A Computational Approach to Feature Selection and Enrollment Forecasting in Brazilian Schools

Lenardo Silva, Gustavo Oliveira, Luciano Cabral, Rodrigo Silva, Luam dos Santos, Thyago de Oliveira, Breno da Costa, Dalgoberto Pinho Júnior, Nicholas da Cruz, Rafael Silva, and Bruno Pimentel

Center for Excellence in Social Technologies
Av. Lourival Melo Mota, S/N, Maceió, 57072-970, Alagoas, Brazil

Abstract. In this study, we used a dataset from the Brazilian school census provided by the Ministry of Education to identify relevant attributes for forecasting the number of students enrolled in a school. This dataset contains 340 characteristic attributes of schools and their respective teaching stages. The large quantity and nature of this data make data analysis more complex, which requires an appropriate method for feature selection to enrollment predictive models. In this sense, this study explores the application of Machine Learning algorithms as a solution to the problem of predicting enrollment, including random forest, multilayer perceptron, linear regression, and support vector regression. We assessed the models' performance using cross-validation, calculating the MAE, MSE, and RMSE metrics and the algorithms' execution time. The results revealed that the Spearman correlation method with thresholds of 0.6 and 0.65 can reduce the dimensionality of the data and the execution time of the predictive models.

Keywords: Feature Selection, Comparative Analysis, Forecasting, Enrollment Schools, Brazil.

1 Introduction

Forecasting enrollment is a critical aspect of any educational institution's annual planning. Accurately predicting the number of students who will attend a school in the following year is essential for numerous reasons, including budget planning and resource allocation, to ensure that the school adapts to its students' new needs [1].

Education problems have become even more critical, especially with the global education crisis caused by the COVID-19 pandemic, severely affecting student learning and resource allocation [2]. This issue becomes even more important for public schools in developing countries, especially in the Global South, due to difficulties related to economic and social issues, which hinder access to information [3].

In Brazil, this scenario is no different, being even more challenging due to its large geographic size and comprehensive educational system [4, 5]. These characteristics, for example, require government logistics to distribute books and teaching materials to all public schools in the country, especially those located in hard-to-reach regions and without adequate infrastructure (e.g., transportation and communication and information technology) so that these materials reach their destinations on time, in sufficient quantities for all students and without financial waste.

Brazilian education includes various levels and institutions, covering Early Childhood, Basic, Secondary, Professional, and Higher Education. Some facts that illustrate the breadth of our education system are: (i) in 2023, Brazil had 47.3 million students across all levels of education, attending 178.5 thousand schools; (ii) approximately 4.1 million enrollments were recorded in Early Childhood Education, which is almost half of the population up to 3 years of age; (iii) Elementary Education, with 26.1 million enrollments and 121.4 thousand schools, represents the majority of basic education students. The municipal network is mainly responsible for offering the initial years by registering

David C. Wyld et al. (Eds): SIGI, CSTY, AI, NMOCT, BIOS, AIMLNET, MaVaS, BINLP – 2025
pp.109-128, 2025.CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.151910

approximately 10 million students in 2023 (69.5%), equivalent to 86.1% of the public network; (iv) in High School, there were 7.7 million enrollments, with the state network having the largest participation at 83.6%, accounting for 6.4 million students [6].

In this context, we defined for this study the following Research Question (RQ): *"Can the data from the Brazilian government's school census be used to predict the number of students for the next year? If so, which variables can be used in this process?"*. For this purpose, we used a dataset provided by the Ministry of Education (MEC), namely microdata, which contains 218,598 schools and 340 characteristic attributes of schools and their respective teaching stages. Considering the large number of attributes, we evaluate feature selection methods to optimize the predictors number for creating an enrollment forecasting model using Machine Learning (ML) algorithms. Therefore, the proposed data analysis is challenging due to the high dimensionality of the data and the particularities of the schools.

By answering this RQ, our goal is to define an experimental protocol and provide a model using ML algorithms to predict the number of primary and secondary school students to enroll in Brazilian public schools each upcoming school year who will benefit from the National Textbook Program (PNLD).

The contributions of this work are (i) an experimental protocol on applying machine learning algorithms to predict the number of public education students; (ii) improved interpretation through feature selection with correlation techniques and algorithm performance evaluation to identify the factors that most influence predictions and their implications, and; (iii) supporting strategic decisions in educational policies, reducing the government cost to provide scholar books.

2 Related Work

School enrollment forecasting models involve many factors to consider and techniques to choose from to ensure an accurate prediction. Factors can vary according to, for example, the type of institution (private *vs.* public), the type of enrollment (full *vs.* partial), and the purpose of the prediction (budget *vs.* staffing) [7]. Therefore, each model identified in the literature has needs specific to the problems addressed and the datasets used in proposing solutions, as evidenced in the works listed below.

Singh (2007) [8] presented an improved method for fuzzy time series forecasting. The model developed uses the differences in the production of the past 3 years and has been considered a fuzzy parameter in framing the fuzzy rules to impose on current year fuzzified enrollment to get a forecast of next year's enrollments. It was implemented based on the historical student enrollment data of the University of Alabama (1971 - 1992). As a result of its performance, was obtained an MSE (mean square error) equal to 87.025 and an Average Error of approximately 1.56.

Stanley (2008) [9] provided an enrollment forecast model that aids decision-making to increase student enrollment and, consequently, revenue, at no additional cost. The authors used prospective student records from all applicants between the fall of 2002 and 2006. A predictive model based on a decision tree algorithm was proposed, and its performance was compared with logistic regression and neural networks. The resulting model achieved an accuracy rate of 93.2% on both the training and test sets.

In the study conducted by Wang, Zhuang, and Liu (2010) [10], the aim was to predict the scale of postgraduate education in Hebei Province to be achieved during the 12th and 13th Five-Year Plans. The data used in the study were admissions to postgraduate programs at the master's and doctoral levels between 2000 and 2008, as well as data

on the average annual Gross Domestic Product (GDP) growth rate in Hebei Province between 2001 and 2007. Two models were proposed: the first used a Weighted Moving Average (WMA), while the second applied Linear Regression. The authors did not report any metrics for evaluating the models. In the reported results they observed a Hebei Province's annual average GDP growth rate was 11.77%, while the average growth rate for postgraduate enrollment was significantly higher at 24.57%.

Borah *et al.* (2011) [11] applied a knowledge-based decision technique to guide the student for admission to the proper branch of engineering. The authors proposed a new Attribute Selection Measure Function based on a heuristic combined with the C4.5 algorithm. This work used a dataset with 65,534 records of the All India Engineering Entrance Examination (AIEEE) 2007 and surveys from different engineering institutions across India. The authors used the Decision tree (C5.0) and Back Propagation algorithm (Artificial Neural Networks - ANN) and the accuracy as performance metrics for evaluating the models. While the ANN model achieved an accuracy of 86.24% in training and 87.28% in the testing set, the model using C5.0 achieved 99.25% during training and 99.05% in testing.

Yang *et al.* (2020) [12] aimed to identify trends in student and teacher numbers in Taiwan to help administrators accurately allocate resources and make informed decisions for the future. The authors applied their proposed methodology to a case study involving student enrollment and teacher statistics from 1999 to 2018 from the Ministry of Education's database. They utilized a combination of the Whale Optimization Algorithm (WOA) and Support Vector Regression (SVR), referred to as the WOASVR algorithm. For public schools, the results showed a MAPE (Mean Absolute Percentage Error) of 2.09% and an RMSE (Root Mean Square Error) of 16171.42.

Abideen *et al.* (2023) [13] developed a model for analyzing student enrollment in secondary schools. Five years of enrollment data from 100 schools in Punjab (Pakistan) have been taken. Five hundred entries are present in the dataset, 120 female genders and 380 male. The convenient features were selected based on related work and the current scenario. Several classification and prediction algorithms were tested: Support Vector Machine (SVM) Linear, SVM Sigmoid, K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, and Decision Tree (DT - classification); Multiple Linear Regression, Random Forest, and DT (prediction). The algorithm that best performed was the Random Forest with an accuracy of 97%, R^2 (determination coefficient) equal to 0.971, and an RMSE of 3.2.

Lojić, Kevrić, and Jukić (2024) [14] used the classifiers Multi-layer Perceptron (MLP), Random Forest, and Random Tree to predict student enrollment in colleges. The dataset used in the study was from participants in the International Exhibition of Ideas, Innovations, and Creativity among Youth collected over five years (2015-2019), including historical data on 607 participants who are high school students from Bosnia and Herzegovina, as well as information about their current career paths. The average accuracy of the models was 46%.

The diversity of studies identified in the literature for the enrollment prediction problem is large, which highlights the specificity of each solution regarding the prediction objective, data context, and applied techniques. In addition, the related works identified do not discuss the interpretability of the predictions for the decision-makers in public policies. These aspects justify our proposal of optimized predictive models to forecast student enrollment in public schools in Brazil, considering the specific characteristics of Education and Economy in our country and the proposal of an innovative solution.

3 Methodology

Since our goal is to predict the number of students enrolled in the next school year in each public school in Brazil using the "MEC-PROSPECCAO" dataset, we will use the *current year + 1* as the target variable.

The "MEC-PROSPECCAO" dataset comprises microdata from the Brazilian government's school census, containing 218,598 schools with their respective educational stages ranging from grades k-12 school to technical and special education (youth and adults). This dataset was provided by the Ministry of Education of Brazil (MEC) and has 340 attributes of binary and categorical types.

To answer the RQ investigated in this work, we will explore the dataset about the following fundamental aspects: (i) the structure of the teaching stages per school (Section 3.1); (ii) the correlation between the other variables (attributes) and the target variable (Section 3.2); and (iii) the importance of the attributes for proposing prediction models (Section 3.3).

3.1 The Dataset

One of the challenges inherent to the "MEC-PROSPECCAO" dataset is the distribution of teaching stages by the school. In this case, each school has different teaching stages, in terms of number and types, ranging from elementary school to high school and technical school. The dataset has 37 types of teaching stages registered, with each school having a specific combination. Figure 1 shows an example of a school record composed of its 11 teaching stages, most of which are related to elementary school.

	CO_ENTIDADE	NO_ENTIDADE
22	11000465	EMEIEF ANTONIO AUGUSTO VASCONCELOS

	CO_ETAPA_ENSINO	NO_ETAPA_ENSINO	ETAPA_ATENDIMENTO
1	2	Educacao Infantil - Pre-Escola (4 E 5 Anos)	EDUCAÇÃO INFANTIL
3	14	Ensino Fundamental De 9 Anos - 1º Ano	ENSINO FUNDAMENTAL - ANOS INICIAIS
4	15	Ensino Fundamental De 9 Anos - 2º Ano	ENSINO FUNDAMENTAL - ANOS INICIAIS
5	16	Ensino Fundamental De 9 Anos - 3º Ano	ENSINO FUNDAMENTAL - ANOS INICIAIS
6	17	Ensino Fundamental De 9 Anos - 4º Ano	ENSINO FUNDAMENTAL - ANOS INICIAIS
7	18	Ensino Fundamental De 9 Anos - 5º Ano	ENSINO FUNDAMENTAL - ANOS INICIAIS
8	19	Ensino Fundamental De 9 Anos - 6º Ano	ENSINO FUNDAMENTAL - ANOS FINAIS
9	20	Ensino Fundamental De 9 Anos - 7º Ano	ENSINO FUNDAMENTAL - ANOS FINAIS
10	21	Ensino Fundamental De 9 Anos - 8º Ano	ENSINO FUNDAMENTAL - ANOS FINAIS
29	41	Ensino Fundamental de 9 Anos - 9º Ano	ENSINO FUNDAMENTAL - ANOS FINAIS

Fig. 1. Example of a school register with its respective teaching stages.

Figure 2 presents a cutout of the distribution of teaching stages by school in the dataset. The cells are CSV (comma-separated values) files that compose the dataset. A green cell represents that the school has the teaching stage, while a red cell indicates that the teaching stage does not exist. Each teaching stage is associated with microdata used to train machine learning algorithms and forecast the number of students expected in the following years for each teaching stage.

In this stage of the methodology, we understand how different schools organize their respective educational structures. For this purpose, we identify how many valid records

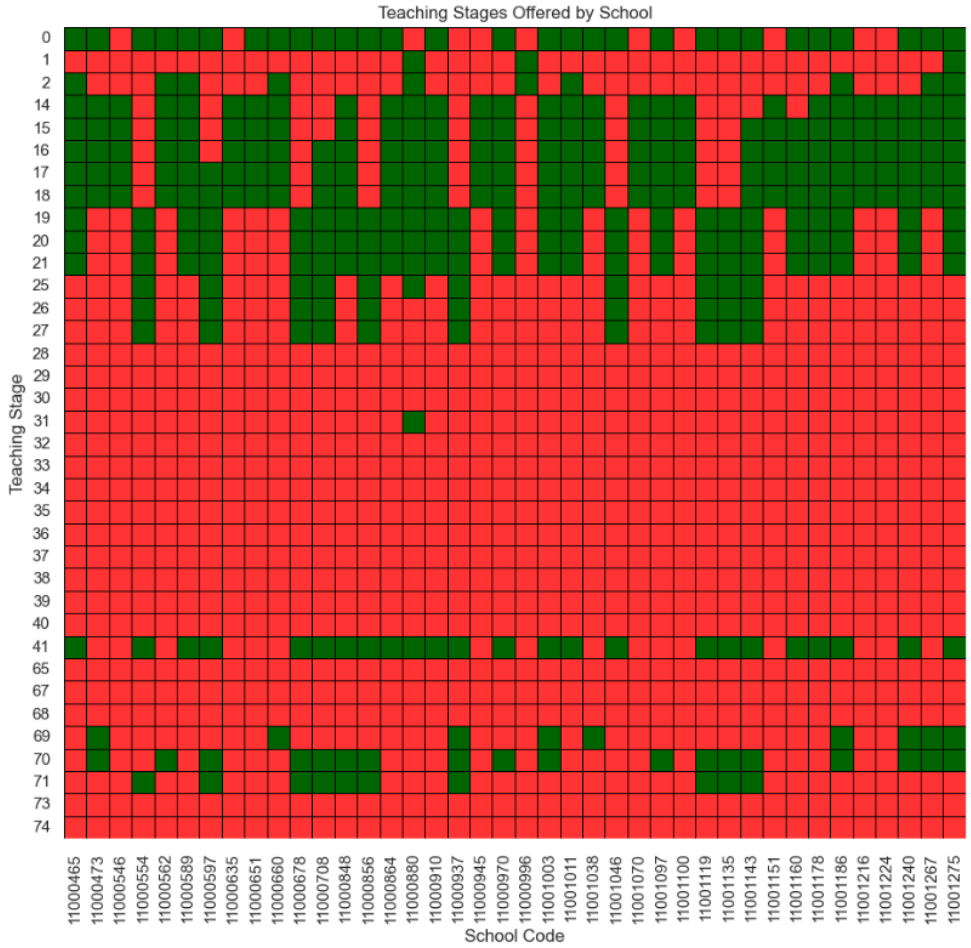


Fig. 2. Distribution of teaching stages per school in the "MEC-PROSPECCAO" dataset. The X-axis represents the school codes. The Y-axis represents the teaching stages.

existed within the dataset, a record being characterized by the combination of the variables school and teaching stage.

Another challenge in using the MEC-PROSPECCAO dataset to make enrollment predictions is the large number of CSV files, which total 1,514,844. Of these, 6.56% (99,374) are empty (part of green cells in Figure 2), making it impossible to calculate forecasts for the respective teaching stages, with only 93.44% of valid records remaining. This problem of missing data can hinder the Brazilian government's decision-making regarding the distribution of books and teaching materials to public schools. For example, several schools may not receive the correct number of books due to the lack of a forecast of the expected number of students for the next year.

As this dataset was exclusively organized to carry out experiments to select the best attributes for the school enrollment prediction problem, it does not have missing values to treat before the model training and testing stage.

3.2 Attributes Correlation

Attribute correlation analysis is a process used to validate whether all variables in the dataset are effective in inducing a good prediction. Figure 3 illustrates the step-by-step process of selecting attributes used in this work, which we will detail throughout this section.

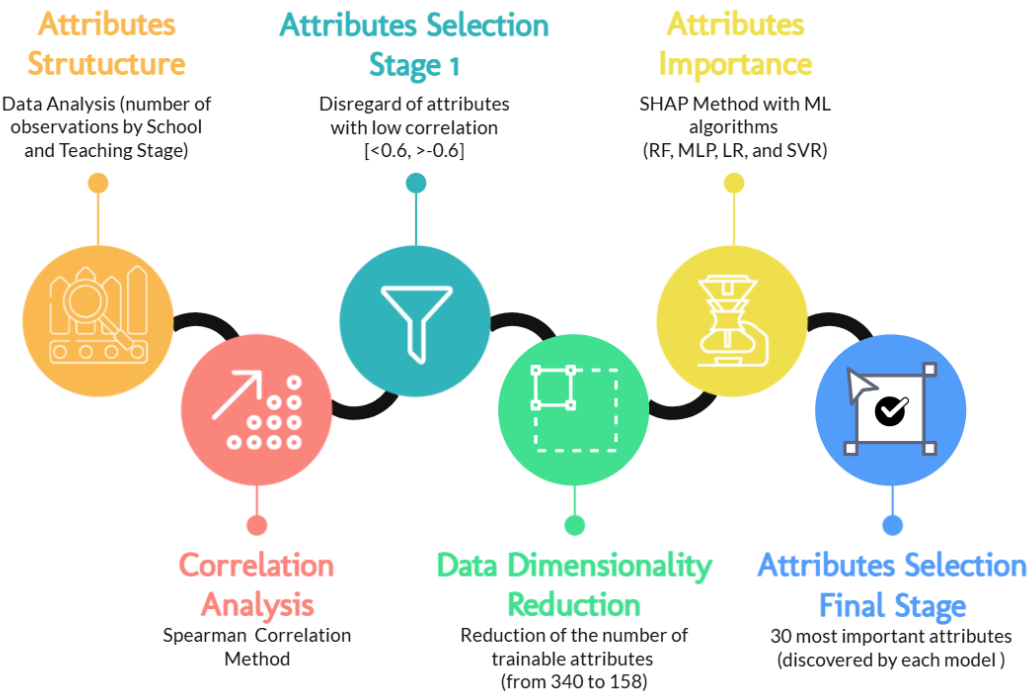


Fig. 3. Attributes Selection Process.

Attributes Structure Initially, it is necessary to understand how the data in each CSV file is structured. The number of observations in each CSV file for each teaching stage in each school is unique. For example, the school with code 11000465 has observations spanning 10 years (2013-2023, except 2022) for stage "2", while for stage "0" there are only four observations (2014-2017). This variation in observations between schools and stages of education makes data analysis even more complex. The number of observations (sample size) is important for analysis because the robustness of the correlation method depends on the dataset. Therefore, the more data there is, the more accurate the statement is about how each variable can impact the prediction of the value for the target variable.

In this methodology stage, we analyzed the correlation between the attributes present in the dataset using the Spearman correlation statistical method [15]. This analysis sought to understand how the attributes existing in the database related to the target variable since such attributes would be potential predictors for the proposed predictive models. The idea is to highlight the importance of the available attributes and eliminate irrelevant attributes, improving the efficiency of the models.

It's important to mention that the proportion of binary attributes varies from one CSV file to another, as shown in Figure 4. In the two examples presented, it's evident that the majority of the attributes are binary, while only a small portion are non-binary.

Feature Selection Protocol Our choice of the Spearman method to calculate the correlation between attributes is justified by the fact that it is a method indicated for data of a discrete, binary, or ordinal nature and that it does not depend on statistical assumptions about the distribution of the data, that is, it is applicable when the data does not follow a normal distribution.

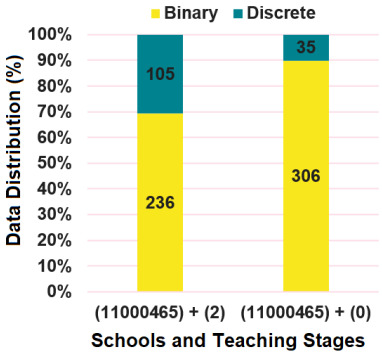


Fig. 4. Number of binary and non-binary attributes for two different teaching stages of school "11000465".

We calculate the correlation (ρ) between two attributes using the Spearman method according to Equation 1:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{1}$$

where d is the pairwise distances of the ranks of the variables (attributes) and n is the number of instances considered in the analysis.

Therefore, values close to 1 indicate that the correlation is positive, which means that when one variable increases, the other variable also increases, while values close to -1 indicate the opposite. The idea is to disregard trainable attributes that have low correlation or do not correlate with the target variable before training the proposed predictive models.

Since the number of trainable attributes was still considered high, we decided to apply a filter to extract only the attributes that correlated greater than a certain range (e.g., $[< 0.6, > -0.6]$), leaving only the most positively and negatively correlated attributes. In Section 4, we will present the experiments performed to define the threshold values for this range.

It is important to note that attributes with a calculated correlation result of *NaN* (Not A Number) were removed during the model's training stage. This occurrence was observed with binary attributes, as their values are constant, which is a behavior documented in the Python library "pandas.DataFrame.corr". For example, for teaching stage 2 from the school "11000465", 47.46% of the binary attributes had a NaN value for correlation.

Figure 5 exemplifies part of this procedure, with the reduction of 340 trainable attributes to 158, less than half of the total attributes of the dataset. For better visualization, we have highlighted positive correlations in green, while negative correlations were highlighted in red.

3.3 Feature Selection

In the third stage of the methodology, we investigated the importance of the attributes for the prediction models using the SHAP (SHapley Additive exPlanations) method [16, 17]. This method provides a detailed analysis of how each attribute contributes to the model predictions, highlighting which characteristics are most influential for the different machine learning algorithms. Despite being a costly method, as it tests all possible combinations of attributes, it is considered advantageous since it is possible to minimize the number of features for a predictive model without any loss in model performance.

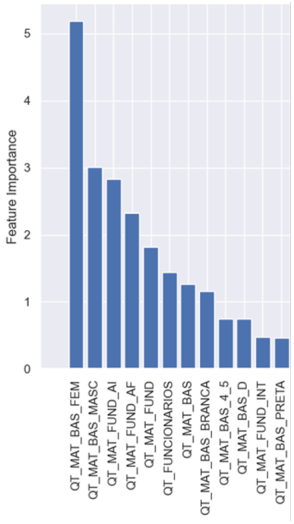
```
QT_TUR_FUND                    0.819286
IN_BANHEIRO_DENTRO_PREDIO      0.811503
IN_EQUIP_FOTO                  0.811503
IN_AGUA_FILTRADA               0.811503
IN_MATERIAL_ESP_NAO_UTILIZA    0.811503
                                   ...
IN_MATERIAL_PED_ARTISTICAS    -0.800152
IN_MATERIAL_PED_MUSICAL       -0.800152
IN_ORGAO_GREMIO_ESTUDANTIL    -0.800152
QT_PROF_SERVICOS_GERAIS       -0.811503
IN_SERIE_ANO                  -0.811503
Name: TARGET, Length: 159, dtype: float64
159
```

	Variable Name	Variable Description
436	QT_TUR_FUND	Número de Turmas de Ensino Fundamental
123	IN_BANHEIRO_DENTRO_PREDIO	Dependências físicas existentes e utilizadas n...
199	IN_EQUIP_FOTO	Equipamentos existentes na escola - Máquina fo...
89	IN_AGUA_FILTRADA	Água consumida pelos alunos
282	IN_MATERIAL_ESP_NAO_UTILIZA	Materiais didáticos específicos para atendimen...
...	...	...
272	IN_MATERIAL_PED_ARTISTICAS	Instrumentos e materiais socioculturais e/ou p...
270	IN_MATERIAL_PED_MUSICAL	Instrumentos e materiais socioculturais e/ou p...
303	IN_ORGAO_GREMIO_ESTUDANTIL	Órgãos colegiados em funcionamento na escola -...
230	QT_PROF_SERVICOS_GERAIS	Quantidade de profissionais que atuam na escol...
260	IN_SERIE_ANO	Forma de organização do ensino - Série/Ano (sé...

158 rows × 2 columns

Fig. 5. Example of variables with the highest correlations to the target variable.

To demonstrate the result of this feature selection process for a predictive model using the SHAP method, we chose the same school used in the previous examples, that is, the code school 11000465 and its respective teaching stage 2. For this demonstration, we chose four machine-learning algorithms from the Python Scikit-Learn library: i. Random Forest Regressor (RF); ii. MLP Regressor (MLP); iii. Linear Regression (LR), and; iv. Support Vector Regression (SVR). For the SHAP method, we selected only the 30 most important features discovered by each model. Figure 6 illustrates the most important features (on the left) identified by the Linear Regression algorithm according to the SHAP method and the respective description (on the right) of each attribute.



	Variable Name	Variable Description
386	QT_MAT_BAS_FEM	Número de Matrículas da Educação Básica - Femi...
338	QT_MAT_FUND_AI	Número de Matrículas do Ensino Fundamental - A...
344	QT_MAT_FUND_AF	Número de Matrículas do Ensino Fundamental - A...
337	QT_MAT_FUND	Número de Matrículas do Ensino Fundamental
226	QT_FUNCIONARIOS	Total de funcionários da escola (inclusive pro...
333	QT_MAT_BAS	Número de Matrículas da Educação Básica\n
389	QT_MAT_BAS_BRANCA	Número de Matrículas da Educação Básica - Cor/...
173	QT_SALAS_EXISTENTES	Número de salas de aula existentes na escola
174	QT_SALAS_UTILIZADAS_DENTRO	Número de salas de aula utilizadas na escola -...
177	QT_SALAS_UTILIZA_CLIMATIZADAS	Condições das salas de aula utilizadas na esco...
215	QT_COMPUTADOR	Quantidade de computadores na escola
391	QT_MAT_BAS_PARDA	Número de Matrículas da Educação Básica - Cor/...
398	QT_MAT_BAS_15_17	Número de Matrículas da Educação Básica - Entr...
436	QT_TUR_FUND	Número de Turmas de Ensino Fundamental

Fig. 6. The most important attributes of the Linear Regression algorithm according to SHAP.

In the example of Figure 6, it can be seen that the LR highlighted different features as the most important. This occurs due to the particularities of each model. In this case, the LR aims to find a linear relationship between the independent variables (340 attributes) and the target variable ('TARGET'). Therefore, those attributes that contribute linearly to the prediction will be considered the most important.

Therefore, we observed that the most important features vary significantly depending on the Machine Learning algorithm. To identify the intersection of the results of applying

the SHAP method to the chosen ML algorithms, we used the *pce* (percent common elements) function. In this way, it was possible to verify the existence of important features commonly identified by the algorithms.

```
def pce(init_list, fin_list):
    comm = [e for e in init_list
                if e in fin_list]
    perc = (len(comm)/len(init_list)) * 100
    return perc, comm
```

Figure 7 shows the similarity matrix that compares all the tested feature selection methods using the *pce* method. Since the order of the vectors has been changed, the diagonals of the similarity matrix are not identical.

	Spearman	RF	MLP	SVR	LR
Spearman	100.00	12.58	11.95	13.21	8.81
RF	66.67	100.00	46.67	33.33	46.67
MLP	63.33	46.67	100.00	50.00	80.00
SVR	70.00	33.33	50.00	100.00	50.00
LR	46.67	46.67	80.00	50.00	100.00

Fig. 7. Similarity of attributes between the selection methods investigated.

It can be seen in this example of the school 11000465 with teaching stage 2, the Spearman correlation method grouped, in most cases, the attributes considered most important for all the algorithms, as highlighted in green. Therefore, only five features were considered important predictors:

- $QT\_MAT\_BAS\_15\_17$ (number of Basic Education enrollments - 15–17 years of age);
- $QT\_MAT\_FUND\_AF$ (number of Elementary Education enrollments - Final Years);
- $QT\_MAT\_BAS\_PARDA$ (number of Basic Education enrollments - Color/Mixed);
- $QT\_MAT\_BAS\_BRANCA$ (number of Basic Education enrollments - Color/White);
- $QT\_MAT\_FUND\_AI$ (number of Elementary Education enrollments - Initial Years).

It is crucial to highlight that, given the unique nature of the records in the dataset, there is generally no overlap of significant attributes between the selection methods.

4 Experiments and Results

In this section, we present two experiments, the first to evaluate and select the most efficient feature selection method for the enrollment prediction problem and the second to assess the performance of machine learning algorithms as a solution using the best feature selection method identified in the first experiment.

For feature selection, we experimented with selection methods that focused on the best features for each machine learning model and the Spearman correlation method with different thresholds. To conceive the predictive models, we experimented with the following ML algorithms: (i) Random Forest (RF) Regressor; (ii) Multi-layer Perceptron (MLP) Regressor; (iii) Linear Regression (LR); (iv) Support Vector Regression (SVR). It is worth noting that the choice of these Machine Learning algorithms with different characteristics is justified by the need to understand the efficiency level of attribute selection methods in all cases.

Table 1 shows how we grouped the models so that was possible to isolate and measure only the impacts that the feature selection methods caused in the predictions carried out by the ML algorithms. To obtain better performance from the predictive models, we used the parameter evaluation technique for machine learning algorithms known as GridSearch [18], also available in the Scikit-learn library. This technique allows us to evaluate all possible combinations of parameters for each model and optimize the settings for each algorithm. Due to the high computational cost of generating the results, we performed this experiment on only three data records by combining a school with some of its respective teaching stages.

Table 1. Setup for the Experiments.

Namespace	Grouping
All	**Training using the 340 existing attributes (All):** (i) RandomForestRegressor; (ii) MLPRegressor; (iii) LinearRegression; (iv) SVR.
SP {0.5, 0.6, 0.65}	**Training using the attributes filtered by the Spearman (SP) method:** (i) RandomForestRegressor; (ii) MLPRegressor; (iii) LinearRegression; (iv) SVR.
M {10, 20, 30}	**Training using the 10, 20, or 30 most important attributes for the model (M):** (i) RandomForestRegressor; (ii) MLPRegressor; (iii) LinearRegression; (iv) SVR.

Table 2 presents the parameters evaluated for each algorithm and those that generated the best prediction results. Therefore, we applied the best values obtained for the evaluated parameters to the other experiments performed in this work. To evaluate the performance of the models, we used the cross-validation, as it is the most suitable approach for databases with few observations, such as those investigated in this document. Cross-validation provides a more robust estimate of the overall performance of the model, as it evaluates it across all observations in the dataset.

For each ML algorithm, runs were performed with 100 and 2500 records from the "MEC-PROSPECCAO" dataset. To evaluate the performance and attest to the efficiency of the models' results, in all cases, we used the following metrics:

– **MAE (Mean Absolute Error) [19]**: a noise-robust metric that calculates the absolute difference (predicted ($\hat{y}$) minus actual (y) value) in forecasts. The closer its value is to zero, the more accurate the estimates are. The MAE is calculated by Equation 2:

$$\mathrm{MAE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{N} \tag{2}$$

where N is the number of observations.

Table 2. GridSearch applied to machine learning models.

Model	GridSearch	Best Parameters
RF	n_estimators:[50, 100, 200]; max_depth:[None, 10, 20]; min_samples_split':[2, 5, 10]; min_samples_leaf:[1, 2, 4].	n_estimators:[100]; max_depth:[None]; min_samples_split:[5]; min_samples_leaf:[1].
MLP	hidden_layer_sizes:[(5,), (10,), (15,), (20,), (30,), (40,)]; alpha:[0.001, 0.01]; max_iter: [50,100, 200, 300, 400]; activation:['relu','tanh','logistic']; learning_rate:'constant','adaptive']; solver':['adam', 'lbfgs', 'sgd'].	hidden_layer_sizes:[(30,)]; alpha:[0.001]; max_iter: [50]; activation:['logistic']; learning_rate:['constant']; solver:['lbfgs'].
LR	C:[0.1, 1, 10, 100]; epsilon:[0.01, 0.1, 0.5, 1]; kernel: ['linear', 'poly', 'rbf','sigmoid']; degree:[3, 8]; coef0:[0.01,10,0.5]; gamma:['auto', 'scale'].	C:100; epsilon:1; kernel:'rbf' degree:3; coef0:0.01; gamma:'auto'.
SVR	fit_intercept: [True, False].	fit_intercept: [False].

– **MSE (Mean Squared Error) [19]:** calculates the mean of the squared errors, that is, the mean of the squared differences between the predicted ($\hat{y}$) and actual (y) values. It is a metric that is sensitive to outliers in which low values indicate that the model performs well in terms of forecast accuracy. The MSE is given by Equation 3:

$$\text{MSE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N} \tag{3}$$

where N is the number of observations.

– **RMSE (Root Mean Squared Error) [19]:** represents the square root of the MSE and has the advantage of being in the same unit as the response variable. This metric provides a measure of the magnitude of the error, allowing for a more direct interpretation compared to the MSE. Like the MSE, lower RMSE values indicate better model adjustments. The RMSE is given by Equation 4:

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \tag{4}$$

where i is the *ith* observed value, y_i is the actual value and $\hat{y}_i$ is corresponding predicted value for y_i, and N is the number of observations.

– **MAPE (Mean Absolute Percentage Error) [20]:** calculates the average percentage difference between the predicted values and the actual value. Despite its ability to interpret errors in percentage terms, it is a sensitive measure to outliers. MAPE values $< 10\%$ indicate high prediction accuracy, while values $> 50\%$ represent poor prediction. The MAPE is given by Equation 5:

$$\text{MAPE}(y, \hat{y}) = \frac{100\%}{N} \sum_{i=0}^{N-1} \frac{|y_i - \hat{y}_i|}{|y_i|} \tag{5}$$

where i, y_i, $\hat{y}_i$, and N are identically as defined for RMSE.

It is essential to note that the metric R^2 (Coefficient of Determination) [20] was not reported in the experimental results due to the high discrepancies in part of the dataset, which impact the model predictions, particularly the linear ones. In addition, we used the Runtime and Number of Features (NoF) selected to assess the quality of the feature selection methods:

- **Execution Time**: training time of each model in each database. We run the experiments on a notebook with a 13th Gen Intel(R) Core(TM) i7-1360P 2.20 GHz processor and 16.0 GB RAM;
- **Mean, Maximum, and Minimum number** of features generated by the selection method.

4.1 Experiment 1 with 100 Records

The main objective of this experiment is to evaluate the feature selection methods and partially answer the Research Question, specifically, regarding the potential use of school census microdata to predict the number of students that will enroll in each school.

In this experiment, we run each machine learning model 10 times on each record to obtain a significant sample of its initialization variations. We chose the records according to their order of presentation in the dataset, that is, the first 100 valid records.

Table 3 presents the medians for each metric defined for the experiment. We omitted the mean and standard deviation due to outliers, which make data analysis difficult, mainly for linear models, such as linear regression and SVR. When observing the results in Table 3, we can see that the Spearman attribute selection method for the 0.5 filter (SP 0.5) was the one that generated the best results.

Table 3. Medians for the evaluated metrics.

Metric	All	SP 0.5	SP 0.6	SP 0.65	M 10	M 20	M 30
MAE	19.79	**16.24**	16.25	16.58	19.22	18.83	19.22
MSE	576.0	**428.16**	430.46	434.89	564.97	553.77	564.11
RMSE	24.0	**20.69**	20.75	20.85	23.77	23.53	23.75
MAPE	32.39	22.42	**21.07**	21.85	29.78	29.78	30.73
Runtime	0.09	**0.04**	**0.04**	**0.04**	38.18	38.5	39.58

To statistically confirm these results, we used the protocol established by the Autorank library [21], which initially assesses the normality of the data and then defines the application of the appropriate statistical test. For all cases, the Friedman test and the Nemenyi post-test were recommended.

Figures 8 to 11 present the Friedman ranking ordered from left to right. The algorithms best positioned in the ranking are those that achieved the best performance in the predictions. We verified whether they were statistically equal using the critical distance (CD) of the Nemenyi post-test. If the ranking of two algorithms is positioned at a distance smaller than CD, they have similar statistical performance, being connected by a line.

As shown in Figures 8 to 11, using the Spearman correlation method for feature selection with thresholds of 0.6 and 0.65 yielded the best prediction performance for the MAE, MSE, RMSE, and MAPE metrics among all evaluated feature selection methods. In several instances, the thresholds of 0.6 and 0.65 exceeded the performance of the 0.5 thresholds, as confirmed by the Nemenyi CD tests. Conversely, the selection methods that

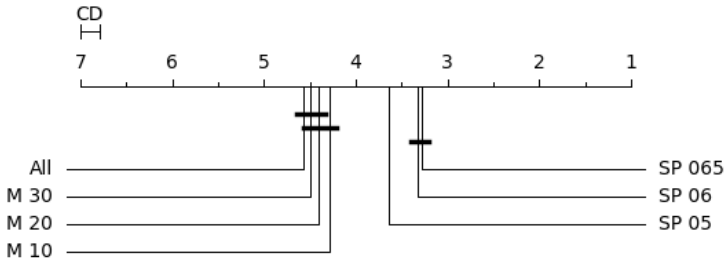


Fig. 8. Friedman test with Nemenyi post-test for MAE.

