# An Experimental Analysis Using Educational Data Mining on INEP Data to Predict the School Profile Regarding Computing Education in Brazil

1st Rômulo Valle
*Programa de Pós Graduação em Informática (PPGI)*
*Universidade Tecnológica Federal do Paraná (UTFPR)*
Cornélio Procópio, Brazil
romulovalle19@gmail.com

***Abstract*—This experimental analysis utilizes Educational Data Mining (EDM) techniques to identify the school profile for schools that have computer science education classes in Brazil, using data provided by the National Institute for Educational Studies and Research (INEP). The study examines key indicators such as infraestructure, regionalization, and also school enrollments and apply Machine Learning methods to develop a comprehensive understanding of the current state of computing classes in Brazil's schools and these profile schools. The results of the analysis provide a comparisson into two methods tested Decision Tree and Random Forest to find if the school have or not the computing class. The findings have significant implications for educators and researchers who are interested in promoting the growth and development of Computing education in Brazil and beyond.**

***Index Terms*—Educational Data Mining, Computing Education, Artificial Inteligence, Instituto Nacional de Estudos e Pesquisas Educacionais**

## I. Introduction

In today's digital age, computer science education has become increasingly important for individuals and societies alike. [5] It is a crucial field that provides learners with the necessary skills to effectively participate in the rapidly changing and technology-driven workforce. The Brazilian educational system has recognized the importance of computer science education and has implemented it in various ways, from primary to tertiary levels. [2] The National Institute of Educational Studies and Research Anísio Teixeira (INEP) is the primary government agency responsible for the development and evaluation of educational policies in Brazil [7].

This article presents an experimental analysis using educational data mining on INEP data to identify the school profile regarding computer science education in Brazil. Educational data mining is a subfield of data mining that deals with the application of data mining techniques to educational data, with the goal of extracting meaningful insights to inform educational policy and practice. [10] The use of educational data mining has been shown to be effective in identifying patterns and trends in student performance, identifying factors that influence student success, and developing personalized learning systems [1].

The analysis presented in this article focuses on data from the INEP School Census, which collects information on schools that have computing classes, these students, and educational in Brazil [7]. The study aims to identify the current state of computing education in Brazilian schools, including factors that influence its delivery and effectiveness. The results of this study can inform policymakers, educators, and researchers on how to improve computer science education in Brazil.

## II. Related Work

Several studies have utilized educational data mining techniques to analyze data from INEP databases, aiming to improve the Brazilian educational system. The first study called Data Mining on INEP Databases: An initial Analysis aiming to improve Brazillian Educational System, this article discusses the use of Educational Data Mining (EDM) to extract useful information from the large datasets provided by the Anísio Teixeira National Institute for Education Research and Studies (INEP) [12]. The article focuses on the Prova Brasil mechanism, which assesses elementary education through tests and questionnaires to students, teachers, and school principals. The Knowledge Discovery in Databases (KDD) process is applied to identify factors that influence Mathematics learning and their positive and negative relationships. The article presents the KDD steps used and analyzes discovered patterns [12].

The second work called Educational Data Mining: A Study on Socioeconomic Indicators in Education in INEP Database is a study that utilizes educational data mining to identify factors that influence student performance and learning. [13]The study uses the database provided by INEP to analyze the relationship between socioeconomic variables and grades obtained by students in the ENEM 2016 exam. The study applies the PCA technique and generates Bayesian networks to analyze performance. The results indicate that income, parental schooling, and school type are strong influencing factors [13].

Another great study in EDM field in Brazil is the Educational data mining: a study on mathematics proficiency in Ceará [14]. The article discusses a study that utilizes the

Knowledge Discovery in Databases (KDD) process and Data Mining to integrate databases from different sources: the Educational Indicators (EI) from INEP and data on Mathematics Proficiency in Ceará's municipalities. The study develops a Knowledge Model based on data mining to classify and predict the performance of Ceará municipalities in the SPAECE/2019 Mathematics Assessment using Educational Indicators. The study constructs a Decision Tree model with an accuracy of 85.86% and identifies the relevant EI for the investigated problem [14].

