Education Data Mining to forecast Essay Score. A case Study about ENEM

***Abstract*—Educational Data Mining is a rising area of study that offers data analysis in different educational studies, improving students' outcomes. National High School Exam (ENEM) is an indicator of schooling effectiveness in Brazil annually and considers students' knowledge in different areas of interest. In this work, we used four different machine learning algorithms to predict the argumentative essay score in the ENEM of 2019 based on various student exams. The algorithms used were Linear Regression, Random Forest, KNN, and Neural Network. We explored different domains using many feature selection techniques. Random Forest followed by Neural Network performed the best results. KNN regressor and Linear Regression algorithms are not able to relevant generalizations.**

Keywords

Regression Models, Educational Data Mining, ENEM.

I. Introduction

SINCE the 1950s, with the birth of artificial intelligence, computers have outperformed humans in several tasks [6]. From the early 2000s, there was the popularization of machine learning, and its use increasingly frequent in the industry [9]. Given this rise and the vital role of education in economic development, an area of research standing out is Educational Data Mining [17].

Educational Data Mining (EDM) consists of using data mining techniques in educational databases [21]. One of the utilities implemented by EDM is the forecasting of grades [25]. In Brazil, it exists an exam that evaluates the performance of the students. this exam is *Exame Nacional do Ensino Médio* (High School National Exam) – ENEM. The challenge ahead is to improve the quality of primary education and select the students into different universities. This work aims to answer the following research question: "How do the different features in the ENEM exam impact the final essay grade?" Carrying out the study, we investigate the ENEM data from 2019. We did a pre-processing and applied four machine learning algorithms: Linear Regression, Random Forest, KNN regressor, and Neural Network, to predict the candidates' grades in some experiment configurations. We can observe Random Forest has the best model generalization in comparison to other algorithms.

This work is organized into seven sections, including this introductory section. Section 2 explains fundamental concepts for understanding the analysis performed. Section 3 reviews the related work similar to our area of interest and details the importance of this topic in the academic community. Section 4 presents the used materials and methods in the experimental analysis. Section 5 presents the results of this study. Section 6 discusses the results of the experiments. Finally, Section 7 is the conclusion of this article, as well as suggestions for future work.

This work was not supported by any organization

II. Fundamental Concepts

A. Regressor Algorithms based on Machine Learning

Regressor algorithms based on Machine Learning (ML) [8] are algorithms that have the task of predicting a continuous number. This type of learning consists of presenting two types of variables into the algorithm. They will be called independent variables, features or attributes X = $(x_1, x_2 \ldots, x_n)$, whereas the other variables are the target variable, or dependent Y = $(y_1, y_2 \ldots, y_n)$. After training, the algorithm will be able to predict new Y values, based on new X values [15]. In the next section, we described the algorithms used in more detail.

1) Linear Regression: Linear Regression (LR) is one of the simplest and oldest models of machine learning regressors. This model sums up the attributes, weighted by a given coefficient, and a bias term, or the intercept, as can be seen in Equation 1 [15].

$$\hat{y} \approx \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{1}$$

Where, $\hat{y}$ is the predicted value. n is the number of attributes. x_i is the $i-th$ attribute. θ_j is the $i-th$ coefficient.

2) Random Forest: Random Forest (RF) is a very popular algorithm due to its ease of implementation, and adjustment [15]. Briefly, the RF is a technique of joining several decorrelated decision trees to reduce the noise, which is characteristic of decision trees [2].

The RF algorithm recursively subdivides the training base according to a given criterion and defines a prediction criterion for that subdivision, extracting from it a certain error [19]. Several decision trees are trained on a database in the regressive RF, and an average is extracted. As each tree in the RF is identically distributed. The created bias by the individual trees is not lost when the number of trees is increased, but the variance of the average is reduced [15].

Random Forest works as follows: First, an N-size S sample is created with V attributes from the training data. Then a decision tree A is made based on S following the process for each node of A [15]:

1) A set of M attributes of V is randomly selected.
2) The heterogeneity of the Y is measured given M.

3) The best variables are selected - less heterogeneity, among the selected M.
4) The leaf is subdivided into two leaves.

This process is repeated until all attributes are selected. By the end of the process, a decision tree is made. The tree creation process is repeated, forming a forest with T trees, which, the randomness of the process, present variations in the divisions of attributes.

