Leveraging String Similarity Algorithms for Educational Data Validation: A Scalable Approach for Digital Governance

Débora Barbosa Leite Silva[a], Emanuel Marques Queiroga[c,a*], Abílio Nogueira Barros[b], Markson Rebelo Marcolino[d,a], Diego Dermeval[a], André Lima[a], Leonardo Brandão Marques[a], Cristian Cechinel[d,a] and Thales Vieira[a]

[a]Center of Excellence for Social Technologies (NEES), Universidade Federal de Alagoas, Maceió 57072, Brazil, {debora.silva, diego.matos, andre.lima, leonardo.marques, thales.vieira}@nees.ufal.br

[b]Departamento de Computação, Universidade Federal Rural de Pernambuco (UFRPE), Recife 52171, Brazil, abilionbarros@gmail.com

[c]Instituto Federal de Educação, Ciência e Tecnologia Sul-Rio-Grandense (IFSul), Pelotas 96.015-360, Brazil, emanuel.marques@nees.ufal.br

[d]Centro de Ciências, Tecnologias e Saúde, Universidade Federal de Santa Catarina, Araranguá 88040, Brazil, cristian.cechinel@ufsc.br

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 19 May 2025

Abstract. Data validation is critical for ensuring the reliability of information in educational systems, particularly in the context of digital governance. In Brazil, fragmented student records across various governmental databases hinder the implementation of educational public policies that rely on massive, high-quality student data. One effective method for validating these records is cross-referencing data across different databases. However, textual data, such as student names, are often prone to errors, including misspellings. Applying overly strict validation rules may result in the exclusion of valid records, while too lenient rules may allow incorrect data to slip through undetected. This study addresses these challenges by proposing a new data validation methodology that uses the Levenshtein distance algorithm. The approach identifies an optimal similarity threshold by taking into account the capacity of the manual validation team since excluded students are manually verified as a subsequent step, allowing for a balanced solution. We applied this methodology to validate student data from the Sistema Gestão Presente (SGP), which manages around 7 million student records and integrated it with the Brazilian Federal Revenue database. Through two key experiments, we demonstrated how an optimal validation threshold could be determined by considering the manual validation team capacity. In this case study, we found an optimal 80% similarity threshold when the manual validation capacity is approximately 950,000.

Keywords. Data Governance, String Similarity, Educational Data Interoperability, Levenshtein distance, public policies.
Research paper, DOI: https://doi.org/10.59490/dgo.2025.942

1. Introduction

Data interoperability emerges as a fundamental concept in the current landscape, characterized by constant technological evolution and the increasing use of data across various sectors (McBride et al., 2022; Wimmer et al., 2018). Data interoperability is primarily defined as the ability of different systems and organizations to exchange, understand, and use data and information securely, efficiently, and in a structured manner. Sectors such as healthcare, education, government systems, and businesses increasingly rely on effective data integration and analysis processes to ensure data-driven decision-making.

In public policy management and implementation, data interoperability between different systems is key, enabling cross-referencing information from various sources (Campmas et al., 2022; Wimmer et al., 2018). However, this process faces technical challenges, such as the lack of data standardization, inconsistent communication protocols, and issues related to data trustworthiness (Almeida et al., 2020; Styrin et al., 2022). In education, these challenges are exacerbated by data fragmentation, often arising from different federative entities, which undermines the quality and reliability of the records used (Ali et al., 2019; Harron et al., 2017; Loureiro et al., 2021).

In this context, data validation is indispensable in ensuring that information is consistent and reliable. Validation is particularly critical in educational settings, where errors and inconsistencies can directly impact the effectiveness of public policies (Cohen, 2000). In Brazil, challenges related to the fragmentation of educational information hinder large-scale data analysis and the formulation and implementation of policies to address educational inequalities (Barbalho et al., 2022; Queiroga, Siqueira, et al., 2024; Tavares & Bitencourt, 2024). This fragmentation is inherent to the country's educational structure, where responsibilities are distributed across federal, state, and municipal levels, each managing their own educational networks and information systems (Costin & Pontual, 2020).

The decentralized nature of Brazil's educational system, while promoting local autonomy, creates significant challenges in data integration and validation processes (Arretche, 2004). School systems often operate with different technological infrastructures, data collection methodologies, and reporting schedules, making it particularly challenging to establish standardized validation protocols and achieve consistent data quality across all educational networks (Segatto et al., 2022). These challenges are further amplified by the substantial regional disparities in technological infrastructure and institutional capacity among different educational administrative units (Queiroz et al., 2020; Rocha et al., 2024; Wanke et al., 2024).

The Sistema Gestão Presente (SGP) was developed to centralize high school student information into a single, integrated system to address the challenges of integrating students' attendance data on a centralized system. This system aims to facilitate data cross-referencing and support the implementation of public policies focused on education, such as the Pé-de-Meia Program (PDM)[1]. Established by the Brazilian government in 2024 (Ministério da Educação, 2025b), the PDM is part of the pay-for-attendance programs and is a Conditional Cash Transfer Program, similar to the Mexican Progresa, a pioneering initiative that links financial incentives to school attendance and health checkups (Fiszbein & Schady, 2009; Kremer, 2003; Parker & Todd, 2017). The PDM provides financial incentives to public high school students to promote school retention, encourage the completion of their studies, and stimulate participation in the National High School Exam (Enem)(Ministério da Educação, 2025a), which was created in 1998 by the Anísio Teixeira National Institute for Educational Studies and Research (INEP).

The PDM operates based on enrollment, attendance, annual approval, and exam participation criteria. However, its effectiveness is heavily dependent on the quality of the data used. The SGP, which manages approximately 7 million unique student records, faces substantial challenges in data validation. Issues such as incorrect names, missing essential information, and registration inconsistencies undermine the database's reliability and hinder payment authorization, directly impacting students who depend on these incentives. These challenges emphasize the complexities of aligning large, heterogeneous databases and the critical need for robust technological solutions and rigorous methodologies to ensure accuracy and minimize student disruptions.

The present paper introduces a case study focused on improving the accuracy of student records in Brazil through data validation processes that leverage text similarity algorithms. The rest of this paper is structured as follows. Section 2 presents the related literature about interoperability in governmental systems and the use of similarity algorithms to improve data quality. Section 3 of the paper details the student data validation process implemented by the SGP, focusing on ensuring data reliability and accuracy under ideal conditions. Section 4 illustrates the proposed approach for the problem of students' data validation. Section 5 describes the experiments conducted to suggest the threshold similarity for validating the records. At last, Section 6 discusses the work results, and Section 7 concludes the paper.

[1] Pé-de-meia is a Brazilian expression that refers to money saved for future use

2. Literature review

Data validation and interoperability are fundamental pillars for efficient information management in government systems. Developing data interoperability platforms that facilitate data exchange, validation, and utilization is a critical priority in the digital transformation of governments (McBride et al., 2022). These efforts represent a significant step toward unifying the databases of complex systems, particularly those in key sectors such as education, health, and governance (Das et al., 2022; Malodia et al., 2021; Saffady, 2021; Styrin et al., 2022).

The challenges in these sectors, such as high data fragmentation and the lack of standardization, pose significant obstacles to the implementation and monitoring of effective public policies (Almeida et al., 2020; Queiroga, Siqueira, et al., 2024; Styrin et al., 2022). Specifically in education, these issues become even more evident due to the high variability of records and the need to integrate local, state, and federal systems (Ali et al., 2019; Almeida et al., 2020; Harron et al., 2017; Loureiro et al., 2021). In Brazil, legislation structures different federative levels of responsibility over public education, fragmenting data on enrollment, attendance, and performance among thousands of municipal, state, and federal education departments. This variability alone presents a significant challenge in the creation of data interoperability platforms, as data often follow different standards or include minor inconsistencies at the source, such as missing surnames or their abbreviations (Almeida et al., 2020; Queiroga, Santana, et al., 2024). Techniques like the Levenshtein distance and the Jaro-Winkler distance effectively address these inconsistencies, ensuring greater reliability in the records (Almeida et al., 2020; Handijono & Suhatman, 2024; Loureiro et al., 2021).

Thus, several authors demonstrate how data interoperability can yield significant results in knowledge discovery and supporting decision-making processes. Queiroga, Siqueira, et al., 2024 proposes a model for extracting key factors from the interoperability of educational data in Brazil. To achieve this, factor analysis techniques are applied to identify eight main factors related to structural equity in the educational system, aiming to provide personalized recommendations supporting digital governance and formulating context-specific public policies. However, the absence of unique identifiers posed a significant challenge, limiting the practical application of the research, particularly in addressing the needs of specific student groups.

In this context, the search for techniques that can help to identify degrees of integrated different databases and enable more automated data validation has been widely explored. One of these methods is string similarity (Kaufman & Klevs, 2022). Entity matching is another recurring challenge in systems aiming to integrate data from different sources. Sakai et al., 2021 investigated the combination of string transformations and similarity metrics to consolidate duplicate records. Similarly, Handijono and Suhatman, 2024 applied the Jaro-Winkler technique to identify strings with a high degree of similarity for detecting duplicate data in records. This approach seeks to optimize databases by reducing inconsistencies and redundancies. By clustering similar records based on names and registration data into single entities, the authors achieved satisfactory results in minimizing data duplication and improving data consistency.

Research on the use of semantic embedding models has also advanced progress. Asadollahi et al., 2024 demonstrated how models adapted to specific contexts can accurately capture textual and semantic nuances. In their study, the authors used construction data to train models aimed at improving the identification of semantic texts, with a focus on data interoperability between different systems and the development of a standardized data pipeline. The results indicate that techniques based on adapted string decomposition outperformed traditional methodologies and even advanced language models while maintaining a significantly lower application cost.

Data modeling, in turn, requires flexible approaches to address the diversity of contexts. Kouremenou et al., 2024 found that among 100 healthcare institutions participating in their study, 23 experienced data management or interoperability issues over the past five years. To address this, the authors proposed a generic and accessible framework designed to facilitate data exchange and mapping. This type of tool proved particularly useful for integrating data from systems where the lack of unified standards often creates redundancies and inconsistencies, enabling interoperability through data modeling prior to transfer.

Similarity metrics, such as TF-IDF and cosine similarity, also play an important role in data validation (Cohen, 2000; Downs et al., 2019). Gómez and Vázquez, 2022 conducted an empirical analysis of different similarity search methods within academic articles and theses. Their results showed that advanced techniques have a significant computational cost, without this cost translating into substantially better outcomes. Conversely, simpler techniques, such as those based on distance measurement, achieved comparable results to advanced methods with a significantly lower computational cost. This is particularly noteworthy as the scalability of solutions becomes critical with the increasing volume of data. Machine learning-based methods offer promising alternatives but still face limitations related to record variability and computational expense. Tools that combine computational efficiency with analytical precision are increasingly necessary to meet the growing demands of governmental systems, especially in the education sector.

Practical approaches have shown significant results. Damasceno et al., 2021 presented SimClear, a tool that leverages similarity functions, such as Levenshtein distance, to standardize educational records during the data cleaning phase. This tool has been successfully applied to correct textual errors, increasing data reliability and reducing the need for manual review.

It is also worth noting that while recent advances in Machine Learning techniques and semantic embedding have significantly expanded the possibilities for data validation and interoperability, traditional methods, such as text similarity algorithms (e.g., Levenshtein and Jaro-Winkler), continue to deliver satisfactory results with a favorable cost-benefit ratio. This is particularly true in scenarios where computational efficiency, scalability, and simplicity of implementation are critical (Gómez & Vázquez, 2022).

3. Student Data Validation

Data validation in educational systems is critical to ensuring the reliability and efficacy of the information used to formulate public policies (Almeida et al., 2020; Cohen, 2000; Tavares & Bitencourt, 2024). The Brazilian educational system is highly decentralized, leading to fragmented data and posing challenges for managing and implementing centralized programs or policies. The data validation flow at SGP is an essential component to assure information integrity, and it was carefully designed to avoid inconsistencies, assure record accuracy, and ease the identification of errors. The general flow of data collection and validation at SGP is presented in Figure 1.

