



UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA APLICADA

ABÍLIO NOGUEIRA BARROS

**APLICAÇÃO DE MINERAÇÃO DE DADOS
EDUCACIONAIS PARA AVALIAÇÃO E EXPLICAÇÃO DE
FATORES ASSOCIADOS À DISTORÇÃO IDADE-SÉRIE**

RECIFE – PE

2025

ABÍLIO NOGUEIRA BARROS

**APLICAÇÃO DE MINERAÇÃO DE DADOS
EDUCACIONAIS PARA AVALIAÇÃO E EXPLICAÇÃO DE
FATORES ASSOCIADOS À DISTORÇÃO IDADE-SÉRIE**

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre em Informática Aplicada.

ORIENTADOR: Prof. Dr. Rafael Ferreira Leite de Mello

RECIFE – PE

2025

Dados Internacionais de Catalogação na Publicação
Sistema Integrado de Bibliotecas da UFRPE
Bibliotecário(a): Auxiliadora Cunha – CRB-4 1134

B277a Barros, Abílio Nogueira.

Aplicação de mineração de dados educacionais para avaliação e explicação de fatores associados à distorção idade-série / Abílio Nogueira Barros. – Recife, 2025.
91 f.; il.

Orientador(a): Rafael Ferreira Leite de Mello.

Dissertação (Mestrado) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, BR-PE, 2025.

Inclui referências e apêndice(s).

1. Distorção idade - Série. 2. Mineração de dados educacionais. 3. Inteligência artificial - Aplicações educacionais. 4. Aprendizagem 5. Política Pública - Educação. I. Mello, Rafael Ferreira Leite de, orient. II. Título

CDD 004

ABÍLIO NOGUEIRA BARROS

**APLICAÇÃO DE MINERAÇÃO DE DADOS
EDUCACIONAIS PARA AVALIAÇÃO E EXPLICAÇÃO DE
FATORES ASSOCIADOS À DISTORÇÃO IDADE-SÉRIE**

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre em Informática Aplicada.

Aprovada em: 28 de Fevereiro de 2025.

BANCA EXAMINADORA

Prof. Dr. Rafael Ferreira Leite de Mello (Orientador)
Universidade Federal Rural de Pernambuco - UFRPE
Departamento de Computação

Prof. Dr. Gabriel Alves de Albuquerque Junior
Universidade Federal Rural de Pernambuco - UFRPE

Prof. Dr. Emanuel Marques Queirogal
Instituto Federal de Educação, Ciência e Tecnologia do Sul
Rio-Grandense - IFSul

Dedico esta dissertação aos meus pais, Maria Aparecida Nogueira Barros e José Erigerson Negromonte de Barros, por me ensinarem que a educação é o maior legado que se pode carregar na vida. Este trabalho também é dedicado aos meus familiares e amigos próximos, que sempre estiveram ao meu lado, celebrando cada conquista e me apoiando em todos os momentos

Agradecimentos

Primeiramente, gostaria de agradecer a Deus por me conceder saúde, força e resiliência para superar os desafios e cumprir mais esta etapa da minha vida acadêmica.

Ao Professor Doutor Rafael Ferreira Mello, expresso minha profunda gratidão por sua orientação não apenas no desenvolvimento desta dissertação, mas ao longo de toda a minha trajetória acadêmica. Sou imensamente grato por todas as oportunidades proporcionadas durante essa jornada.

Aos professores Doutora Andrêza Leite de Alencar e Doutor Gabriel Alves de Albuquerque Junior, agradeço pela valiosa contribuição ao meu desenvolvimento como pesquisador, através de nossas colaborações e apoio contínuos.

Na pessoa da Professora Doutora Taciana Pontual Falcão, atual coordenadora do Programa de Pós-Graduação em Informática Aplicada da Universidade Federal Rural de Pernambuco, estendo meus agradecimentos a todos os membros do corpo docente, equipe administrativa e profissionais de serviços gerais, pelo apoio, conhecimento e suporte que enriqueceram minha experiência acadêmica ao longo do mestrado.

Aos meus pais, Maria Aparecida Nogueira Barros e José Erigerson Negromonte de Barros, sou eternamente grato por terem me ensinado, desde cedo, que a educação é o melhor caminho a ser seguido na vida. Agradeço por todo suporte pessoal, emocional e financeiro que me ofereceram durante esta jornada. Estendo minha gratidão também a todos os familiares próximos, que sempre estiveram ao meu lado apoiando e celebrando a cada conquista.

Por fim, agradeço aos meus amigos, tanto os de longa data quanto os que tive o privilégio de conhecer durante esta caminhada, por compartilharem momentos de aprendizado, motivação e alegria.

*“A alfabetização é uma ponte da miséria para a
esperança.”*

(Darcy Ribeiro)

Resumo

Esta dissertação explora o uso de mineração de dados educacionais e inteligência artificial explicável (XAI) para tornar indicadores educacionais mais interpretáveis e acessíveis, com foco na Taxa de Distorção Idade-Série (TDI). O objetivo é apoiar a formulação de políticas públicas baseadas em evidências e promover uma gestão educacional mais equitativa e estratégica por meio da utilização de dados educacionais.

O trabalho demonstra a aplicação prática de dados abertos educacionais na tomada de decisão, iniciando com a proposição de uma arquitetura de dados construída a partir de fontes públicas, como INEP, IBGE e DATASUS, estruturadas em um banco relacional PostgreSQL para o monitoramento das metas do Plano Nacional de Educação (PNE). Em seguida, apresenta como pequenos times podem consolidar bases de dados públicas para acompanhamento local das metas do PNE, preparando o terreno para a análise de um dos principais indicadores educacionais brasileiros: a TDI.

Após introduzir o indicador, é desenvolvido um conjunto de dados que permite traçar um panorama nacional da TDI entre 2018 e 2022, utilizando ferramentas para análise e classificação de municípios em quartis. A análise é aprofundada com o uso de Learning Analytics (LA) e algoritmos explicáveis, como o Decision Tree Regressor, para classificar escolas e prever indicadores relacionados. Além disso, a técnica SHAP é empregada para identificar os fatores mais influentes sobre a TDI, destacando variáveis como conectividade à internet e disponibilidade de equipamentos multimídia nas escolas.

Os resultados evidenciam que a integração de dados educacionais com técnicas de mineração de dados e inteligência artificial explicável pode oferecer insights estratégicos e inovadores para o aprimoramento do sistema educacional brasileiro. O trabalho destaca como os dados podem ser utilizados de maneira flexível e contextualizada, permitindo uma aproximação mais fiel às diversas realidades do sistema educacional, e oferecendo subsídios para uma gestão mais equitativa e adaptada às necessidades locais.

Palavras-chave: Taxa de Distorção Idade-Série (TDI); Mineração de Dados Educacionais; Inteligência Artificial Explicável; Learning Analytics; Políticas Públicas Educacionais; Dados Educacionais Abertos; SHAP.

Abstract

This dissertation explores the use of educational data mining and explainable artificial intelligence (XAI) to make educational indicators more interpretable and accessible, focusing on the Age-Grade Distortion Rate (TDI). The objective is to support evidence-based public policymaking and promote more equitable and strategic educational management through the use of educational data.

The study demonstrates the practical application of open educational data in decision-making, beginning with the proposal of a data architecture built from public sources such as INEP, IBGE, and DATASUS, structured in a PostgreSQL relational database to monitor the goals of the National Education Plan (PNE). It then shows how small teams can consolidate public databases for local monitoring of PNE goals, setting the stage for the analysis of one of Brazil's key educational indicators: the TDI.

