Monitoramento Escalável da Qualidade Cadastral via Similaridade Nominal no Sistema Gestão Presente: Uma Abordagem Analítica com Painel Interativo

Tiago Paulino[1], Leonardo Marques[1], Diego Matos[1], Elthon Oliveira[1], Flavia Galvani[4], Emanuel Marques Queiroga[1,2], Cristian Cechinel[1,3], Anita Gea Martinez Stefani[5], Thales Vieira[1]

[1] Núcleo de Excelência em Tecnologias Sociais
Universidade Federal de Alagoas (NEES–UFAL)

[2]Instituto Federal de Educação, Ciência e Tecnologia Sul-Rio-Grandense (IFSul)

[3]Centro de Ciências, Tecnologias e Saúde Universidade Federal de Santa Catarina (UFSC)

[4]Blavatnik School of Government, University of Oxford

[5]Ministério da Educação (MEC)

tiagopaulino1989@gmail.com, emanuelmqueiroga@gmail.com,
leonardo.marques@nees.ufal.br,diego.matos@nees.ufal.br,
thales.vieira@nees.ufal.br,elthon.oliveira@nees.ufal.br,
cristian.cechinel@ufsc.br,flavia.galvani@bsg.ox.ac.uk,
anitastefani@mec.gov.br

Abstract. *The study evaluates the performance of data validation in the Present Management System (SGP), focusing on record authentication through name similarity, using 2025 school enrollment data. The analysis identified 98.9% of automatic authentications, highlighting the algorithm's effectiveness in correcting spelling discrepancies. The work proposes an interactive dashboard to monitor indicators, configure parameters such as similarity threshold, and analyze results by school, network, and region. The solution aims to improve data quality, support strategic decision-making, and strengthen data-driven public policies.*

Resumo. *O estudo avalia o desempenho da validação de dados no Sistema Gestão Presente (SGP), com foco na autenticação de registros por similaridade nominal, usando dados de matrículas escolares de 2025. A análise identificou 98,9% de autenticações automáticas, evidenciando a eficácia do algoritmo na correção de divergências de grafia. O trabalho propõe um painel interativo para monitorar indicadores, configurar parâmetros, como limiar de similaridade, e analisar resultados por escola, rede e região. A solução visa aprimorar a qualidade cadastral, apoiar decisões estratégicas e fortalecer políticas públicas baseadas em dados.*

1. Introdução

A gestão eficiente de dados educacionais é um dos pilares fundamentais para a formulação e implementação de políticas públicas que promovam equidade [Queiroga et al. 2024], permanência e qualidade no ensino. No Brasil, a diversidade entre redes de ensino, sistemas de registro e contextos regionais impõe desafios significativos à consolidação e padronização das informações de matrícula e frequência escolar em âmbito nacional.

DOI: 10.5753/wapla.2025.16231

Nesse contexto, o Sistema Gestão Presente[1] (SGP) é a plataforma federada do MEC que integra os registros de matrícula e frequência das redes públicas de ensino, produzindo a base de referência para monitoramento da assiduidade dos estudantes. Esses dados são utilizados para verificar condicionalidades de políticas atreladas à presença escolar, como o Programa Pé-de-Meia[2], programa federal de transferência direta de recursos a estudantes do ensino médio. Ao consolidar e padronizar a informação de presença, o SGP viabiliza a apuração de elegibilidade e a correta identificação dos beneficiários.

Para garantir a qualidade dos dados da base, a autenticação confiável dos alunos registrados é um requisito essencial [Tavares and Bitencourt 2024] [Valeriano 2024]. A vinculação de estudantes a benefícios sociais, abertura de contas bancárias e rastreamento da frequência dependem de chaves consistentes, como CPF ou código INEP, e da conciliação dessas informações entre as bases governamentais e a instituições pagadoras. Contudo, variações na grafia dos nomes, resultantes de erros de digitação, abreviações ou inconsistências cadastrais, dificultam a validação baseada em correspondência exata.

Este trabalho propõe o desenvolvimento de um painel de monitoramento voltado à avaliação da funcionalidade de validação por similaridade no SGP, apresentando indicadores para mensurar seu desempenho na reconciliação de registros. O objetivo central é avaliar o impacto da funcionalidade na melhoria da qualidade dos dados e na efetividade das políticas públicas que deles dependem.

2. Análise do Fluxo de Validação de dados

A compreensão do fluxo de dados no Sistema Gestão Presente (SGP) é essencial para avaliar a confiabilidade das informações utilizadas em políticas sociais condicionadas à frequência escolar. Após o envio dos dados, oriundos de planilhas ou por envios via API, o SGP realiza o armazenamento dos registros em seu banco de dados. A partir daí, inicia-se o processo de verificação junto à Receita Federal do Brasil (RFB), utilizando os dados de CPF, nome completo, data de nascimento e nome da mãe.

Quando todos os campos coincidem com os registros da RFB, o CPF é automaticamente autenticado e o cadastro liberado. No entanto, em casos de divergência apenas no campo do nome do estudante, o registro é submetido a um processo de validação por similaridade. Esse processo é dividido em duas etapas: primeiro, realiza-se um pré-processamento para remover caracteres inválidos, palavras irrelevantes e espaçamentos inconsistentes. Em seguida, é aplicada a análise de Similaridade de Levenshtein [Silva et al. 2025] para mensurar o grau de similaridade entre o nome informado e o nome constante na base oficial da RFB.

A partir de análises empíricas conduzidas durante a implementação [Silva et al. 2025], foi definido um limiar mínimo de similaridade de 80%, a partir do qual a correspondência é considerada válida, mesmo com grafias ligeiramente diferentes. Essa abordagem permite conciliar variações comuns em grandes bases de dados, sem comprometer a integridade das informações.

[1] https://www.gov.br/mec/pt-br/mec-gestao-presente
[2] https://www.gov.br/mec/pt-br/pe-de-meia

2.1. Análise Exploratória dos Dados

Para avaliar o desempenho do sistema de autenticação por similaridade nominal, foi realizada uma extração estruturada do banco de dados do SGP, contemplando apenas matrículas válidas do ano letivo de 2025, com abrangência nacional. Registros cancelados, duplicados ou provenientes de ambientes de teste foram excluídos, assegurando a integridade e a representatividade da amostra.

A Tabela 1 apresenta a estrutura da base utilizada, com destaque para os principais campos relacionados ao processo de validação e autenticação dos registros, incluindo informações institucionais, dados pessoais e indicadores cruzados com a base da RFB.

Tabela 1. Descrição dos campos extraídos do banco de dados do SGP

Coluna	Tipo de Dado	Descrição
co-escola	Int64	Código da escola
co-matricula	Int64	Código da matrícula
cpf	String	Número de Cpf do Aluno (Parcial)
created-at	DateTime	Data da validacao
st-cadastro-validado	Int64	Status geral da validação
fl-dt-nascimento-validado	Boolean	Flag de validação da data de nascimento
fl-no-pessoa-validado	Boolean	Flag de validação do nome do aluno
fl-no-mae-validado	Boolean	Flag de validação do nome da mãe do aluno
vl-rate-no-pessoa-validado	Int64	Score de similaridade do nome do aluno
vl-rate-no-mae-validado	Int64	Score de similaridade do nome da mãe do aluno

A base consolidada reuniu 9.058.447 estudantes com matrículas registradas em 2025. Dentre esses, 98,9% apresentaram status de autenticação automática concluída com sucesso. Os demais 1,1% referem-se a casos em que a autenticação não pôde ser realizada de forma automática, geralmente em razão de inconsistências de dados cadastrais ou da ausência de uma correspondência confiável.

Esses resultados iniciais demonstram a robustez do processo de validação, com alto número de autenticações bem-sucedidas. No entanto, para compreender plenamente o impacto da funcionalidade sobre as políticas públicas, é necessário aprofundar a análise, considerando, por exemplo, aspectos regionais e operacionais das validações. Essa análise será conduzida a partir dos indicadores propostos no painel de monitoramento, conforme detalhado nas seções seguintes.

3. Metodologia

3.1. Infraestrutura de Integração e Atualização de Dados

Para viabilizar a análise contínua e massiva dos dados de autenticação, propõe-se a estruturação de um processo de replicação quase em tempo real entre o banco de dados operacional do Sistema de Gestão Presente (SGP) e um ambiente analítico dedicado. A arquitetura sugerida tem como objetivo garantir que as consultas analíticas e a atualização periódica dos painéis de monitoramento possam ser realizadas sem comprometer o desempenho do sistema transacional principal.

Uma alternativa para replicação dos dados é por meio de uma estratégia de captura de alterações (*Change Data Capture* – CDC), que permita o registro de inserções, atualizações e exclusões em tempo real no banco de origem [Seenivasan and Vaithianathan 2023]. Esses eventos são transferidos de forma incremental para um banco de dados analítico, mantendo-o sincronizado com baixa latência. Essa abordagem de integração desacoplada favorece a escalabilidade do sistema e evita sobrecarga sobre o ambiente de produção do SGP.

Uma vez replicados, os dados passam por um processo de transformação, a fim de consolidar, padronizar e preparar os registros para análise. O dataset resultante é estruturado de forma a otimizar o desempenho das consultas analíticas e alimentar continuamente um painel interativo com indicadores relevantes sobre o processo de autenticação, permitindo também o monitoramento de padrões de inconsistência e indicadores de qualidade. Com isso, gestores e analistas técnicos podem dispor de indicadores para orientar decisões no âmbito da gestão do SGP.

3.2. Desenvolvimento do Painel de Monitoramento

Com o objetivo de apoiar o monitoramento contínuo da qualidade dos dados cadastrais e da efetividade de algoritmos de similaridade na autenticação de estudantes, propõe-se o desenvolvimento de um painel interativo de indicadores. O painel, em seu desenho conceitual, visa proporcionar uma visualização clara e objetiva dos resultados processados pelo sistema de validação do SGP, oferecendo recursos fundamentais para a tomada de decisão técnica e estratégica por parte de analistas e gestores educacionais.

Para o painel, foi planejada uma divisão em sessões. A primeira apresenta uma visão geral da base de dados, destacando o total de estudantes processados, os números absolutos e relativos de validações realizadas com sucesso e os registros não validados. Como exemplificado na Figura 1, a simulação com dados da análise exploratória ilustra que, de um total de 9.058.447 registros, 98,9% foram validados e 1,1% não atenderam aos critérios estabelecidos. Essa abordagem inicial proporciona uma métrica clara da abrangência e da eficiência do sistema.

Ainda nesta seção, é possível observar a diferença entre validações realizadas por correspondência exata dos nomes e aquelas dependentes do algoritmo de similaridade. Na Figura 1, uma avaliação indica que cerca de 13,03% dos registros dependem do uso do algoritmo para serem validados, considerando um limiar de similaridade de 80%, evidenciando sua relevância na automatização da autenticação de dados cadastrais. A modulação desses parâmetros, aliada à sua visualização, permite aos gestores técnicos [Azzone 2018] [Suominen and Hajikhani 2021] uma análise mais refinada do processo de autenticação e a retroalimentação da informação para melhoria do mesmo.

Entre os recursos inovadores sugeridos para o painel, destaca-se a inclusão de filtros interativos, que permitem a parametrização do processo de análise, como a definição de um limiar de similaridade e a possibilidade de considerar ou não o nome da mãe como critério de validação. Estudos prévios [Gali et al. 2019], [Libuy et al. 2021], [Yu et al. 2016] reforçam a importância de tais parâmetros, especialmente para melhorar a assertividade da vinculação entre registros nominais em bases extensas e heterogêneas.

Figura 1. Visão geral da validação dos dados e uso da similaridade

A proposta também inclui um componente visual que exibe uma matriz de erros por tipo de campo (Figura 2), voltada ao diagnóstico de falhas nos processos de validação. A funcionalidade permitiria identificar quais atributos, como nome do estudante, nome da mãe ou data de nascimento, foram responsáveis pelas reprovações, contribuindo para ajustes nos padrões de coleta ou de pré-processamento dos dados. A título de exemplo, em uma simulação, os maiores índices de inconsistência estariam associados ao campo ”nome da mãe”, possivelmente em razão de variações de grafia ou abreviações.

Figura 2. Matriz de erros de validação por tipo de campo

Para ampliar a capacidade de monitoramento em nível institucional, o painel traz uma seção voltada à análise desagregada por escola (Figura 3). Nessa visualização, seriam exibidas as validações totais por unidade, o número de registros que exigiram o uso da similaridade textual e os respectivos percentuais de efetividade. Com filtros por UF, rede e unidade escolar, essa funcionalidade permitiria análises comparativas regionais e a identificação de padrões de uso e dependência do algoritmo, retroalimentando políticas de correção ou capacitação localizadas.

Outro componente essencial desta proposição é a análise da distribuição acumulada de validações por faixa de similaridade (Figura 4), que serviria de base para o ajuste fino do algoritmo. A curva de distribuição permitiria identificar faixas críticas, como validações muito próximas ao limiar, e orientar decisões sobre a política de aceitação de registros no fluxo da validação por similaridade, com vistas a maximizar a eficiência.

Dep. Administrativa	Escola	UF	Validacoes	Validadoes p/ Similaridade	Efetividade (%)
Estadual	[redacted]	[redacted]	1,392	194	13.94
Estadual	[redacted]	[redacted]	1,252	180	14.38
Estadual	[redacted]	[redacted]	1,000	176	17.60
Estadual	[redacted]	[redacted]	1,293	213	16.47
Estadual	[redacted]	[redacted]	732	352	48.09
Total			**28,119**	**3,817**	**13.57**

Figura 3. Validações por similaridade em nível de escola, com filtros por UF, rede e unidade

A Figura 5 representa uma seção do painel voltada ao monitoramento analítico individualizado. Nela, propõe-se a inclusão de filtros por nome, CPF, UF, rede e escola, além de uma tabela contendo dados anonimizados, com status de validação codificados por cores. Também são apresentados indicadores agregados, como total de estudantes validados e distribuição percentual de alertas amarelos e vermelhos, além de gráficos horizontais que representam a distribuição desses alertas por município, permitindo cruzamentos entre a granularidade e o contexto territorial.

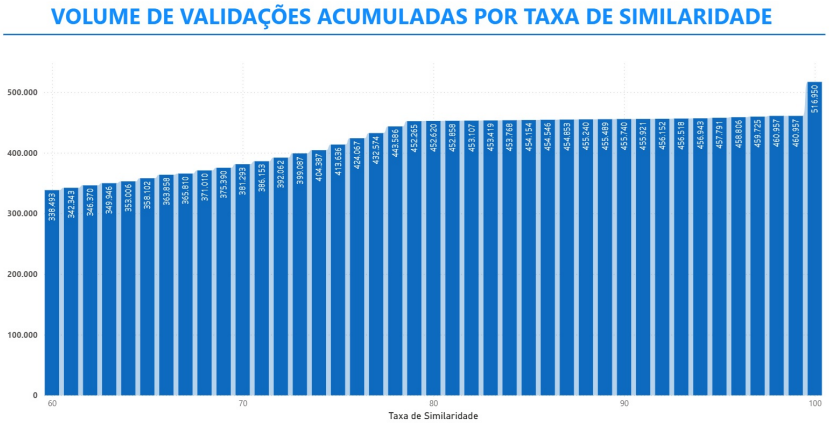

Figura 4. Distribuição de validações acumuladas por taxa de similaridade

A Figura 5 ilustra a seção do painel dedicada à análise das validações, com foco no rastreamento individualizado de registros. A interface contempla filtros dinâmicos por nome, CPF, unidade federativa (UF), rede de ensino e escola, permitindo a segmentação precisa da base de dados conforme critérios operacionais. A tabela principal apresenta informações parcialmente anonimizadas dos estudantes, incluindo nome, CPF, instituição de ensino, data da validação e status de alerta codificado por cores (verde, amarelo e

vermelho). Essa estrutura possibilita a identificação rápida de registros inconsistentes e fornece subsídios para auditorias técnicas e intervenções pontuais nos dados cadastrais.

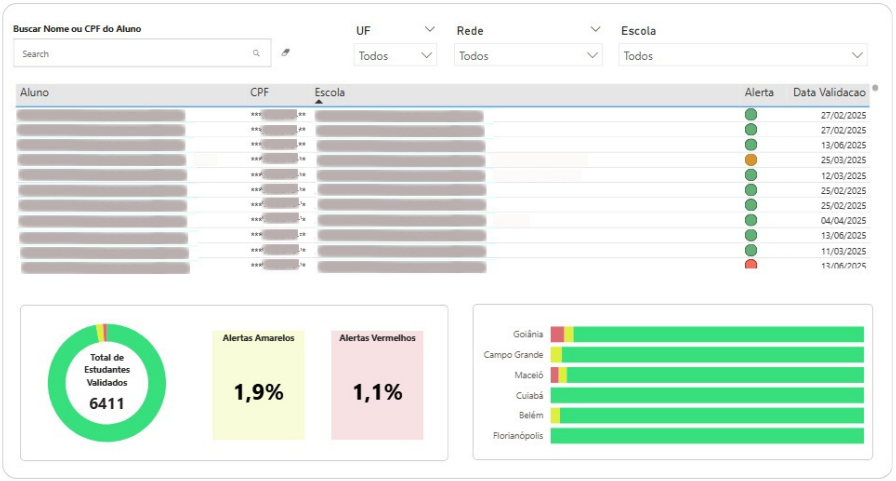

Figura 5. Tabela analítica das validações com alertas, por Estudante, Escola, Rede e UF

Na parte inferior do painel, são disponibilizados indicadores agregados do desempenho geral da validação. Um gráfico de rosca exibe o total de estudantes validados (6.411), com destaque para a distribuição percentual dos alertas amarelos (1,9%) e vermelhos (1,1%). Em paralelo, visualizações em barras horizontais representam a incidência de alertas por município, permitindo análises comparativas entre contextos regionais. Essa abordagem combina granularidade e visão agregada, sendo especialmente útil para diagnósticos de qualidade de dados, priorização das correções junto às redes, alinhando-se às boas práticas de governança e monitoramento de dados educacionais.