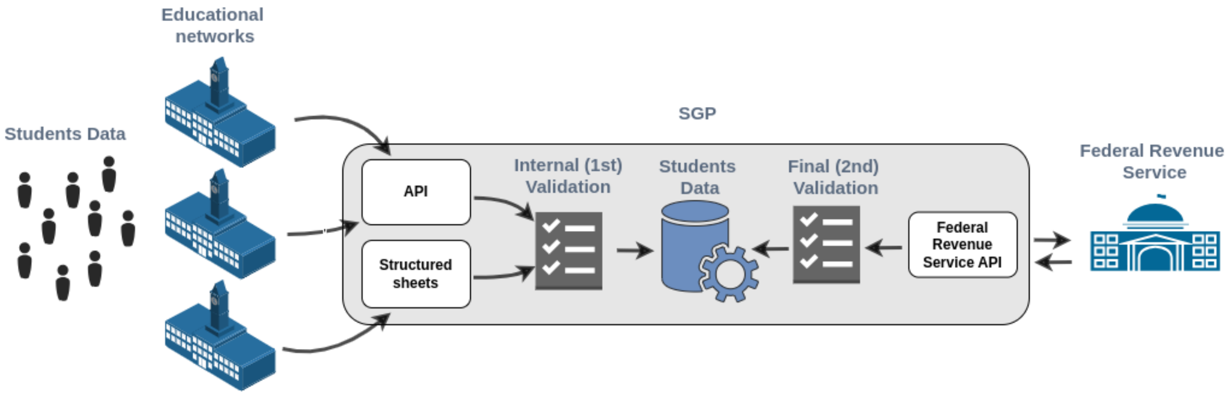

Fig. 1 – Data Validation Flow.

As seen in the figure, Educational networks (formed by the state, federal, and municipal entities) are responsible for initially submitting student data to the SGP through two methods: an integrated API for automated communication between their systems and SGP and direct submission of spreadsheet files on the platform. SGP administration encourages using the API to enhance security, improve efficiency, and reduce errors. However, as educational networks may not yet possess the required infrastructure for API integration, the submission of structured file sheets provides a viable alternative.

After receiving the data from educational networks, the system processes a first validation step that involves some basic verifications, such as ensuring the correct format of the fields, fulfilling all mandatory information, and identifying duplicates. Once verified, the data is preprocessed and stored in the system's database for the next validation stage. The tax identification numbers of students are then sent to the Federal Revenue Service of Brazil to verify consistency with government records.

The primary key for this validation with the Federal Revenue Service is the Brazilian Tax Identification Number (CPF, from the Portuguese Cadastro de Pessoa Física). The SGP sends the student's CPF to the Federal Revenue Service and receives the following information associated with him/her: day of birth, full name, and mother's name. These data are then compared to the information stored in the SGP database; if identical, the register is considered valid. This represents the ideal scenario where no additional intervention is needed. However, the decentralized nature of Brazil's educational system, combined with differences in data structure and update timeframes among educational departments, introduces substantial challenges. Initial checks already indicate an 11.97% inconsistency in student records, further emphasizing the need for robust validation methodologies. Thus, for those cases where only the date of birth coincides, but the full name and the mother's name present differences, we propose to apply similarity algorithms to assess the correspondence between the data.

4. Proposed Approach

Figure 2 provides a general overview of the proposed students' data validation approach. The process begins with data preprocessing, where names are normalized, abbreviations are standardized, and uniformity is ensured. This is followed by the data understanding phase, identifying patterns, inconsistencies, and duplicated or incomplete records. The output of these steps is fed into the similarity threshold evaluation, which is designed to learn the optimal threshold that ensures high validation accuracy, minimizes false positives, while still being tolerant to subtle differences in names that usually do not invalidate the identification of the students. The suggestion of a similarity threshold is possible after the conduction of two key experiments: (E1) Analyzes the impact of different similarity thresholds on the validation of records, considering both the student's name and the mother's name, to determine the optimal threshold (80% in our case study) for balancing accuracy and operational efficiency; and (E2) demonstrating the optimal threshold's ability to capture expected variations and to reduce matching errors.

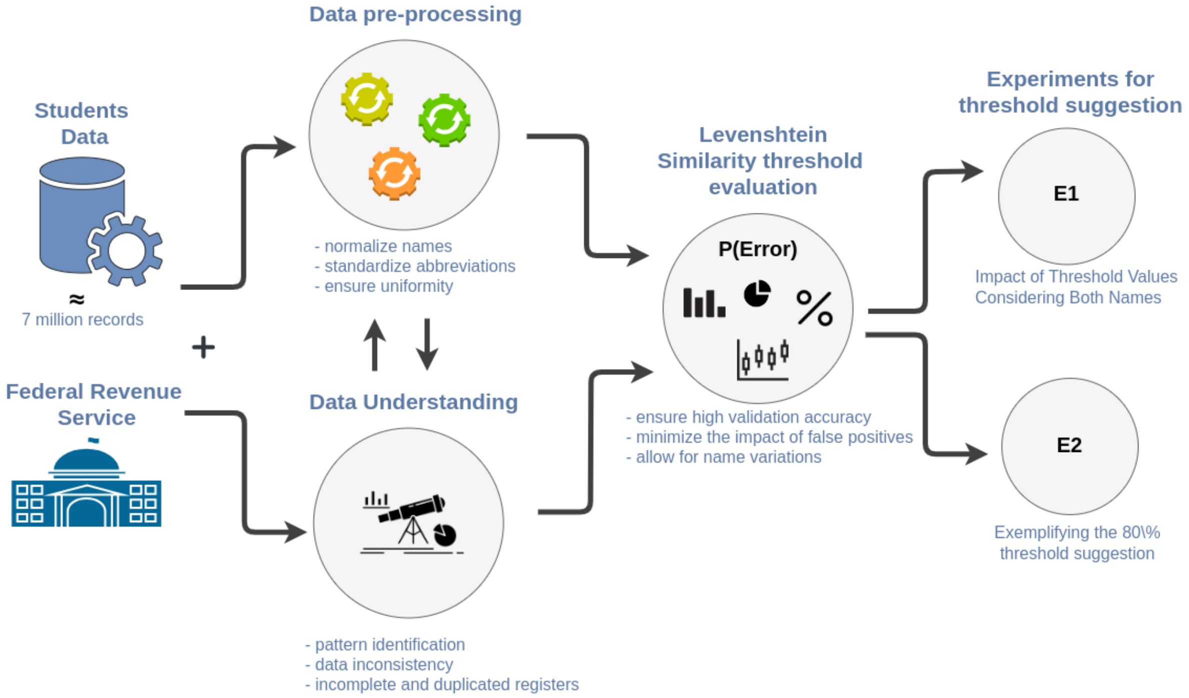
Data pre-processing
Students Data
≈ 7 million records
+
Federal Revenue Service
- normalize names
- standardize abbreviations
- ensure uniformity
Data Understanding
- pattern identification
- data inconsistency
- incomplete and duplicated registers
Levenshtein Similarity threshold evaluation
P(Error)
%
- ensure high validation accuracy
- minimize the impact of false positives
- allow for name variations
Experiments for threshold suggestion
E1
Impact of Threshold Values Considering Both Names
E2
Exemplifying the 80\% threshold suggestion