After introducing the indicator, a dataset is developed to provide a national overview of the TDI from 2018 to 2022, using tools for analysis and municipal quartile classification. The analysis is deepened with the use of Learning Analytics (LA) and explainable algorithms, such as the Decision Tree Regressor, to classify schools and predict related indicators. Additionally, the SHAP technique is used to identify the most influential factors affecting the TDI, highlighting variables such as internet connectivity and the availability of multimedia equipment in schools. The results show that integrating educational data with data mining and explainable artificial intelligence techniques can offer strategic and innovative insights to improve the Brazilian education system. The study highlights how data can be used flexibly and contextually, allowing for a more accurate understanding of the diverse realities within the education system and providing support for more equitable and locally adapted management.

Keywords: Age-Grade Distortion Rate (TDI); Educational Data Mining; Explainable Artificial Intelligence (XAI); Learning Analytics; Educational Public Policies; Open Educational Data; SHAP.

Sumário

1	Introdução	11
1.1	Caracterização do Problema	12
1.2	Justificativa	13
1.3	Perguntas de Pesquisa	14
1.4	Objetivos	15
1.4.1	Objetivo Geral	15
1.4.2	Objetivos Específicos	15
1.5	Organização do Trabalho	16
1.5.1	Arquitetura de dados educacionais como plataforma para governo inteligente: utilizando dados abertos para apoio à gestão educacional baseada em evidência	16
1.5.2	Workflow for the acquisition, processing, and dissemination of Brazilian public data focused on education	17
1.5.3	Criação de um Panorama da Taxa de Distorção Idade-Série com base nos dados abertos do INEP	17
1.5.4	Aplicação de Learning Analytics para identificação de tomada de decisão sobre a distorção idade-série no Brasil	18
1.5.5	Entendendo os Fatores de Previsão do TDI nas Escolas Públicas Brasileiras: Uma Abordagem Usando a Técnica SHAP	19
2	Considerações Finais	21
2.1	Artigos aceitos	22
2.2	Limitações da pesquisa	22
2.3	Trabalhos Futuros	23
	Referências	24
3	Apêndice A - Arquitetura de Dados Educacionais como Plataforma para Governo Inteligente Utilizando dados abertos para apoio à gestão educacional baseada em evidências	25
4	Apêndice B - <i>Workflow</i> para aquisição, processamento e disseminação de dados públicos brasileiros voltados à educação	37

5	Apêndice C - Criação de um panorama da Taxa de distorção idade-série com base nos dados abertos do INEP	58
6	Apêndice D - Aplicação de Learning Analytics para Identificação de Tomada de Decisão sobre a Distorção Idade-Série no Brasil	65
7	Entendendo os Fatores de Previsão do TDI nas Escolas Públicas Brasileiras: Uma Abordagem Usando a Técnica SHAP	77

1 Introdução

A análise de dados educacionais tem se consolidado como uma ferramenta estratégica para o planejamento e a implementação de políticas públicas efetivas (FIALHO et al., 2024). No Brasil, órgãos como o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) desempenham um papel crucial ao disponibilizar informações abrangentes sobre o sistema educacional brasileiro, como o Censo Escolar e indicadores de desempenho, que subsidiam o monitoramento do Plano Nacional de Educação (PNE) (INEP, 2024). Contudo, a utilização plena desses dados ainda enfrenta desafios relacionados à acessibilidade e à interpretação por parte de gestores educacionais (OLIVEIRA et al., 2017).

As taxas de abandono escolar, distorção idade-série e os índices de aprendizagem estão entre os principais indicadores utilizados para medir a qualidade da educação no Brasil. Estudos mostram que uma análise detalhada desses indicadores pode revelar disparidades regionais e identificar áreas críticas que necessitam de intervenção (CHIAPINOTO; LUNARDI, 2024). No entanto, a carência de ferramentas de fácil acesso e interpretação dificulta a larga utilização desses dados e suas informações geradas.

