Forecasting Student Enrollments in Brazilian Schools for Equitable and Efficient Education Resource Allocation

Lenardo Chaves e Silva[a,c*], *Luciano* de Souza Cabral[e,c], *Jário* José dos Santos Júnior[b,c], *Luam* Leiverton Pereira dos Santos[f,c], *Thyago* Tenório Martins de Oliveira[b,c], *Breno* Jacinto Duarte da Costa[g,c], *Joana* Fusco Lobo[d], *Dalgoberto* Miguilino Pinho Júnior[b,c], *Nicholas* Joseph Tavares da Cruz[b,c], *Rafael* de Amorim Silva[b,c], *Bruno* Almeida Pimentel[b,c]

[a]Federal Rural University of the Semi-Arid, Av. Francisco Mota, 572, Mossoró, 59625-900, Rio Grande do Norte, Brazil, e-mail: lenardo@ufersa.edu.br

[b]Federal University of Alagoas, Av. Lourival Melo Mota, S/N, Maceió, 57072-970, Alagoas, Brazil

[c]Center for Excellence in Social Technologies, Av. Lourival Melo Mota, S/N, Maceió, 57072-970, Alagoas, Brazil, e-mail: {luciano.cabral, jario.junior, luam.leiverton, thyago.tenorio, breno.duarte, dalgoberto.pinho, nicholas.cruz, rafael.amorim, bruno.pimentel}@nees.ufal.br

[d]National Education Development Fund, Setor Bancário Sul, Quadra 2, Bloco F, Brasília, 70070-929, Distrito Federal, Brazil, e-mail: joana.fusco@fnde.gov.br,

[e]Pernambuco Federal Institute of Education, Science, and Technology, Av. Barão de Lucena, 251, Jaboatão dos Guararapes, 55014120, Pernambuco, Brazil

[f]Federal University of the San Francisco Valley, Av. José de Sá Maniçoba, S/N, Petrolina, 56304-917, Pernambuco, Brazil

[g]Alagoas Federal Institute of Education, Science, and Technology, Av. do Ferroviário, 530, Maceió, 57020-600, Alagoas, Brazil

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 19 May 2025

Abstract. In recent years, there has been growing scientific interest in developing effective techniques for forecasting student enrollment across the school spectrum (i.e., primary, secondary, and higher education). Enrollment forecasting is crucial in shaping public education policies by guiding resource allocation and ensuring equitable access to educational opportunities. In this sense, Machine Learning (ML) models emerge as a promising approach to forecasting the number of students that should be enrolled in a given school term by considering the high complexity of grouping and identifying useful patterns in the prediction process. In this work, we develop a predictive model based on the Random Forest (RF) algorithm to forecast the enrollment of students across the entire spectrum of Brazilian education. We use a database provided by the National Education Development Fund (FNDE), a Brazilian government body responsible for purchasing and distributing textbooks to all public schools. We generate 1,531,185 time series to serve as an input to RF processing. Our training dataset utilized data between 2010 and 2020, and our testing dataset utilizes data from 2021. As a result, RF obtains a higher performance in all the investigated scenarios concerning the Exponential Smoothing (ES) baseline algorithm. Since RF demonstrated acceptable performance, the Brazilian government could benefit from this forecasting technique for student enrollment in school environments and to ensure equitable access to essential resources, such as didactic materials, for the students.

Keywords. Machine Learning, Random Forest, forecasting, enrollment, Brazilian education
Research paper, DOI: https://doi.org/10.59490/dgo.2025.939

1. Introduction

Brazilian education is extensive and consists of several levels and institutions. As an instance, this system encompasses Early Childhood, Basic, Secondary, Professional, and Higher Education. Some facts highlight the width of our education system: (i) Brazil has around 178.3 thousand basic education schools in operation, serving students in the initial years (1st to 5th), with a total of 105.4 thousand schools, and in the final years (6th to 9th), totaling 61.8 thousand schools; (ii) The municipal network is mainly responsible for offering the initial years, registering approximately 10.1 million students in 2022 (69.3%), equivalent to 85.5% of the public network; (iii) In the final years of primary education, the municipal network serves 5.3 million students (44.4%), while the state network serves 4.8 million (39.9%) (Andrade, 2023).

In this sense, millions of students are enrolled yearly in public or private schools to ensure access to quality education. Government bodies must have as an essential part of their public policies the ensurence of the availability of educational resources (e.g., textbooks and other didactic materials) for schools and students. In Brazil, the National Education Development Fund (FNDE) is responsible for delivering textbooks for these schools and ensuring that the content of these textbooks is used exclusively for educational purposes.

However, it is challenging to precisely forecast the number of students being enrolled each year. Currently, FNDE distributes textbooks according to projections from the school census for the two years prior to the program year, as this is the information available at the time of processing the choice made by the schools. Such forecasts are carried out using the Exponential Smoothing method (Billah et al., 2006). Therefore, there may be small fluctuations between the number of books and the number of students. The greater this fluctuation, the greater the financial cost in the process, due to the excess of books that will be purchased and distributed. Furthermore, the lack of a precise number implies that many students do not receive textbooks during the school period, causing great harm to their education and hindering fair and equal educational development.

In recent literature, the context of the problem of predicting job enrollments is mainly in university environments (Ayasi et al., 2023; Khademi and Nakhkob, 2016; Shao et al., 2022) and has a significant number of related studies. In the school context, there is a minimal number of studies (e.g., Abideen et al., 2023) and the scope of the experiments is not representative, due to the small sample size. Our work has a much more comprehensive and representative scope, in addition to being applied to a complex and challenging scenario.

As highlighted (Scholl, 2024), the Digital Government Research (DGR) field has evolved significantly in recent years, especially in developing solutions for public administration. These solutions aim to bring government closer to society through technologies, such as the services offered to the population, social media as a means of communication, and Artificial Intelligence as a tool to support decision-making. In this sense, this paper investigates using Machibe Learning (ML) models to forecast the number of enrolled students (Sarker, 2021). We develop a predictive model based on the Random Forest (RF) algorithm (Pichler and Hartig, 2023) able to forecast the enrollment of students across the entire spectrum of Brazilian education.

The research methodology adopted for this work is based on the Design Science Research (DSR) paradigm (Gregor and Hevner, 2013; Hevner et al., 2004). To apply the DSR methodology, we followed the guidelines of (Peffers et al., 2007), which specify the process in a sequence of six activities: 1. Identify the problem and motivation; 2. Define the objectives for the solution; 3. Design and develop; 4. Demonstrate; 5. Evaluate; 6. Communicate. Figure 1 presents the DSR methodology process defined for this work.

Therefore, considering the complexity (Sobrinho et al., 2023) and the high cost (FNDE, 2022) of the Brazilian government to the processes involving the acquisition and distribution of books to students in public schools (Activity 1), our motivation is to improve decision-making by FNDE, proposing predictive models based on machine learning to more assertively and automatically infer the number of students at each stage of education in each school (Activity 2), minimizing the financial losses of the government (e.g., excessive purchases) and social losses (e.g., students without access to teaching materials).

We used a database provided by the FNDE to develop the proposed predictive model (Activity 3). We generate 1,531,185 times series to serve as an input to the RF processing. We also use a data mining process with an interpolate imputation method that mitigates null values in the dataset and a hold-out method that splits the dataset into training and testing sets. Our training dataset utilizes data proceeding from 2010 to 2020.

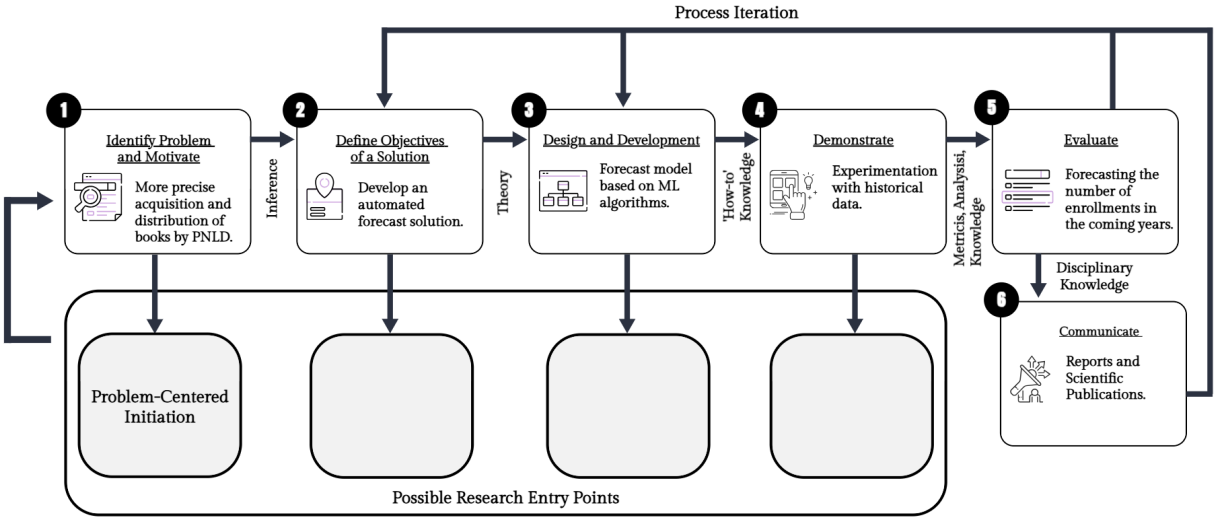

Fig. 1 – DSR Process. Adapted from Peffers et al., 2007.

For demonstration purposes (Activity 4), we used historical data from 2021 as our testing dataset and considered in the experiments the aspects of school grade, geographic location, Brazilian regions, federative unit, population size, dependence type, and school size. As a result, RF obtains higher performance in all the investigated scenarios in relation to the Exponential Smoothing (ES) baseline algorithm. The ES algorithm was chosen as a comparative baseline with the RF because in the related literature (Chen, 2022) it achieved better accuracy and stability in forecasting the number of high school enrollments and vocational education enrollment.

Since the RF demonstrated acceptable performance in forecasting student enrollments in school environments, the next step is to integrate the model into the system used by FNDE to forecast the enrollments in the coming years (Activity 5). Finally, for communication purposes (Activity 6), we wrote technical reports about the entire research and development process of the proposed predictive model and provided them to FNDE. Additionally, we carry out the process of scientific writing and publishing articles based on the results achieved with the experiments performed.

The main contributions of this study are two-fold:

1. the proposition of an ML model for automatic prediction of enrollment in public schools in Brazil aims to increase efficiency and reduce the time needed to perform this challenging task of forecasting the number of student enrollments in each class of each public school in a country of continental geographic dimension;
2. the provision of a solution to support FNDE's decision-making in effective planning for the distribution of textbooks in the country, with the possibility of reducing financial costs and social impacts on Education.

The justification for using ML algorithms as an alternative solution to the Exponential Smoothing method used by FNDE to forecast student enrollment is based on the ability of such algorithms to make predictions without the need for explicit programming, learning automatically from historical data (Sarker, 2021). This flexible and dynamic nature of learning makes ML algorithms the ideal solution for complex prediction problems, such as predicting student enrollment.

Studies such as those by (Fischer-Abaigar et al., 2024) and (Amarasinghe et al., 2023) emphasize the alignment of ML models with real-world complexities, ensuring that their outputs are actionable, explainable, and equitable. These works argue for a shift from purely predictive goals to decision-making frameworks that prioritize policy impact.

The work structure is presented as follows. Section 2 highlights some literature works that investigate the usage of ML techniques for enrollment forecasting-related issues. Section 3 describes the RF model that it

utilized as a baseline for running our experiments. Section 4 presents scenarios and methodology utilized in our experiments. Section 5 presents the results extracted from our experiments and discussion. In Section 6 we present our concluding remarks.

2. Related Work

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed, hence the term learning (Sarker, 2021). ML deals with the design and construction of systems that can automatically learn and improve from experience, typically by analyzing and extracting patterns from large amounts of data. The RF algorithm belongs to the ensemble learning category of ML algorithms using bagging and boosting techniques (Pichler and Hartig, 2023). They are renowned for their robustness and versatility in both classification and regression tasks. Within the ensemble framework, RF operates by constructing multiple decision trees during training and then aggregating their predictions to make a final prediction. With their ability to deliver accurate predictions and interpretability, RF has become a popular choice for various predictive modeling tasks across different domains, including Education (Sarker, 2021).

The task of accurately forecasting school enrollment rates is critical for school management. It allows to minimize unnecessary administrative costs as well as burdens to both students and faculty. In addition, it is possible to adequately plan for the number of classes, teachers, and resources needed to meet students' demands. The use of ML models to forecast student enrollment in schools has shown to be a promising approach to dealing with this problem, as observed in the literature (Abideen et al., 2023; Masini et al., 2023; Shao et al., 2022).

Several related works have emerged in the literature in recent years. (Feng et al., 2011) analyzed enrollment data from 25 randomly selected Chinese provinces as training data to predict enrollment rates in another six provinces using an Artificial Neural Network (ANN) model combined with association rules. The authors stated that the proposed model is viable and effective in supporting decision-making in university admissions.

(Khademi and Nakhkob, 2016) used classical ML models (Logistic Regression - LR, Naive Bayes - NB and Decision Tree - DT) and Artificial Neural Networks - ANN (MultiLayer Perceptron - MLPs) combined (*bagging* and *boosting*) to predict the increase in enrollment in universities in Iran. The results showed that the model was able to accurately predict the increase in enrollments based on historical data, student demographic information, and socioeconomic data from the region, reaching almost 96% accuracy with the best model.

A study designed by (Soltys et al., 2021) used an AI services framework from Amazon Web Services (AWS), SageMaker, along with standard Python tools for data analysis, including Pandas, NumPy, Matplotlib, and Scikit-Learn [1]. Based on three years of enrollment history, a model was built to calculate, individually or in batch mode, enrollment probabilities for certain candidates. The idea was to use these probabilities during admissions periods to target undecided students.

More recent research by (Ayasi et al., 2023) delves into the application of ML and neural network algorithms in forecasting future student enrollments in higher education courses. Utilizing real data from the Arab American University in Palestine (AAUP), the study employed eight ML algorithms alongside the Multilayer Perceptron (MLP) neural network model to anticipate the enrollment likelihood of students in specific courses. Findings indicate that ensemble-based and bagging algorithms outshine other classifiers, including neural network models, in forecasting individual-level student enrollments. Notably, the RF achieves a remarkable accuracy of 94% and an F1 score of 79% following the application of undersampling techniques to address dataset imbalance. The study suggests future research endeavors to develop a universal model for forecasting enrollment across all courses at AAUP, underscoring the efficacy of these techniques in enhancing resource allocation and student support.

(Shao et al., 2022) developed a study to replicate course enrollment forecasts through a conditional probability analysis utilizing student data from San Diego State University. Subsequently, the study sought to enhance these predictions by employing classification and regression trees (CART) and the RF algorithm. By integrating student demographic and academic information into the algorithms, the authors aimed to evaluate their

[1] https://scikit-learn.org/

impact on refining the accuracy of course enrollment predictions. Furthermore, the authors conducted an analysis to determine the most influential factors affecting General Chemistry enrollment numbers, utilizing a variable importance metric derived from tree-based algorithms.

Finally, (Abideen et al., 2023) analyzed five years of enrollment data from 100 schools within Punjab, Pakistan. Essential features have been extracted and analyzed using various ML algorithms, including Multiple Linear Regression, RF, and Decision Tree. These algorithms play a crucial role in forecasting future school enrollments and classifying target levels for each school. The findings enable a brief analysis of forthcoming registrations and target levels. Moreover, the proposed model aids in identifying solutions to address low enrollment rates in schools, thereby contributing to the enhancement of literacy rates within the region.

Therefore, the main differences between our work and the aforementioned related works are the following: 1. our scope is broader and is related to public schools in Brazil and their respective educational stages, while the cited literature has a more specific scope with a primary focus on universities around the world (e.g., USA, China, Iran, and Palestine); 2. our work involves a complex, challenging and socially impactful context, while the contexts of the aforementioned works do not present these characteristics; 3. we propose a solution using Random Forest, a traditional Machine Learning algorithm, simple to implement and that does not require as many computational resources to execute, while most of the solutions identified in the literature, despite making use of modern and robust algorithms, such as Artificial Neural Networks, require a better computational infrastructure.

3. Forecasting Student Enrollments

This section presents the methodological step-by-step for conceiving the predictive model based on Random Forest algorithm as a potential solution for projecting enrollment in classes in elementary education schools in Brazil.

3.1. Dataset

In this work, we designed a predictive model using a dataset provided by FNDE. The dataset comprised 10,980,537 enrollment records of Brazilian Basic Education schools. To filter the dataset, we used the year and school grade code criteria, which resulted in 33 teaching stages and 12 years of school census data, spanning from 2010 to 2021. In addition, the dataset contains a total of 221.771 schools distributed in 5.570 Brazilian cities. The school records has six variables, as illustrated in Table 1: (a) Census Collection Year (YEAR); (b) Municipal Code (MUN_CO); (c) School Code (CD_SCH); (d) School Grade Code (GRADE_CD); and (e) Number of Students (STD_N).

Tab. 1 – A random sample with five records from the investigated dataset.

YEAR	MUN_CO	CD_SCH	GRADE_CD	STD_N
2010	43016413	4301107	15	1
2013	21097690	2105500	19	78
2012	23165820	2307304	20	11
2014	22134484	2210938	21	16
2019	21001219	2101301	16	2

In the predictive analysis proposed in this study, we only used the variables YEAR and STD_N. The other variables were used only to compare the performance results of the models according to the scenarios defined for the experiments.

Table 2 presents the descriptive statistics of the main variables present in the dataset, providing an idea of the data profile.

As seen in Table 2, the variable STD_N (target) presents an average of 49.29 students enrolled, with a median of 26 and standard deviation of 75.36. The minimum number of registered enrollments is equal to 1, while the maximum number is 31,542. These numbers characterize a discrepant variation in enrollments among Brazilian schools.

Tab. 2 – Descriptive Statistics of the dataset provided by the FNDE.

Variable	Mean	Median	Std. Dev.	Min.	Max.
YEAR	-	-	-	2010	2021
STD_N	49.29	26	75.36	1	31,542

3.2. Model

The biggest challenge in time series forecasting with traditional ML methods is setting up the problem correctly. There are basically three different possibilities by which we can classify a time series forecasting problem as a Supervised ML problem, namely:

1. Predict the next time step using previous observations;
2. Predict the next time interval using a sequence of past observations;
3. Predict a sequence of future time steps using a sequence of past observations.