Fig. 2 – Approach followed for student's data validation.

Given two pairs of strings representing a student's name and their mother's name, extracted from two different databases (SGP and the Federal Revenue database), their similarity is measured separately using the Levenshtein distance algorithm. This helps quantify differences due to typos, formatting inconsistencies, or minor spelling variations, allowing for more accurate data reconciliation. If the data are completely identical, no alert is issued, and the student's registration is successfully validated. If the similarity rate equals or exceeds the minimum similarity threshold for both pairs, the record is classified as valid with an inconsistency alert, which is saved to enable future monitoring and auditing. If, on the other hand, the similarity rate of one of the pairs is below the minimum similarity threshold, the record is classified as invalid, and an alert is issued on the SGP platform, notifying the education network to review and correct the submitted data (see Figure 3). This alert promotes continuous database improvement, preventing inconsistent information from hindering the program's progress.

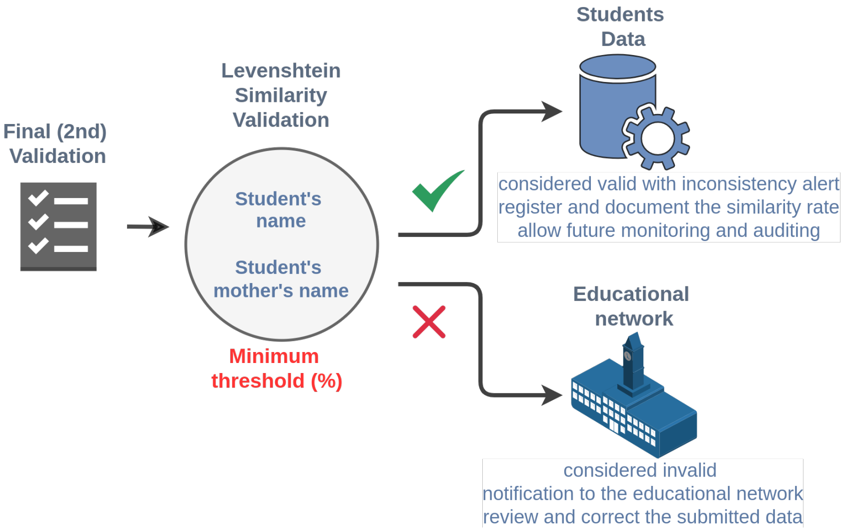
Final (2nd) Validation
Levenshtein Similarity Validation
Student's name
Student's mother's name
Minimum threshold (%)
Students Data
considered valid with inconsistency alert
register and document the similarity rate
allow future monitoring and auditing
Educational network
considered invalid
notification to the educational network
review and correct the submitted data

Fig. 3 – Using Levenshtein distance algorithm.

4.1. Data preprocessing and understanding

Data preprocessing was performed to ensure a standardized format for the data, minimizing the impact of irrelevant variations and increasing the precision of the similarity algorithm, as follows:

- Removal of accents and special characters to normalize the spelling of names.
- Standardization of abbreviations and removal of stopwords, such as "de," "da," and "do," which are common in Brazilian names and can interfere with similarity calculations.
- Elimination of extra spaces and conversion to lowercase, ensuring greater uniformity in the data.

Exploratory data analysis was conducted to understand structural features and underlying patterns existing in the data. The database, with approximately seven million registers of students from the high school level, presented a high variability in the names of the students and their mothers. The initial analysis focused on the frequencies of the most common students' names registered in the database. Around 5,73% of the students' full names correspond to records with a high probability of repetition; the top five most common names had 1,150; 685; 603; 582; and 405 occurrences, respectively. This high frequency of repetition in names highlights the possibility of ambiguities, making it difficult to distinguish unique records without using complementary variables, such as the date of birth and the mother's name.

In this sense, the probability of coincidences rate was estimated that approximately 7% of the students in the SGP database have a name that exactly matches another student's name. A similar percentage was found for mother's names. Consequently, the probability of a validation error when simultaneously considering a student's name and mother's name is estimated to be 0.49%.

A second challenge involves identifying students across different databases when their names exhibit slight spelling differences due to typographical errors, abbreviations, or orthographic variations. The analysis identified inconsistencies such as differences in accentuation (e.g., "María Eduarda" versus "Maria Eduarda"), standard abbreviations (e.g., "José Antônio" versus "José A."), and typos (e.g., "João Vítor" versus "Jão Vitor"). These variations complicate the process of reconciling records between databases and require robust similarity measures to ensure accurate data integration.

A correlation analysis was conducted among the variables available in the database to minimize false positives (acceptance of incorrect records) and false negatives (rejection of valid records) in validation. The date of birth stood out as a robust variable less prone to errors, making it an excellent candidate for use as a reference point in validation. The combination of names and dates of birth proved effective in distinguishing duplicate records or identifying inconsistencies, particularly in cases of prevalent names. The analysis revealed the following most common problems related to data quality:

- **Empty fields**: Approximately 2.8% of the records lacked complete information about the mother's name.
- **Formatting errors**: Presence of special characters and multiple spaces in some records.
- **Duplicate data**: Around 0.9% of the records showed evident duplications, such as two entries with the same CPF and similar names.

