EduVizBR: A decision support system for Brazilian high school students' performance analysis

Chrystinne Fernandes[1], Dienert Vieira[1], Tassiane Barros[1], Aian Shay[1], Nicksson Freitas[2] and Tiago Vinuto[2]

[1] Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife – PE – Brazil

[2]Samsung Development Institute for Informatics (SiDi)
Recife – PE – Brazil

{cof2, dav, mtbl, asbdc}@cin.ufpe.br

{nicksson.a, t.vinuto}@sidi.org.br

Abstract. *Due to the explosive growth of educational data, more assistant tools are demanded by Educational managers to mine and filter strategic knowledge into massive databases. In this paper, we present the EduVizBR, a decision support tool designed to assist managers in analyzing the students' performance in the Brazilian national exam (ENEM). The EduVizBR allows managers to explore a massive volume of integrated educational data by narrowing their analyses according to the most relevant impact factors presented in the literature, year of interest, and school subjects. In our case study, we analysed how gender impacts students' grades in all Brazilian states and the influence of parents' educational attainment and parents' professions on their children's grades.*

1. Introduction

As public datasets become available in the Education scenario, more assistant tools are demanded to integrate and filter strategic knowledge into massive databases [M. A. P. Terrin and Bugatti. 2014]. The Educational Data Mining (EDM) research area addresses different methods for exploring the increasingly large-scale educational data to better understand students' context [Romero and Ventura 2020], while the Educational Process Mining (EPM) emerging field aims to make unexpressed knowledge explicit and facilitate a better understanding of the educational process [Bogarín et al. 2017].

Educational managers tend to be overwhelmed by the massive volume of data available during their decision-making process. Therefore, Decision Support Systems (DSS) are essential to help education professionals focus on the relevant subjects and gain the necessary knowledge to make faster and more reasonable decisions [Velasco et al. 2020, Faria et al. 2019]. In this process, consistent methods are required to treat, group, and prepare all data in a single data warehouse.

Finding a set of factors that influence students' performance that can be applied to a specific education scenario of investigation is a difficult task. This is mainly because the set of factors presented in the literature is too subjective to be measured through the available datasets in the field of education, such as personal motivation, parents' expectations and involvement, home affective environment, discipline, and structure for learning.

In this work, we investigated factors that affect students' performance in assessments worldwide. The studies we found come mainly from areas such as psychology. These works met their expecting results of finding the factors and presenting interesting analyses regarding the education scenario. However, as a drawback to these studies, we can point out the lack of any technological resource utilized to present their results through interactive data visualization tools. Thus, we utilized these factors (collected from the literature) as the main targets to generate data visualization in our system to assist educational managers in quickly finding the most important information. Through the use of the EduVizBR, managers have the flexibility to filter the most relevant information according to the factors, desired year (2015-2020), and even Brazilian states.

We defined two Research Questions (RSQ) to guide our study: RSQ1- Which factors influence the performance of high school students? RSQ2- Is the creation of a DSS to guide managers viable?

Our Research Goals are defined as follows: i- To investigate the main factors that impact Brazilian high school students performance using the ENEM [1] dataset; ii- To design and develop a dashboard to concentrate the information needed to support the decision-making process of education managers.

This paper is organized as follows: Section 2 presents our related work, describing strategies to deal with the problem of analysing students' performance. Section 3 presents the EduVizBR with its architecture scheme. Our main contributions are presented in Sections 3 and 4, where we discuss our research questions. Finally, we conclude this paper and present possible future work in Section 5.

2. Related Work

According to [Lopes et al. 2020], several aspects can affect student performance in assessments, including the family, and classes or social strata. They can be divided into hierarchical levels organized into three distinct levels: i) at the student level, comprising personal and family aspects such as socio-demographic characteristics and family socioeconomic and cultural capital; ii) at the class level, related to aspects of the classroom (i.e., teacher characteristics, peer effects, and pedagogical styles and practices), and iii) at the school level, which encompasses factors related to the educational institution (i.e. violence, school policies or practices) [Lopes et al. 2020].

The research carried out in [Chaia et al. 2017] listed the five factors that have the greatest impact on the academic performance of students in the exam of the Programme for International Student Assessment (PISA): i) personal motivation, ii) the proper combination of teacher guidance and self-research in teaching practice, iii) the use of information and communication technology (ICT) as a pedagogical tool by teachers, iv) the increase in school hours and v) early childhood education.