Técnicas de Mineração de Dados Educacionais (Educational Data Mining) têm sido amplamente exploradas na literatura nacional e internacional para compreender padrões e prever tendências no desempenho educacional (HUSSAIN et al., 2021; YAĞCI, 2022; KHOUDIER et al., 2023). Essas técnicas, quando combinadas com abordagens de Inteligência Artificial Explicável (XAI), que é um conjunto de técnicas e métodos que trazem a explicabilidade sobre seu processo decisório a humanos(??), permitem que a aplicação dessas técnicas sejam mais transparentes e fundamentadas, tornando os resultados compreensíveis para gestores e tomadores de decisão (MARCH; SMITH, 1995).

Em especial, o indicador Taxa de Distorção Idade-Série (TDI) apresenta grande potencial para apoiar políticas que combatam a repetência e a evasão escolar. Pesquisas recentes indicam que análises interpretáveis da TDI podem oferecer proposições valiosas para reduzir as desigualdades no sistema educacional, especialmente em contextos socioeconômicos vulneráveis (PORTELLA et al., 2017).

A integração de bases de dados abertas com ferramentas analíticas é fundamental para a democratização do acesso à informação. No Brasil, a ampla disponibilização de microdados pelo INEP representa um avanço significativo, mas sua utilização ainda requer conhecimentos

técnicos avançados, limitando o alcance dos benefícios esperados (BARROS et al., 2022).

Este trabalho buscou explicar como técnicas de mineração de dados e inteligência artificial explicável podem ser aplicadas a indicadores educacionais, com foco na criação de estudos interpretáveis que auxiliem na tomada de decisões estratégicas. Particularmente, a pesquisa se concentra no indicador Taxa de Distorção Idade-Série (TDI), propondo abordagens que promovam acessibilidade e compreensão dos dados.

Ao abordar lacunas existentes na utilização de dados educacionais no Brasil, esta pesquisa almeja contribuir para o fortalecimento de políticas públicas baseadas em evidências. Por meio do desenvolvimento e combinações de métodos atuais e interpretáveis, o estudo visa ampliar o impacto dos dados divulgados pelo INEP e promover uma gestão educacional mais equitativa e eficiente.

1.1 Caracterização do Problema

A crescente disponibilização de dados educacionais pelo INEP oferece um grande potencial para análises estratégicas, mas também levanta desafios significativos quanto ao uso efetivo dessas informações. Embora os dados estejam acessíveis, a plena utilização ainda é dificultada por barreiras relacionadas aos processos de integração, às ferramentas analíticas disponíveis e à falta de capacitação técnica de gestores e educadores para interpretá-los. A ausência de um fluxo claro e eficiente de como proceder as etapas de processamento e análise limita o impacto que esses dados podem ter nas decisões e políticas educacionais, restringindo sua utilidade prática. Outro ponto crucial é a aplicação de técnicas avançadas, como Mineração de Dados Educacionais e Inteligência Artificial Explicável (Explainable AI), que poderiam evidenciar padrões, tendências e percepções estratégicos para indicadores do INEP. Essas técnicas não apenas facilitam análises mais aprofundadas, mas também tornam os resultados mais compreensíveis e utilizáveis para tomadores de decisão. No entanto, sua aplicação aos indicadores educacionais, como os disponibilizados pelo INEP, ainda é limitada e carece de estudos que explorem suas potencialidades no contexto educacional brasileiro. Um exemplo que ilustra esse desafio é o indicador Taxa de Distorção Idade-Série (TDI). Apesar de sua relevância para compreender problemas estruturais da educação, como reprovação e evasão escolar, sua análise pouco é acompanhada de abordagens interpretáveis que auxiliem gestores educacionais a transformarem os dados em ações estratégicas. O desenvolvimento de métodos interpretáveis

para esse indicador poderia facilitar o entendimento das causas subjacentes à distorção idade-série e apoiar intervenções mais direcionadas e eficazes. Ademais, a apresentação dos dados configura-se como uma questão central. Muitos indicadores, embora possuam elevado valor informativo, são frequentemente divulgados de maneira técnica e pouco acessível a públicos não especializados. Isso reforça a necessidade de desenvolver formas simplificadas de acesso, promovendo a democratização da informação. Tal abordagem não apenas facilita o uso efetivo dos indicadores, mas também amplia a aplicação dos resultados de análises e experimentos, potencializando sua otimização e impacto. O uso desse indicador deve ser combinado com outras análises e resultados obtidos de forma integrada, a fim de possibilitar uma avaliação mais abrangente do contexto em que os dados foram coletados. Isso se deve ao fato de que, isoladamente, um único indicador pode não refletir com precisão a complexidade da realidade educacional. Portanto, o problema a ser investigado reside na lacuna entre a disponibilidade dos dados educacionais e sua plena utilização para gerar impacto positivo no planejamento e na execução de políticas educacionais. Essa lacuna inclui desafios técnicos, como a aplicação de métodos de análise e interpretação, e desafios de comunicação, como a acessibilidade e clareza dos resultados para diferentes tomadores de decisão.

1.2 Justificativa

A utilização de indicadores educacionais como base para a tomada de decisão é uma prática indispensável para o aprimoramento do sistema educacional e para o fortalecimento das políticas públicas no setor. Essa abordagem justifica-se pela capacidade de traduzir dados brutos em informações significativas, permitindo uma compreensão ampla e detalhada dos desafios e das oportunidades no ambiente educacional. Indicadores como taxas de aprovação, reprovação, abandono escolar, distorção idade-série e desempenho em avaliações externas são instrumentos fundamentais para diagnosticar a situação da educação em diferentes contextos e propor soluções baseadas em evidências.

A tomada de decisões fundamentadas em dados possibilita o direcionamento estratégico de investimentos para áreas com maior necessidade de intervenção, como escolas situadas em regiões vulneráveis ou grupos populacionais em risco de exclusão escolar. Por meio de relatórios e painéis de acompanhamento baseados em indicadores, os gestores educacionais podem monitorar os impactos das políticas implementadas e ajustar suas estratégias de uma forma

mais atualizada e, a depender de sua elaboração, de forma mais simples. Essa abordagem favorece uma alocação mais eficiente de recursos e esforços, permitindo a priorização de programas e políticas com maior potencial de impacto e promovendo, assim, uma gestão educacional mais eficaz e equitativa.

A análise de indicadores também contribui significativamente para identificar desigualdades educacionais, que permanecem muitas vezes ocultas sem uma avaliação detalhada. Ao desagregar os dados por fatores como gênero, raça, território ou dependência administrativa, é possível evidenciar disparidades históricas e estruturais no acesso e na qualidade da educação. Esse mapeamento fundamenta ações específicas e direcionadas para reduzir essas lacunas, promovendo a equidade no sistema educacional.

Relacionado a isso está o aspecto relevante ao papel dos indicadores no acompanhamento e avaliação de políticas públicas. Medir o desempenho de programas como a ampliação do ensino integral, trazendo métricas claras e comparáveis, que permitam avaliar sua efetividade e justificar a continuidade ou reformulação dessas ações.

Os indicadores educacionais também são ferramentas essenciais para o planejamento estratégico ao nível local, regional e nacional. Eles oferecem uma visão integrada do sistema, permitindo que gestores antecipem tendências, como o aumento da demanda por matrículas em determinadas regiões, ou a situação inversa, buscando entender a causa de algumas localidades possuírem baixo nível de matrículas no sistema educacional. Indicadores esses utilizados por até mais de uma política pública fundamentada como a relação direta do indicador 1B do Plano Nacional da Educação e o Bolsa Família, onde ambos buscam que crianças entre 4 a 5 anos já estejam inclusas no fluxo escolar.

1.3 Perguntas de Pesquisa

PERGUNTA DE PESQUISA 1 (PP1):

Que processos e ferramenta são necessários para conseguir a plena utilização dos dados divulgados pelo INEP?

A pergunta PP1 visa identificar e descrever os processos e ferramentas essenciais para a utilização de dados e indicadores, principalmente aqueles fornecidos pelo INEP. Busca-se definir, ao menos, uma versão de quais processos e ferramental são necessários para o uso eficaz desses dados, promovendo uma aplicação mais estruturada e consistente das informações disponíveis.

PERGUNTA DE PESQUISA 2 (PP2):

Como técnicas de Mineração de Dados Educacionais e Inteligência Artificial Explicável podem ser aplicadas a indicadores do INEP ?

A PP2 pretende identificar técnicas de mineração de dados educacionais que possam ser combinadas com indicadores já disponibilizados pelo INEP, visando aprimorar diretamente a qualidade no âmbito escolar.

PERGUNTA DE PESQUISA 3 (PP3):

Quais abordagens interpretáveis podem ser desenvolvidas a partir da análise do indicador Taxa de Distorção Idade-Série (TDI) para apoiar decisões estratégicas de gestores educacionais?

A PP3 visa desenvolver metodologias que utilizem a Taxa de Distorção Idade-Série como ferramenta analítica, permitindo que, a partir dos resultados desse indicador, sejam apontadas áreas de melhorias nos entes escolares.

1.4 Objetivos

1.4.1 Objetivo Geral

Desenvolver abordagens baseadas em Mineração de Dados Educacionais e Inteligência Artificial Explicável para tornar a Taxa de Distorção Idade-Série (TDI) mais interpretável e acessível.

1.4.2 Objetivos Específicos

Para atingir o objetivo geral foram definidos os seguintes objetivos específicos:

- Analisar indicadores educacionais disponíveis e sua relevância para o mapeamento da educação brasileira.
- Definir processos para utilização e computo dos dados brutos e indicadores.
- Explorar e aplicar técnicas de Mineração de Dados Educacionais para identificar padrões relevantes na Taxa de Distorção Idade-Série.
- Construir modelos interpretáveis que permitam entender como diferentes variáveis impactam a Taxa de Distorção Idade-Série.

- Promover a integração dos resultados dos indicadores educacionais, facilitando o uso contínuo dos dados.
- Implementar uma metodologia de análise da TDI voltada para apoiar a formulação de mudanças estruturais e pedagógicas.
- Disponibilização de conjunto de dados para utilização de dados e indicadores educacionais.

1.5 Organização do Trabalho

A organização desta dissertação fundamenta-se na publicação de artigos que compõem o documento e respondem diretamente às perguntas de pesquisa previamente formuladas. Durante o período de investigação, foram desenvolvidos trabalhos acadêmicos que, além de abordarem os objetivos centrais da pesquisa.

1.5.1 Arquitetura de dados educacionais como plataforma para governo inteligente: utilizando dados abertos para apoio à gestão educacional baseada em evidência

O artigo apresenta uma arquitetura de dados educacionais abertos projetada para fortalecer a transparência e a avaliação no monitoramento do Plano Nacional de Educação (PNE). Integrando e processando dados públicos provenientes do INEP, IBGE, Ipea e DATASUS, a solução proposta viabiliza o acesso democrático a indicadores educacionais, permitindo que gestores públicos, pesquisadores e a sociedade civil realizem análises baseadas em evidências. A arquitetura emprega ferramentas open-source, como o PostgreSQL, e adota princípios de dados abertos, assegurando replicabilidade, sustentabilidade e baixo custo de implementação. Por meio de um estudo de caso, o artigo demonstra como o cruzamento dos dados do PNE com indicadores municipais (MUNIC) pode revelar correlações relevantes entre políticas locais e o atingimento de metas educacionais, evidenciando o potencial dos dados abertos para qualificar o debate público e orientar de maneira mais precisa as políticas educacionais.

A realização desta etapa foi fundamental para a pesquisa, pois possibilitou identificar que tipo de arquitetura mínima e gratuita poderia ser utilizada nas fases seguintes do estudo, respeitando as limitações de infraestrutura e o arcabouço de dados públicos disponíveis. Além disso, o trabalho destaca a relevância de soluções acessíveis para promover um Governo Inteligente, onde a disponibilização estruturada de dados abertos amplia a participação social,

melhora o acompanhamento de políticas públicas e fortalece o processo de tomada de decisão baseado em evidências.

1.5.2 Workflow for the acquisition, processing, and dissemination of Brazilian public data focused on education

O artigo descreve o processo de criação de bases de dados públicas voltadas às áreas educacional e populacional, a partir de fontes oficiais como o INEP e o IBGE. A coleta, adaptação e otimização dos dados para construção de séries históricas foram realizadas com foco em boas práticas e na geração de metadados qualificados. O artigo destaca as especificidades e desafios de cada temática, educação e população, bem como as particularidades de seus respectivos conjuntos de dados. Os resultados já disponíveis permitem o acompanhamento direto de indicadores do Plano Nacional de Educação (PNE), evidenciando o potencial dessas bases combinadas para o monitoramento de uma das principais políticas públicas educacionais do Brasil.

A realização deste trabalho foi fundamental para a compreensão do processamento dos dados públicos e da forma de interconexão entre eles, etapa indispensável para o avanço da pesquisa. Isso se deve ao fato de que tanto os dados dos indicadores quanto os dados escolares seguem o mesmo modelo de publicação e disponibilização adotado pelas instâncias públicas responsáveis.

1.5.3 Criação de um Panorama da Taxa de Distorção Idade-Série com base nos dados abertos do INEP

O estudo analisa a Taxa de Distorção Idade-Série (TDI) no Brasil entre 2018 e 2022, utilizando dados abertos do INEP para mapear desafios educacionais. A TDI, que mede a defasagem entre a idade do estudante e a série escolar frequentada, foi processada utilizando ferramentas open-source, como Python e Pandas, garantindo transparência metodológica, replicabilidade e acesso aberto aos resultados. Os dados revelam uma redução significativa da TDI nos anos iniciais do Ensino Fundamental, com uma queda aproximada de 50%, enquanto nos anos finais a melhoria, embora mais moderada, manteve-se consistente. Um estudo de caso conduzido em Pernambuco destacou avanços mesmo durante a pandemia, com municípios alcançando, pela primeira vez em 2022, a faixa mais baixa de distorção idade-série. A

utilização de quartis dinâmicos na classificação dos resultados permitiu comparações mais justas entre diferentes municípios, evidenciando desigualdades educacionais persistentes no território nacional. A base de dados gerada foi disponibilizada publicamente, incentivando a realização de futuras pesquisas e reforçando o potencial dos dados abertos para o monitoramento de políticas públicas, a identificação de disparidades regionais e o estímulo a ações baseadas em evidências.

A execução deste trabalho foi essencial para o avanço da pesquisa, pois constituiu o primeiro contato com o indicador de distorção idade-série, possibilitando uma avaliação inicial de seu potencial analítico. A análise proporcionou uma compreensão mais aprofundada sobre a aplicação do indicador em estudos anteriores e permitiu a realização de um mapeamento temporal da flutuação dos valores da TDI especificamente para o estado de Pernambuco. Tal etapa se revelou estratégica para estruturar investigações futuras, como a inclusão de novos indicadores (tais como evasão escolar e resultados do SAEB), ampliando o escopo analítico e fortalecendo a base empírica da pesquisa.

1.5.4 Aplicação de Learning Analytics para identificação de tomada de decisão sobre a distorção idade-série no Brasil

Este estudo aplicou técnicas de Learning Analytics para identificar fatores determinantes da Taxa de Distorção Idade-Série (TDI) no Brasil, utilizando dados abertos do INEP referentes ao período de 2018 a 2022. Por meio da implementação de um modelo de Decision Tree Regressor otimizado, foram analisados três cenários distintos: dados completos, apenas dados de matrículas e atributos qualitativos. Os resultados evidenciaram que escolas com maior número de matrículas na faixa etária de 6 a 10 anos, acesso à internet e infraestrutura básica (como disponibilidade de energia elétrica) apresentaram TDI abaixo da média nacional (11,71%). Em contrapartida, instituições públicas, que representam cerca de 90% da amostra, e aquelas com menor acesso a recursos básicos, registraram indicadores de distorção idade-série significativamente piores. O modelo de regressão obteve alta correlação (Pearson = 0,79) no experimento baseado em dados de matrículas, reforçando a robustez dos achados. Além disso, a análise destacou disparidades raciais, observando que alunos brancos apresentaram, em média, uma TDI inferior à de outros grupos. As implicações práticas dos resultados incluem o direcionamento de políticas públicas voltadas para a redução de desigualdades, como a expansão da infraestrutura digital e a revisão

dos critérios de alocação de turmas escolares. O estudo reforça o potencial dos dados abertos educacionais para embasar decisões estratégicas e aponta caminhos para futuras análises, como a extensão da metodologia para outras etapas de ensino.

Este trabalho representou o primeiro experimento da pesquisa utilizando técnicas de Learning Analytics aplicadas a dados educacionais abertos, sendo pioneiro na utilização de medidas estatísticas para identificar características associadas a um desempenho abaixo da média nacional da TDI. A realização deste experimento aprofundado foi crucial para levantar um panorama inicial sobre as variáveis mais influentes no contexto da distorção idade-série, orientando as etapas seguintes da investigação.

1.5.5 Entendendo os Fatores de Previsão do TDI nas Escolas Públicas Brasileiras: Uma Abordagem Usando a Técnica SHAP

Este estudo aplicou técnicas de Learning Analytics e análise de valores SHAP para identificar fatores determinantes da Taxa de Distorção Idade-Série (TDI) em escolas públicas brasileiras, utilizando dados abertos do INEP referentes ao período de 2018 a 2023. Para a modelagem preditiva, foi empregado o algoritmo CatBoost, reconhecido por seu desempenho robusto, que alcançou um coeficiente de determinação (R^2) de 0,83 e um erro absoluto médio (MAE) de 5,83. A análise permitiu extrair as dez principais variáveis associadas à TDI. No âmbito pedagógico, a existência de grêmio estudantil e a disponibilidade de internet para atividades de aprendizagem foram identificadas como fatores que contribuem para a redução da distorção. Em relação à infraestrutura, a presença de quadras cobertas e o acesso à energia pública mostraram-se positivamente relacionados a menores níveis de TDI, assim como o maior número de equipamentos multimídia e computadores disponíveis nas escolas. A análise temporal evidenciou mudanças nas variáveis mais relevantes ao longo dos anos: antes de 2020, o ensino organizado em ciclos era o principal fator associado à TDI, enquanto em 2023 fatores relacionados à participação estudantil e infraestrutura digital passaram a predominar. Além disso, a introdução do transporte público escolar, avaliada em 2023, correlacionou-se a TDIs mais elevadas, apontando novos desafios no campo da mobilidade escolar. Os resultados reforçam a necessidade de investimentos estratégicos em infraestrutura física e tecnológica, bem como em práticas de gestão democrática, como caminhos para a redução de desigualdades educacionais. A metodologia baseada em SHAP demonstrou alta eficácia na interpretabilidade dos resultados,

e o pipeline de análise desenvolvido é replicável para outros indicadores educacionais. Todo o processamento de dados foi realizado de forma aberta, consolidando o compromisso com a transparência e o fortalecimento de políticas públicas fundamentadas em evidências.

Este trabalho representou a etapa final da pesquisa, incorporando técnicas mais visuais de interpretabilidade dos modelos de regressão, o que permitiu uma compreensão aprofundada dos fatores críticos que influenciam a TDI. