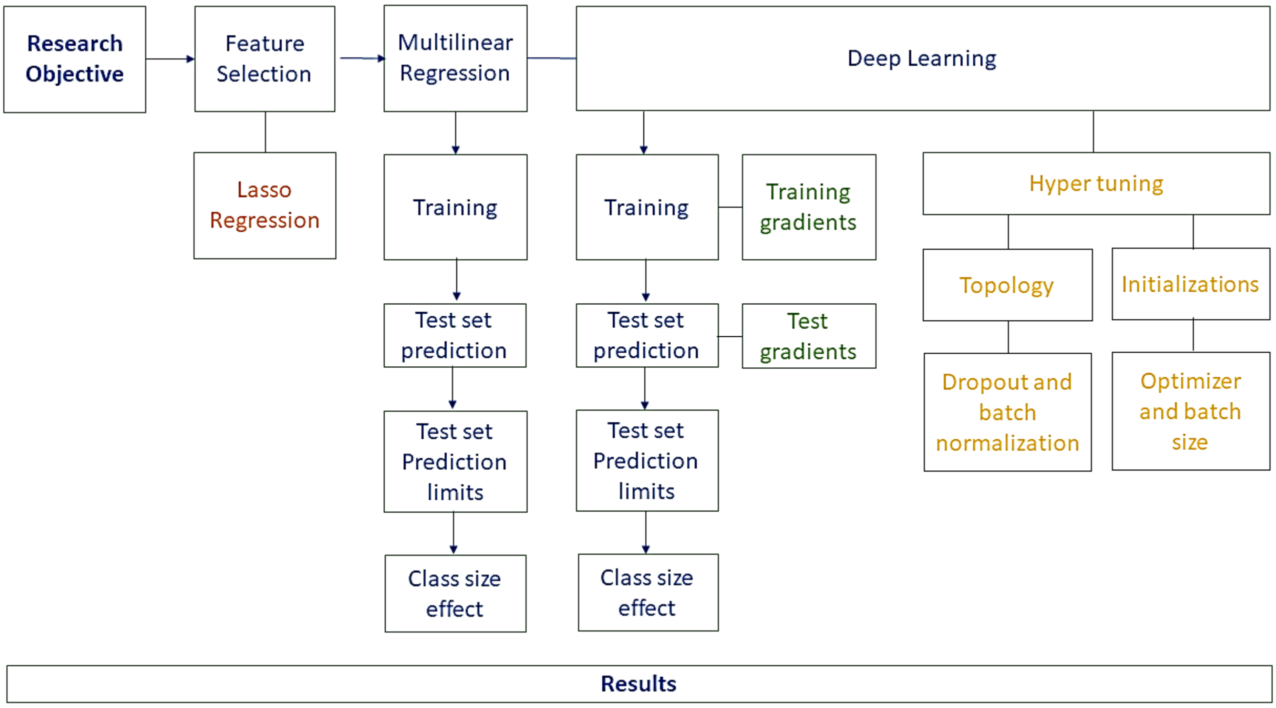

Figure 1. Methodological steps

3-1- Multilinear Regression

MLR establishes a linear relationship between a dependent variable to be explained and predicted and a set of independent variables. It provides easily interpretable results by imposing important restrictions. The error terms are assumed to be independent of one another, homoscedastic, and with a null mean. The model and the individual statistical significance tests of the coefficients β_i presuppose that the error term follows a gaussian distribution. The ordinary least squares method was used in the learning phase, and the model parameters were estimated from the training and validation set. The point and interval predictions of the test set followed the standard practice [58]. The implementation was based on the statsmodels python library [59].

3-2- Deep Multilayer Perceptron

The MLP stems from the perceptron model [60], capable of solving linearly separable classification problems. The MLP architecture adds hidden layers between the input and output layers. The number of nodes of the input layer equals the number of the input variables. It is a feed-forward topology, as the connections between nodes are established from lower to upper layers, and no connections exist between nodes of the same layer. Each connection is assigned a weight. The input of every node in any hidden layer or the output layer is a weighted average of the nodes' outputs of the preceding layer plus a bias. The input is transformed through a nonlinear activation function in a new signal propagated forward up to the output layer [61]. The theoretical analysis of an MLP is not an easy task as the nonlinearity of the distributed processing and the high connectivity enlarge the optimization search space to numerous possible representations of the input patterns by the hidden nodes. The task becomes even more difficult in the case of a deep MLP with several large hidden layers. The learning phase of an MLP consists of optimizing the weights and biases to minimize the gap between the network output and the target. This optimization is carried out by the backpropagation algorithm [62] combined with gradient descent techniques. The learning process has two phases. In the forward phase the signals are propagated from lower to upper layers up to the output layer, and the weights and biases remain unchanged. In the backward phase the network error is first computed and then propagated backward layer by layer, inducing the weights and the biases to change in the direction determined by the gradient of the loss function. The learning phase is considered successful as it reaches a configuration of the weights that results in an acceptable value of the loss function [61, 63, 64].

The implementation was developed using Keras [65], which is a deep learning API written in Python, running on top of TensorFlow (an end-to-end machine learning platform). It was developed with a focus on enabling fast experimentation and is characterized by flexibility and scalability. In fact, as stated in the Keras documentation, it is possible to run Keras on large clusters of GPUs, and export Keras models to run in the browser or on a mobile device.

3-2-1- Layer weight initializers

The assignment of initial node weights is done just before the learning phase. Insignificant initial weights tend to produce vanishing backward propagated weights. Large initial weights can induce exploding gradients [66]. The hyper-tunning procedure encompassed four weight initialization alternatives: the random normal initializer with mean zero and standard deviation of 0.05, the random uniform initializer between -0.05 and 0.05, the normalized random uniform initializer, and the normalized random normal initializer [67]. In terms of biases, the ones initializer, activating every node in the deep MLP, and the most commonly used zeros initializer were included.

3-2-2- Activation Function

The ANN can learn very complex patterns due to both the nonlinearity of the activation function and the existence of hidden layers. The activation function of the hidden layers is the Rectified Linear Unit function $(f(x) = \max(x, 0))$ as it typically enhances learning in networks with many layers [68]. The output layer has no activation function.

3-2-3- Dropout

Dropout is a regularization technique that randomly and temporarily stops training some nodes and their interconnections. Dropout regularization can be compared to model ensembles without the explicit need of creating multiple learners [69]. The ANN generalization ability is enhanced because it avoids adapting weights to overfit the training set. To keep the mean weight unchanged between training and testing, the weights for unseen data come as follows [70]:

$$Wud_i = W_i \,.\, p \tag{2}$$

where p is the dropout rate, the probability of not training the node.

The hyper-tuning phase evaluated a dropout layer with different dropout rates after any hyper-tuned dense layer. A dropout layer sets the inputs to zero according to the dropout rate.

3-2-4- Batch Size and Batch Normalization

The batch size corresponds to the number of observations considered in the forward step of the backpropagation before updating the network's weights [71]. There is a trade-off between the computation cost and the ANN accuracy in batch size. A larger batch size induces a more straightforward computation but poorer ANN accuracy. The default batch size of 32 examples was used [72]. The batch size search space was built from powers of two [71].

Batch normalization consists of normalizing the layer inputs for each training batch to maintain its mean close to zero and its standard deviation close to 1. With the distributions of the layer inputs stabilized, the optimizer is less prone to lead to layer saturations, accelerating learning, reducing the importance of the weight initialization, and eliminating the need for dropout [73]. The option of having a batch normalization layer before the activation was evaluated in the hyper-tuning phase.

3-2-5- Optimizers

The hyper-search includes a representative set of various versions of the gradient descent optimizer to analyse which better suits the data's convergence and pattern.

3-2-5-1- Mini-Batch Gradient Descent with Momentum

The batch gradient descent [3] with moment updates the weights and biases w_t for a learning rate η and a momentum hyperparameter β are as following:

$$w_t = w_{t-1} - \eta \, . V_{dw_t} \tag{2}$$

where;

$$V_{dw_t} = \beta . V_{dw_{t-1}} + (1 - \beta) \, . \frac{\partial L}{\partial w_{t-1}} \tag{3}$$

The momentum addition in the gradient function makes the actual gradient dependent on the previous gradient, accelerating convergence and avoiding excessive oscillation.