3) KNN Regressor: K-Nearest Neighbor Regressor, or KNN, is a supervised learning technique capable of working with classification and regression. The technique is using a number of neighbors near the k observations X of the training base to predict $\hat{Y}$. The nearby neighborhood N_k is defined by a measure of proximity, which consists of the shortest Euclidean distance from the neighbors. To make a prediction $\hat{Y}(x)$, you calculate an average of response values y_i, related to each observation x_i in each of the neighbors N_k (Eq. 2) [15].

$$\hat{Y}(X) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{2}$$

KNN is a relatively simple and non-parametric algorithm. This algorithm does not depend on prior knowledge of the relationship between the attributes and the response variable. With effect, this algorithm can adapt very well to all types of databases [16].

The definition of the k value is crucial for the performance of this algorithm. For minimal k values, there is a high variance of the forecast. On the other hand, the bias is very low. In other words, small k values imply overfitting, failing to adapt to new values. Conversely, as k increases, the bias increases when the forecast variance decreases, approaching linearity, therefore more adaptable to new data [16].

4) Neural Network: The basic structure of a neural network can include one or more nodes, in one or more layers of nodes, preceded by one or more inputs and followed by one or more outputs [9]. With effect, each node can be compared to neurons, which communicate through synapses (vectors), being able to form a network of neurons that learn and improve [20].

The artificial neural network can be very complex. However, it consists of relatively simple units - Perceptrons, the simplest model of a Neural Network. It consists of n inputs $(x_1, x_2 \dots x_n)$ and their respective weights $(w_1, w_2 \dots w_n)$, a node that computes the z-weighted sum of the inputs and applies a given function to that sum $f(z)$. Finally, there is an output layer that represents the result of applying $f(z)$ [9].

$$w_{ij}^{*} = w_{ij} + \eta(y_j - \hat{y}_j)x_i \tag{3}$$

Where w_{ij}^{*} represents the weight of the next step. w_{ij} represents the weight of the current step. η represents the learning rate. $y_j - \hat{y}_j$ represents the predictive error. x_i represents the value of the variable in the current step.

Given that Perceptron faces some problems that it cannot solve, the solution can be mitigated by stacking several layers of Perceptrons, the so-called Multi-layer Perceptrons (Multi-Layer Perceptron - MLP). MLP (Equation 6) can also be divided into layers. We have that a hidden layer of Perceptrons mediates the input and output layer, plus biases [9]. The number of hidden layers can vary from one to hundreds. Although an exact number has not been defined, a large number of hidden layers is what is called Deep Learning [15].

Still on MLP, one of the improvements was the advent of Backpropagation, which implements a global optimum, or close to it, through these adjustments in the weights and biases made backward. The method uses a cost function and, using the chain rule, punishes or strengthens weights and biases to reduce a given cost [15].

$$\begin{aligned} &Z_m = \sigma(\alpha_{0m} + \alpha_W^T X), m = 1, ...m \\ &W = \beta_0 + \beta^W Z \\ &f(X) = g(W) \end{aligned} \tag{4}$$

Where M represents the number of vectors that connect neurons from the input layer to the hidden layer of the Neural Network. Z_m represents the transformation of the input variables with an activation function, as among those mentioned above. W is a weight assigned to the output of Z, and finally, the function $g(W)$ returns the output of the model, in this case, an identity function $g(T) = T$ [15].

B. Educational Data Mining

Educational Data Mining (EDM) allows us to develop many tools to reduce poor academic results, such as high dropout rates, absences, low grades, and failure. Early identification of students who have learning problems can be an important advance in personalizing education policy and allocating public resources [5].

There is excellent relevance concerning EDM: the use of machine learning to predict grades. In this context, Machine Learning is a strategy for action in educational management. As shown [21], the students' grade represents an extension of the achievement of long and short-term goals in the educational system. The grade can be presented as the final grade of a course, specific subjects, or an arithmetic average of several tests. As it represents an objective metric of academic results, grade prediction can feed intelligent tutoring systems and automated progress monitoring.

In this way, our studies contributed to EDM, allowing the developers to build new studies through our academic results.

C. Performance Metrics

The main metrics to measure the assertiveness of the regression was Mean Square Error (MSE) (Eq. 5), and Root Mean Square Error (RMSE) (Eq. 6), Mean Absolute Error (MAE)(Eq. 7) and the R-square (R^2) (Eq. 8). Mean Square Error (MSE) confers a measurement of the predictive error of the total of the predicted values [8, 6]. As the errors measured by the MSE are squared, which means out of the scale of the observed values, RMSE gives us the advantage of observing the error in the same scale and unit of points as the test note,

in addition to being the most used metric in EDM works, according to [21]. MAE does not have this mathematical representation of bias and variance. Still, it is less sensitive to outlier values when compared to RMSE and MSE. R-square (R^2) is an excellent indicator of the model's adjustment to the data. This metric suggests how much percentage a model can explain the variance of the observations, given by values from 0 to 1 (Equation 12). Equation 12 is composed by: the ratio between the sum of squares and squared sum of the difference between the mean and the observed values [3, 11]. [3, 11].

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{5}$$

$$RMSE = \sqrt{MSE} \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2 \tag{7}$$

$$R^2 = \frac{\sum_i(\hat{y}_i - \bar{y}_i)^2}{\sum_i(y_i - \bar{y}_i)^2} = 1 - \frac{\sum_i(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y}_i)^2} \tag{8}$$

D. ENEM

Exame Nacional do Ensino Médio (ENEM), or High School National Exam, is a standardized national exam, not mandatory, and it is one of the largest in the world in terms of the number of participants [4]. This exam is great potential for studies and research, especially within the scope of EDM. ENEM measures the ability of candidates to master the set of scientific and technological principles about Secondary Education [7].

ENEM is today one of the main methods of entry into Public Higher Education in Brazil. Currently, the ENEM, applied in two test days, consists of 180 questions involving knowledge in 5 areas of competence, namely: Natural Sciences and its Technologies; Human Sciences and its Technologies; Languages, Codes, and their Technologies; Mathematics and their Technologies and an Argumentative Essay [7].

III. Related Works

In a review of works in Portuguese in the area of Educational Data Mining, [25], it was possible to identify the most used algorithms, such as J48, Neural Networks, Random Forest, and Naive Bayes. Authors also noticed that most works were dedicated to performance evaluation in subjects - 15 (50%), or in specific courses - 10 (33.33%), and much less in schools - 2 (6.7%).

Golino and Gomes [10] studied the use of algorithms - Learning Trees, Bagging, Random Forest, and Boosting, in the prediction of grades divided into classes (high and low) of students of the medical course of a private university.

In work dedicated to EDM, Sorgatto *et al.* [26] used the INEP database of the School Census to predict grades through ENEM data by school. Several classification algorithms were used, considering that the ENEM score was categorized according to the statistical criteria for distributing values.

Rodrigues et al. [4] also divided grades into categories and used classification algorithms. The data used were those of ENEM 2011, and the response variable was an arithmetic mean of the objective notes and the essay.

Stearns et al. [27] also used two algorithms based on CART: AdaBoost and Gradient Boosting to predict the math score of the ENEM 2014 test. The attributes used were personal information, information from the school of origin, socioeconomic factors, all components of the ENEM registration form.

Using data from ENEM 2016, Santos et al. [24] used the PCA technique to reduce the number of attributes, and the authors applied a Bayesian network to understand the most significant factors in the construction of the note.

Our study selected Neural Network, Random Forest, Linear Regression, and KNN regressor algorithms to predict the essay grade. Using these algorithms will make it possible to compare the prediction made concerning the grade of tests and the prediction of other disciplines' grades. The work will also make it possible to compare the predictive power between the algorithms used, especially the comparison with Linear Regression, which is the simplest [18].

IV. Materials and Methods

A. Database

To carry out this work, we used the database of Microdata from ENEM - National High School Examination of 2019, the most recent available at the time of this work, made available by INEP. The data consists of 5,095,270 rows and 136 columns, with each row represented by one of the registered candidates. The data brings socioeconomic information from each candidate, which is answered when he or she enrolls, and which were used to make predictions in several works [13, 4, 26, 10, 11, 14].

In addition to the socioeconomic data, the columns also list data about the essay test, the objective test, the location of the test application, requests for specialized and specific resources to carry out the tests. There are also columns about requests for specific attendance, requests for specialized attendance, school data, and participant information [7].

B. Pre-Processing

Based on Gomes et al. [12]. and Rodriguez et al. [4], we selected 135 columns. The first criterion for deleting columns was its relevance, which comprises columns with repeated data, such as code and municipality name. The second criterion for deleting it was data that do not vary as the year of the test and others that do not have relevance for predicting the grade as registration number [24] [22]. We removed columns such as the code of the test notebook, participant's response, or proof template.

