Data Handling: Cleaning, Biases and Privacy

Juan David Nieto García, Giovani de Almeida Valdrighi

1 Introduction

With the motivation to address the problem outlined in Activity 1, which aims to confirm through our ENEM and SAEB score prediction models that schools with better economic conditions are more likely to achieve higher scores, while those with poorer conditions are more likely to rank lower. We chose the Basic Education Assessment System 2019 (SAEB) database[1] authored by National Institute of Educational Studies and Research Anisio Teixeira (INEP). Throughout this activity, we focused on key aspects related to data collection planning, data preparation and privacy assurance.

2 Related Works

Since 2019, the evaluation of the inputs offered in basic education schools has been implemented through SAEB. This initiative, which was initially carried out on a sample basis, was introduced with years of delay as it had been planned since 2010 in the National Curriculum Guidelines for Early Childhood Education. The evaluation is also included in the National Education Plan (PNE) and was supposed to be implemented in 2016.

To provide context, not so much on data preprocessing, but on data protection and privacy in education in Brazil, the following references were found:

The General Data Protection Law (LGPD) requires principals to review a series of procedures, such as privacy policies and terms of use for their platforms, to ensure that students personal data is protected and both their privacy and their results are maintained. Likewise, these processes must be followed by national-level evaluating entities [1].

On the other hand, we found that both SAEP and INEP data have already been processed in accordance with privacy policies. Privacy control in educational censuses conducted by INEP was analyzed through a Decentralized Execution Agreement (TED) signed between the Institute and the Federal University of Minas Gerais (UFMG). Based on the analysis results, the Directorate of Educational Studies (DEED) at INEP issued a technical note suggesting actions to address privacy treatment in public microdata [2].

Thus, through tests and questionnaires applied every two years in the public network and a sample in the private network, the SAEB seeks to determine the learning levels of the students surveyed, with the aim of determining the quality of Brazilian education.

1. www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb

3 Data Description

Dataset name: Basic Education Assessment System 2019 (SAEB).

Author: National Institute of Educational Studies and Research Anisio Teixeira (INEP).

The dataset is organized into related into 5 tables: student, school, teacher, principal and municipal secretary. Each table contain a questionnaire with answers about social conditions of the involved. To ensure the privacy of each school, the information regarding the school and municipality is anonymized, making it impossible to discern the identity of any specific school from the results.

Below are some important points aimed at addressing certain questions posed in the study. These will help us better comprehend the structural nature of the database.

- **Data Structure:** The data is organized in a structured static format (tabular), with each row representing an individual or school.
- **Data Source:** The data is available as open access on the government's official webpage.
- **Data construction Phase:** The data is in the raw data space, *i.e.*, they have been collected from a sampled population and organized into tables.
- **Data Privacy:** The data was published already with privacy considerations. The name of the school and the name of the municipality is masked with a random code, this random code cannot be used in reverse engineering to obtain the original names. One problem that arise with this anonymization is that other surveys about school, such as the census of schools that is not anonymized in the same manner can't be related with the SAEB results.

Regarding the information provided by the questionnaires, while cognitive tests conducted on students allow evaluating their performance, tests administered to the principals provide access to data related to their profile as leaders, activities carried out, available resources, and infrastructure of the educational institutions.

Currently, students from all over Brazil respond to printed questionnaires during the application of the SAEB cognitive tests. As secretaries, directors and teachers, they answer electronic questionnaires remotely, through response links sent to INEP by email, always supporting their confidentiality and privacy.

4 Data Treatment

It was necessary to relate two different tables, the table from schools **`df_escola`** and the table from principals **`df_director`**, as the dataset for schools alone had very few variables. Some directors did not answer the questionnaire, so these were subsequently removed from the dataframe.

A detail is that there were more than one principal for each school (just a few of them), to deal with that, it was selected only one principal from each school, and the relationship was made.

Next, columns related to educational levels other than the final year of basic education were removed from the table (*e.g.*, average grade of high school students).

To quantitatively explore the structure of the final dataset, some important information was printed, which is displayed below.

- **Total rows:** 39164
- **Total columns:** 269
- **Unique schools:** 39164
- **Unique cities:** 5518

Subsequently, for data processing, we aimed to manage data dependencies within the dataset based on specified rules and replace unanswered values according to these dependencies.

First, we iterated through the columns of the dataset and checked if the column exists in the predefined dictionary **`questions_dict`** (Questions asked to the principal). Then, we checked if the column has a specified dependency rule in **`question_dict`** and requested to process and extract the dependency rules. Finally, we iterated through the extracted dependencies, identifying rows where the specified dependencies are not met and updating the dataset by marking them as "invalid".

Continuing with the processing of data, columns with more than 30% missing values were selected and treated by replacing these values with predefined values ('0' or 'B') based on the column names. Finally, columns with more than 1% missing values were removed from the dataset, followed by the removal of rows containing at least one missing value in any of the remaining columns.

Finally, Figure 1 show some distributions afer treatment: (a) Distribution of NaN values in columns. (b) Distribution dtypes of columns. (c) Histogram of unique values per column. We can see that for Figure 1 (a) there were no values with NanN afer treatment, for (b) we have "object" type values that must be processed in the modeling and for (c) we have numbers of unique between 0 and 25, the histogram decreasing as this value increases.

- **Total rows:** 23940
- **Total columns:** 240
- **Memory usage:** 44.02 MBs

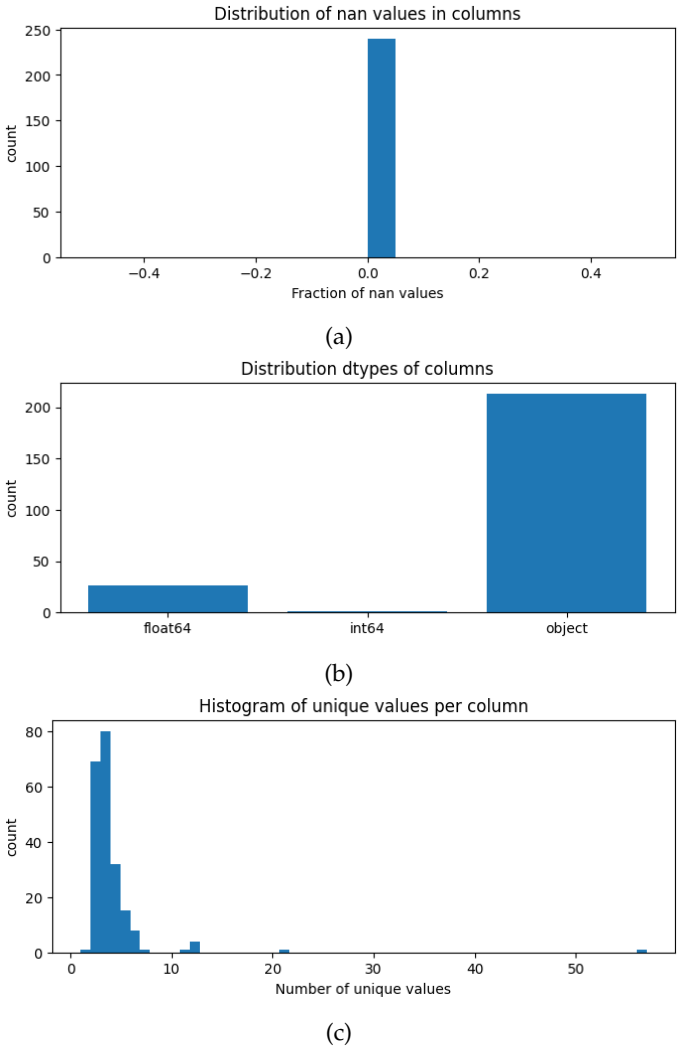

Fig. 1: Some distributions afer treatment: (a) Distribution of NaN values in columns. (b) Distribution dtypes of columns. (c) Histogram of unique values per column.

5 Data Usage

Clustering Schools Based on Economic Groups: The questionnaire variables from the survey as features will be used to cluster schools based on economic groups. This indicates an analytical approach where schools are grouped or clustered together based on the similarity of their economic characteristics or responses to the questionnaire. Clustering can provide insights into economic patterns among schools and assist in targeted policymaking or interventions.

Predicting SAEB Average of Schools: the questionnaire variables will be used to build a predictive model that estimates the SAEB average for each school. The SAEB average serves as an educational performance metric. By using the questionnaire responses as features, a predictive model can be trained to estimate a school's SAEB average, which can be valuable for educational planning and understanding the factors influencing academic performance.

References

[1] (2020) Lgpd na educaÇÃo: Entenda seus impactos e como as secretarias de educaÇÃo devem se adaptar. [Online]. Available: https://cieb.net.br/lgpd/

[2] (2019) Microdados -. [Online]. Available: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados