A New Data Mining and Visualization Tool for Brazilian Educational Indicators

Rodrigo Silveira de Pinho, Danilo Borges da Silva, Suzana Matos França de Oliveira

[1] Universidade Estadual do Piauí

rodrigopinho@aluno.uespi.br,{danilo,suzana.matos}@frn.uespi.br

Abstract. *The National Institute of Educational Studies and Research Anísio Teixeira provides open data that help to understand Brazilian Education through Educational Indicators. Despite the easy access to the data, it is hard to manipulate and analyze it to understand the educational scenario in different contexts. This work presents a new visualization tool, inspired by the data mining process, which aims to allow the extraction of knowledge through these indicators using selections provided by users in a very simple manner. To show the tool's potential, graphics of several works were recreated and new graphics are presented.*

1. Introduction

Recording educational data is of fundamental importance to a nation's development. For this, several indicators were created to measure the educational level of students and institutions such as: SAT (Scholastic Aptitude Test), GPA (Grade Point Average), ENEM (High School Educational Examination), from Brazil, among others [Heiskala et al. 2021, Nagarathinam et al. 2021, Gomes and Borges 2009]. A Brazilian institute concentrates data capable of measuring the educational at all levels.

The National Institute of Educational Studies and Research Anísio Teixeira (INEP) has made available, as open data since 2015 [Vitelli et al. 2018], several Educational Indicators which belong to schools across the country [INEP 2021]. This large amount of data is accessible in raw form. Therefore, to acquire information and analyze this data it is necessary to filter, integrate and to group, summarizing the information as condensed representations using more simplified tables or graphs [Munzner 2015].

The analysis of these data is interesting for various groups such as public agencies, researchers from different areas, or even companies [Penteado et al. 2019]. There are researchers in education who carry out studies on the teaching-learning process is in a given region to achieve different purposes, especially to assess the quality of education [Caetano 2018, Lacruz et al. 2019, Costa 2019, Vitelli et al. 2018]. However, these researchers often present the results in simple tables, which do not explore the human visual system as a means of communicating this information [Munzner 2015]. That is, a set of visualizations could be used to better understand the data [Yau 2013].

This work presents a new tool for visualization of Educational Indicators that aims to help researchers in the field of education and others to analyze teaching-learning in Brazil using various filtering options. This tool is based on the process of Data Mining in the manipulation of data [Goldschmidt et al. 2015]. The extraction of knowledge about the quality of the education is left to be made by educational researchers. In this tool, we

use data from elementary and secondary education in ten educational indicators between the years of 2015 and 2020. To demonstrate the tool's potential, we reproduce several works and possibilities for new analyses.

2. Related Work

This section presents works that use the Educational Indicators from INEP open data [INEP 2021] to assess education in different locations.

[Costa 2019] analyzed the initial years of 549 schools focusing in the schools of the Municipal Education network of São Paulo in 2015. She detailed the Indicators showing how they work and some information about the emergence of new indicators and their use. In one of the results of this work, a reproduction of a correlation matrix of the Indicators similar to the one presented in [Costa 2019] was used.

The correlation between Educational Indicators is also addressed in [Bezerra et al. 2020] selecting public high schools in the city of Rio de Janeiro between 2010 and 2016. Despite the different time interval from the one used in this work (between 2015 and 2020), it is still a valid comparison due to the possibility of examining how these data have changed over the years.

The Indicators of high school students in state schools in Porto Alegre in 2014 are the core of the study of [Vitelli et al. 2018]. Interestingly, the source work is similar to the ones in the technical notes of the Indicators on the INEP website [INEP 2021]. Some scatter plots are provided and we also reproduced them in our tool.

[Pereira et al. 2018] highlights higher education throughout Brazil between 2014 and 2016 , using the Tableau[1] software for the creation of graphs and tables showing the differences that occurred during each of the years discussed for both public and private universities. This same software was used to build our visualization tool.

In most of the works presented, the authors tend to limit themselves to certain regions and schools, making it impossible for a comparison among different economic realities to be carried out. However, the same procedure can be used to analyze the performance of education in other locations. One possible reason could be that the process of processing raw data from INEP is exhaustive, making future comparisons unfeasible.

In the Section 3, the construction steps of the tool are presented, that allowing us to reproduce these works easily.

3. Data Mining and Visualization Tool for Brazilian Educational Indicators

This tool aims to organize and manipulate some Educational Indicators from INEP in a simple way. For this we use as part of the process the fundamentals of Data Mining, inspired in the model process KDD (Knowledge Discovery Databases) [Shafique and Qaiser 2014], that aims to extract the hidden knowledge according to the database. The Figure 1 shows the construction process. Briefly, three steps are performed:

1. Acquisition of Educational Indicators that possess all information referring to elementary, middle, and high school between 2015 and 2020, from the INEP [2];

[1] https://www.tableau.com/

[2] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais

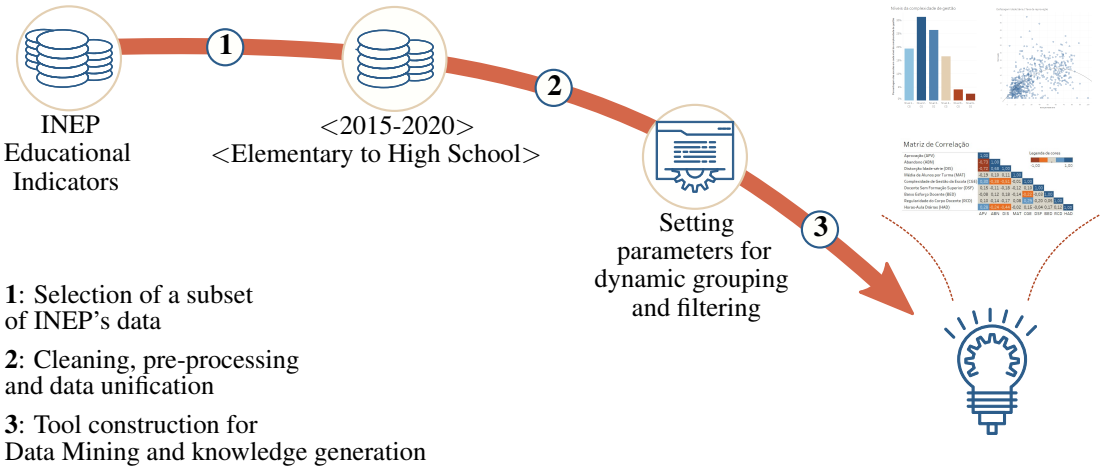


Figure 1. Process of building the visualization tool with data mining. This scheme was inspired by KDD model where part of the INEP's data was extracted, processed, and made available in a tool where the user can do data mining and generate knowledge from ten Educational Indicators.

2. Cleaning, pre-processing and unification of data in a single file, removing unnecessary data; and
3. Construction of visualization tool using parameters that allow to dynamic filter and user interactivity.

Each one of these steps is detailed in Sections 3.1 to 3.3.

3.1. Data from INEP

As the selection criteria of which Indicators to use, it was chosen only Educational Indicators of schools that have information between 2015 and 2020, with it, the used Indicators were listed in the Table 1.

Due to the absence of a selected year from some indicators, this work does not use all the Educational Indicators from INEP website, such as: Educational Financial Indicators, Average of Teacher Remuneration and Transition rates. Once that the Social-Economic Level Indicator is outdated, it was not used, although it is one of the most important indicators for visualization to make correlations between educational and social context.

The analysis of how each one of these Educational Indicators was created is not in the scope of this work.

3.2. Data processing

Data processing was performed using a language and an environment specialized in data, the R language [R Core Team 2020], integrated with Google's notebook environment. For this, it was necessary to configure the Google Colaboratory since it is configured with Python programming language by default. In addition, it was essential that all indicators were in the cloud and we choose hosting them on Google's own storage service (Google Drive).

We begin removing the header and footnote information, letting only the columns. Furthermore, the names, columns positions, and the city name column in 2015, of all the indicators, are not following the other years' model, hence it was necessary to perform adjustments. After this stage, it was realized a crossing of every referent indicator for each corresponding year, which generate a table with all information for each year, resulting in six different tables. Up next, all of these tables were unified into one single table which contains all of the data referents to the indicators of the years from the chosen interval. The key that joins all this data was the *School Code* column.

Another adjustment was in the cells, which must be empty, had a worthless fill pattern: $--$. Therefore, that cells had their values were altered to *null* . In the final stage, we delimited that rows with at least one null value , related to the education level specific (elementary, middle, or high school), would be discarded .

The influence of the data processing may be evaluated with the comparison between the number of rows and columns which each Indicator had, as well as the original file size, compared with the post-processed file. Table 1 presents that values, regarding the lines of the years which will be added, the columns are not dependent of the year and the size, too, will be added between 2015 and 2020. It was possible to obtain a single file whose size is about 24% of the size of the original set and a reduction of about 27% of columns. The code used in this process can be accessed on: https://bit.ly/DMVTBEI-code.

Table 1. Comparison between the size of the indicator's original files and the size of the created file. The number of rows and the file size take into account the sum between 2015 and 2020.

Educational Indicators	Rows	Columns	MB
Adequacy of Teacher Training	1079637	44	238
School Management Complexity	1098845	10	66,9
Teaching Effort	826706	33	135
Average of Students per Classroom	1065527	31	183
Daily Average Hour-Classroom	940714	30	134
Percentage of Teachers with Higher Education	1096035	19	110
Teachers Regularity	1008651	10	71,1
Age-grade Distortion Rates	854498	26	125
Non-response Rates	816190	27	112
School Performance Rates	816191	63	254
TOTAL:		293	1429
Table file after the pre-processing	747000	212	344

3.3. Parameters setting and Visualization Tool

With the unique file created which contains all the information, the process to manipulate the data through the different configurations has become simpler because the user just needs to be focused in applying the filters to get the desired result.

The Tableau software was used to create the charts, this is a visual analysis platform that allows the user to create and share visualizations in a simple and dynamic way. For the filters to work as expected it was necessary to manipulate the software more internally. This software allows you to create new parameters that are responsible for

manipulating a data or data set using conditional commands, for example. We create a parameter for teaching stages, which are divided into Elementary School, Middle School, and High School.

Among the different possibilities, the user may visualize statistically-based in the year, geographical region, the complexity level of management. However if you want to go deeper, you may set filters to the different stages of teaching, being on their role choose the tool which is most appropriated and apply the desired filters. The visualization tool is composed with:

- **Correlation Matrix**: to show the correlation between several columns from the Indicators;
- **Bar chart**: to show the management complexity and to show the level of teaching effort;
- **Dispersion plot**: to show some correlations between columns from the indicators with the tendency curve.
- **Line plot**: to show the evolution of some indicators over the years.

We show these visualizations with studies cases in the Section 4.

4. Results

To proceed with the steps presented in Section 3 some software were used. Data processing was done using Google's Colaboratory[3] integrated with the R language while data visualization was done with Tableau. It is noteworthy that these steps can be performed in other tools.

Some of the visualizations created will be presented and, when possible, we compare them with visualizations from other works. To show the tool's potential, watch the reference video: https://bit.ly/DMVTBEI-reference-video; to see the interactivity to reproduce this section static results. The tool can be accessed in its entirety on: https://bit.ly/DMVTBEI-tool.

Correlation Matrix. [Costa 2019] used a multivariate correlation matrix that contains the correlation coefficient between different indicators. Figure 2 shows the correlation matrix, where data from schools in the city of São Paulo were used for elementary school in 2015 to compare with what was analyzed by [Costa 2019].

The color palette used in the generated correlation matrix makes it more prominent in the identification of indicators that have some correlation. Thus, it is easily observed that the Indicators that have a negative correlation with Approval, for example, are Dropout and Age-grade Distortion, while many indicators have low or no correlation (values closer to 0).

Bar Chart. Figure 3 presents two graphics. Figure 3a shows the percentage of schools at each level of Management Complexity and Figure 3b shows the percentage of teachers in a school who fall at each level of teaching effort. Both are based on the graphic present in [Vitelli et al. 2018]. While the reference graphic shows the information in 2013, which was added to the tool shows the average between 2015 and 2020. In this case, a direct comparison cannot be made due to the difference in the chosen interval.