Another study [Farooq et al. 2011] examined different factors influencing the academic performance of secondary school students in a metropolitan city of Pakistan. A survey was conducted by using a questionnaire for gathering information from 600 10^{th} grade students (300 male and 300 female). The factors brought by this work were: i) age, ii)

[1]ENEM dataset available at: `https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem`

gender, iii) geographical belongingness, iv) ethnicity, v) marital status, vi) language, vii) parents' educational level, viii) parental profession, ix) income, x) socioeconomic status (SES), and xi) religious affiliations. The findings of this study are the following: 1- SES and parents' education have a significant effect on students' overall academic achievement; 2- The high and average socio-economic level affects the performance more than the lower level; 3- Parents' education means more than their occupation in relation to their children's academic performance; 4- Girls perform better than the male students.

In our work, we defined a set of eight factors based on the factors showed in [Farooq et al. 2011]. We chose this set mainly because they were measurable through our ENEM dataset that contains information about gender, parents' education level, and parental profession. Additionally, we conducted our analysis based on the analysis made by [Farooq et al. 2011] to compare our results with those listed in their research.

3. EduVizBR

In this section, we present the architecture of our tool and describe each of its layers.

3.1. Architecture

In Figure 1 we present the architecture of our developed tool.

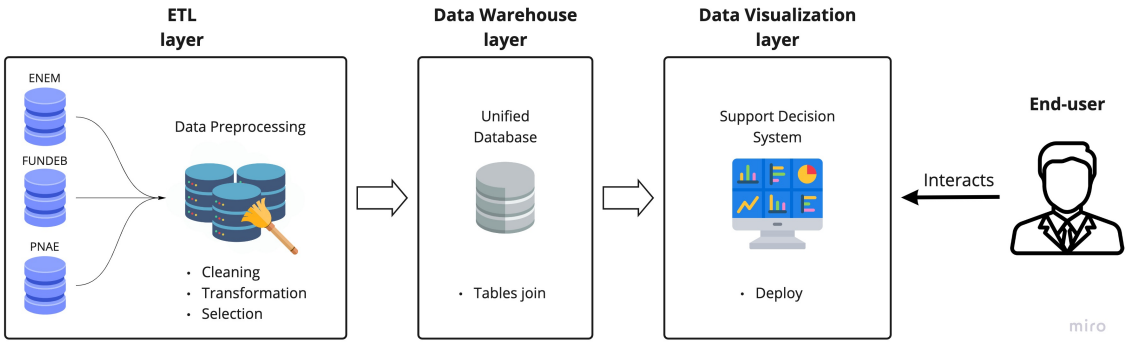

Figure 1. EduVizBR Architecture

The EduVizBR[2] comprises the following interconnected layers (L1-L3):

L1- Extract Transformation and Load (ETL) Layer: this layer gets raw input data from our dataset to preprocess, clean, and transform this data by selecting specific columns of interest, and removing rows with no information. The tidied data is sent to the next layer.

L2- Data Storage Layer: this layer stores this unified version using Parquet and processes it using Dask. This unified version makes it easier for us to explore data of different contexts according to specific needs, depending on which kind of visualization the manager wants, and send it to L3;

L3- Data Visualization Layer: this layer consumes the unified data from L2 and plots the visualizations to the end-user (educational managers). We utlized the Streamlit technology to build our DSS, and Matplotlib, Seaborn, and Plotly for the graphs.

[2]Code available at: `https://github.com/Chrystinne/enem/enem`

4. Case Study: Analyzing high school student's performance in the ENEM

In this section, we present a case study and investigated factors that affect students' performance in assessments worldwide to evaluate EduVizBR,

4.1. Data

In our study, we utilized the ENEM dataset from 2015 to 2020 containing students' personal information, students grades, and the student's answers to a socioeconomic questionnaire with 25 questions.

In data processing, we removed from the dataset all data related to candidates who missed at least one day of the exam. Since there are millions of participants each year, we filtered only the necessary columns to generate the views with Dask[3] to reduce data loading and processing time. And to reduce storage space, we saved the database using Parquet[4].

We created groups based on the candidates' region and school type, then we replaced missing grades with the group mean. Thus, we observed that some states had a lot of null data, and using the mean could influence the results. We solved this problem by excluding the data where one of the grades was null.

4.2. Case Study Results

We chose to use the mean scores in all our analyses. We found that, in 2020, the mean score for male students was higher than the mean score for female students in Mathematics (Figure 2). We used the ztest statistical test from the statsmodels library to compare the two sets of samples (i.e., men's and women's mean scores in math for 2020). We defined the following null hypothesis (H0): Men's and women's mean grades in mathematics were the same for 2020.

Since the ztest method returned a p_value equal to zero, we can refute the null hypothesis which states that the grades for men and women were the same for 2020. As the p_value is less than 0.05, there is a statistically significant difference between the grades of the two genders, where the average of the men's grades is higher than the average of women's grades in 2020. This difference is within a confidence interval between 49.88 and 50.46 in the averages of these two groups, as observed in the result of the zconfint method from statsmodels library.

Regarding parents' educational attainment and parents' profession, we also perceived the impact of these factors in students' performance, as also shown in [Farooq et al. 2011]. Students with parents who had higher educational attainment showed better grades for all the five school subjects in comparison to the ones whose parents had lower educational attainment (Figure 3). Similarly, students whose parents had less rewarded professions performed worse than the ones whose parents are better rewarded (Figure 4). Even though we showed these results using 2020 data, the same pattern occurred for all the range considered in this paper (2015-2020). Students whose parents had higher educational attainment and are better rewarded performed better in all study areas.

[3]Dask available at `https://dask.org/`

[4]Parquet available at `https://pypi.org/project/parquet/`

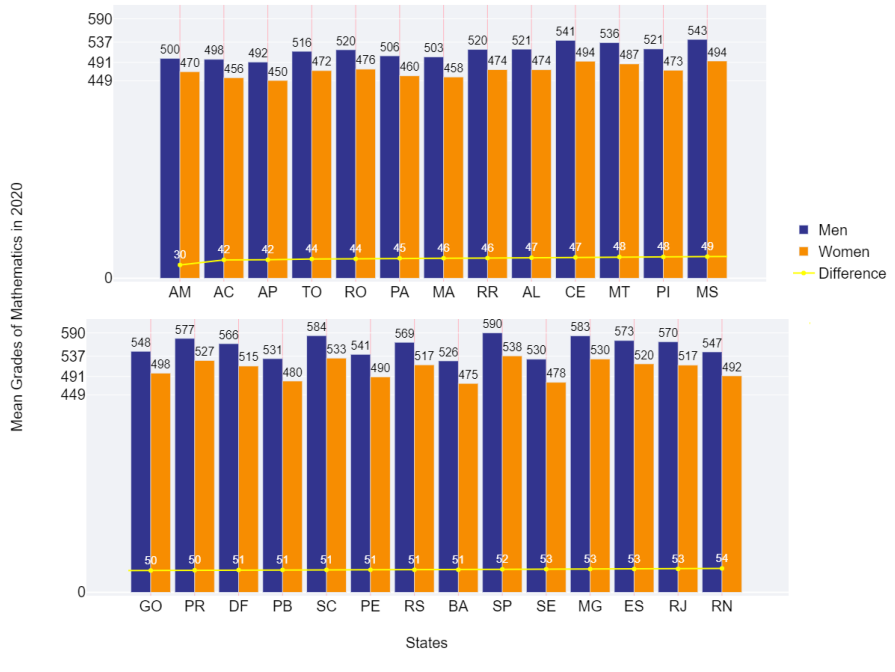

Figure 2. Mean of mathematics grades per Brazilian State, in 2020.

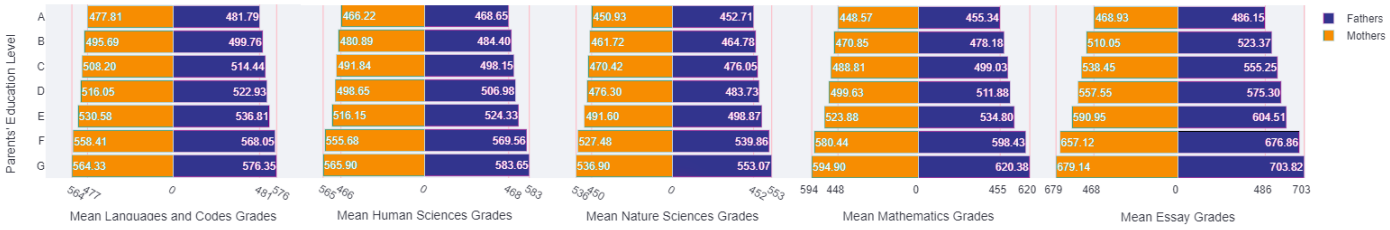

Figure 3. Mean grades per parents' attainment in 2020.

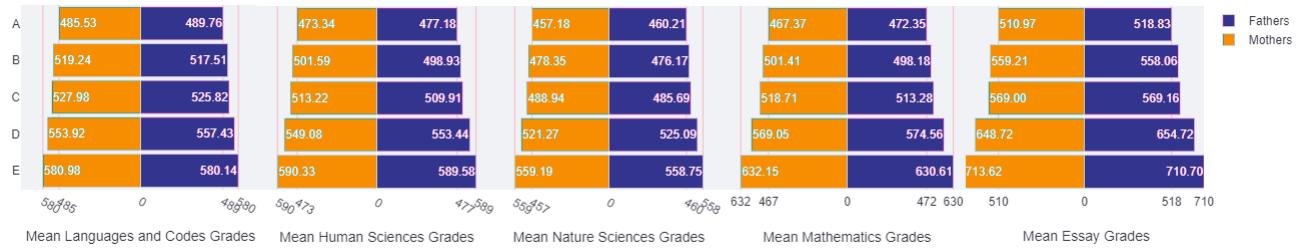

Figure 4. Mean grades per parents' professions, in 2020.

4.2.1. Discussion

We answered our RQ1 by presenting the set of factors that impact Brazilian students for the ENEM scenario. This set can be utilized by other researchers to replicate similar analyses in the Educational scenario. At last, by answering our RQ2, we reached our most important result, which was the design and development of a DSS to guide managers in their decision-making process.

5. Conclusion and Future Work

The main goal of this work was to assist managers in analyzing Brazilian students' performance. We considered students' performance analyses by gender for all Brazilian states, as well as analyses by parents' educational attainment, and parents' profession. As shown in our case study, we met our expected goals of demonstrating that these factors had some influence on students' performance considering the ENEM scenario. The EduVizBR can facilitate the decision-making process of Education managers. By using the tool, managers have access to public datasets and interactive visualizations. Our tool can assist managers to make faster and more reasonable (data-driven) decisions.

We plan to evolve this work by adding Key Performance Indicators to investigate whether actions taken in this scenario are generating the expected results. We aim at extending the EduVizBR with a Machine Learning module to predict students' performance for the years ahead. We also aim to validate our DSS using different use cases such as SAEB data, PNAE, and FUNDEB. At last, we plan to analyze the Brazilian investments made in Education over time using FUNDEB and PNAE datasets.

Acknowledgement

The results presented in this paper have been developed as part of a project between SiDi, financed by Samsung Eletrônica da Amazonia Ltda., under the auspices of the Brazilian Federal Law of Informatics no. 8248/91.

References

Bogarín, A., Cerezo, R., and Romero, C. (2017). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(1):e1230.

Chaia, A., Child, F., Dorn, E., Frank, M., Krawitz, M., and Mourshed, M. (2017). What drives student performance in latin america.

Faria, C. R., Júnior, M. M. C., and Barbosa, C. R. S. C. D. (2019). Sistema de apoio à decisão por PLN para consultas de pragas na cultura da soja. In *Anais do Seminário Integrado de Software e Hardware (SEMISH)*. Sociedade Brasileira de Computação - SBC.

Farooq, M. S., Chaudhry, A. H., Shafiq, M., and Berhanu, G. (2011). Factors affecting students' quality of academic performance: a case of secondary school level. *Journal of quality and technology management*.

Lopes, S. G., de Carvalho Xavier, I. M., and dos Santos Silva, A. L. (2020). Rendimento escolar: um estudo comparativo entre alunos da área urbana e da área rural em uma escola pública do piauí. *Ensaio: Avaliação e Políticas Públicas em Educação*.

M. A. P. Terrin, C. N. S. J. and Bugatti., P. H. (2014). Utilizando técnicas de mineração de dados para apoiar a busca ativa de famílias em situação de vulnerabilidade e risco social. Master's thesis, Universidade Tecnológica Federal do Paraná.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3).

Velasco, R. B., Carpanese, I., Interian, R., Neto, O. C. G. P., and Ribeiro, C. C. (2020). A decision support system for fraud detection in public procurement. *International Transactions in Operational Research*, 28(1):27–47.