With these first transformations, the database ended up with 1,000,256 rows and 95 columns. The response variable (argumentative essay score) had an arithmetic mean of 573.96 points, a standard deviation of 197.53 points, a kurtosis of 1.27, and a skewness of -0.77. The database is available at the link: https://bit.ly/34KgkwR.

Also, in the process of preparation of the database, a normalization of the values was performed. The Min-Max normalization consists of reducing all the original values in values comprising a maximum value of 1 and a minimum value 0 [27].

Finally, we did the hold-out evaluation, where we split the data into training data for fitting the model, and test data, where we have previously not seen [11] data by ML models. The division chosen was 70% for training data and 30% for test data, both randomly distributed [1].

C. Algorithms

We used four machine learning algorithms to predict the scores: Linear Regression, Random Forest, KNN regressor, and MLP. We implemented all through the Scikit-learn module based on the Python language, whose names are, respectively, on the platform: Linear Regression, Random Forest Regressor, K-Neighbors Regressor, and MLP Regressor [23].

The Linear Regressor algorithm implements a traditional linear regression method by predicting coefficients to minimize the residual sum of the squares. Random Forest Regressor was trained with 250 trees with the limit of computational resources and processing time.

K-Neighbors Regressor implements a KNN-based regression. The k value was performed in the training set, obtaining the best results with 70 neighbors. We used uniform weights for neighbors.

The Regressor MLP implements a regression by a Multi-Layer Perceptron using Backpropagation. We used 25 neurons and one hidden layer. The activation function used was ReLU with 0.001 of initial learning rate and 200 maximum iteration limit.

V. Results

In the present section, we elaborate on four different configurations for experiments resulting from the exploratory analysis of the database. The results presented the impact of each configuration algorithm quality and a comparative analysis of the different configurations.

A. Configuration 1

Initially, we split the database into train and test, which we will call Configuration 1. We gathered 1,000,256 rows and 95 columns, including the column with the essay grades (label). Table 2 shows the comparative result of the algorithms in this configuration. Random Forest obtains the best performance for three measured metrics. In worst cases, KNN has these results in R-squared and RMSE. And linear regression has the worst MAE.

B. Configuration 2

Sequentially, the experiment continued by removing the columns with lower explanatory power: we called Configuration 2. We used the correlation between each attribute and the response variable in ranking form. The selection of attributes was also used by [12, 11, 24, 14] even though these justified the deletion of the columns due to the nature of the attributes.

TABLE I
Configuration 1 Performance

Algorithm	RMSE	MAE	R^2
Linear Regression	176,23	129,12	0,2034
Random Forest	169,17	124,71	0,2659
KNN	177,02	128,99	0,1962
Neural Network	170,77	125,48	0,2520

TABLE II
Configuration Performance 2

Algorithm	RMSE	MAE	R^2
Linear Regression	176,24	129,13	0,2033
Random Forest	169,23	124,78	0,2654
KNN	176,72	128,87	0,1990
Neural Network	171,71	128,41	0,2437

We applied a test to remove different percentiles of these columns in training set to establish an optimal value for the attribute extraction method. We analyzed between the 10th and 90th percentile, varying in every ten values of percentile. The range from the 60th to the 70th percentile demonstrated lower MAE in the performance of all algorithms based on the attribute extraction method. Thus, we adopted the 60th percentile for Configuration 2.

Table II demonstrates the results of algorithms with Configuration 2. As can be seen, there was virtually no significant performance increase. Overall, all algorithms have some performance loss on all metrics, except KNN, which achieved a slight reduction in predictive error. Figure 2 confirms the virtual maintenance of predictive power compared to the performance of Configuration 1.

C. Configuration 3

In Configuration 3, we only the transformed target variable based on the ENEM Microdata [7]. As it turns out, the argumentative essay score contains objective criteria for assigning value 0 (annulled essay). Since those criteria prevent a grade from being attributed to the essay, we decided to remove the lines corresponding to those occasions. Thus, Configuration 3 only has students who have not had their annulled essays. The arithmetic mean of the argumentative essay score became 599.62, standard deviation 159.27, kurtosis -0.19, and skewness of 0.07. In addition, the scores started to have an amplitude of 40 to 1000 points.