The exploratory analysis provided a comprehensive view of the database's challenges and characteristics. The results highlight the importance of using additional variables, such as the date of birth, to enhance name validation. Observations on name repetition and spelling variations were crucial for fine-tuning the similarity algorithm and adjusting the matching threshold to improve accuracy. The analysis also underscored the need for data cleaning and standardization techniques to improve database quality and facilitate interoperability with other governmental systems.

4.2. Levenshtein Distance Algorithm

The Levenshtein distance algorithm, also known as edit distance, is a widely used technique for measuring the similarity between two strings. It calculates the minimum number of operations needed to transform one string into another, which includes insertion (adding a character at any position), deletion (removing a character from any position), and substitution (changing one character to another at a specific position) (Gómez & Vázquez, 2022; Ouarda et al., 2023).

For example, to transform the string *"cat"* into the string *"rat"*, only one substitution is required (changing "c" to "r"). In this case, the Levenshtein distance is equal to 1. To transform *"par"* into *"parallel"*, five insertion operations are needed, resulting in a distance of 5.

The calculation of the Levenshtein distance is based on a dynamic programming model, where a two-dimensional matrix is constructed to represent the cumulative cost of transforming prefixes of one string into another. The steps include:

1. Inicialization of matrix D of size $(m+1) \times (n+1)$, where m and n are the lenghts of the strings A and B, respectively.
2. Filling the boundary conditions:

$$D[i][0] = i \quad \text{and} \quad D[0][j] = j$$

   representing the cost of transforming a string into an empty one.
3. Iteratively filling the matrix using the formula:

$$D[i][j] = \min \begin{cases} D[i-1][j] + 1 & \text{(deletion)} \\ D[i][j-1] + 1 & \text{(insertion)} \\ D[i-1][j-1] + c & \text{(substitution, where } c = 0 \text{ if } A[i] = B[j], \text{ or 1 otherwise)} \end{cases}$$

4. The final value $D[m][n]$ indicates the Levenshtein distance between A and B.

The interpretation of the distance can be normalized into a percentage similarity metric, called *Edit Similarity* (EDS):

$$\text{EDS} = \left(1 - \frac{\text{Levenshtein Distance}}{\text{Maximum possible length between the two strings}}\right) \times 100$$

5. Defining the optimal similarity threshold

This work proposes a series of experiments to establish the optimal similarity threshold necessary for validating student records. These tests were aimed to tackle the following tradeoff:

- a higher threshold diminishes the probability of erroneous instances being validated, but requires a larger amount of manual checks - possibly for valid students.
- a lower threshold would increase the probability of erroneous instances being validated, but requires a smaller amount of manual checks.

Thus, we propose to consider as input the capacity of the manual validation team, since excluded students are manually verified in a subsequent step. The optimal similarity threshold is defined as the larger value that results in an amount of manual checks that can be handled by the manual validation team.

This approach allows the identification of an effective balance, ensuring that the system accurately identifies and corrects mistakes in student records without overloading the process with unnecessary manual checks. In this section, we describe two experiments performed on the SGP and Federal Revenue databases to test the effect of different thresholds and better understand how variability in the threshold influences the detection of potentially invalid records. Each experiment addresses a specific aspect of the issue, as follows:

- **Experiment 1 (E1):** Analyzes the impact of different similarity thresholds on the validation of records, simultaneously considering both the student's name and the mother's name, to determine the optimal threshold (80% in our case study) for balancing accuracy and operational efficiency;
- **Experiment 2 (E2):** Examines simulated examples from the database to demonstrate if the optimal threshold found in our case study for both attribute captures expected name variations, minimizing false positives and false negatives.

5.1. Experiment 1: Impact of Threshold Values Considering Both Names Simultaneously

This experiment was designed to investigate how different threshold values influence the validation process, assessing their impact on the rejection or acceptance of records. The initial goal was to calibrate a threshold that balances the number of invalidated records and the probability of erroneous instances being validated, while accounting for the variability of both the student's name and the mother's name, two critical identification variables in the Brazilian context.

For each record of the SGP database, we use its CPF as the primary key to retrieve its potential match record from the Federal Revenue database. Then, we perform a hard validation by verifying whether its day of birth, full name and mother's name are identical, in which case the SGP record is considered valid. Otherwise, it is considered for experiment 1, composing a subset named $\mathcal{M} = \{m_{,1}, m_2, \ldots, m_n\}$ that is comprised of pairs of records $m_i = (s_i, r_i)$, where s_i is a record from SGP and r_i is a record from the Federal Revenue database.

We aim to assess how many records in M are rejected based on a soft-validation process with varying threshold values. A pair m_i is considered soft-validated if it meets the following conditions:

1. the day of birth of s_i and r_i matches;
2. the EDS score for the student names in s_i and r_i exceeds the threshold;
3. he EDS score for the student's mother's names in s_i and r_i exceeds the threshold.

The methodology involved testing thresholds ranging from 60% to 100% in units increments.

Figure 4 shows the number of records rejected per threshold. As expected, lower thresholds (e.g. 60%-70%) resulted in fewer rejections but increased the risk of false positives, where inconsistent records were incorrectly validated. However, higher thresholds (e.g., 90%-100%) significantly increased the number of rejections, potentially leading to false negatives, where valid records were improperly rejected. To address this, we propose calibrating the threshold based on the capacity of the manual validation team, considering that it can handle a maximum of 950,000 manual verifications. This led to the selection of 80% as the optimal threshold for the student's name.