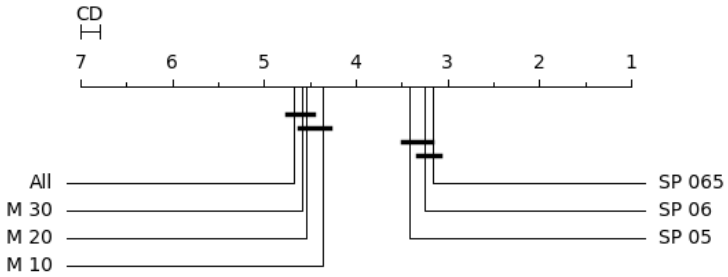


Fig. 9. Friedman test with Nemenyi post-test for MSE.

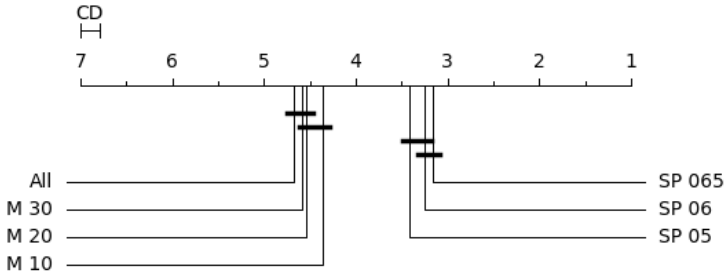


Fig. 10. Friedman test with Nemenyi post-test for RMSE.

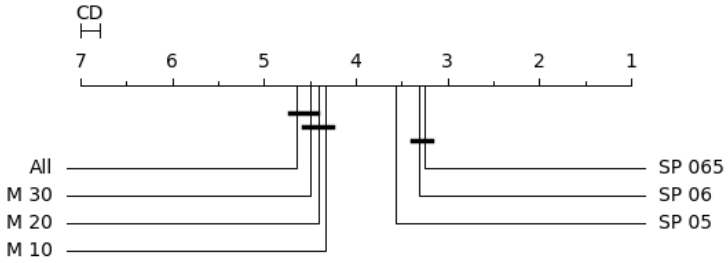


Fig. 11. Friedman test with Nemenyi post-test for MAPE.

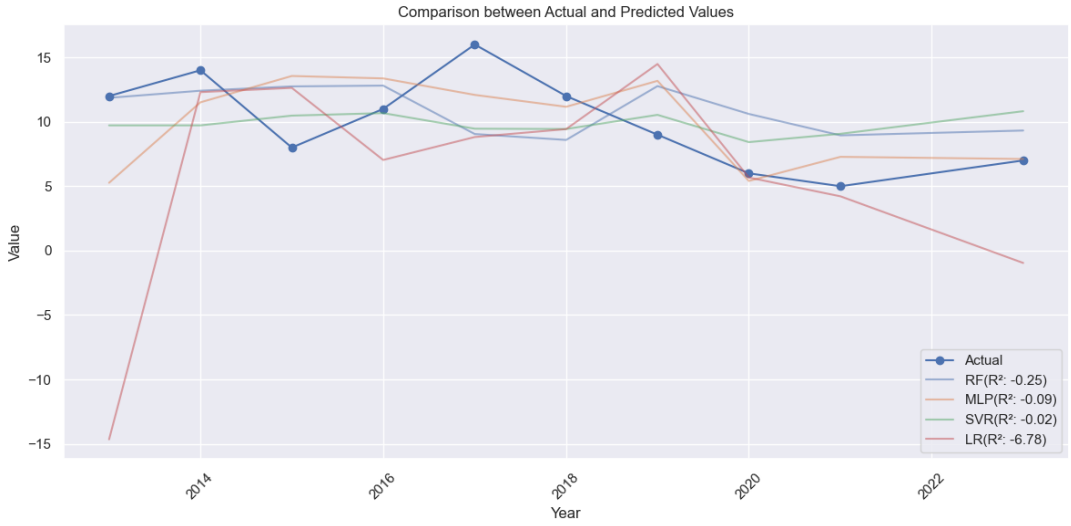


Fig. 12. Actual values *vs.* predictions for each ML model using the Spearman selection filter for 0.6.

focused on the best features for each machine learning model produced the least favorable results according to the tests.

We can see in Figure 12 that the majority of models can follow the trend of the time series, indicating that the use of the attributes filtered by the Spearman method generated, in fact, the best adjustments to the target variable. This is evidenced by comparing the results closest to zero with the data presented in Table 3.

One explanation for these results is related to the number of attributes that machine learning models need to evaluate during the training. For example, when trained with all attributes (ALL), models need to investigate, through trial and error, which of the 340 attributes are most important for prediction. According to the results, this attempt often results in training attributes that, in practice, harm the algorithms' prediction performance. On the other hand, when trained with attributes suggested by the Spearman correlation method, machine learning models can focus only on the data that influence the prediction of the target variable, which results in superior prediction performance.

Another positive factor regarding attribute selection can be seen in Table 4, which presents the average, maximum, and minimum number of attributes used in each experiment. As we can see, the higher the Spearman correlation threshold, the lower the number of attributes used for training.

Table 4. Average, maximum, and minimum number of attributes.

Experiment	Mean	Maximum	Minimum
All	340	340	340
SP 05	99.6	189	3
SP 06	72.0	185	6
SP 065	**53.5**	**184**	**2**
M 10	10.0	10	10
M 20	20.0	20	20
M 30	30.0	30	30

As a result of reducing the number of attributes, we had a statistical impact on the training time of the models, as shown in Figure 13.

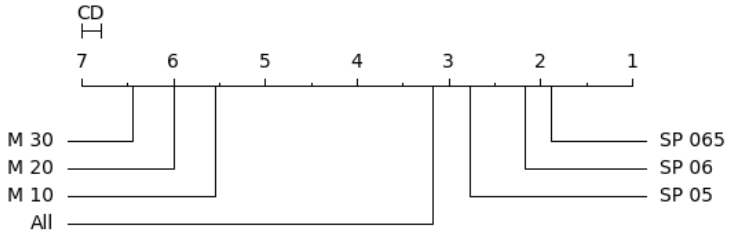


Fig. 13. Friedman test with Nemenyi post-test for Runtime.

It can be seen that the fewer attributes used in training the models, the faster this process becomes. It is worth noting that the experiment with the best attributes per model (M 10, M 20, and M 30) resulted in a longer execution time. This occurred because the SHAP method trained each evaluated model by applying all possible combinations for the 340 attributes existing in the database.

4.2 Experiment 2 with 2500 Records

For this experiment, we ignored the approaches that performed poorly in the experiment in Section 5.1, which were:

- attributes with Spearman correlation above 0.5 and 0.65; and
- the 10, 20, and 30 most important attributes per model according to the SHAP method, because they generated the worst results due to the high computational cost associated with them.

Therefore, in this experiment, we compared only the machine learning models using (i) training with 340 attributes and (ii) attributes with Spearman correlation above 0.6. Our purpose was to verify whether the best results from the previous experiment remain with the same distribution when such algorithms are trained with a larger sample of records. In this case, we selected 2,500 data records (a combination of schools and respective teaching stages) according to their order of presentation in the dataset, that is, the first valid records.

In this experiment, unlike the previous one, each machine-learning model was executed only once for each data record. This decision was made to allow for experimentation with a larger number of records. To ensure the reproducibility of the results, an execution seed was employed in Python, ensuring that the model would consistently generate the same outcome. We applied the seed value 42 to all machine learning models used in this study.

Table 5 presents the medians for the evaluated metrics. To ensure the veracity of the results, we followed the statistical experimentation protocol outlined by Herbold [21]. In this case, the protocol recommended using the Wilcoxon test to compare the results, as it is suitable for comparing two samples that do not follow a Normal distribution. An asterisk (*) next to the best result indicates statistically significant comparisons, demonstrating that the null hypothesis was rejected at a significance level of $p = 0.05$.

When analyzing the results in Table 5, it is clear that in all the metrics evaluated, the performance of the machine learning models' predictions improved when using only the attributes selected by the Spearman correlation method. Furthermore, it is worth noting that the execution time decreased by approximately 50% due to the smaller number of attributes used, as highlighted in Table 6.

Table 5. Results of the comparison between using all attributes and those filtered by the Spearman method with a threshold higher than 0.6.

Metric	All	SP 0.6
MAE	15.35	**12.11***
MSE	355.27	**224.52***
RMSE	18.85	**14.98***
MAPE	40.98	**33.65***
Runtime	0.06	**0.03***

Table 6. Average, maximum, and minimum number of attributes used for 2500 records.

Experiment	Mean	Maximum	Minimum
All	340	340	340
SP 06	**47.9**	**201**	**6**

Table 7 presents the performance of the machine learning algorithms that were evaluated, considering all attributes. To demonstrate the importance of the feature selection method on the performance of the predictive models, we added a column in the table with the result of the algorithm that performed best using the Spearman correlation method with its threshold hyperparameter set to 0.6. In this way, we highlight the improvement in Random Forest predictions in all metrics analyzed. For the Brazilian government, this result would represent greater savings in acquiring teaching materials for schools without compromising the availability of these materials for students. In the supplementary material, we present complementary results about the performance of the machine learning algorithms evaluated, such as the Friedman tests with the Nemenyi post-test for each metric analyzed.

Table 7. Medians for the evaluated metrics in each machine learning algorithm.

Metric	RF	MLP	SVR	LR	RF + SP 0.6
MAE	**13.34**	16.45	17.84	63.23	**12.5**
MSE	**296.64**	466.66	486.43	6675.73	**237.94**
RMSE	**17.22**	21.6	22.6	81.71	**15.43**
MAPE	**16.03**	19.69	22.81	93.82	**14.95**
Runtime	0.37	0.71	**0.02**	**0.02**	**0.27**

Given the results, this approach allows us to select the appropriate attributes for predicting enrollments at each teaching stage in each school, considering its particularities to create an optimized prediction model.

5 Discussion

Considering the Research Question defined in this work *"Can the data from the Brazilian government's school census be used to predict the number of students for the next year? If so, which variables can be used in this process?"*, we discuss how this question was answered based on the results achieved with our experiments.

Thus, with experiment 1, we analyzed two approaches for feature selection, the first using the Spearman correlation method with different thresholds and the second using the best features for each machine learning algorithm evaluated. As a result of this experiment,

we can state that the Spearman correlation method with thresholds of 0.6 and 0.65 was the approach with the best performance in dimensionality reduction and feature selection to be used as predictors for the problem of predicting the number of school enrollments. It is important to highlight that the attributes selected as predictors are easy to interpret, facilitating decision-making.

As part of proposing an optimized model for predicting school enrollments in experiment 2, the Spearman correlation method with a threshold of 0.6 was applied. This approach allowed us to demonstrate that the microdata from the Brazilian school census was sufficient to select the appropriate predictor attributes for enrollment prediction using the machine learning algorithms tested.

Therefore, the experiments performed in this work demonstrated the efficiency of the proposed experimental protocol, improving predictions regardless of the machine learning model. In this way, the Brazilian government can use our protocol, adopting it as a natural solution to the problem of predicting enrollment in public schools in Brazil. Thus, managers and the government are assisted in decision-making, especially in the acquisition and distribution of teaching materials, ensuring equity and the right of every student to quality education.

6 Conclusion

Aiming to choose the approach that provides the best features for proposing predictive models for forecasting the number of enrollments (target variable) in public schools in Brazil, we investigated two approaches for feature selection in the "MEC-PROSPECCAO" dataset. The first approach used the SHAP method, which evaluates the best features for each machine learning model. The second feature selection approach used the Spearman correlation statistical method. We used the machine learning algorithms Random Forest, MLP, Linear Regression, and SVR for the experiments.

Through our investigations, we showed that the SHAP method was effective in identifying relevant features for predicting the target variable, even improving the performance of the analyzed machine learning models compared to using all available features. However, due to its high computational cost, this approach proved to be unfeasible for large-scale real-world applications.

Considering the associated trade-offs, we obtain the best results using the Spearman correlation method, using thresholds of 0.6 and 0.65. This approach allowed us to significantly reduce the initial number of 340 attributes to an average of 50, improving the computational efficiency and predictive performance of the machine learning models. We validated these results through statistical analyses, demonstrating the superiority of the Spearman correlation method using the metrics MAE, MSE, RMSE, and MAPE.

With these results, it is possible to demonstrate the ability of the Spearman correlation method to reduce the dimensionality of the data and qualitatively indicate the most relevant attributes for each school in order to obtain good predictions. In practical terms, the predictions are calculated immediately after the input of the data. Thus, the choice of this method is promising for optimized predictive models, ensuring good performance and efficiency in data processing.

Therefore, the proposed method can be scaled to all public schools in Brazil and implemented by the Brazilian government so that education agencies such as the MEC and the FNDE (National Education Development Fund) can use the results of enrollment predictions to support decision-making in their public education policies, such as the PNLD.

In future work, we intend to carry out an experiment to evaluate the performance of the feature selection method using Spearman's correlation with all records of the "MEC-PROSPECCAO" dataset. The feasibility of this task is being analyzed, since it requires a significant computational effort and, consequently, a more robust infrastructure for execution.

Acknowledgements

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.

References

1. C. Lazar and M. Lazar, "Forecasting methods of the enrolled students' number," *Economic Insights – Trends and Challenges*, vol. IV, no. LXVII, pp. 41–51, 2015.
2. UNICEF, "The state of the global education crisis: A path to recovery," tech. rep., International Bank for Reconstruction and Development, Dec. 2021. Accessed: Nov. 19, 2024. [Online.] Available:.
3. UNESCO, "Latin american education systems in response to covid-19: Educational continuity and assessment: analysis from the evidence of the latin american laboratory for assessment of the quality of education (llece): programme document, july 2020," tech. rep., "Latin America Laboratory for Assessment of the Quality of Education (LLECE / UNESCO Santiago)", 2020. Accessed: Nov. 1, 2024. [Online.] Available: https://unesdoc.unesco.org/ark:/48223/pf0000374018_eng.
4. OCDE, *A Caminho da Era Digital no Brasil*. Paris: OECD Publishing, 2020. Accessed: Dec. 02, 2024. [Online.] Available: https://doi.org/10.1787/45a84b29-pt.
5. IPEA, "Agenda 2030: objetivos de desenvolvimento sustentável: avaliação do progresso das principais metas globais para o brasil: Ods 4: assegurar a educação inclusiva e equitativa e de qualidade, e promover oportunidades de aprendizagem ao longo da vida para todas e todos.," in *Cadernos ODS, 4*, Brasília: Instituto de Pesquisa Econômica e Aplicada, 2024. Accessed: Nov. 20, 2024. [Online.] Available: http://dx.doi.org/10.38116/ri2024ODS4.
6. INEP, "MEC e Inep divulgam resultados do Censo Escolar 2023," Feb. 2024. Accessed: Apr. 7, 2024. [Online.] Available: https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/mec-e-inep-divulgam-resultados-do-censo-escolar-2023.
7. UMASBoston, "Factors and techniques for projecting enrollment," tech. rep., University of Massachusetts Boston, Nov. 2017. Accessed: Nov. 19, 2024. [Online.] Available: https://www.umb.edu/media/umassboston/content-assets/documents/Factors_and_Techniques_Affecting_Enrollment.pdf.
8. S. Singh, "A robust method of forecasting based on fuzzy time series," *Applied Mathematics and Computation*, vol. 188, no. 1, pp. 472–484, 2007.
9. C. J. Stanley, "A data mining study of the matriculation of covenant college applicants," in *Proceedings of the 46th Annual ACM Southeast Conference*, ACMSE '08, (New York, NY, USA), p. 209–214, Association for Computing Machinery, 2008.
10. L. Wang, W.-W. Zhuang, and Y.-F. Liu, "An empirical study on the prediction model of postgraduate education in hebei province," in *2010 International Conference on Machine Learning and Cybernetics*, vol. 3, pp. 1327–1331, 2010.
11. M. D. Borah, R. Jindal, D. Gupta, and G. C. Deka, "Application of knowledge based decision technique to predict student enrollment decision," in *2011 International Conference on Recent Trends in Information Systems*, pp. 180–184, 2011.
12. S. Yang, H.-C. Chen, W.-C. Chen, and C.-H. Yang, "Student enrollment and teacher statistics forecasting based on time-series analysis," *Computational Intelligence and Neuroscience*, vol. 2020, no. 1, p. 1246920, 2020.
13. Z. u. Abideen, T. Mazhar, A. Razzaq, I. Haq, I. Ullah, H. Alasmary, and H. G. Mohamed, "Analysis of enrollment criteria in secondary schools using machine learning and data mining approach," *Electronics*, vol. 12, no. 3, 2023.
14. A. Lojić, J. Kevrić, and S. Jukić, "Predictive analysis of student enrollment on a faculty base on innovation research," in *2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1–5, 2024.

15. W. W. Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences*. New Jersey: John Wiley & Sons, Inc., 9 ed., 2009.
16. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 4768–4777, Curran Associates Inc., 2017.
17. S. Lundberg, "Shap documentation v0.46.0," 2024. Accessed: Oct. 10, 2024. [Online.] Available: `https://https://shap.readthedocs.io`.
18. J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, p. 281–305, Feb. 2012.
19. M. Shcherbakov, A. Brebels, N. Shcherbakova, A. Tyukov, T. Janovsky, and V. Kamaev, "A survey of forecast error measures," *World Applied Sciences Journal*, vol. 24, pp. 171–176, 01 2013.
20. S. Hiregoudar, "Ways to evaluate regression models," Aug. 2020. Accessed 23 Feb. 2023.
21. S. Herbold, "Autorank: A python package for automated ranking of classifiers," *Journal of Open Source Software*, vol. 5, no. 48, p. 2173, 2020.

Authors

Lenardo Silva received his PhD in Computer Science in 2015. He earned his Master's degree in Computer Science in 2011. He did his Bachelor's degree in Computer Science in 2009. Currently, he is a Adjunct Professor from the Computing Department at the Federal University of the Semi-Arid Region (UFERSA) and Reseacher at NEES-UFAL. His research interests include Computational Modelling, Data Science, and Software Development.

Gustavo Oliveira received his Master's and PhD degrees in Computer Science. He received his Bachelor's degree in Computer Science. He is currently an adjunct professor in the Undergraduate Information Systems program. His research interests include Software Development and Artificial Intelligence, with a focus on Web and Mobile Programming, Data Science, and Machine Learning.

Luciano Cabral received his PhD in Electrical Engineering from UFPE, with a focus on Communications. He holds a Master's degree in Computer Science from UFPE, specializing in Artificial Intelligence, and a Bachelor's degree in Information Systems. He completed two postdoctoral fellowships in Artificial Intelligence. Currently, he is an Associate Professor at IFPE and a Data Scientist at NEES-UFAL. His research interests include Applied Artificial Intelligence in Education, Health, and Security.

Luam dos Santos holds a Bachelor's degree in Information Systems from the Federal University of Alagoas, postgraduate specializations in Data Science and Public Management, and a Master's degree in Computational Knowledge Modeling from the Federal University of Alagoas (2016). Currently an Information Technology Analyst at the Federal University of the São Francisco Valley. Has experience in the field of Computer Science, with an emphasis on Information Systems, Software Engineering, Data Engineering, Machine Learning Engineering, and Data Science.

Rodrigo Silva received his M.Sc. degree in Informatics from the Graduate Program in Informatics (PPGI) at the Institute of Computing, Federal University of Alagoas (IC/-UFAL). He earned his B.Sc. degree in Computer Science from the Federal University of

Alagoas (UFAL). He is currently a Ph.D. student in Computer Science at the Department of Computer Science, Federal University of Minas Gerais (UFMG). His research interests include Smart Cities, Artificial Intelligence applied to Medicine, and Gamification.

Thyago de Oliveira received a Master's Degree in Computer Science from the Federal University of Alagoas (UFAL). He did a Bachelor's degree in Computer Science. Currently, he is pursuing his PhD in Computer Science and Computational Mathematics from the University of São Paulo (USP). His research interests include Computers in Education, Artificial Intelligence, and Collective Intelligence applied to educational technologies.

Breno da Costa received his MSc. in Computer Science in 2009. He did his Bachelor degree in Information Systems (2006) and an MBA in Project Managament (2015). Currently, he is pursuing his PhD in Industrial Engineering, with emphasis on Educational Technology, from the Federal University of Bahia. His research interests include Web3, Educational Technology, Learning Analytics, Active Learning and Interactive Textbooks.

Dalgoberto Pinho Júnior holds a PhD in Technologies and Public Policies, a Master's degree in Computational Modeling, and a degree in Information Systems. He is a researcher in areas related to Technology in Educational Public Policies, Project Management, Educational Technologies, and usability. He is a Professor at the Federal University of Alagoas - Arapiraca Campus - Penedo Unit and works on spin-off projects with the MEC and FNDE. His main interests are in agile project management, information architecture, and project management.

Nicholas da Cruz holds a degree in Administration from the Federal University of Alagoas (2003), a Master's degree in Administration from the Federal Rural University of Pernambuco (2007), and a PhD in Production Engineering from the Federal University of Santa Catarina (2018). He works mainly on the following topics: entrepreneurship, opinion and market research, educational technologies, public policies, and environmental impacts.

Rafael Silva studied Information Systems at the Faculty of Alagoas (2006), a Master's degree in Computer Science at the Federal University of Pernambuco (2008), and a PhD in Electronic and Computer Engineering at the Technological Institute of Aeronautics (2014). He holds a Postdoctorate in Electrical and Computer Engineering from Mackenzie Presbyterian University (2018) and in Computer Science from the ICMC of the University of São Paulo (2019). He is an Adjunct Professor at the Institute of Computing at the Federal University of Alagoas and a Specialist in Intelligent Systems Engineering, with experience in the following areas: Location Systems, Communication Systems, IoT/IoE/IoA Systems, System of Systems, Educational Systems 4.0, and Intelligent Systems in Healthcare.

Bruno Pimentel studied for a Bachelor's Degree in Computer Science at the Computer Science Center of UFPE (2010). He has a Master's Degree in Computer Science from the Computer Science Center of UFPE (2013), with Computational Intelligence as his research area. He has a PhD in Computer Science from UFPE (2017) in the area of Computational Intelligence, with Cluster Analysis as his research topic. He developed research as a postdoctoral fellow at the Institute of Mathematical and Computer Sciences (ICMC-USP), investigating algorithm recommendations through Meta-learning. He joined the Federal University of Alagoas (UFAL) in 2019 as a professor.