## III. Problem Definition

The Brazillian Law 5.692, of August 11, 1971, which establishes the guidelines and bases of national education, has as its general objective to provide students with the necessary training for the development of their potential as elements of self-realization, qualification for work, and preparation for the conscious exercise of citizenship [3].

An important milestone for Brazilian Basic Education is the Opinion CNE/CEB No. 2/2022 which establishes standards for the inclusion of Computing in the school curriculum. Published on January 21, 2022, it complements the National Common Curricular Base (BNCC) and aims to provide guidance for the implementation of computing at all levels of education [4]. Based on last BNCC Complement, computing enables the exploration and experience of playful interactions with peers, which relate to several fields of experience in early childhood education [4]. This point meets the necessary training for the development, qualification and also the preparation students for digital age.

Understanding the profile of schools that have computing education comparing which one do not have computing education in Brazil is essential for the expansion of this type of learning. This is necessary to ensure that schools comply with current legislation and to promote educational equity throughout the country and using Educational Data Mining (EDM) is one of the measures adopted to combat deficits in education, through the analysis of data collected regularly in different cycles of the educational context, both in-person and online [1]. In addition, an in-depth analysis of the school profile can help identify gaps in computer science education and develop effective strategies to fill these gaps.

## IV. Educational Data Mining

Data Mining is the process of discovering hidden patterns and knowledge from large datasets through various methods such as statistics, machine learning, and artificial intelligence. [6] It involves the extraction of previously unknown, valuable information from complex data, which can be used to make informed decisions and gain insights into a wide range of applications, from business and finance to healthcare and education. Data Mining can uncover patterns and relationships in data that might not be immediately apparent to humans, and can help to reveal underlying trends and factors that affect outcomes [8].

One application of Data Mining is in Educational issues. The EDM field focuses on the development, research, and application of computer-based techniques to identify patterns in vast collections of educational data. This type of data would be otherwise difficult or impossible to analyze due to its massive volume. [10] Educational data includes not only individual student interactions with the educational system, such as navigation behavior and quiz responses, but also collaboration data, administrative data like school district and teacher information, and demographic data like age, gender, and school grades. Finally EDM, which stands for Educational Data Mining, is classified as both a learning science and a field of data mining. Below are some of the key uses of EDM [11]:

- Analysis and Visualization of Data: To analyze students' course activities and usage information, providing useful insights for decision making. Statistics and visualization are the two main techniques used for this task. Statistical analysis can provide information such as where students enter and exit, popular pages, and usage summaries. Visualization uses graphics to help people understand and analyze data, with studies focused on visualizing educational data such as patterns of user behavior and student tracking data. These techniques can help educators gain knowledge on student progress, performance, and engagement.
- Predicting Student Performance: To predict unknown variables that describe the student, such as their performance, knowledge, score, or marks. The prediction can be numerical or categorical. Regression analysis finds the relationship between dependent and independent variables, while classification groups items based on their characteristics or labeled items. Predicting student performance is a popular application of data mining in education, using techniques such as neural networks, Bayesian networks, rule-based systems, and regression analysis. Different regression techniques are applied to predict student marks, including linear regression, stepwise linear regression, and multiple linear regression. These techniques help educators predict student success and final grades based on logged data and other extracted features.
- Grouping Students: To create groups of students based on their customized features and personal characteristics. These groups can be utilized by instructors/developers to build a personalized learning system that promotes effective group learning. Classification and clustering techniques are used in this task. Clustering algorithms, such as hierarchical agglomerative clustering, K-means, and model-based clustering, are used to group students based on their similar learning characteristics. Hierarchical clustering algorithms are used in intelligent e-learning systems to group students according to their individual learning style preferences. These techniques help to create effective personalized learning environments for students.
- Enrollment Management: A term used in higher edu-

cation to describe the strategies and tactics designed to shape the enrollment of an institution and meet established goals. These strategies include marketing, admission policies, retention programs, and financial aid awarding, which are informed by the collection, analysis, and use of data to project successful outcomes. Competitive efforts to recruit students are a common focus of enrollment managers. The numbers of universities and colleges with offices of "enrollment management" have increased in recent years. These offices provide direction and coordination of efforts of multiple offices such as admissions, financial aid, registration, and other student services.