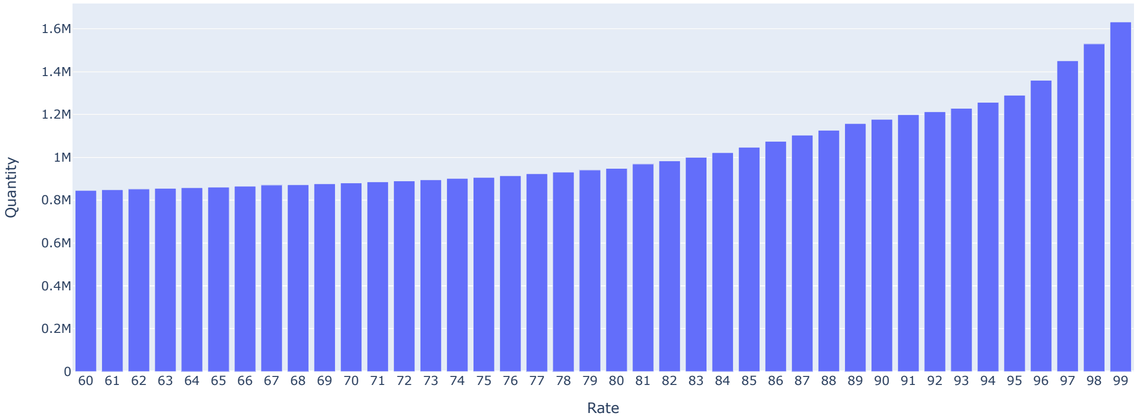

Fig. 4 – Amount of students that are invalidated for different thresholds.

5.2. Experiment 2: Exemplifying the 80% threshold suggestion

Experiment 2 was designed to illustrate, through practical examples, how the 80% threshold plays a crucial role in classifying records as valid or invalid. This experiment complements the quantitative analyses of the previous experiments by providing a qualitative and detailed view of how the model operates in real-world scenarios. The motivation for this experiment stems from the need to validate the suitability of the 80% threshold in real situations, highlighting representative cases where the choice of this value directly impacts the accuracy and scope of the validation process. Data such as the student's and mother's names are particularly sensitive to subtle variations, such as typographical errors, abbreviations, and marital status changes. Thus, understanding how the model handles these variations in real-world contexts is essential for refining its practical applicability.

To conduct this experiment, we present a simulated example and calculate the similarity between string pairs using the Levenshtein distance algorithm. Two cases were analyzed, representing scenarios of positive and negative validation, as described below:

- **Positive Example:** ”Maria Eduarda da Silva Rodrigues” and ”Maria Eduarda da Silva Rodigues” have a similarity of 96.43%, passing the 80% threshold. This high similarity is due to a single discrepancy: the absence of the letter ”r” in the second ”Rodrigues” last name. The Levenshtein algorithm identified this minimal difference as a substitution, resulting in a successful validation. This example highlights the robustness of the model in handling minor variations that do not compromise individual identification.
- **Negative Example:** ”Maria Eduarda da Silva Rodrigues” and ”Maria Eduarda Silva Antunes” have a similarity of 60.71%, falling below the 80% threshold and being rejected. The discrepancies include the absence of ”da” and the substitution of the last name ”Rodrigues” with ”Antunes,” totaling an edit distance of 11 characters. These significant differences justify the rejection of the record, demonstrating the model's sensitivity to variations that may compromise data integrity.

These practical examples reinforce that the 80% threshold provides an appropriate balance for capturing acceptable variations while rejecting more significant discrepancies that could compromise record integrity. Additionally, the examples highlight the effectiveness of the Levenshtein distance algorithm in identifying subtle patterns of difference, making it a suitable choice for applications in educational data validation.

6. Discussion

The results obtained in this study reinforce the importance of adequately calibrating the similarity threshold in educational data validation systems. The experiments demonstrated how an optimal threshold can be chosen quantitatively and qualitatively. The threshold value of 80% was capable of keeping the required number of manual validations below 950,000 while maintaining robustness in the challenging cases presented in Experiment 2. This finding is particularly relevant for large-scale systems such as the SGP, which operates with a database of approximately 7 million records.

Based on the results of Experiment 1, the 80% similarity threshold was determined as the optimal balance between precision and recall. This threshold minimizes false positives by rejecting records with significant discrepancies while tolerating minor variations such as typos, abbreviations, and orthographic inconsistencies.

The effectiveness of using similarity thresholds above 70% in string comparison has also been validated in previous studies (Gali et al., 2019; Libuy et al., 2021; Yu et al., 2016). For instance, the work by Valeriano, 2024 tested different thresholds for string matching and consistently found the best results with values above 70%. This aligns with our findings, where the 80% threshold provided a robust trade-off between validation accuracy and operational efficiency.

Compared to the work of Libuy et al., 2021, our study highlights the scalability of the proposed approach. Reducing the similarity threshold from 99% to 80% resulted in a reduction of over 40% in the number of rejected records without compromising data integrity. On the other hand, in relation to the scope of operations discussed in Cohen, 2000, it was necessary to adopt stricter criteria to ensure that minor name discrepancies did not result in the acceptance of invalid records. This decision reflects the specificities of the educational context, where imprecise data can significantly impact the distribution of government benefits.

The choice of an 80% threshold also had a profound impact on the workload of the teams responsible for data validation. With thresholds lower than 80%, the likelihood of false positives increased, requiring additional manual checks to ensure data consistency. Conversely, thresholds above 80% rejected a greater number of valid records, creating an excessive demand for manual review. By adopting the 80% threshold, the system optimized human resources, allowing teams to focus on more complex cases instead of addressing minor inconsistencies.

Practical examples reinforced the effectiveness of this threshold. For instance, a similarity score of 96.43% successfully validated minor discrepancies (e.g., “Rodrigues” vs. “Rodigues”), while a score of 60.71% appropriately rejected significant mismatches (e.g., “Rodrigues” vs. “Antunes”). These results demonstrate that the 80% threshold ensures reliable validation without overburdening the system with excessive manual checks.

Additionally, the results demonstrate that the Levenshtein distance algorithm is a practical solution for the analyzed context, offering simplicity and computational efficiency. While more advanced techniques, such as machine learning, could provide higher accuracy, their computational cost makes them impractical for a database the size of the SGP. This highlights the importance of solutions tailored to operational constraints and the need for scalability.

Although the presented results demonstrate the proposed approach’s effectiveness, some limitations must be highlighted. The reliance on auxiliary variables, such as the CPF and date of birth, ensures accuracy but creates dependency. The absence or inconsistency of these variables can introduce challenges, particularly in regions with less rigorous data collection practices. Future research could explore complementary variables, such as school history or address data, to enhance the robustness of the model. In addition, the optimal threshold found in this work may not be the optimal one in other scenarios. Thus, we recommend conducting similar experiments to calibrate the threshold according to the manual validation team capacity, and perform qualitative experiments to assure that the optimal threshold is robust.