Table III shows the results Linear Regression. LR performed worse among all but almost imperceptibly after KNN, the most important difference being the MAE of RL 109.18 versus 108.94 of KNN. The best performances were from Random Forest and the Neural Network. Random Forest was the best of all metrics with the explanatory power of 30.32% versus 27.58% of the Neural Network.

D. Configuration 4

Configuration 4 was implemented, adding the Configuration 2 and Configuration 3 procedures. Configuration 4 represents both procedures for the attribute selection and target variable. Table IV also shows no increase in the performance of

TABLE III
CONFIGURATION PERFORMANCE 3

Algorithm	RMSE	MAE	R^2
Linear Regression	140,26	109,18	0,2236
Random Forest	132,87	103,59	0,3032
KNN	140,24	108,94	0,2238
Neural Network	135,46	105,75	0,2758

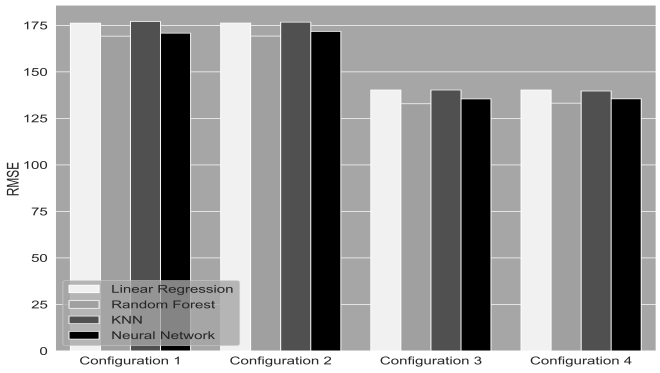

Fig. 1. Comparison of RMSE between settings.

algorithms except for KNN, which, as noted in Configuration 2, benefited from the attribute selection.

TABLE IV
CONFIGURATION PERFORMANCE 4

Algorithm	RMSE	MAE	R^2
Linear Regression	140,27	109,18	0,2235
Random Forest	133,16	103,84	0,3002
KNN	139,68	108,58	0,2300
Neural Network	135,48	105,88	0,2755

VI. DISCUSSION

Chart analysis helps us understand the comparison between settings. Figures 1, 2, and 3 show RMSE, MAE, and R-square evolution according to each configuration, respectively. We can observe Configuration 2 has virtually no difference from Configuration 1. Eliminating columns with little explanatory potential can be a factor.

Configuration 3 is the only one that provided significant improvement to reduce predictive error. As can be seen, the elimination of students who had their essay grades zero reduced the skewness of the distribution of grades (before -0.77, then 0.07). As pointed out [1], this improvement can be noted by the difference between RMSE and the MAE. We can conclude a greater explanatory power over the grades in Configuration 3. This same gain cannot be observed when implementing Configuration 2 or Configuration 4, compared to Configuration 1 and Configuration 3, respectively.

It was also possible to identify that RF and MLP perform at all times better than RL and KNN. The present work suggests that in the educational database used and under the regression task, the RF and MLP algorithms perform better than a statistical model such as RL.

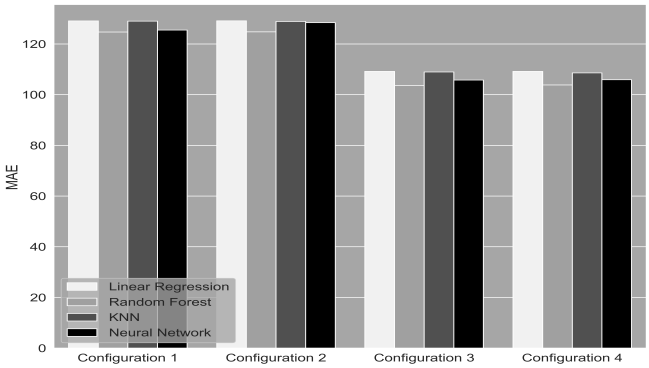

Fig. 2. Comparison of MAE between Configurations.

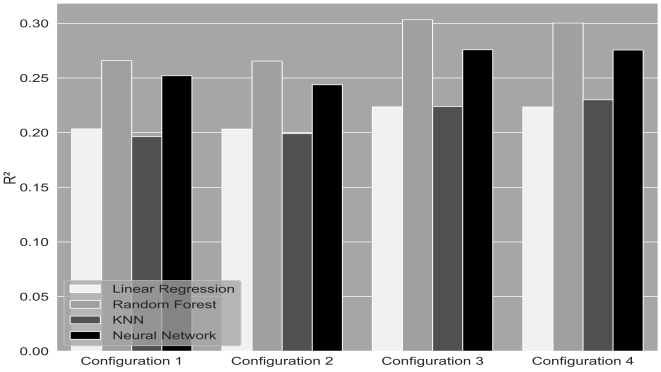

Fig. 3. Comparison of R^2 between settings.

Also, the difference in performance between RF and MLP may have been due to the size of the database and a greater amount of predictors similar to Cornell et al. [5]. The algorithms benefited from this beyond the linear capacity of a statistical model. The best adaptation to RF data before RL corroborates the conclusion since tree algorithms benefit from their exploratory behavior [12].

Gomes *et al.* [13] used the 2011 ENEM database to also predict the essay score. According to the authors, the mean score was 545.21, a standard deviation of 146.45, and amplitude ranging from 40 to 1000 points in the training set. This distribution is closer to that obtained in Configuration 3, where zero values are eliminated. The skewness of the response variable measured by the authors in the training base was 0.09, close to 0.07 obtained in Configuration 3, and, far away from the distribution of Configuration 1 (-0.77), when zero values are not removed.

The worst performance in R-square scans for the present study was the KNN in Configuration 1 (0.1962). As demonstrated, Configuration 1 has impaired learning by algorithms.

Using the CART algorithm, [12] they predicted the arithmetic mean of the objective disciplines of the ENEM. The best R-square obtained by the authors was 0.4029. It is important to highlight that they used the [12] ENEM 2011 database, which had different attributes compared to the 2019 database. With this same database and algorithm, [11] they predicted

the Natural Science score and its Technologies using CART, obtaining the best R-square of 0.3252. Table VI shows all comparisons.

In [14], the authors predicted the grade of Mathematics and its Technologies of ENEM 2011 with the same algorithm, obtaining an R-square of 0.3842. Also, using the grade of mathematics in 2014 ENEM base, Stearns et al. [27] got the best R-square of 0.35 and MAE of 65.90 with Gradient Boosting. The best R-squared and MAE to predict the writing score of the present study were 0.3032 and 103.59, respectively. Thus, we also the conclusion since the R-squared did not reach one-third of the variance of the response variable [11].

VII. Conclusion

This work carried out, through four different algorithms, the prediction of ENEM 2019 argumentative essay score based on the attributes of its microdata. The experiment allowed us to conclude that the prediction obtained satisfactory results when compared to others performed for the same variable and the same [11] database. When compared to the results of the prediction of other disciplines, or the arithmetic mean of these, the predictive power was not as efficient [12] [11] [14]. With the algorithms used, the attributes were not good to explain how much essay grade they were to explain the grade of other disciplines.

It was also possible to conclude that Random Forest was the algorithm that best came out among all configurations, followed by the Neural Network. Linear Regression and KNN had almost identical performances. The selection of attributes was not effective in reducing the predictive error of any of the algorithms. The treatment of the response variable showed an important improvement in the prediction of the argumentative essay scores.

One limitation of the work was the fact that the analysis did not incorporate data with time lag. The need for better computational resources was also an impediment. Another limitation, also pointed out by [11], is that the database provided by the ENEM does not provide important variables already identified in other studies [5] [11].

As future work, the influence of students' history will be explored to predict their results. In addition, forecasts will be made on all years made available by, testing the models in the following years. Deep learning algorithms will also be implemented to investigate whether they achieve better results by relating attributes to the writing grade [7].

References

[1] Seth Adjei et al. "Clustering students in assistments: exploring system-and school-level traits to advance personalization". In: *The 10th International Conference on Educational Data Mining*. ERIC. 2017, pp. 340–341.

[2] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[3] A Colin Cameron and Frank AG Windmeijer. "An R-squared measure of goodness of fit for some common nonlinear regression models". In: *Journal of econometrics* 77.2 (1997), pp. 329–342.

[4] Diego de Castro Rodrigues et al. "A Data Mining Approach Applied to the High School National Examination: Analysis of Aspects of Candidates to Brazilian Universities". In: *EPIA Conference on Artificial Intelligence*. Springer. 2019, pp. 3–14.

[5] Sarah Cornell-Farrow and Robert Garrard. "Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia". In: *Communications in Statistics: Case Studies, Data Analysis and Applications* 6.2 (2020), pp. 228–246.

[6] Matthew F Dixon, Igor Halperin, and Paul Bilokon. *Machine Learning in Finance*. Springer, 2020.

[7] MEC/INEP. Microdados do Enem. "Braslia". In: *Disponivel em: <Acesso em:* 23 (Mar. 2019). URL: http://portal.inep.gov.br/web/guest/microdados%3E.

[8] Peter Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

[9] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[10] Hudson Golino and Cristiano Gomes. "Four Machine Learning Methods to Predict Academic Achievement of College Students: A comparison study". In: *Revista E-Psi* 4 (July 2014), pp. 68–101.

[11] CMA Gomes, A Amantes, and EG Jelihovschi. "APPLYING THE REGRESSION TREE METHOD TO PREDICT STUDENTS'SCIENCE ACHIEVEMENT". In: *Trends in Psychology. doi* 109788 (2020).

[12] Cristiano Mauro Assis Gomes and Enio Jelihovschi. "Presenting the regression tree method and its application in a large-scale educational dataset". In: *International Journal of Research & Method in Education* 43.2 (2020), pp. 201–221.

[13] Cristiano Mauro Assis Gomes, Gina C Lemos, and Enio G Jelihovschi. "Comparing the Predictive Power of the CART and CTREE algorithms". In: *Avaliação Psicológica* 19.1 (2020), pp. 87–96.

[14] Cristiano Mauro Assis Gomes, Denise de Souza Fleith, and Claisy Maria. "Predictors of Students' Mathematics Achievement in Secondary Education". In: *PSICOLOGIA: TEORIA E PESQUISA* 36 (2020), e3638.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[16] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[17] Maniam Kaliannan and Suseela Devi Chandran. "Empowering students through outcome-based education

TABLE V

Work	Database Year	Response Variable	Algorithm	RMSE	MAE	R^2
Configuration 3	2019	Essay	Linear Regression	140.26	109.18	0.2236
			Random Forest	**132.87**	**103.59**	**0.3032**
			KNN	140.24	108.94	0.2238
			Neural Network	135.46	105.75	0.2758
[13]	2011	Essay	CTREE			0.0331
[13]	2011	Essay	CART			0.1657
[12]	2011	Arithmetic mean of objective tests	CART			0.4029
[11]	2011	Natural Science and its Technologies	CART			0.3252
[14]	2011	Mathematics and its Technologies	CART			0.3842
[27]	2014	Mathematics and its Technologies	Gradient Boosting		65.90	0.3500
			AdaBoost		72.66	0.1800

(OBE)". In: *Research in Education* 87.1 (2012), pp. 50–63.

[18] Sotiris B Kotsiantis. "Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades". In: *Artificial Intelligence Review* 37.4 (2012), pp. 331–344.

[19] Wei-Yin Loh. "Classification and regression trees". In: *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1.1 (2011), pp. 14–23.

[20] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *Bulletin of mathematical biology* 52.1-2 (1990), pp. 99–115.

[21] Abdallah Namoun and Abdullah Alshanqiti. "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review". In: *Applied Sciences* 11.1 (2021), p. 237.

[22] Malini M Patil and Basavaraj N Hiremath. "A systematic study of data wrangling". In: *Int. J. Inf. Technol. Comput. Sci.(IJITCS)* 1 (2018), pp. 32–39.

[23] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research, n.* 12 (2011), pp. 2825–2830.

[24] Aurea TB Santos et al. "Educational Data Mining: A Study on Socioeconomic Indicators in Education in INEP Database". In: *Advances in Data Science and Management*. Springer, 2020, pp. 51–65.

[25] Rodrigo Santos et al. "Análise de trabalhos sobre a aplicaçao de técnicas de mineraçao de dados educacionais na previsao de desempenho acadêmico". In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. Vol. 5. 1. 2016, p. 960.

[26] Doglas Sorgatto et al. *Predição de indicadores educacionais utilizando técnicas de aprendizado de máquina*. July 2020.

[27] B. Stearns et al. "Scholar Performance Prediction using Boosted Regression Trees Techniques". In: *ESANN 2017 proceedings*. Ed. by European Symposium. on Artificial Neural Networks, Computational Intelligence. Bruges: [s.n, 2017.