An epoch is a complete pass through the entire training set. In the batch gradient descent, there is one update per epoch. In the case of the mini-batch version, as the internal parameters are updated for every successive subsample of the training data, there are several updates per epoch. In the case of the stochastic version, the update is undertaken for every single example. The mini-batch version comes up as a good compromise between the large gradient oscillation of the stochastic version that demands lower learning rates and more time to converge and the computation cost of the batch version that computes the gradients for the entire training set at once.

In the hyper-tuning phase, its adoption was evaluated for different learning rates and momentum coefficients.

3-2-5-2- Root Mean Square Propagation (RMSprop)

The gradients of each weight or bias w_t can differ substantially, making it hard to find a proper single learning rate that fits every case. Higher gradients should correspond to lower learning rates in terms of convergence and efficiency. The RMSprop is based on the mini-batch gradient descent and introduces adaptive learning rates. It divides the actual mini-batch gradient by the moving average of the square of the consecutive mini-batch gradients, resulting in different learning rates for each weight:

$$w_t = w_{t-1} - \frac{\eta}{\sqrt{S_t}} \, . \frac{\partial L}{\partial w_{t-1}} \tag{4}$$

$$S_t = \varrho . S_{t-1} + (1 - \varrho) \, . \left(\frac{\partial L}{\partial w_{t-1}} \right)^2 \tag{5}$$

In the hyper-tuning phase, its inclusion was evaluated for different ϱ values.

3-2-5-3- Adaptive Moment Estimation (Adam)

The Adam optimizer [74] uses adaptive learning rates and momentum. As sparse features are bound to generate sparse gradients, their learning rates should be higher. The adaptive learning rates allow different feature learning rates based on the sum of squares of their previous gradients.

The Adam optimizer updates the weights and biases w_t for a core learning rate η, a momentum hyperparameter β_1 and an adaptive learning rate hyperparameter β_2:

$$w_t = w_{t-1} - \frac{\eta}{\sqrt{\hat{S}_t + \varepsilon}} \, . \hat{V}_{dw_t} \tag{6}$$

$$V_{dw_t} = \beta_1 . V_{dw_{t-1}} + (1 - \beta_1) \, . \frac{\partial L}{\partial w_{t-1}} \tag{7}$$

$$\hat{V}_{dw_t} = \frac{V_{dw_t}}{1 - \beta_1{}^t} \tag{8}$$

$$S_t = \beta_2 . S_{t-1} + (1 - \beta_2) \, . \left(\frac{\partial L}{\partial w_{t-1}} \right)^2 \tag{9}$$

$$\hat{S}_t = \frac{S_t}{1 - \beta_2{}^t} \tag{10}$$

where $\varepsilon > 0$ avoids the null denominator case and $\hat{S}_t$ is different for each feature. As the initial values $V_{dw_0}$and S_0 are zero, the rectifications in Equations 8 and 10 recentre the exponential averages.

In the hyper-tuning phase, the default optimizer was Adam. When tuned, the core learning rate η, the momentum $\beta_1$and the adaptive β_2 hyperparameters were taken-into-account.

3-2-5-4- Learning Rate Schedule

Scheduling the learning rate consists of reducing it as the training goes. Sometimes it is called annealing rate because it allows both higher weights variance in the beginning to avoid local minima and lower variance in the final epochs, enhancing the likelihood of convergence [72, 75].

In the hyper-tuning phase, an exponential learning rate schedule was put forward:

$$\eta_s = \eta_{s-1}.\varepsilon \tag{11}$$

where $\varepsilon \subset]0,1]$ and s is the number of steps completed every $d\%$ of total batches.

3-3- Feature Selection

Lasso multilinear regression [76] introduces an L1 regularization in the MLR model, penalizing the magnitude of the regression coefficients. As the shrinkage pressure increases, the resulting model is likely to be simpler and sparser. In a feature selection procedure, the variables that have null $\widehat{\beta_j}$ are dropped as they are considered unimportant for the explanation of the target variable.

$$\left(\widehat{\alpha},\bar{\bar{\beta}},\lambda\right) = arg\ min\left\{\sum_{i=1}^{N}\left(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda.\sum_{j=1}^{p}\left|\beta_j\right|\right\} \tag{12}$$

n the feature selection phase, the choice of the regularization factor λ was carried out through a four-fold cross-validation search grid. The λ of the feature selection model is the highest, which allows the loss function to be less than the optimum plus its standard deviation.

3-4- Hyper-Tuning

The hyperparameters selection was divided into three steps. The first was based on the hyperband optimization method, whereas the other two were based on the Bayesian optimization method.

The hyperband optimization algorithm was used to select the deep MLP topology. The aim was to include as many deepness and width combinations as possible and simultaneously follow a reasonable computation budget.

3-4-1- Hyperband Optimization

The hyperband optimization [77] speeds up the random search algorithm [78] (commonly used for hyper-parameter optimization) by introducing an adaptive mechanism and an early stopping system. For the same computation budget, these two components allow the algorithm to look at more possible configurations with respect to traditional hyper-parameter optimization approaches. The hyperband undertakes a grid search for n possible configurations. Each grid search iteration is called a bracket and includes a complete run of the Successive Halving algorithm [79].

The schedule used is described in Table 2. The Max-epochs refer to the maximum iterations per configuration and the Factor to the configuration down-sampling rate.

Table 2. Hyperband Schedule

Brackets									
4		3		2		1		0	
n n_i	r_i	n_i	r_i	n_i	r_i	n_i	r_i	n_i	r_i
81	1	27	2	9	6	3	15	1	37
27	3	9	6	3	18	1	45		
9	9	3	18	1	54				
3	27	1	54						
1	81								

Hyperparameters

Max-epochs	200	**Factor**	3
n_i	# Configurations		
r_i	# Units of computation		

3-4-2- Bayesian Optimization

Bayesian optimization uses the Bayes Theorem to estimate an acquisition function that determines the spatial location of the following search. The acquisition function represents a formal trade-off between exploration, high variance areas of the surrogate objective function with insufficient posterior information, and exploitation, areas for which posterior information points to adequate objective function values. It is cost-efficient because it minimizes the demand for configuration evaluations and suits non-convex optimization problems [80].

In the hyper-tuning phase the Bayesian optimization maximum of trials was set to 200.

3-5- Deep MLP Prediction Intervals

Let us suppose a target random variable y as follows [81]:

$$y_i = F(x_i;\theta) + \varepsilon_i \tag{13}$$

where x is a vector of independent variables and ε is a term of stochastic noise, mean μ, and finite variance δ^2.

Supposing a statistical learning model $F\left(x_i;\hat{\theta}_{ts}\right)$ and a training set TS, the Equation 13 comes as,

$$y_i = F\left(x_i;\hat{\theta}_{ts}\right) + F(x_i;\theta) - F\left(x_i;\hat{\theta}_{ts}\right) + \varepsilon_i \tag{14}$$

$$y_i = F\left(x_i;\hat{\theta}_{ts}\right) + M_{ts}(x_i;\theta) + \varepsilon_i \tag{15}$$

where $M_{ts}(x_i;\theta)$ is the model error.

The prediction for any unseen example x_0 is:

$$y_0 = F(x_0;\theta) + \varepsilon_0 \tag{16}$$

$$y_0 = F\left(x_0;\hat{\theta}_{ts}\right) + F(x_0;\theta) - F\left(x_0;\hat{\theta}_{ts}\right) + \varepsilon_0 \tag{17}$$

$$y_0 = F\left(x_0;\hat{\theta}_{ts}\right) + M_{ts}(x_0;\theta) + \varepsilon_0 \tag{18}$$

To arrive at the prediction interval of y_0 the model prediction $F\left(x_0;\hat{\theta}_{ts}\right)$ is first computed and then both the distribution of the model error $M_{ts}(x_0;\theta) = F(x_0;\theta) - F\left(x_0;\hat{\theta}_{ts}\right)$ and the distribution of the stochastic error ε_0 are studies.

3-5-1- Model Error

A bootstrap can approximate the model distribution. Let us draw with replacement m random successive subsamples j of the training set TS and fit a model on each of them. The bootstrap predictions on a validation set VS can be denoted as follows:

$$B_{i,j} = F\left(x_{vs,i};\hat{\theta}_j\right) \tag{19}$$

The mean of the bootstrap distribution $\bar{B}_{i,m}$ converges to the true mean of the model:

$$\bar{B}_{i,j} = \frac{\sum_{j=1}^{m} F\left(x_{vs,i};\hat{\theta}_j\right)}{m} \tag{20}$$

In turn, the empirical distribution of the centred bootstrap samples converges to the distribution of the model error $M_{ts}\left(x_{vs,i};\theta\right)$:

$$m_{ij} = B_{i,j} - \bar{B}_{i,j} \tag{21}$$

3-5-2- Stochastic Error

The distribution of the stochastic error can be approximated by the distribution of the residuals ε projected on the validation set:

$$\hat{\varepsilon}_{vs,i} = y_{vs,i} - F\left(x_{vs,i};\hat{\theta}_{ts}\right) \tag{22}$$

3-5-3-Total Error

The set T of the empirical distribution of the total error projected in the validation set comes as:

$$T = \{t_{i,j} = m_{ij} + \hat{\varepsilon}_{vs,i}, i = 1,2, \dots, h; j = 1,2, \dots, m\} \tag{23}$$

where h is the validation set dimension.

For a level of significance of $\alpha = 5\%$, the 2.5% and the 97.5% quantiles Q of the T set were taken to build the prediction intervals,

$$PI_{\alpha}(x_0) = F(x_0; \hat{\theta}_{ts}) + (Q_{2,5\%}, Q_{97,5\%}) \tag{24}$$

In the experimental phase the trained deep MLP was used as the statistical learning model $F(x_i; \hat{\theta}_{ts})$ and the bootstrap was carried out as a subsequent fine-tuning. The number of bootstraps samples was 200 and the number of epochs was 30.

The accuracy and adequacy of the Prediction limits were inferred from the Prediction Interval Coverage Probability (PICP) and Mean Prediction Interval Width (MPIW) as follows [82]:

$$PICP = \frac{c}{n} \tag{25}$$

where c is the number of samples of the test set whose target falls inside the prediction interval and n is the total of the test set samples.

$$MPIW = \frac{1}{n} \sum_{i}^{n} (PLu_i - PLl_i) \tag{26}$$

where PLu and PLl are the upper and the lower limit, respectively

4- Data and results

4-1- Data

The dataset comprises 673,992 grades (from 0 to 20) from Portuguese upper secondary students. The data refer to the three final high school years (10th, 11th, and 12th) and comprise 27 subjects from Portuguese and English to Physics and Math for the 2018-19 academic year (see Appendix I). A dummy variable was associated with each subject.

Regarding the proportion of years, 29% are 10th grades, 36% 11th grades, and 35% 12th grades. 53% of the grades are from girls. 14% are from half-scholarship students and 12% from those holding a full scholarship. The dataset was built from a Microsoft® SQL Server Management Studio series of queries. There are 34 features, 7 of which are from Statistics Portugal and the remainder from the Directorate-General for Statistics of Education and Science of the Portuguese Ministry of Education. The latter are essentially categorical variables creating a sparse dataset in terms of measurements of AA critical factors (see Appendix II for more details).

The one-hot encoding ended with 131 independent variables to be selected to the input space via the Lasso Regression selection procedure. The dataset was split into 404,394 observations for training, 134,799 for validation, and 134,799 for testing. The test set is a complete holdout set that did not participate in any step of the learning phase, replicating unseen data.

4-2- Results

4-2-1- Feature Selection

The feature selection Lasso procedure picked 85 of 131 available predictive variables for an optimized shrinkage pressure of 0.004 (Figure 2). The distance between the student's home and the school was considered irrelevant. The dummy variables concerning students whose fathers' nationality is from either wealthy Western countries or poor Eastern European ones were discarded as their behaviour is not significantly different from nationals. The dummy variables related to employment situation and student guardians' education level were largely discarded because they are strongly correlated with the same dummy variables that correspond to the parents. In terms of both parents' job situation, several dummies were considered irrelevant and indifferent from the base status of being employed. Here, the unemployed situation dummy passed the LASSO filter in both situations. Almost every dummy associated with the parents' education level went to the input space. In terms of scholarship, the half support dummy was discarded, not differing significantly from the non-scholarship situation. Among the Statistics Portugal socioeconomic variables, the illiteracy rate, unemployment rate, and the primary sector importance were also discarded.

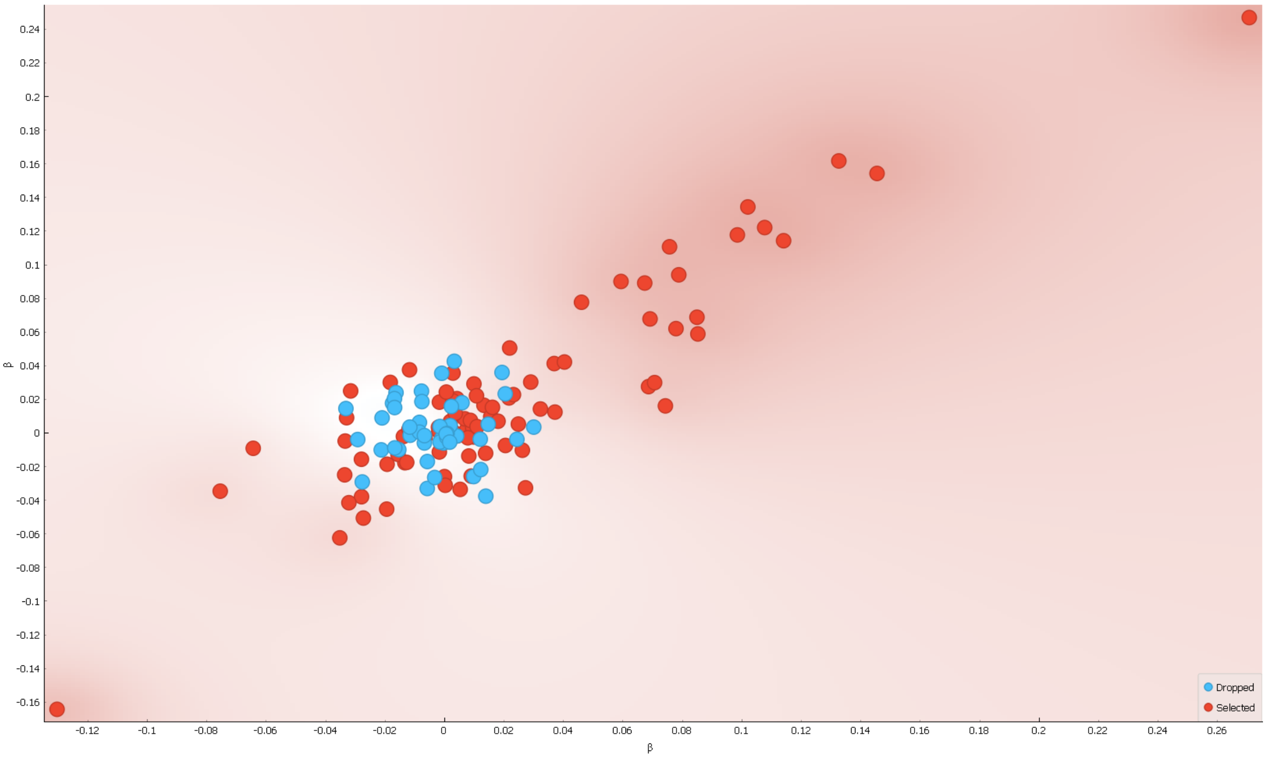

Figure 2. Feature selection

4-2-2- Hyper-Tuning

The initial deep MLP changes throughout the hyper-tuning process because it incorporates the tuning optimization of the preceding steps. For reference, the base MLP is shown in Table 3.