Another relevant point in the process is the non-use of more advanced techniques based on artificial intelligence or machine learning and the motivations behind this choice. Although such techniques were initially considered, the Levenshtein distance algorithm was strategic due to its simplicity, computational efficiency, and scalability. However, it is essential to acknowledge that using artificial intelligence-based models could identify more complex patterns of inconsistency. Nevertheless, the computational cost generated by these techniques is a considerable barrier, especially in massive datasets like the SGP, which manages around 7 million records. Furthermore, implementing these solutions requires advanced technical infrastructure and specialized teams, which are not always available in educational systems with limited budgets.

At the same time, we do not rule out the use of machine learning techniques in the future, combined with traditional methods, such as dynamic adjustment of similarity thresholds, metric learning, name preprocessing, and automated alerts to system operators about potential inconsistencies. Thus, although this study focused on a low computational cost solution, there is considerable room for expanding and improving the developed practices, aiming for greater efficiency in the system and ensuring increasingly reliable data.

Finally, data interoperability proved to be a powerful tool for improving the quality of information within the context of the SGP. Integration with the Federal Revenue database enabled the identification of inconsistencies and error reduction, ensuring greater reliability for public policies such as the Pé-de-Meia Program. Future work can explore expanding this model to other government databases, increasing the system's scope and reliability.

7. Final Remarks

This study consolidated the importance of digital transformation and data interoperability in the context of digital governments, with an emphasis on large-scale educational systems such as the SGP. The adoption of the Levenshtein distance algorithm, combined with a similarity threshold of 80%, proved to be an effective solution for addressing discrepancies in records, ensuring the quality and integrity of processed data. By prioritizing scalable and accessible solutions, this work exemplifies how simple technologies can be adapted to tackle complex operational challenges, contributing to the modernization of the SGP and the implementation of more efficient public policies, such as the Pé-de-Meia program.

Furthermore, the interoperability between the SGP and the Federal Revenue database highlighted the potential of data integration to optimize processes and increase the reliability of governmental systems. This model, although developed for the Brazilian educational context, presents principles that can be adapted to other sectors and countries, especially those facing similar challenges of data heterogeneity and fragmentation.

For future work, it is proposed to expand interoperability with other governmental databases, such as the Cadastro Único and the National Basic Education Assessment System (SAEB). Additionally, adopting hybrid techniques that combine traditional methods with machine learning can open new opportunities to address even more complex contexts, preserving efficiency and scalability as priorities. Thus, this study not only meets a practical demand but also makes a significant contribution to the literature on data interoperability and educational management.

Acknowledgement

This work was supported by the Ministry of Education (MEC), Brazil, under Grant TED13914. The funders had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript. Additionally, the funders had no influence on the interpretation of data and the final conclusions drawn.

References

Ali, M. S., Ichihara, M. Y., Lopes, L. C., Barbosa, G. C., Pita, R., Carreiro, R. P., Dos Santos, D. B., Ramos, D., Bispo, N., Raynal, F., et al. (2019). Administrative data linkage in brazil: Potentials for health technology assessment. *Frontiers in pharmacology*, *10*, 984.

Almeida, D., Gorender, D., Ichihara, M. Y., Sena, S., Menezes, L., Barbosa, G. C., Fiaccone, R. L., Paixão, E. S., Pita, R., & Barreto, M. L. (2020). Examining the quality of record linkage process using nationwide brazilian administrative databases to build a large birth cohort. *BMC medical informatics and decision making*, *20*(1), 173.

Arretche, M. (2004). Federalismo e políticas sociais no brasil: Problemas de coordenação e autonomia. *São Paulo em perspectiva*, *18*(2), 17–26.

Asadollahi, H., Meouche, R. E., Zheng, Z., Eslahi, M., & Farazdaghi, E. (2024). Semantic text similarity in the civil engineer domain to enhance data interoperability: A domain-specific embedding approach. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5064591

Barbalho, I. M. P., Fernandes, F., Barros, D. M. S., Paiva, J. C., Henriques, J., Morais, A. H. F., Coutinho, K. D., Coelho Neto, G. C., Chioro, A., & Valentim, R. A. M. (2022). Electronic health records in brazil: Prospects and technological challenges. *Frontiers in Public Health*, *10*. DOI: https://doi.org/10.3389/fpubh.2022.963841.

Campmas, A., Iacob, N., & Simonelli, F. (2022). How can interoperability stimulate the use of digital public services? an analysis of national interoperability frameworks and e-government in the european union. *Data & Policy*, *4*, e19. DOI: https://doi.org/10.1017/dap.2022.11.

Cohen, W. W. (2000). Data integration using similarity joins and a word-based information representation language. *ACM Trans. Inf. Syst.*, *18*(3), 288–321. DOI: https://doi.org/10.1145/352595.352598.

Costin, C., & Pontual, T. (2020). Curriculum reform in brazil to develop skills for the twenty-first century. In F. M. Reimers (Ed.), *Audacious education purposes: How governments transform the goals of education systems* (pp. 47–64). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-41882-3_2.

Damasceno, C. D. N., Lobato, F. M. F., Moutinho, E. R., França, A. S. d., Oliveira, I. I. d., & Santana, Á. L. d. (2021). Simcleaner – sistema de padronização de bases de dados utilizando funções de similaridade. *arXiv preprint arXiv:2107.12884*. https://arxiv.org/abs/2107.12884

Das, M., Tao, X., Liu, Y., & Cheng, J. C. (2022). A blockchain-based integrated document management framework for construction applications. *Automation in Construction*, *133*, 104001.

Downs, J. M., Ford, T., Stewart, R., Epstein, S., Shetty, H., Little, R., Jewell, A., Broadbent, M., Deighton, J., Mostafa, T., Gilbert, R., Hotopf, M., & Hayes, R. (2019). An approach to linking education, social care and electronic health records for children and young people in south london: A linkage study of child and adolescent mental health service data. *BMJ Open*, *9*(1). DOI: https://doi.org/10.1136/bmjopen-2018-024355.

Fiszbein, A., & Schady, N. R. (2009). *Conditional cash transfers: Reducing present and future poverty*. World Bank Publications.