In this work, we will frame the problem as non-prediction of school enrollment for the coming year using historical data from previous years. In this configuration, the model will generate a prediction for the next time step, considering only the previous observation.

For our analysis, it was necessary to organize the dataset in such a way that previous observations (historical time series) become a resource to predict the next observation (forecasting of enrollments in subsequent years). Therefore, we added a second column (y) that shifts the STD_N column so that the value in one year (red line, e.g., 2019) becomes a predictor of the value for the next year (blue line, e.g., 2020), allowing the comparison shown in Figure 2.

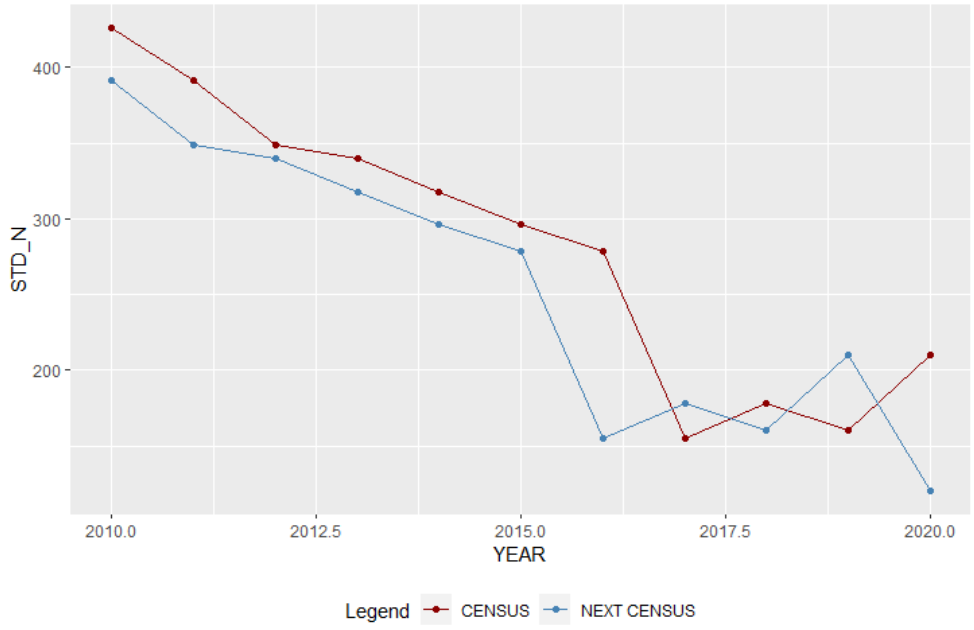

Fig. 2 – Example of viewing the series 15042537.

After arranging the data for the analysis proposed in the work, we started the experiments, as described in Section 4.

4. Experiment

In this work, we develop an experiment to validate the efficiency of RF into scenarios of forecasting Student Enrollments. The following metrics are utilized to measure this efficiency: (i) R^2 (R-squared) (Hiregoudar,

2020); (ii) MAE (Mean Absolute Error) (Penteado, 2021); (iii) MAPE (Mean Absolute Percentual Error) (Penteado, 2021); (iv) MASE (Mean Absolute Scaled Error) (Hyndman and Koehler, 2006); (v) MSE (Mean Squared Error) (Penteado, 2021); (vi) RMSE (Root Mean Squared Error) (Penteado, 2021); (vii) RAE (Relative Absolute Error) (Hiregoudar, 2020); and (viii) U-Theil (Bliemel, 1973).

4.1. Scenarios

The scenario utilized in this work is described as follows. We adopt Brazilian primary schools as our baseline scenario to perform forecasting techniques. In that sense, we investigate the dataset with the following concerns:

- **Global:** it encompasses the application of metrics along all the considered dataset with no filtering utilization. In this case, we consider previously considered filters, being deleted null data.
- **Teaching Stage:** it encompasses grouped data in according with the teaching stage, thus being defined by 36 categories;
- **Geographic Location:** it encompasses grouped data in according with the location in the urban perimeter of schools, thus being defined by Urban and Rural categories;
- **Brazilian Regions:** it encompasses grouped data in according with the location of schools in Brazilian Regions, thus being defined by the categories: (a) North; (b) Northeast; (c) East Center; (d) Southeast; and (e) South;
- **Federative Unit:** it encompasses grouped data according to the location of 27 Brazilian states;
- **Population Size:** it encompasses grouped data according to the population ranges of cities and municipals that encompass such schools, thus being defined by the following categories: (i) Up to 20 thousand inhabitants; (ii) Between 20 and 50 thousand inhabitants; (iii) Between 50 and 100 thousand inhabitants; (iv) Between 100 and 500 thousand inhabitants; (v) Between 500 and 1 million inhabitants; and (vi) Over 1 million inhabitants;
- **Dependency Type:** it encompasses grouped data by school sector, and its categorization is: (1) Federal, (2) State, (3) Municipal, (4) Private, and (5) All Public.
- **School Size:** categorized into five classes (1 to 5), the first being the smallest (smallest range of total number of students per school) and the last being the largest, the entities/schools according to their statistical distribution in the population (quantiles) for each year analyzed.

To carry out our experiments, we grouped the data according to each scenario mentioned above. In this way, we were able to discuss the performance results of RF and ES algorithms (baseline) according to the context analyzed in each experiment.

4.2. Design Methodology

Figure 3 presents the steps of the design methodology defined for the model development and evaluation.

Starting from a dataset provided by FNDE, we generate 1.531.185 times series in order to serve as an input to RF processing. The composition of these series is performed by grouping the following columns: (a) *CD_SCH*; (b) *MUN_CO*; and (iii) *GRADE_CD*. Likewise, we estimate the regression value for the reference year 2021.

We utilize the *interpolate* imputation method (Robinson and Hamann, 2011) to mitigate absent values in the investigated dataset since it presents acceptable performance to reduce dataset-related prediction errors. Moreover, we also use the *hold-out* method (Sammut and Webb, 2017) to slice the dataset into two subsets: a training one, used to tune the model, and a testing one, used to evaluate the model's performance in forecasting the data not belonging to the training subset. We have partitioned the training dataset with data proceeding from 2010 to 2020 to each entity, and we define the testing dataset with data proceeding from 2021.

Each school and teaching stage is represented by the regression estimated value. Therefore, such value is obtained for each time series and concatenated, thus generating a new dataset. Therefore, the model was evaluated by processing error measurement metrics and model evaluation. We utilize a *random state* in the enrollment forecasting model using the RF in order to ensure scientific-level repeatability. Although the RF model was designed based on the pre-processed dataset, the prediction results were filtered by each scenario to allow for a more in-depth discussion.

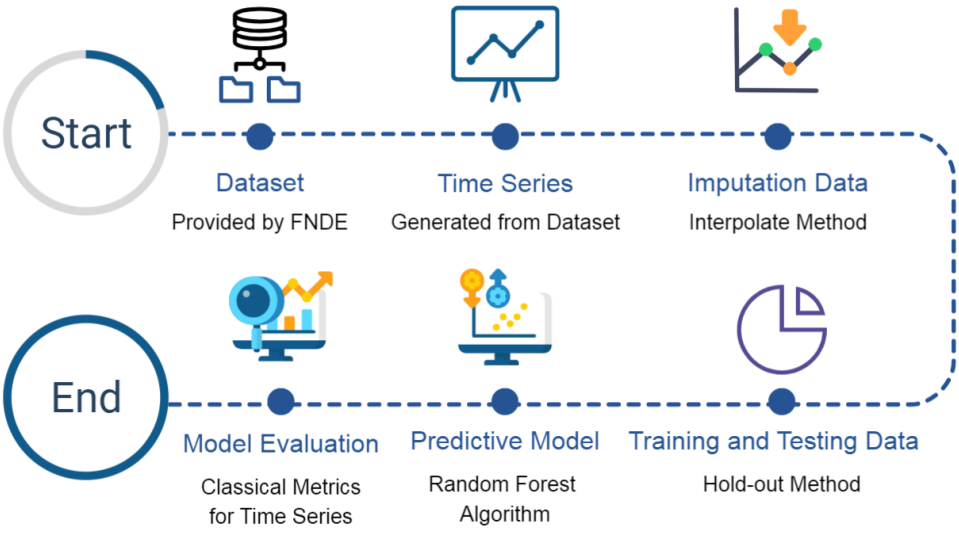

Fig. 3 – Development and Evaluation Process.

Finally, the computing environment used to perform the experiments consists of a computer with an Intel(R) Xeon(R) E-2244G processor, with CPU cores of 3.80GHz, and 32.0 GB of RAM memory. We understand this information is relevant since processing time is an important factor in the practical application.

5. Results

This section presents the results and discussion on predicting enrollment in Brazilian schools using the Random Forest (RF) algorithm and comparing its performance with the baseline model that uses the Exponential Smoothing (ES) technique.

Table 3 presents the *benchmarking* results of the designed model by contrasting the RF algorithm with the ES model in the "Global" scenario with complete dataset. In this scenario, except for U-Theil, we observed a better performance of the RF in all other metrics evaluated.

Tab. 3 – Evaluation of RF *vs.* ES (baseline) models for the "Global" scenario.

Model	R^2	MAE	MAPE	MSE	RMSE	MASE	RAE	U-Theil
ES	0.83	12.52	46.97	941.63	30.69	0.44	0.29	**0.25**
RF	**0.87**	**11.62**	**42.01**	**745.04**	**27.30**	**0.41**	**0.27**	0.26

In other investigated scenarios, we first have calculated the values of each metric in each subcategory and then we have calculated the average. For example, in the "Teaching Stage" scenario, the comparative results between the ES (baseline) model and the RF algorithm are presented in Table 4. As a result, the performance of the models in this scenario is similar to that of the "Global" scenario, but with a reduction in R^2 and an increase in the values of the error metrics, with the exception of MASE.

Tab. 4 – Evaluation of RF *vs.* ES (baseline) models for the "Teaching Stage" scenario.

Model	R^2	MAE	MAPE	MSE	RMSE	MASE	RAE	U-Theil
ES	0.76	21.70	57.45	2952.4	41.49	0.35	0.41	**0.507**
σ ($\pm$)	0.20	14.00	45.70	5975.0	35.60	0.14	0.18	0.230
RF	**0.79**	**20.20**	**52.80**	**2537.18**	**38.43**	**0.332**	**0.38**	0.514
σ ($\pm$)	0.18	12.70	36.20	5313.00	33.10	0.130	0.17	0.230

We applied the paired-samples t-test (Xu et al., 2017), where the null hypothesis (H_0) assumes that the true mean difference (μd) is equal to zero. Table 5 shows the *p-values* calculated for each metric used to evaluate

the model's performance. Considering the statistical significance of $p \leq 0.05$, when analyzing the *p-values* of all metrics, we found that they are lower than 0.05. Therefore, we confirm the hypothesis that there is a statistical difference between the performance results of the RF and ES models.

Tab. 5 – T-test - Paired Two Sample for Means: "Teaching Stage" scenario.

Metric	***p-value***
R^2	0.0005579
MAE	9.865728e-06
MAPE	0.02979593
MSE	0.02354106
RMSE	0.0009053389
MASE	1.362274e-06
RAE	3.476034e-07
U-Theil	0.0012871

In relation to "Geographic Location", the results presented in Table 6 demonstrate a superiority of RF, with emphasis on the R^2 metric.

Tab. 6 – Evaluation of RF *vs.* ES (baseline) models for the "Geographic Location" scenario.

Model	R^2	**MAE**	**MAPE**	**MSE**	**RMSE**	**MASE**	**RAE**	**U-Theil**
ES	0.771	11.57	48.2	841	25.6	0.499	0.389	**0.212**
σ ($\pm$)	0.042	9.31	12.3	989	19.3	0.066	0.014	0.009
RF	**0.81**	**10.75**	**43.3**	**666**	**22.9**	**0.466**	**0.363**	0.222
σ ($\pm$)	0.042	8.58	12.8	774	16.9	0.066	0.016	0.007

We conducted a paired t-test to compare the means of each metric of each model across different scenarios, similar to what we did for the "Teaching Stages" scenario. The results for the "Geographic Location" scenario are presented in Table 7. The obtained *p-values* for the metrics R^2, MAPE, MASE, and RAE were below 0.05, indicating a statistical difference between the means of these metrics for the ES model compared to the RF model. However, there was no statistical difference observed for the MAE, MSE, RMSE, and U-Theil metrics, suggesting similar performance of both models for these metrics.

Tab. 7 – T-test Paired Two Sample for Means: "Geographic Location" scenario.

Metric	***p-value***
R^2	0.003233613
MAE	0.3573881
MAPE	0.04610538
MSE	0.4554039
RMSE	0.3571983
MASE	0.0002570144
RAE	0.04314648
U-Theil	0.1019322

In the "Brazilian Regions" scenario (see Table 8), the RF has achieved a better performance than the ES model (*baseline*). This behavior is repeated in scenarios that encompass data grouped by "Federation Unit", "Population Size" and "Type of Dependency", as shown in Tables 10, 12, 14, and 16, respectively.

We show the results of the paired t-test for the "Brazilian Regions" scenario in Table 9. The *p-values* calculated for all metrics allow us to reject the null hypothesis and conclude that the means are statistically different when comparing the performance of the ES and RF models. A similar result is also observed in the "Federative Unit" scenario, as shown in Table 11.

The results of the paired t-test for the "Population Size", "Dependency Type", and "School Size" scenarios are shown, respectively, in Tables 13, 15, 17. In these scenarios, there were many variations in results. For example, in the "Population Size" scenario (see Table 13), except the *p-value* obtained for the R^2 metric, the other

Tab. 8 – Evaluation of RF *vs.* ES (baseline) models for the "Brazilian Regions" scenario.

Model	R^2	MAE	MAPE	MSE	RMSE	MASE	RAE	U-Theil
ES	0.808	12.85	46.68	935	30.37	0.444	0.314	**0.272**
σ ($\pm$)	0.075	2.75	7.51	227	3.99	0.022	0.039	0.021
RF	**0.841**	**11.90**	**41.63**	**760**	**27.35**	**0.411**	**0.291**	0.281
σ ($\pm$)	0.076	2.60	7.03	199	3.83	0.021	0.038	0.020

Tab. 9 – T-test Paired Two Sample for Means: "Brazilian Regions" scenario.

Metric	*p-value*
R^2	0.00267542
MAE	0.0006234588
MAPE	3.944342e-05
MSE	0.01473016
RMSE	0.01258553
MASE	0.0001404595
RAE	8.05893e-05
U-Theil	0.0002031214

Tab. 10 – Evaluation of RF *vs.* ES (baseline) models for the "Federative Unit" scenario.

Model	R^2	MAE	MAPE	MSE	RMSE	MASE	RAE	U-Theil
ES	0.783	13.02	47.46	988	29.8	0.443	0.319	**0.350**
σ ($\pm$)	0.095	4.18	7.91	860	10.1	0.047	0.049	0.074
RF	**0.824**	**11.99**	**42.21**	**805**	**27.01**	**0.409**	**0.294**	0.357
σ ($\pm$)	0.072	3.79	7.18	683	8.85	0.044	0.046	0.072

Tab. 11 – T-test Paired Two Sample for Means: "Federative Unit" scenario.

Metric	*p-value*
R^2	0.0001219464
MAE	9.750767e-12
MAPE	2.814718e-15
MSE	0.0006719472
RMSE	1.774589e-05
MASE	1.220686e-18
RAE	3.434326e-18
U-Theil	7.375808e-12

p-values had values lower than 0.05, indicating the rejection of the null hypothesis. Therefore, in this scenario, we can conclude that there is a statistical similarity between the means of the ES and RF models only for the R^2 metric.

Tab. 12 – Evaluation of RF *vs.* ES (baseline) models for the "Population Size" scenario.

Model	R^2	MAE	MAPE	MSE	RMSE	MASE	RAE	U-Theil
ES	0.778	14.99	44.66	1232	33.5	0.441	0.335	**0.270**
σ ($\pm$)	0.097	6.26	5.57	880	11.3	0.008	0.041	0.012
RF	**0.84**	**13.95**	**39.74**	**1027**	**30.1**	**0.409**	**0.31**	0.279
σ ($\pm$)	0.03	5.94	5.51	862	12.2	0.009	0.04	0.011

For the "Dependency Type" scenario (see Table 15), the *p-values* for the error metrics MSE, MASE, RAE, and U-Theil exceeded the value of 0.05, which leads us to accept the null hypothesis that the means of these metrics, for both ES and RF models, are statistically equal. For the other error metrics and the R^2 metric, the means

Tab. 13 – T-test Paired Two Sample for Means: "Population Size" scenario.

Metric	***p-value***
R^2	0.1235081
MAE	0.0006663735
MAPE	1.503282e-05
MSE	0.005347984
RMSE	0.0182089
MASE	3.38188e-06
RAE	2.196989e-06
U-Theil	7.119195e-06

Tab. 14 – Evaluation of RF *vs.* ES (baseline) models for the "Dependency Type" scenario.

Model	R^2	**MAE**	**MAPE**	**MSE**	**RMSE**	**MASE**	**RAE**	**U-Theil**
ES	0.791	17.02	45.71	1991	38.88	0.450	0.35	0.453
σ ($\pm$)	0.053	8.81	18.60	2504	24.50	0.060	0.06	0.303
RF	**0.851**	**15.11**	**40.84**	**1512**	**33.75**	**0.390**	**0.30**	**0.446**
σ ($\pm$)	0.025	8.54	17.6	1841	21.60	0.052	0.05	0.270

are statistically different.

Tab. 15 – T-test Paired Two Sample for Means: "Dependency Type" scenario.

Metric	***p-value***
R^2	0.02513679
MAE	0.03603456
MAPE	0.0005633837
MSE	0.1921146
RMSE	0.0353659
MASE	0.09901843
RAE	0.09540237
U-Theil	0.6418034

Tab. 16 – Evaluation of RF *vs.* ES (baseline) models for the "School Size" scenario.

Model	R^2	**MAE**	**MAPE**	**MSE**	**RMSE**	**MASE**	**RAE**	**U-Theil**
ES	0.659	12.53	46.9	943	24.9	0.493	0.46	0.507
σ ($\pm$)	0.204	9.60	14.1	1393	20.2	0.063	0.09	0.203
RF	**0.70**	**11.64**	**41.9**	**746**	**22.3**	**0.461**	**0.43**	**0.453**
σ ($\pm$)	0.21	8.88	14.1	1079	17.6	0.066	0.09	0.247

Finally, for the "School Size" scenario (see Table 17), in addition to the *p-value* of R^2, the *p-values* for the MAPE, MASE and RAE metrics were less than 0.05, allowing us to conclude that the means are statistically different for the ES and RF models. Differently, the averages for the MAE, MSE, RMSE, and U-Theil metrics are statistically similar, as the *p-values* were greater than 0.05.

Our experiments have presented relevant results, thus enabling the comparison of the ES (*baseline*) model in time series with the performance of the RF algorithm. Given the experiment results, we understand that there is a variation in performance of the RF-based model in accordance with the scenario. For example, in the "Teaching Stages" and "School Size" scenarios, both models performed less than 0.8 for the R^2 metric. In the scenarios "Geographic Location", "Federative Unit", "Population Size" and "Dependency Type", only the RF model performed better than 0.8 for the R^2 metric. Only in the Global and "Brazilian Regions" scenarios did both models exceed the value of 0.8 for the R^2 metric.

When it comes to the error metrics MAE, MAPE, MSE, RMSE, MASE, and RAE, the RF model reduced the value of

Tab. 17 – T-test Paired Two Sample for Means: “School Size” scenario.

Metric	***p-value***
R^2	8.62111e-05
MAE	0.05070706
MAPE	5.064793e-05
MSE	0.234543
RMSE	0.09082856
MASE	0.0004177525
RAE	0.0004358726
U-Theil	0.3324351

the prediction error in all scenarios: “Global”, “Teaching Stages”, “Geographical Location”, “Brazilian Regions”, “Federative Units”, “Population Size”, and “Type of Dependency”.

For the U-theil metric, specific to evaluating predictive models for time series, the baseline model outperformed RF only in the “Dependency Type” and “School Size” scenarios.

Considering all the variations of the analyzed scenarios, we obtained 86 results in total. Analyzing the frequency of the best performing models (ES *vs.* RF), the supremacy of the RF algorithm was evident in all scenarios, with a processing time of 160934.6 seconds.

As mentioned in Section 2, this work focuses on public schools in Brazil. However, our solution can be adopted in other countries with different realities. While it is most suitable for countries similar to Brazil—particularly developing nations with heterogeneous income distribution and economies influenced by major world powers—our approach can also benefit underdeveloped countries, especially those in the Global South. These countries, despite having low socioeconomic indicators and the Health and Education sectors are mostly the government’s responsibility, can utilize our solution due to its simplicity and minimal reliance on computational resources.

As mentioned in Section 2, this work focuses on public schools in Brazil. However, our solution can be adopted in other countries with different realities. While it is most suitable for countries similar to Brazil—particularly developing nations with heterogeneous income distribution and economies influenced by major world powers—our approach can also benefit underdeveloped countries, especially those in the Global South. These countries, despite having low socioeconomic indicators and , can utilize our solution due to its simplicity and minimal reliance on computational resources.

6. Concluding Remarks

The Brazilian National Textbook Program (Programa Nacional do Livro e do Material Didático — PNLD) is one of the largest and most democratic government programs for education in Brazil. It was created in 1938, but was entitled of PNLD in 1985, to distribute textbooks, pedagogical and literary books, and other educational materials to teachers and students in public schools throughout the country (Sobrinho et al., 2023). Given the vast geographic, economic, and cultural diversity of the country, it is not easy to estimate the exact number of students in each class at each school in every Brazilian municipality.

This paper highlights the untapped potential of machine learning in transforming enrollment forecasting into a powerful tool for public policy. The study investigated a RF-based ML Model to forecast student enrollment in Brazilian schools. We used as a scenario the public school, and we created a dataset to carry out this forecasting. RF with ES as baseline to verify the performance of enrollment forecasts in public schools. The results show an improvement in the distribution process, allowing children to have access to books and improving equity. As a consequence, implementing a solution that is less likely to have prediction errors can significantly reduce the issue of having too many books in schools, which leads to financial waste. Additionally, this solution can also solve the problem of students not receiving books due to their unavailability, which results in inestimable losses in terms of their education.

This type of experiment seeks to provide more effective solutions compared to the ES model used by the Brazil-

ian government, as it is not just about forecasting the number of enrollments in classes in basic education schools in Brazil. The use of this information becomes even more important for government bodies responsible for Brazilian Education to plan the acquisition of teaching resources, such as books, to meet the demands of society.

In this context, having an effective solution for predicting the number of students who will enroll in each class in each school the following year allows the government to make more efficient decisions regarding purchasing books and other didactic materials to be distributed in schools. Thus, regarding the state of the practice, we characterize the advancement of our work by proposing a predictive model based on a machine learning algorithm that does not require assumptions (i.e., Random Forest), unlike the Exponential Smoothing method, which requires data to be stationary and predictions to be short-term. Therefore, our solution assists the Brazilian government in optimizing resources for education, ensuring equal access to didactic materials for all students in public schools.

As limitations, we identified the following points: (i) During the descriptive and graphical analysis stage of the data present in the dataset, we observed that there were observations with values that deviated from the standards of the large mass of data. However, we decided not to analyze in this study the influence and impact of these data on the performance of the models, baseline, and RF, therefore characterizing a limitation of this study. (ii) We only use historical data relating to the number of enrollments in previous years as predictors to forecast the number of enrollments in the following year. This decision limited the solution to a reduced set of variables that, perhaps alone, were insufficient to explain the value of the target variable.

In order to work on the limitations, we suggested the following future works: (i) investigate the characteristic observations of outliers and influential points in the dataset that may impact the performance of the models; (ii) enhance the performance of our proposed predictive model, adding more predictor variables. For this purpose, we can use the microdata from the Brazilian basic education census provided by INEP. This dataset contains over 300 attributes, which should provide us with ample options to choose other variables. Once we have added these variables, we need to evaluate the impact of this change on the model's performance; (iii) optimize the model parameters through *tunning* techniques such as *GridSearch* or *RandomSearch*, which have the potential to improve prediction and evaluation results of predictive models; (iv) In addition, other ML models will be tested, aiming to present a fairer comparative analysis and seek even better solutions to the enrollment forecast problem; (v) Finally, we will present the results of this work to the FNDE and propose integrating our solution into the enrollment prediction system used by the Brazilian government.

Acknowledgement

We thank the members of the Center for Excellence in Social Technologies (NEES), who collaborated with this research. We also thank the National Education Development Fund from Brazil for supporting this research through the Decentralized Execution Term (TED) 12244.

Funding

We thank the FNDE for supporting this research through the Decentralized Execution Term (TED) (Grant Number: TED 12244).

Data Access Statement

We also thank FNDE for providing the datasets that supported our research.

Contributor Statement

We thank the members of the Center for Excellence in Social Technologies (NEES) who collaborated with this research.

Conflict Of Interest (COI)

Author Bruno Pimentel has received research grants from the FNDE in Brazil. The remaining authors are members of the NEES.

References

Abideen, Z. u., Mazhar, T., Razzaq, A., Haq, I., Ullah, I., Alasmary, H., & Mohamed, H. G. (2023). Analysis of Enrollment Criteria in Secondary Schools Using Machine Learning and Data Mining Approach [Publisher: MDPI]. *Electronics*, *12*(3), 694. Retrieved February 15, 2024, from https://www.mdpi.com/2079-9292/12/3/694

Amarasinghe, K., Rodolfa, K. T., Lamba, H., & Ghani, R. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy*, *5*, e5. DOI: https://doi.org/10.1017/dap.2023.2.

Andrade, J. (2023, February). Censo escolar: Matrículas na educação básica cresceram em 2022 [Last accessed 19 Feb. 2024].

Ayasi, B., Saleh, M., García-Vico, A., & Carmona, C. J. (2023, December). Predicting course enrollment with machine learning and neural networks: A comparative study of algorithms. ISTES Organization Monument, CO, USA.

Billah, B., King, M. L., Snyder, R. D., & Koehler, A. B. (2006). Exponential smoothing model selection for forecasting. *International Journal of Forecasting*, *22*(2), 239–247. DOI: https://doi.org/https://doi.org/10.1016/j.ijforecast.2005.08.002.

Bliemel, F. (1973). Theil's forecast accuracy coefficient: A clarification. *Journal of Marketing Research*, *10*(4), 444–446. Retrieved January 5, 2024, from http://www.jstor.org/stable/3149394

Chen, Q. (2022). A comparative study on the forecast models of the enrollment proportion of general education and vocational education. *International Education Studies*, *15*(6), 109–126.

Feng, S., Zhou, S., & Liu, Y. (2011). Research on data mining in university admissions decision-making [Last accessed 20 Jun. 2023.]. *International Journal of Advancements in Computing Technology*, *3*(6), 176–186. https://doi.org/10.4156/ijact.vol3.issue6.21

Fischer-Abaigar, U., Kern, C., Barda, N., & Kreuter, F. (2024). Bridging the gap: Towards an expanded toolkit for AI-driven decision-making in the public sector. *Government Information Quarterly*, *41*(4), 101976. DOI: https://doi.org/10.1016/j.giq.2024.101976.

FNDE. (2022, January). Em 2021 foram investidos r$1,9 bilhoem livros e material didtico do PNLD$. https://www.gov.br/fnde/pt-br/assuntos/noticias/em-2021-foram-investidos-1-9-bilhao-em-livros-e-material-didatico-do-pnld

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Q.*, *37*(2), 337–356. DOI: https://doi.org/10.25300/MISQ/2013/37.2.01.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Q.*, *28*(1), 75–105.

Hiregoudar, S. (2020, August). Ways to evaluate regression models [Accessed 23 Feb. 2023]. https://towardsdatascience.com/ways-to-evaluate-regression-models%20-77a3ff45ba70

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. DOI: https://doi.org/https://doi.org/10.1016/j.ijforecast.2006.03.001.

Khademi, M., & Nakhkob, B. (2016). Predicted increase enrollment in higher education using neural networks and data mining techniques. *Journal of Computer Research and Development*, *7*, 125–140.

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76–111. DOI: https://doi.org/10.1111/joes.12429.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & and, S. C. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. DOI: https://doi.org/10.2753/MIS0742-1222240302.

Penteado, K. (2021, June). Métricas de avaliação para séries temporais [Accessed 22 Feb. 2023]. https://www.alura.com.br/artigos/metricas-de-avaliacao-para-se-%20ries-temporais

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, *14*(4), 994–1016. DOI: https://doi.org/10.1111/2041-210X.14061.

Robinson, A. P., & Hamann, J. D. (2011). Imputation and interpolation. In *Forest analytics with r: An introduction* (pp. 117–151). Springer New York. DOI: https://doi.org/10.1007/978-1-4419-7762-5_4.

Sammut, C., & Webb, G. I. (Eds.). (2017). *Encyclopedia of machine learning and data mining*. Springer. DOI: https://doi.org/10.1007/978-1-4899-7687-1.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. DOI: https://doi.org/10.1007/s42979-021-00592-x.

Scholl, H. J. (2024). Digital government research: Evolution of topical directions. *Proceedings of the 25th Annual International Conference on Digital Government Research*, 423–433. DOI: https://doi.org/10.1145/3657054.3657106.

Shao, L., Ieong, M., Levine, R. A., Stronach, J., & Fan, J. (2022). Machine Learning Methods for Course Enrollment Prediction. *Strategic enrollment management quarterly*, *10*(2). Retrieved February 15, 2024, from https://par.nsf.gov/servlets/purl/10389427

Sobrinho, A., Bittencourt, I. I., Silveira, A. C. M., Silva, A. P., Dermeval, D., Marques, L. M., Rodrigues, N. C. I., Souza, A. C. S., Ferreira, R., & Isotani, S. (2023). Towards digital transformation of the validation and triage process of textbooks in the brazilian educational policy. *Sustainability*, *15*(7). DOI: https://doi.org/10.3390/su15075861.

Soltys, M., Dang, H. D., Reyes Reilly, G., & Soltys, K. (2021). Enrollment predictions with machine learning [Last accessed 20 Jun. 2023]. *Strategic Enrollment Management Quarterly*, *9*(2), 11–18. https://eric.ed.gov/?id=EJ1311204

Xu, M., Fralick, D., Zheng, J., Wang, B., Tu, X., & Feng, C. (2017). The differences and similarities between two-sample t-test and paired t-test. *Shanghai Arch Psychiatry*, *29*(3), 184–188.