Table 3. Base MLP topology and hyperparameters

# layers		1
# nodes		10
Initializers	Weight	Glorot uniform
	Bias	Zeros
Dropout	Boolean	False
Batch normalization	Boolean	False
Batch size		32
Optimizer		Adam
Learning rate	η	0.001
Adam	β_1	0.9
	β_2	0.999
Learning schedule		False

4-2-2-1- Deep MLP Topology

The first step of the hyper-tuning phase was to optimize the topology of the deep MLP through a hyperband search. The search space was built according to Table 4.

Table 4. Topology search space

	Minimum	Maximum	Step
# Layers	2	20	1
# Nodes	2	50	1

The optimization results show a clear preference for topologies with depth lower than 10 hidden layers and a global size of fewer than 250 nodes (see Figure 2). On the other hand, topologies with a depth deeper than 12 tend to have a higher MAE. This outcome arises from the pattern of the data itself and not from possible divergence issues, as some deep topologies reach fair values for MAE (size and colour of the dots in Figure 3).

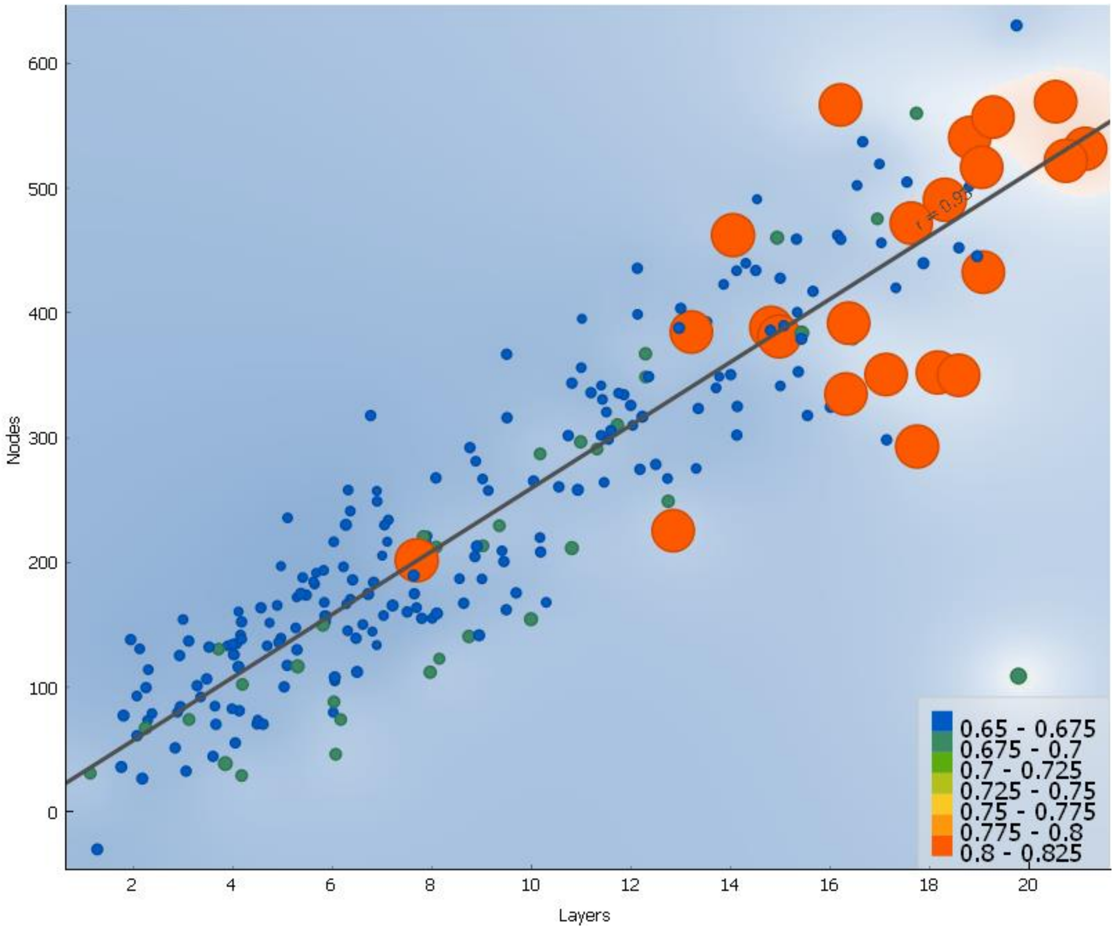

Figure 3. **Hidden layers, size, and validation MAE**

The selected topology consists of 6 hidden layers with widths of 45, 45, 3, 22, 48, and 19. The deep MLP seems to allocate the first three layers to condense the data and then the latter ones to search for the universal approximator. The MLP reduces the dimensionality of the data with an edge: it does not follow a predefined linear or kernelized mathematical transformation.

4-2-2-2- Weight and Bias Initializations, Dropout Layer, and Batch Normalization

The second step of the hyper-tuning phase consists of choosing the weight and bias initialization method, the existence of a dropout layer after each dense layer, and a batch normalization before every activation. The search space was built according to Table 5.

Table 5. **Initializations, dropout, and batch normalization search space**

Initializers	Choices		
Weight	Random uniform	Random normal	Glorot uniform
Bias	Zeros	Ones	

Dropout	Boolean
Batch normalization	Boolean

	Minimum	Maximum	Step
Dropout rate	0	0.9	0.1

The selected combination was random normal and ones for weight and bias initializations, the existence of batch normalization but no dropout. The Bayesian optimization directed the search toward areas where the weight and bias initializations were random normal (68%) and zeros (91%), respectively, and neither dropout nor batch normalization existed (83.50%). Thus, only the choices of random normal for the weight initializer and the dropout inexistence can be said to have a robust decision basis. The other choices were substantiated on a weaker stand.

4-2-2-3- Optimizer and Batch Size

The last step of hyper-tuning encompasses the batch size optimization and the optimizer choice along with the tuning of its hyperparameters: learning rate and schedule, momentum, and adaptive learning rate factors. The search space is described in Table 6.

Table 6. Optimizer and batch size

	Choices			
Optimizer	SGD	Adam	RMSprop	
Batch size	32	64	128	256
Learning schedule	Boolean			
		minimum	**maximum**	**Step**
Learning rate	η	1.E-05	0.001	
Learning schedule	d	0.1	0.5	0.1
	ε	0.7	0.95	
SGD	β	0.6	0.99	
Adam	β_1	0.6	0.99	
	β_2	0.6	0.99	
RMSprop	ϱ	0.1	0.99	

The Bayesian choices of Adam as optimizer and 64 as batch size (see Table 7) are robust because they are present in 55% and 45% of the 20 best combinations, respectively. However, the choice for a learning rate schedule is weak as half of the 20 best combinations have no learning schedule. The learning rate and the batch size increased from the default value of 0.001 and 32 to 0.00553 and 64, respectively. The surge of the learning rate is not unexpected given the existence of a learning schedule and the increase in the batch size.

Table 7. Optimizer and batch size tuning

Batch size		64
Optimizer		Adam
Learning rate	η	0.0055
Adam	β_1	0.7503
	β_2	0.8226
Learning schedule		True
Learning schedule	d	0.5
	ε	0.7514

4-2-3- Learning Results

The deep MLP presents better results than the MLR in training and testing (Table 8). The MLP MAE in training and test are 0.6357 and 0.6484, respectively, better than the MLR 0.6944 and 0.6910. The multilinear regression training set includes the validation set. Regarding the MLP, using the validation set allowed for saving the best combination of epoch weights used further to compute predictions limits and gradients. The MLP suffers from some overfitting as the test results are poorer than both the validation and the training results. Several variance reduction techniques were foreseen when optimizing the architecture and the hyperparameters, so the MLP overfitting should be interpreted as a virtuous cost associated with achieving a better generalization error.