A proposta busca aliar inteligência computacional, transparência analítica e flexibilidade de parametrização, visando não apenas fortalecer a governança de dados educacionais, mas também permitir ações preventivas e corretivas mais eficazes. Etapas futuras deste projeto podem envolver testes em ambientes controlados e ajustes de usabilidade.

4. Considerações Finais

O painel de indicadores proposto neste trabalho configura-se como uma solução estratégica para o monitoramento contínuo da qualidade dos dados cadastrais de estudantes e para a avaliação da efetividade de mecanismos automatizados de autenticação baseados em similaridade textual. A concepção da ferramenta prevê a integração de recursos de visualização interativa, filtros parametrizáveis e métricas operacionais próximas de tempo real, com o objetivo de viabilizar diagnósticos granulares e orientar tecnicamente as ações de gestores educacionais [Azzone 2018].

A possibilidade de configurar variáveis críticas, como o limiar de similaridade e o uso de parâmetro para simular a validação do nome da mãe, ampliaria significativamente o potencial analítico da ferramenta, permitindo a simulação de diferentes cenários, projetando decisões mais precisas na escolha de um limiar de similaridade. Além disso, a

proposta inclui a segmentação dos dados por rede, escola, UF e faixa de similaridade, o que facilitaria reconhecer padrões regionais e aprimorar práticas de registro.

Funcionalidades previstas, como a matriz de erros por campo e a análise da efetividade do algoritmo por unidade escolar, seriam elementos centrais para promover transparência, rastreabilidade e fomentar uma cultura de melhoria contínua na gestão da informação educacional [Suominen and Hajikhani 2021]. Em contextos onde a confiabilidade dos dados é essencial para subsidiar políticas públicas educacionais e a alocação dos recursos, a proposta apresentada neste texto mostra elementos fundamentais para elevar o grau de maturidade analítica das instituições.

Num próximo passo, sugerimos a implementação do protótipo da ferramenta e sua validação em ambientes de teste com bases reais, além da expansão do modelo para outros sistemas de informação educacional. Recomenda-se também a integração futura com mecanismos de retroalimentação automatizada, como a correção assistida de registros inconsistentes e notificações de erro em contextos de envio massivo de dados com baixa taxa de validação, potencializando a eficiência do processo de autenticação do SGP.

5. Agradecimentos

Este trabalho foi apoiado pelo Ministério da Educação (MEC), Brasil, sob a concessão TED11476.

Referências

Azzone, G. (2018). Big data and public policies: Opportunities and challenges. *Statistics & Probability Letters*, 136:116–120.

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

Libuy, N., Harron, K., Gilbert, R., Caulton, R., Cameron, E., and Blackburn, R. (2021). Ligação de dados de educação e hospitalares em inglaterra: Processo de ligação e qualidade. *Jornal Internacional de Ciência de Dados Populacionais*, 6(1).

Queiroga, E. M., Siqueira, E. S., Portela, C. D. S., Cordeiro, T. D., Bittencourt, I. I., Isotani, S., Mello, R. F., Muñoz, R., and Cechinel, C. (2024). Data-driven strategies for achieving school equity: Insights from brazil and policy recommendations. *IEEE Access*, 12:101646–101659.

Seenivasan, D. and Vaithianathan, M. (2023). Real-time adaptation: Change data capture in modern computer architecture. *ESP International Journal of Advancements in Computational Technology (ESP-IJACT)*, 1(2):49–61.

Silva, D. B. L., Queiroga, E. M., Barros, A. N., Marcolino, M. R., Dermeval, D., Lima, A., Marques, L. B., Cechinel, C., and Vieira, T. (2025). Leveraging string similarity algorithms for educational data validation: A scalable approach for digital governance. In *Conference on Digital Government Research*, volume 1.

Suominen, A. and Hajikhani, A. (2021). Research themes in big data analytics for policymaking: Insights from a mixed-methods systematic literature review. *Policy & Internet*, 13(4):464–484.

Tavares, A. A. and Bitencourt, C. M. (2024). Evaluation of public policies and interoperability from the perspective of digital public governance. In *E-Government Digital Frontiers-Transforming Public Administration through Technology*. IntechOpen.

Valeriano, E. S. (2024). Deduplication methods using levenshtein distance algorithm. *Journal of Electrical Systems*, 20(73):997–1006.

Yu, M., Li, G., Deng, D., and Feng, J. (2016). String similarity search and join: a survey. *Frontiers of Computer Science*, 10(3):399–417.