Gali, N., Mariescu-Istodor, R., Hostettler, D., & Fränti, P. (2019). Framework for syntactic string similarity measures. *Expert Systems with Applications*, *129*, 169–185.

Gómez, J., & Vázquez, P.-P. (2022). An empirical evaluation of document embeddings and similarity metrics for scientific articles. *Applied Sciences*, *12*(11), 5664.

Handijono, A., & Suhatman, Z. (2024). Meningkatkan deduplikasi data melalui kesamaan teks dalam pembelajaran mesin: Pendekatan komprehensif. *AKADEMIK: Jurnal Mahasiswa Humanis*, *4*(2), 602–615.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big data & society*, *4*(2), 2053951717745678.

Kaufman, A. R., & Klevs, A. (2022). Adaptive fuzzy string matching: How to merge datasets with only one (messy) identifying field. *Political Analysis*, *30*(4), 590–596.

Kouremenou, E., Kiourtis, A., & Kyriazis, D. (2024). A data modeling process for achieving interoperability. *Advances in Digital Health and Medical Bioengineering*, 711–719. https://link.springer.com/chapter/10.1007/978-3-031-62502-2_80

Kremer, M. (2003). Randomized evaluations of educational programs in developing countries: Some lessons. *American Economic Review*, *93*(2), 102–106.

Libuy, N., Harron, K., Gilbert, R., Caulton, R., Cameron, E., & Blackburn, R. (2021). Linking education and hospital data in england: Linkage process and quality. *International Journal of Population Data Science*, *6*(1). DOI: https://doi.org/10.23889/ijpds.v6i1.1671.

Loureiro, A., Cruz, L., & Mello, U. (2021). Brazil case study. *The Role of Intergovernmental Fiscal Transfers in Improving Education Outcomes*, *201*, 201–233.

Malodia, S., Dhir, A., Mishra, M., & Bhatti, Z. A. (2021). Future of e-government: An integrated conceptual framework. *Technological Forecasting and Social Change*, *173*, 121102.

McBride, K., Kamalanathan, S., Valdma, S.-M., Toomere, T., & Freudenthal, M. (2022). Digital government interoperability and data exchange platforms: Insights from a twenty country comparative study. *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*, 90–97. DOI: https://doi.org/10.1145/3560107.3560123.

Ministério da Educação. (2025a). ENEM: Sua porta de entrada para a educação superior [Acessado em: 20 jan. 2025]. https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

Ministério da Educação. (2025b). Pé-de-Meia: Ministério da Educação Brasileiro [Acessado em: 19 jan. 2025]. https://www.gov.br/mec/pt-br/pe-de-meia

Ouarda, L., Malika, B., & Brahim, B. (2023). Towards a better similarity algorithm for host-based intrusion detection system. *Journal of Intelligent Systems*, *32*(1), 20220259.

Parker, S. W., & Todd, P. E. (2017). Conditional cash transfers: The case of progresa/oportunidades. *Journal of Economic Literature*, *55*(3), 866–915.

Queiroga, E. M., Santana, D., da Silva, M., de Aguiar, M., dos Santos, V., Mello, R. F., Bittencourt, I. I., & Cechinel, C. (2024). Anticipating student abandonment and failure: Predictive models in high school settings. *International Conference on Artificial Intelligence in Education*, 351–364.

Queiroga, E. M., Siqueira, E. S., dos Santos Portela, C., Cordeiro, T. D., Bittencourt, I. I., Isotani, S., Melo, R. F., Muñoz, R., & Cechinel, C. (2024). Data-driven strategies for achieving school equity: Insights from brazil and policy recommendations. *IEEE Access*.

Queiroz, M. V. A. B., Sampaio, R. M. B., & Sampaio, L. M. B. (2020). Dynamic efficiency of primary education in brazil: Socioeconomic and infrastructure influence on school performance. *Socio-Economic Planning Sciences*, *70*, 100738. DOI: https://doi.org/https://doi.org/10.1016/j.seps.2019.100738.

Rocha, J. C., Ramos, V., Cechinel, C., Hernández-Leal, E. J., Muñoz, R., & Primo, T. T. (2024). Data interoperability in learning analytics - review of literature, 1–8. DOI: https://doi.org/10.1109/clei64178.2024.10700464.

Saffady, W. (2021). *Records and information management: Fundamentals of professional practice*. Rowman & Littlefield.

Sakai, K., Dong, Y., Oyamada, M., Takeoka, K., & Okadome, T. (2021). Entity matching with string transformation and similarity-based features. *International Workshop on Software Foundations for Data Interoperability*, 76–87.

Segatto, C. I., Santos, F. B. P. d., Bichir, R. M., & Morandi, E. L. (2022). Inequalities and the covid-19 pandemic in brazil: Analyzing un-coordinated responses in social assistance and education. *Policy and Society*, *41*(2), 306–320. DOI: https://doi.org/10.1093/polsoc/puac005.

Styrin, E., Mossberger, K., & Zhulin, A. (2022). Government as a platform: Intergovernmental participation for public services in the russian federation. *Government Information Quarterly*, *39*(1), 101627.

Tavares, A. A., & Bitencourt, C. M. (2024). Evaluation of public policies and interoperability from the perspective of digital public governance. In D. K. Kang & D. L. Li (Eds.), *E-government digital frontiers - transforming public administration through technology*. IntechOpen. DOI: https://doi.org/10.5772/intechopen.1007596.

Valeriano, E. S. (2024). Deduplication methods using levenshtein distance algorithm. *Journal of Electrical Systems*, *20*(7s), 997–1006.

Wanke, P., Lauro, A., dos Santos Figueiredo, O. H., Faria, J. R., & Franklin G. Mixon, J. (2024). The impact of school infrastructure and teachers' human capital on academic performance in brazil [PMID: 37610037]. *Evaluation Review*, *48*(4), 636–662. DOI: https://doi.org/10.1177/0193841X231197741.

Wimmer, M. A., Boneva, R., & di Giacomo, D. (2018). Interoperability governance: A definition and insights from case studies in europe. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. DOI: https://doi.org/10.1145/3209281.3209306.

Yu, M., Li, G., Deng, D., & Feng, J. (2016). String similarity search and join: A survey. *Frontiers of Computer Science*, *10*, 399–417.