Table 8. Learning results

	Training			**Validation**			**Test**		
	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2
MLP	0.6357	0.6414	32.68%	0.6503	0.6710	29.56%	0.6484	0.6686	29.64%
MLR	0.6944	0.7414	25.86%				0.6910	0.7357	26.10%

The deep MLP training optimization converged smoothly and reached a low variance plateau around epoch 600 (Figure 4), the weight combination that prevailed corresponds to the 799th epoch.

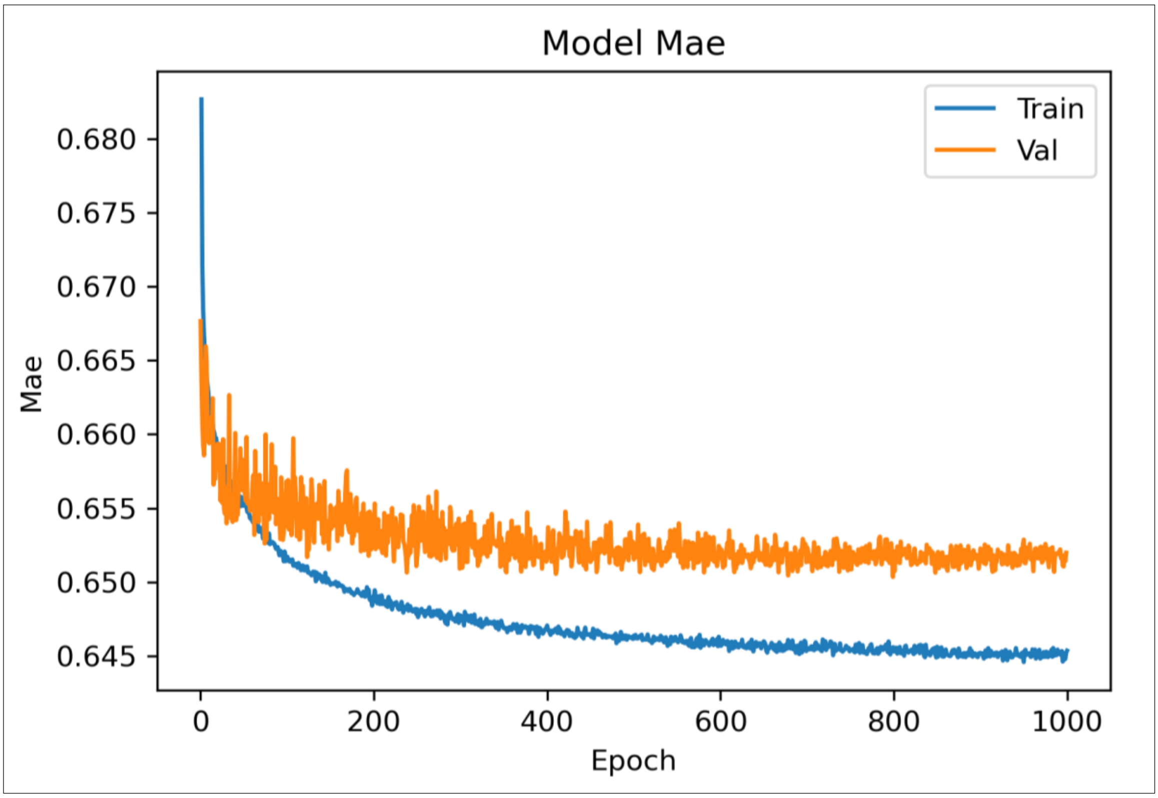

Figure 4. Deep MLP training convergence (Depicted from the Keras-Tensorflow learning history)

The predictions limits were built for a $\alpha = 5\%$, and both deep MLP and MLR have a $PICP$ greater than 95%. The deep MLP has a prediction interval 5% shorter than MLR (Table 9).

Table 9. Prediction limits

	Test	
	PICP	**MPIW**
MLP	0.9504	3.2140
MLR	0.9562	3.3760

4-2-4- Gradients Analysis

The gradients correspond to the first derivatives with respect to the input variables. In the MLR the gradients are the data-invariant βs. In deep MLP the gradients vary from data point to data point, forming a vector of βs for each input variable. The analysis consists of comparing the MLP mean $\hat{\beta}s$ with the MLR $\hat{\beta}s$ in the light of what is expected from the literature.

Inconsistencies between the MLR $\hat{\beta}s$ and the mean MLP $\hat{\beta}s$ were found in only 8 out of 85 input variables (see and follow Table 10). The guardian not being a parent or a close relative can indicate a dysfunctional family background, which is detrimental to AA. However, the MLP $\hat{\beta}s$ relative to both *Guardian is not a relative* and *Guardian is a relative but not parent* input variables have a contradicting positive signal. Regarding internet usage, the MLP $\hat{\beta}$ signal is negative and different from the positive $\hat{\beta}$ in the MLR. In the literature, the use of the Internet is reported as having both positive and negative effects on AA, depending on being directed to school activities or entertainment. Regarding the professional teacher category, it might be expected that any category below being a *definitive permanent staff member of a school* would be detrimental to AA. However, in two such cases the MLR has positive $\hat{\beta}s$ and in one such case the MLP also has a positive mean $\hat{\beta}$. The MLR and the MLP $\hat{\beta}s$ disagree in signal once again in the *collective dwellings* input variable. As some collective dwellings such as hotels and state buildings are bound to be found in high-income urban zones and others like shopping centres and hospitals can be placed in some suburban areas, the literature does not indicate a specific signal on this β.

Table 10. Differences in the gradients ($\widehat{\beta}s$)

Input variables	MLP$_{mean}$	MLP$_{std}$	MLR	Literature	MLP|MLR
Guardian is not a relative	0.0025	0.0303	-0.0056	-	+-
Guardian is a relative but not parent	0.0030	0.0425	-0.0134	-	+-
Student uses Internet	-0.0005	0.0747	0.0092	?	-+
Pedagogic zone no definitive permanent staff members	-0.0002	0.0358	0.0077	-	-+
School cluster definitive permanent staff members	-0.0097	0.1040	0.0068	-	-+
School cluster no definitive permanent staff members	0.0012	0.0197	-0.0063	-	+-
Guardian has a university degree	0.0001	0.0263	-0.0016	+	+-
Percentage of collective dwellings	-0.0034	0.0590	0.0076	?	-+

More inconsistencies between the results and the literature are noted in Table 11. The results show that students belonging to the Chinese community tend to have better grades than the natives, contradicting the literature. Other examples are the negative effect of *teacher age* and the positive effect of *teacher years to retirement* on AA. Indeed, it could be expected that more lecturing experience would result in higher grades. Curiously, female teachers are bound to assign lower grades, a finding not explicitly addressed by the literature. Fixed contract teachers tend to assign higher grades than fully permanent teachers, contradicting the notion that a teacher with a stable career is more efficient in lecturing, thereby yielding higher AA levels. The results also show that the mother being an employer is detrimental to the student's AA, contradicting the positive association between parental SES and AA to some degree. Lastly, the results also show an unequivocal negative class size effect on AA even though the literature is non-conclusive in this regard.

Table 11. Gradients ($\widehat{\beta}s$) and literature

Input variables	MLP$_{mean}$	MLP$_{std}$	MLR	Literature	MLP|MLR
Father nationality is Chinese	0.0028	0.0211	0.0092	-	++
Teacher age	-0.0244	0.0995	-0.0257	+	--
Teacher gender is female	-0.0051	0.0816	-0.0018	?	--
Teacher years to retirement	0.0056	0.1073	0.0046	-	++
Fixed term staff	0.0044	0.0808	0.0134	-	++
Mother is an employer	-0.0086	0.0269	-0.0062	+	--
Class size	-0.0196	0.1296	-0.0088	+-	--

4-2-5- Class Size Effect

It is necessary to change the test set to analyse an increase of five students in the size of the classes accordingly. In this case, the impacts on grades arise naturally from the difference between the modified and the original test set predictions.

In the MLR the impact is the same whichever test example is considered and is driven by the $\hat{\beta}$ associated with the class size. In this case, the grades of every example were down by 0.0282.

In the deep MLP every test example has assigned a specific $\hat{\beta}_i$ and the impacts on grades varied accordingly. The mean impact is down in 0.1047, and the standard deviation is 0.3747. The deep MLP anticipates, on average, a more substantial effect on grades than MLR does. The test set is split into three clusters regarding the type of impact on grades: a first cluster in which grades are predicted to improve, a second cluster in which grades are predicted to worsen, and a third cluster in which grades are predicted to remain unchanged (see Table 12).

Table 12. Class size impact clusters

Mean	-0.1047
Sd	0.3747
# Negative impact	82,872
# Positive impact	51,751
# Null impact	176
# Test set	134,799

The formation of the clusters closely followed the test set gradients even though there is a clear difference between first derivatives and differentials in the MLP framework. The analysis of the confusion matrix of Table 13 highlights that 78.46% of the impacts on grades are *per* the respective gradient signal.

Table 13. Confusion matrix: gradients and impacts on grades

	Grade variation		
Gradient	+	-	Null
+	40,797	17,995	0
-	10,845	64,794	0
Null	109	83	176

5- Discussion

The feature selection procedure solved multicollinearity problems concerning the variables measured for the guardian, the parents simultaneously, and the socioeconomic variables retrieved from Statistics Portugal. On the other hand, the regularization techniques such as dropout and batch normalization had only a minor rule in the deep MLP hyper-tuning optimization, seemingly coherent with the high-bias knowledge-intensive model in question and an inherent low variance trait [6].

In terms of efficiency, the deep MLP has better results than MLR, whichever the metrics or approach. The deep MLP generalization error is more minor in the student grades prediction, and its prediction intervals are more accurate. This added generalization ability is a hallmark of machine learning, particularly deep learning. Furthermore, the deep MLP gradients empirical distributions are primarily in line with the regression coefficients estimates of the MLR, pointing to a satisfactory MLR fit to the pattern embedded in the data. The relationship between the structure of the MLR regression coefficients and the deep MLP gradients empirical distributions corroborates the absence of significant specification distortions in the MLR, strengthening its results and inferences. There is no doubt that in the presence of a strong nonlinear pattern, the divergence between the gradient structures would undoubtedly be accentuated. In fact, the deep MLP implementation turns out to be an extremely robust way to assess the adequacy and soundness of the MLR fit.

In terms of discrepancies between the resulting gradients and what would be expected according to the literature, it should be highlighted that teachers with fixed-term contracts tend to assign higher grades than teachers with a permanent contract and school. AA seems to be negatively associated with lecturing experience, as older teachers closer to retirement tend to assign lower grades. However, care should be taken when interpreting this empirical result. Perhaps, with more lecturing experience, teachers tend to increase their stringency for excellence concerning student performance, resulting in lower grades for comparable attainments of AA. Female teachers also tend to assign lower grades, which can also be associated with stricter evaluation criteria. The AA of students belonging to the Chinese community highlights successful integration, sound economic and social endowments, and efficient support networks [83]. The mother being an employer does not seem to be a positive factor in the student's AA. This is an important empirical result because it is essential to ensure that women’s empowerment in their aspirations, objectives, undertakings, and civic participation is followed by a benign paradigm change in terms of the balance between home and career affairs for both genders. Therefore, the father should reinforce his role at home, and career demands should not follow the more aggressive patterns of Western patriarchal society.

Deep MLP broadens the spectrum of possibilities and greets each individual specificity as a core element of the phenomenon by providing a quantum solution hinged on a universal approximator. For example, there is room for a critical factor with an average positive impact on the student’s grades to have a detrimental effect in a hypothetical individual example. In the case of a critical AA factor such as class size, for which the literature is unanimous regarding neither its importance nor its direction, the MLP formed three distinct clusters per the individual gradients. The first cluster is formed by the students most likely to benefit from the increase. This cluster is followed by those most likely to be indifferent to it. The third cluster is for students most likely to be harmed by the increase. The gradients anticipate the likely response to a change in class size and therefore must be considered in decision-making processes and policy design.

Deep MLP can have a revolutionary effect on the social sciences in general and the educational sciences in particular. Deterministic mathematical functions cannot formalize social science conceptual relationships without an evident loss of explanatory and predictive power. The heterogeneity of responses to social phenomena is a pattern that should be accepted into social conceptual frameworks. Forging a quantitative basis that does not need a deterministic functional assumption and welcomes high levels of heterogeneity is a decisive breakthrough clearly adequate for the complexity of social phenomena. The aim is not to increase complications. The objective is to use a quantum method of empirical inference and prediction that can anticipate the conceptual behavior of phenomena, extending it to the complexity and heterogeneity that have always been the hallmark of the social sciences. Moreover, in this heterogeneity, it is possible to achieve the character of "new normality" in the presence of relational divergence between concepts and enhance the ex-ante tools that can explain, anticipate, and resolve concrete inequities and discrepancies.

6- Conclusion

The high school grades attributed by teachers appeared to be negatively associated with lecturing experience. However, drawing conclusions about AA is not straightforward. For instance, a simple increase in the teaching stringency as teachers grow older should result in lower grades for comparable AA attainments. Female teachers are also bound to attribute lower grades. This can also be linked to stricter evaluation criteria. The mother being an employer is detrimental to student AA. It is of utmost importance to ensure that women's empowerment in their aspirations, objectives, undertakings, and civic participation is followed by an appropriate balance between home and career affairs for both genders.

Deep MLP is more efficient than other methods in predicting students' grades. However, the adoption of deep learning as an experimental approach in educational and social sciences also has remarkable advantages beyond its predictive capacity. We are dealing with a paradigm that does not depend on a specific mathematical form to express relationships between concepts with a particular aptitude to represent social phenomena whose heterogeneity is paramount. The treatment of conceptual heterogeneity is undertaken naturally and spontaneously. By widening the spectrum of possibilities, deep learning introduces a capacity to anticipate nonconformities, which is an inducement to the search for fairer and more equitable policies. In deep learning, any policy measure that induces changes in the critical factors of AA is evaluated within the heterogeneous spectrum of both the possible outcomes and the underlying gradient structure. Deep learning recreates a quantum space of representation and explanation of phenomena that promotes a diversity of leads and accurate predictions. On the other hand, in the presence of more uniform realities, it establishes an intelligible relationship with the MLR and the classic meaning of its coefficients. The absence of a strong empirical relationship between deep learning and classic MLR is a robust means to assess the correctness of implementing the latter.

6-1- Limitations

Like any other study of this nature, some limitations need to be acknowledged. The vast majority of the variables under consideration are categorical and do not directly measure the critical factors of AA. They are proxy variables with measurement biases. There is no variable associated with parent involvement and school environment and design. The target variable itself, being teacher-attributed grades rather than exam scores, is susceptible to issues such as differences in the stringency of teachers' assessment criteria. Adopting a data-driven approach for policy definition and design needs a substantial improvement in the quantity and quality of the data to forge a capable and reliable education data system.

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

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- Informed Consent Statement

Not applicable.

7-5- 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] Haykin, S. (2009). Neural networks and learning machines (3rd Ed.). Pearson, New Yoork City, United States.

[2] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. doi:10.1126/science.aaa8415.

[3] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE Access, 7, 53040–53065. doi:10.1109/ACCESS.2019.2912200.

[4] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3). Portico. doi:10.1002/widm.1355.

[5] Namoun, A., & Alshanqiti, A. (2020). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. Applied Sciences, 11(1), 237. doi:10.3390/app11010237.

[6] Costa-Mendes, R., Cruz-Jesus, F., Oliveira, T., & Castelli, M. (2021). Machine learning bias in predicting high school grades: A knowledge perspective. Emerging Science Journal, 5(5), 576–597. doi:10.28991/esj-2021-01298.

[7] Hanushek, E. A., & Wößmann, L. (2010). Education and Economic Growth. International Encyclopedia of Education, 245–252, Elsevier Science, Amsterdam, Netherlands. doi:10.1016/b978-0-08-044894-7.01227-6.

[8] Lei, D., Chen, X., & Zhao, J. (2018). Opening the black box of deep learning. arXiv preprint arXiv:1805.08355. doi:10.48550/arXiv.1805.08355.

[9] Golovko, V. A. (2017). Deep learning: an overview and main paradigms. Optical Memory and Neural Networks (Information Optics), 26(1), 1–17. doi:10.3103/S1060992X16040081.

[10] 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(6). doi:10.1016/j.heliyon.2020.e04081.

[11] Jensen, A. R. (1999). The G factor: The science of mental ability. Psycoloquy, 10(04), 36–2443–36–2443,. doi:10.5860/choice.36-2443.

[12] Georgiou, G. K., Guo, K., Naveenkumar, N., Vieira, A. P. A., & Das, J. P. (2020). PASS theory of intelligence and academic achievement: A meta-analytic review. Intelligence, 79, 101431. doi:10.1016/j.intell.2020.101431.

[13] Frey, M. C., & Detterman, D. K. (2004). Scholastic Assessment or g? Psychological Science, 15(6), 373–378. doi:10.1111/j.0956-7976.2004.00687.x.

[14] Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. Intelligence, 35(1), 83–92. doi:10.1016/j.intell.2006.05.004.

[15] Francis, B., & Skelton, C. (2005). Reassessing Gender and Achievement. Routledge, London, United Kingdom. doi:10.4324/9780203412923.

[16] 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(1), 25–42. doi:10.1080/1359813042000225320.

[17] Mensah, F. K., & Kiernan, K. E. (2010). Gender differences in educational attainment: Influences of the family environment. British Educational Research Journal, 36(2), 239–260. doi:10.1080/01411920902802198.

[18] 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. doi:10.1016/j.lindif.2016.10.006.

[19] 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(8), 2095–2104. doi:10.1016/j.paid.2007.06.025.

[20] 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. doi:10.1016/j.appdev.2018.07.001.

[21] Perreira, K. M., Harris, K. M., & Lee, D. (2006). Making it in America: High school completion by immigrant and native youth. Demography, 43(3), 511–536. doi:10.1353/dem.2006.0026.

[22] 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(5), 835–853. doi:10.1177/000312240807300507.

[23] Lei, J., & Zhao, Y. (2007). Technology uses and student achievement: A longitudinal study. Computers and Education, 49(2), 284–296. doi:10.1016/j.compedu.2005.06.013.

[24] 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. doi:10.1016/j.chb.2016.06.024.

[25] Kubey, R. W., Lavin, M. J., & Barrows, J. R. (2001). Internet use and collegiate academic performance decrements: Early findings. Journal of Communication, 51(2), 366–382. doi:10.1111/j.1460-2466.2001.tb02885.x.

[26] Fan, X., & Chen, M. (2001). Parental Involvement and Students' Academic Achievement: A Meta-Analysis. Educational Psychology Review, 13(1), 1–22. doi:10.1023/A:1009048817385.

[27] 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. doi:10.1016/j.cedpsych.2019.101780.

[28] 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(6), 1053–1064. doi:10.1007/s10964-016-0431-4.

[29] Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417–453. doi:10.3102/00346543075003417.

[30] Steinmayr, R., Dinger, F. C., & Spinath, B. (2010). Parents' education and children's achievement: The role of personality. European Journal of Personality, 24(6), 535–550. doi:10.1002/per.755.

[31] Tomul, E., & Savasci, H. S. (2012). Socioeconomic determinants of academic achievement. Educational Assessment, Evaluation and Accountability, 24(3), 175–187. doi:10.1007/s11092-012-9149-3.

[32] Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. Quarterly Journal of Economics, 115(4), 1239–1285. doi:10.1162/003355300555060.

[33] Krueger, A. B. (1999). Experimental estimates of education production functions. Quarterly Journal of Economics, 114(2), 497–532. doi:10.1162/003355399556052.

[34] 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(3), 695–736. doi:10.1016/j.euroecorev.2004.11.005.

[35] Leithwood, K., & Jantzi, D. (2009). A review of empirical evidence about school size effects: A policy perspective. Review of Educational Research, 79(1), 464–490. doi:10.3102/0034654308326158.

[36] Schneider, M. (2002). Do School Facilities Affect Academic Outcomes?. Information Analysis, National Clearinghouse for Educational Facilities, Washington, United States.

[37] 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(1), 47–70. doi:10.1080/03054980601094693.

[38] Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. Journal of Labor Economics, 25(1), 95–135. doi:10.1086/508733.

[39] Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. Review of Educational Research, 73(1), 89–122. doi:10.3102/00346543073001089.

[40] Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. American Economic Review, 94(2), 247–252. doi:10.1257/0002828041302244.

[41] Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. Econometrica, 73(2), 417–458. doi:10.1111/j.1468-0262.2005.00584.x.

[42] Baashar, Y., Alkawsi, G., Ali, N., Alhussian, H., & Bahbouh, H. T. (2021). Predicting student's performance using machine learning methods: A systematic literature review. 2021 International Conference on Computer &amp; Information Sciences (ICCOINS). doi:10.1109/iccoins49721.2021.9497185.

[43] Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49-64.

[44] Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. Computational Intelligence and Neuroscience, 2022, 1–11. doi:10.1155/2022/4151487.

[45] Feng, G., Fan, M., & Chen, Y. (2022). Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. IEEE Access, 10, 19558–19571. doi:10.1109/ACCESS.2022.3151652.

[46] Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks. IEEE Access, 9, 140731–140746. doi:10.1109/ACCESS.2021.3119596.

[47] Al-Tameemi, G., Xue, J., Ajit, S., Kanakis, T., Hadi, I., Baker, T., Al-Khafajiy, M., & Al-Jumeily, R. (2021). A Deep Neural Network-Based Prediction Model for Students' Academic Performance. 2021 14th International Conference on Developments in ESystems Engineering (DeSE). doi:10.1109/dese54285.2021.9719552.

[48] Costa-Mendes, R., Oliveira, T., Castelli, M., & Cruz-Jesus, F. (2021). A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. Education and Information Technologies, 26(2), 1527–1547. doi:10.1007/s10639-020-10316-y.

[49] Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: a machine-learning approach. Higher Education, 80(5), 875–894. doi:10.1007/s10734-020-00520-7.

[50] 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. doi:10.1109/ACCESS.2020.2981905.

[51] Aydoğdu, Ş. (2020). Predicting student final performance using artificial neural networks in online learning environments. Education and Information Technologies, 25(3), 1913–1927. doi:10.1007/s10639-019-10053-x.

[52] Li, F., Zhang, Y., Chen, M., & Gao, K. (2019). Which Factors Have the Greatest Impact on Student's Performance. Journal of Physics: Conference Series, 1288(1). doi:10.1088/1742-6596/1288/1/012077.

[53] Altaf, S., Soomro, W., & Rawi, M. I. M. (2019). Student Performance Prediction using Multi-Layers Artificial Neural Networks. Proceedings of the 2019 3rd International Conference on Information System and Data Mining-ICISDM 2019. doi:10.1145/3325917.3325919.

[54] Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. SN Applied Sciences, 1(9), 1–10,. doi:10.1007/s42452-019-0884-7.

[55] Arunachalam, A. S., & Velmurugan, T. (2018). Analyzing student performance using evolutionary artificial neural network algorithm. International Journal of Engineering and Technology(UAE), 7(2.26), 67–73. doi:10.14419/ijet.v7i2.26.12537.

[56] Mondal, A., & Mukherjee, J. (2018). An Approach to Predict a Student's Academic Performance using Recurrent Neural Network (RNN). International Journal of Computer Applications, 181(6), 1–5. doi:10.5120/ijca2018917352.

[57] Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT Press, Cambridge, Massachusetts, United States.

[58] Hastie, T., Friedman, J., & Tibshirani, R. (2001). The Elements of Statistical Learning. Springer Series in Statistics, New York, United States. doi:10.1007/978-0-387-21606-5.

[59] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference. doi:10.25080/majora-92bf1922-011.

[60] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408. doi:10.1037/h0042519.

[61] Ramchoun, H., Idrissi, M. A. J., Ghanou, Y., & Ettaouil, M. (2017). Multilayer Perceptron. Proceedings of the 2nd International Conference on Big Data, Cloud and Applications. doi:10.1145/3090354.3090427.

[62] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning Internal Representations by Error Propagation. Readings in Cognitive Science, 399–421, Elsevier, Amsterdam, Netherlands. doi:10.1016/b978-1-4832-1446-7.50035-2.

[63] Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods, 43(1), 3–31. doi:10.1016/s0167-7012(00)00201-3.

[64] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning. Springer, New York, united States.

[65] Chollet, F. (2015). About Keras. Available online: https://keras.io/about/ (accessed on June 2022)..

[66] Narkhede, M. V., Bartakke, P. P., & Sutaone, M. S. (2021). A review on weight initialization strategies for neural networks. Artificial Intelligence Review, 55(1), 291–322. doi:10.1007/s10462-021-10033-z.

[67] Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, 249-256, 13-15 May, 2010, Chia Laguna Resort, Sardinia, Italy.

[68] Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. doi:10.1038/nature14539.

[69] Garbin, C., Zhu, X., & Marques, O. (2020). Dropout vs. batch normalization: an empirical study of their impact to deep learning. Multimedia Tools and Applications, 79(19–20), 12777–12815. doi:10.1007/s11042-019-08453-9.

[70] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(56), 1929–1958.

[71] Kandel, I., & Castelli, M. (2020). The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. ICT Express, 6(4), 312–315. doi:10.1016/j.icte.2020.04.010.

[72] Bengio, Y. (2012). Practical Recommendations for Gradient-Based Training of Deep Architectures. Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, 7700, Springer, Berlin, Germany. doi:10.1007/978-3-642-35289-8_26.

[73] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32nd International Conference on Machine Learning, 448–456, 6-11 July, 2015, Lille, France.

[74] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 doi:10.48550/arXiv.1412.6980.

[75] Bottou, L. (2012). Stochastic Gradient Descent Tricks. Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, 7700, Springer, Berlin, Germany. doi:10.1007/978-3-642-35289-8_25.

[76] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. Statistics in medicine, 16(4), 385-395. doi:10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

[77] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. Journal of Machine Learning Research, 18, 1–52.

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

[79] Jamieson, K., & Talwalkar, A. (2016). Non-stochastic best arm identification and hyperparameter optimization. Proceeding of the 19th International Conference on Artificial intelligence and Statistics (AISTATS), 240-248, 9-11 May, Cadiz, Spain.

[80] Brochu, E., Cora, V. M., & De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint. doi:10.48550/arXiv.1012.2599

[81] Kumar, S., & Srivistava, A. N. (2012). Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection. 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 12-16 August, 2012, Beijing, China.

[82] Pearce, T., Brintrup, A., Zaki, M., & Neely, A. (2018). High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. Proceeding of the 35th International Conference International Conference on Machine Learning, 10-15 july, 2018, Stockholm, Sweden.

[83] Rocha-Trindade, M. B. (2020). Chinese Community in Portugal: History , Migration , and Business. Gaudium Scendi Universidade Católica Portuguesa, 19, 15–42. doi: doi:10.34632/gaudiumsciendi.2020.10065.

Appendix I: Number of Observations per Subject

	Portuguese	Biology	Biology & Geology	Chemistry	Communication Nets
10th	29,105	775	11,372	486	483
11th	27,765	974	18,084	400	496
12th	50,441	12,956	78	4,795	1,894
Total	107,311	14,705	29,534	5,681	2,873

	Design	Economics	English	Geography	History
10th	1,094	3,805	26,382	9,536	7,422
11th	986	5,809	41,236	14,556	6,386
12th	1,765	5,395	15,809	6,963	11,593
Total	3,845	15,009	83,427	31,055	25,401

	Informatic Applications	ICT	Math	Math A	Math to Social Sciences
10th	0	8,554	5,506	15,355	5,099
11th	0	2,170	5,430	13,582	7,201
12th	10,168	1,122	8,824	22,412	33
Total	10,168	11,846	19,760	51,349	12,333

	Physical Education	Physics	Physics and Chemistry	Psychology	Sociology
10th	25,019	284	13,828	1,509	254
11th	23,851	464	22,408	2,083	321
12th	49,028	5,340	2,403	14,991	5,548
Total	97,898	6,088	38,639	18,583	6,123

	Descriptive Geometry	History of Culture and Arts	Philosophy	Spanish
10th	2,274	1,566	24,017	2,841
11th	3,526	2,627	38,340	3,900
12th	41	2,267	22	943
Total	5,841	6,460	62,379	7,684

Appendix II: Features

Feature	Description	Literature AA Critical Factor	Data Type
Enrolments	Number of student enrolments	Cognitive ability	Integer
Retentions	Number of student retentions estimated by age in excess	Cognitive ability	Integer
Gender	Feminine and masculine gender	Gender	Categorical
Father nationality	Portugal, Africa, Brazil, China, East Europe, developed countries, and others	Ethnicity	Categorical
Computer	Student owns a personal computer	Computer usage	Binary
Internet	Student has access to the Internet	Internet usage	Binary
Parish	Student's home is located in the school parish	SES	Binary
County	Student's home is located in the school county	SES	Binary
Guardian	Mother, father, the student representing themselves, close relative, and guardian	SES	Categorical
Job situation	Student works	SES	Binary
Responsible job situation	Unknown, employee, unemployed, self-employed, employer, home affairs, retired, student, and other	SES	Categorical
Father job situation	Unknown, employee, unemployed, self-employed, employer, home affairs, retired, student, and other	SES	Categorical
Mother job situation	Unknown, employee, unemployed, self-employed, employer, home affairs, retired, student, and other	SES	Categorical
Guardian educational level	Unknown, no formal education, elementary I (grades 1- 4), elementary II (grades 5 - 6), Middle III (Junior-high, grades 7 – 9), secondary (Senior-high, grades 10 – 12), undergraduate degree, university degree, post-graduation, master, Ph.D., and other	SES	Categorical
Father educational level	Unknown, no formal education, elementary I (grades 1- 4), elementary II (grades 5 - 6), Middle III (Junior-high, grades 7 – 9), secondary (Senior-high, grades 10 – 12), undergraduate degree, university degree, post-graduation, master, Ph.D., and other	SES	Categorical
Mother educational level	Unknown, no formal education, elementary I (grades 1- 4), elementary II (grades 5 - 6), Middle III (Junior-high, grades 7 – 9), secondary (Senior-high, grades 10 – 12), undergraduate degree, university degree, post-graduation, master, Ph.D., and other	SES	Categorical
Scholarship	No support, half support, and full support	SES	Categorical
Family non-classic dwellings	Percentage of family non-classic dwellings that exist in the student's home parish	SES	Percentage
Collective dwellings	Percentage of collective dwellings that exist in the student's home parish	SES	Percentage
Feature	Description	Literature AA critical factor	Data Type
Illiteracy rate	Student home parish Illiteracy rate	SES	Percentage
Post-secondary schooling rate	Student home parish post-secondary schooling rate	SES	Percentage
Primary sector importance	Student home parish primary sector activities importance	SES	Percentage
Secondary sector importance	Student home parish secondary sector activities importance	SES	Percentage
Unemployment rate	Student home parish unemployment rate	SES	Percentage
School size	Number of school students	School size	Integer
Class size	Number of class students	Class size	Integer
Teacher age	the age of the teacher	Lecturing quality	Integer
Teacher gender	Feminine and masculine gender	Lecturing quality	Categorical
Lecturing time	Teacher time dedicated to lecturing in hours	Lecturing quality	Integer
Non-lecturing time	Teacher time not dedicated to lecturing in hours	Lecturing quality	Integer
Teacher's years to retirement	Years until retirement age	Lecturing quality	Integer
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	Lecturing quality	Categorical
Teacher educational level	Bachelor, university degree, master and Ph.D., and other	Lecturing quality	Categorical
Grade year	10th, 11th, and 12th high school grades	n.a	Categorical
Subjects		n.a	Categorical
Teacher mark	Teacher end of the year mark	Target variable	Integer