## V. Experimental analysis methodology considering Educational Data Mining:

This text outlines the steps taken in an experimental analysis that utilized both data mining and visualization techniques. The goal of this analysis was to effectively use INEP datasets that contain information about Brazilian schools and classes in 2020. The school dataset contained 700 megabytes of data, 224229 records, and 388 attributes, while the class dataset contained 1.5 gigabytes of data, 2353351 records, and 76 attributes. One of the key differences in this analysis compared to other literature works was the first stage, which involved refining the dataset through the use of data visualization techniques. By using visualization techniques such as heatmaps and boxplots, the analysis could identify highly correlated attributes and exclude or transform them without losing important data. The visualization techniques used were particularly useful in reducing the computational effort needed for this analysis.

In the second stage, it utilized the results obtained in the previous phase, which involved evaluating, comprehending, normalizing, and reconstructing a database that meets the desired needs through preprocessing techniques that help eliminate null and out-of-scope records and recover missing information whenever possible without changing the original data.

During the third stage of the experiment, the data was transformed as a preliminary step for applying machine learning models. The OneHotEncoder function from sklearn library was used to transform qualitative variables, such as regions and school types, into true or false values that could be applied to the dataset. Machine learning algorithms were then selected and applied to create predictive models, with the most commonly used algorithms being DecisionTree, RandomForest, and LogisticRegression. To evaluate the quality of the predictors, metrics such as Accuracy, Precision, Recall, and F1-Score were utilized. These metrics measure the performance of the models in predicting the presence of computer classes in schools, allowing for a comprehensive analysis of the predictive quality.

Accuracy, Precision, Recall, and F1-Score are evaluation metrics commonly used in machine learning to assess the performance of predictive models. Accuracy is a metric that measures the proportion of correct predictions made by the model in relation to the total number of predictions. It is a simple and intuitive metric, but it may not be the best choice when the dataset is imbalanced or when the cost of false positives and false negatives is different. Precision measures the proportion of true positive predictions made by the model in relation to the total number of positive predictions. It is a metric that assesses the model's ability to correctly identify positive cases, but it may generate a large number of false negatives. Recall, also known as sensitivity, measures the proportion of true positive predictions made by the model in relation to the total number of actual positive cases in the dataset. It is a metric that assesses the model's ability to correctly identify all positive cases, but it may generate a large number of false positives. Finally, F1-Score is a metric that combines Precision and Recall, providing a balance between them. It is the harmonic mean of Precision and Recall, ranging from 0 to 1, with higher values indicating better performance. F1-Score is a useful metric to evaluate the overall performance of a predictive model, especially when the dataset is imbalanced.

## VI. Results

Statistical and visual analyses were performed on the data used in the study to reduce data dimensionality, identify trends, and evaluate data density, aiming to aggregate attributes with high significance to direct relevant information to this study and facilitate data analysis. This study had been analyse the effects of infraestructure, students enrollments, regions and schools position, using a Machine Learning methods to fit and find the schools profile that have computing classes.

First, it is possible to see a couple of points about this carachteristics above from the Census data. Figure 1 illustrates a data visualization constructed from the Brazilian regionalization with school profiles, showing the volume of classes per school type throughout Brazil. This heatmap visualization shows a higher volume of classes in municipal schools, with emphasis on the Northeast, Southeast, and South regions, respectively. Federal model functional schools are not as present in this visualization.

Figure 2 goes in line with the context presented in the previous figure. Still using the same heatmap visualization of class volume in Brazil, a considerable data divergence is found for classes that have computer education. In this visualization, it is possible to note that classes that have computer education are mostly in the Southeast region, first and foremost in municipal and private schools. Therefore, with this initial observation, we can highlight the importance of the region for the presence or absence of computer education in Brazilian classes.

The presented stacked bar chart on Figure 3 shows the percentage of schools with and without computer science education in different regions of Brazil. It is clear that the South and Southeast regions have a higher percentage of schools with computer science education, while the Northeast region has a higher concentration of schools without this type of education.

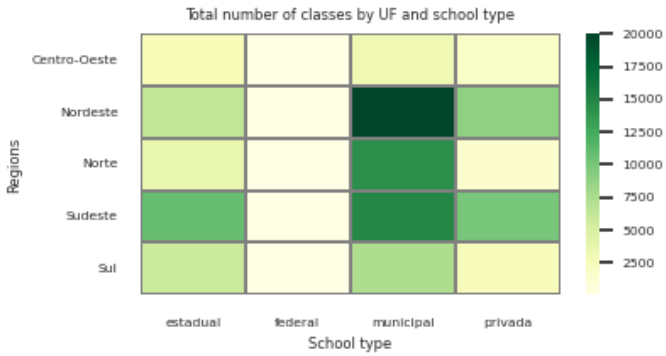


Fig. 1. Total number of classes without Computing by Region and school type.

Source: Elaboreted by the author.

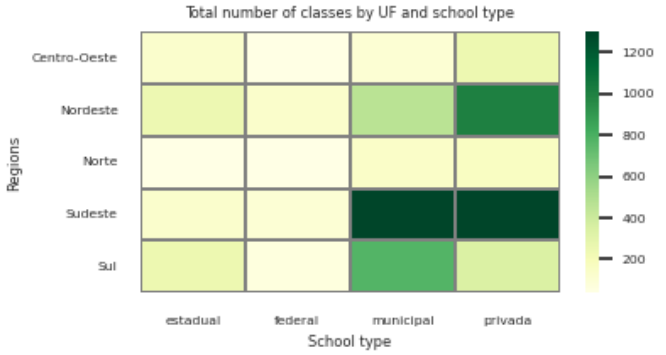


Fig. 2. Total number of Computing classes by Region and school type.

Source: Elaboreted by the author.

This result highlights the disparity between regions regarding access to technology and technological education, which can directly impact socioeconomic development in these regions. Therefore, public policies promoting equity in the distribution of educational resources and opportunities throughout the country are necessary.

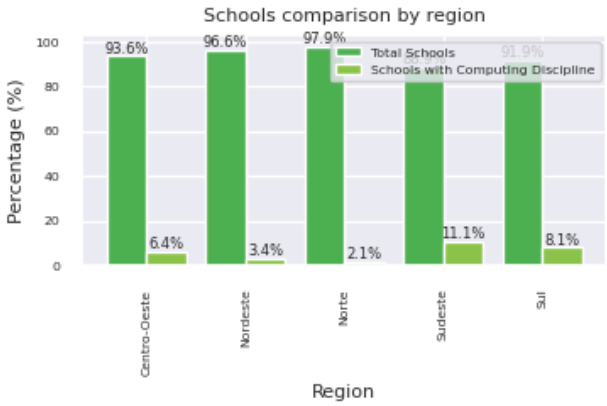


Fig. 3. Comparison of schools with and without computing classes by region.

Source: Elaboreted by the author.

The Figure 4 shows the percentage enrollments of whole brasillian schools in 2020. It is clear that there is a standart in the three sections of education, only the Nordeste region have more then 10% in Basic Education for kids.

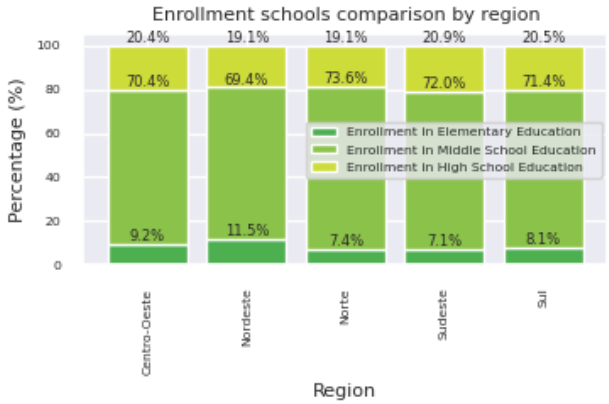


Fig. 4. Total number of Enrollments by Region and school level.

Source: Elaboreted by the author.

From the presented data, a combination of the various attributes shown in the Images was made, along with the application of the OneHotEncoder() library to transform the region attribute, which is a determining factor in the presence of classes. The next step was to apply the DecisionTree and RandomForest predictive models from the scikit-learn library and compare the two methods. 80% of the data set was randomly selected for training and 20% for testing. The analyzed parameters were: school id, computing classes, region, bathroom, non-existent water, non-existent energy, non-existent sewer,library, high school stage, elementary school stage, infant school stage, eja school stage, garbage service, sports court, quantity female enrollment, quantity male enrollment, quantity high school enrollment, quantity elementary enrollment, quantity high school enrollment.

The metrics below show the performance evaluation of a binary classification model on a test set. The model DecisionTree classes 0 or 1 for each sample in the test set, and the metrics were calculated by comparing the predictions with the true classes. The accuracy is a general metric that measures the percentage of correct predictions out of the total predictions. In this case, the accuracy is 0.90, which means that the model correctly predicted the class in around 90% of the test set samples.

The precision measures the proportion of correct positive predictions out of the total positive predictions, including true and false positives. The precision for class 0 is high, at 0.95, indicating that the model correctly predicted the majority of positive predictions for this class. However, the precision for class 1 is low, at 0.26, indicating that the model had difficulty identifying the true positive samples for this class. The recall measures the proportion of true positive samples correctly predicted out of the total positive samples. The recall for class 0 is high, at 0.94, indicating that the model found the majority

of true positive samples for this class. However, the recall for class 1 is low, at 0.29, indicating that the model missed many true positive samples for this class. Finally, the F1-score provides a harmonic mean of precision and recall, giving a general measure of model performance on both metrics. The F1-score for class 0 is high, at 0.95, while for class 1, it is low, at 0.27.

TABLE I
DECISIONTREECLASSIFIER MODEL EVALUATION METRICS

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.95 | 0.94 | 0.95 | 24726 |
| 1 | 0.26 | 0.29 | 0.27 | 1713 |

The RandomForestClassifier classification model had an accuracy of 0.94 on the test set, indicating that it correctly predicted the class in around 94% of the samples. The precision for class 0 was high, at 0.94, but for class 1, it is better than the first one: 0.60. The recall for class 0 was also really high, at 1, while for class 1, it was low, at 0.11. The F1-score for class 0 was high, at 0.97, while for class 1, it was low, at 0.18. Overall, the model performed well for class 0 and with a good perception of precision for 1.

TABLE II
RANDOMFORESTCLASSIFIER MODEL EVALUATION METRICS

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.94 | 1.00 | 0.97 | 24726 |
| 1 | 0.60 | 0.11 | 0.18 | 1713 |

## VII. CONCLUSION

Considering the significant advances in the Brazilian National Common Curricular Base (BNCC) in 2021, which included the subject of informatics and computing in the school curriculum, this study proposes an experimental analysis using Experimental Data Modeling (EDM) based on data visualization techniques, attribute selection, and predictive model construction using machine learning algorithms. The goal is to identify the most relevant factors that determine the presence or absence of this subject in current schools. Attribute selection was guided by data visualization techniques, which allowed for accelerated understanding and reduced data complexity.

It is crucial to emphasize that the development of new data visualization techniques and the improvement of models were crucial to the prediction process. There were significant improvements in the performance of attribute selection and machine learning algorithms, resulting in reduced computational costs. Additionally, it was found that the transformation of attributes from different tables allowed for an expanded view of the influencing factors since there is still no governmental database with information on the offering of subjects.

It is important to highlight that the study is not solely focused on the use of attribute selection techniques but also on the understanding of the factors that influence the teaching of informatics in 2020. The use of bivariate visualizations and machine learning algorithms allowed for the identification of a significant relationship between regional, infrastructure, school stages, race, and the presence of the subject. New studies could be conducted by reviewing the selected parameters and also the suggested visualizations. This would allow for a more comprehensive analysis of the factors that influence about schools and could lead to the discovery of new insights about computing classes. Additionally, future research could also explore the impact of different teaching methodologies and resources on students' performance and interest in the subject.

One limitation of the study was the use of only 2020 school census data due to the volume and computational impact. It is important to note that experimental analysis is an extensive process, as indicated by authors in the general data mining area. Additionally, comparative analyses with different discretizations in the target class result can be performed. Several discretizations were made, and the one that obtained the best results was used. Finally, cross-referencing the data with information from the School Census by Teachers and the School Census by Enrollment can be a next step to investigate whether the teaching staff and enrollment volume are also influential factors in this analysis.

## REFERENCES


[1] Banni, Maicon Ribeiro, Marcos Vinicius dos P. Oliveira, and Flavia Cristina Bernardini. "Uma análise experimental usando mineração de dados educacionais sobre os dados do enem para identificação de causas do desempenho dos estudantes." Anais do II Workshop sobre as Implicações da Computação na Sociedade. SBC, 2021.

[2] Baker, Ryan, Seiji Isotani, and Adriana Carvalho. "Mineraçao de dados educacionais: Oportunidades para o brasil." Revista Brasileira de Informática na Educação 19.02 (2011): 03.

[3] BRAZIL. 1971. Lei nº 5.692, de 11 de agosto de 1971. Fixa Diretrizes e Bases para o ensino de 1° e 2º graus, e dá outras providências.

[4] BRAZIL. 2022a. Computação: complemento à BNCC. Available http://portal.mec.gov.br/docman/fevereiro-2022-pdf/236791-anexo-ao-parecer-cneceb-n-2-2022-bncc-computacao/file.

[5] Greenhow, Christine, Beth Robelia, and Joan E. Hughes. "Learning, teaching, and scholarship in a digital age: Web 2.0 and classroom research: What path should we take now?." Educational researcher 38.4 (2009): 246-259.

[6] Han, Jiawei, Jian Pei, and Hanghang Tong. Data mining: concepts and techniques. Morgan kaufmann, 2022.

[7] INEP. (2021). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Retrieved from http://www.inep.gov.br/

[8] Mining, What Is Data. "Data mining: Concepts and techniques." Morgan Kaufinann 10 (2006): 559-569.

[9] Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews) 40.6 (2010): 601-618.

[10] Romero, Cristobal, et al., eds. Handbook of educational data mining. CRC press, 2010.

[11] Namratha, B., and Niteesha Sharma. "Educational data mining–applications and techniques." International journal of latest trends in Engineering and Technology 7.2 (2016): 484-488.

[12] Fonseca, Stella Oggioni da, and Anderson Amendoeira Namen. "Data mining on inep databases: An initial analysis aiming to improve brazilian educational system." Educação em Revista 32 (2016): 133-157.

[13] Santos, Aurea TB, et al. "Educational data mining: A study on socioeconomic indicators in education in inep database." Advances in Data Science and Management: Proceedings of ICDSM 2019. Springer Singapore, 2020.

[14] Araújo, Herlane Martins, et al. "Mineração de dados educacionais: um estudo sobre a proficiência em matemática no Ceará: Educational data mining: a study on mathematics proficiency in Ceará." Brazilian Journal of Development 8.10 (2022): 68289-68303.