[3]https://colab.research.google.com/

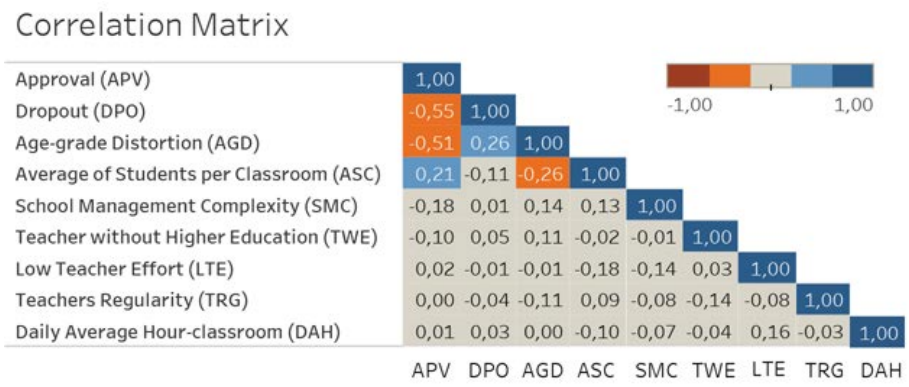
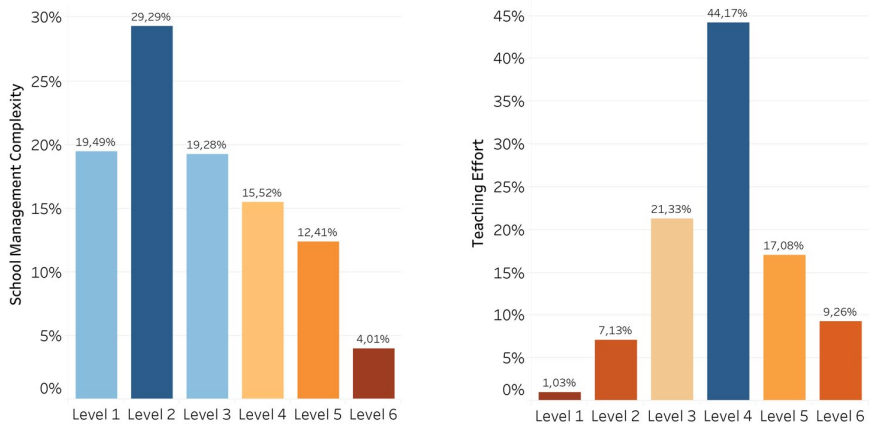


Figure 2. Correlation matrix based on data used by [Costa 2019].



(a) Percentage of schools according to the level of management complexity.

(b) Distribution of teachers according to the level of teaching effort.

Figure 3. Graphs were generated with data from high schools throughout Brazil between 2015 and 2020.

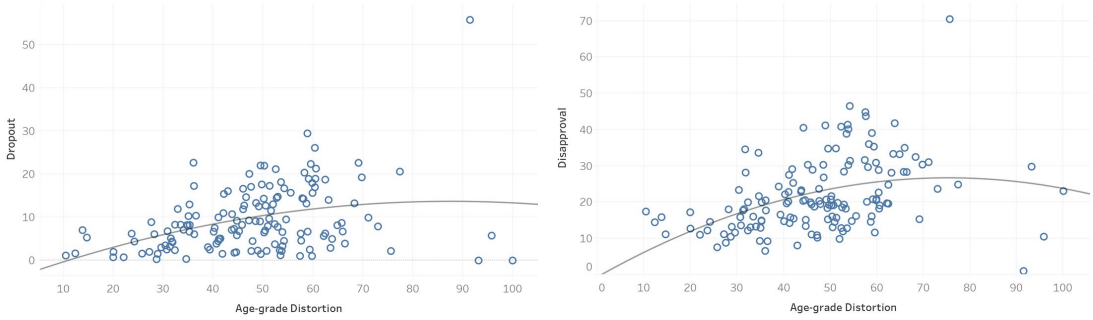
Dispersion plot. [Vitelli et al. 2018] also presents two dispersion plots: a) between the Age-grade and Disapproval rate; and b) between Age-grade and Dropout rate. Both on high schools from Porto Alegre in 2014. Despite the difference in the years as previously, this graph was replicated between the chosen interval (Figure 4).

Therefore, it's possible to have another way to evaluate the correlation between the data which were showed on the Correlations Matrix (Figure 2), where is possible to see how each school influences this result, once the information is showed when an icon is pointed with the mouse (reference video).

Line graph. Based on the dispersion plots, two line plots were created (Figure 5) showing the rate of Dropout and Approval over the years. With these graphics, are possible to analyze that the approval has increased during the years and the Dropout has decreased.

5. Conclusion and Future Work

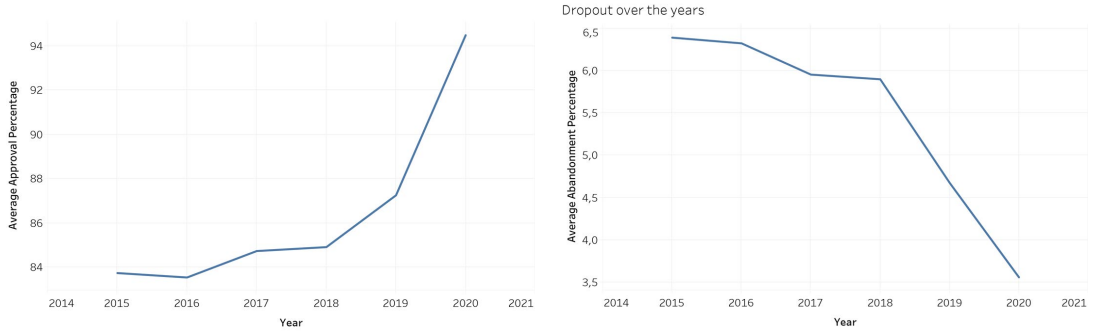
This work was inspired by the KDD model and in order to create a tool for visualizing Educational Indicators data. For this, it was necessary to select and clean the data provided by INEP. The created plots were based on several works that use this data to analyze



(a) Dispersion plot between the rate of Age-grade distortion and the rate of Dropout.

(b) Dispersion plot between the rate of Age-Grade distortion and the rate of Disapproval.

Figure 4. Generated graphs with the data of the high school of the state schools of Porto Alegre between 2015 and 2020. The curve represents a quadratic regression.



(a) Approval percentage average.

(b) Abandonment percentage average.

Figure 5. Generated graphs with the data of the high schools from Brazil between 2015 and 2020.

teaching in a given place (state or city). In order to validate these plots, it is possible to compare the information with the original works. This comparison was only possible because the tool is able to filter and generate interactive visualization according to certain requirements, such as place, year period, teaching stage and administrative dependence.

As a result, several researchers may use this tool to compose their conclusions in their respective contexts about education. As future work, we intend to integrate into the tool socio-economic data to allows us to observe that there may be several factors that influence education in a nation as large as Brazil.

References

Bezerra, L. F., Gonçalves, C. P., da Cunha, D. d. O., and de Oliveira, F. L. (2020). Análise da correlação entre a média de alunos por turma na taxa de rendimento de alunos nas escolas públicas de ensino médio no Município do Rio de Janeiro. *Revista Educação Pública*, 20(36).

Caetano, L. L. (2018). Docentes Do Ensino Médio: Análise De Alguns Indicadores Educacionais Relacionados Aos Docentes Do Ensino Médio Nos Estados Da Região Sul Do Brasil No Período De 2013 A 2017. Technical report, Universidade do Sul de Santa Catarina.

Costa, V. A. (2019). Os novos indicadores educacionais brasileiros: um estudo sobre a Rede Municipal de Ensino de São Paulo. Master's thesis, Universidade de São Paulo, São Paulo.

Goldschmidt, R., Passos, E., and Bezerra, E. (2015). *Data Mining*. Elsevier Brasil.

Gomes, C. M. A. and Borges, O. (2009). O enem é uma avaliação educacional construtivista? um estudo de validade de construto. *Estudos em avaliação educacional*, 20(42):73–87.

Heiskala, L., Erola, J., and Kilpi-Jakonen, E. (2021). Compensatory and multiplicative advantages: Social origin, school performance, and stratified higher education enrolment in finland. *European Sociological Review*, 37(2):171–185.

INEP (2021). Indicadores educacionais. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais. Acesso em: 19 de jul. 2021.

Lacruz, A. J., Américo, B. L., and Carniel, F. (2019). Indicadores de qualidade na educação: análise discriminante dos desempenhos na prova brasil. *Revista brasileira de educação*, 24.

Munzner, T. (2015). *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press.

Nagarathinam, T., Elangovan, V., Obaid, A. J., Akila, D., and Tuyen, D. Q. (2021). E-learning in data analytics on basis of rule mining prediction in dm environment. In *Journal of Physics: Conference Series*, volume 1963, page 012166. IOP Publishing.

Penteado, B., Bittencourt, I. I., and Isotani, S. (2019). Modelo de referência para dados abertos educacionais em nível macro. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1808.

Pereira, P. F. O., da Conceição, K. R., Lemos, R. R., Fiuza, P. J., and Gonçalves, A. L. (2018). Uma análise temporal de dados abertos do ensino superior utilizando o software de visualização tableau. In *II Simpósio Ibero-Americano de Tecnologias Educacionais – SITE 2018*.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shafique, U. and Qaiser, H. (2014). A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12(1):217–222.

Vitelli, R. F., Fritsch, R., and Corsetti, B. (2018). Indicadores educacionais na avaliação da educação básica e possíveis impactos em escolas de Ensino Médio no município de Porto Alegre, Rio Grande do Sul. *Revista Brasileira de Educação*, 23(0):1–25.

Yau, N. (2013). *Data points: visualization that means something*. John Wiley & Sons.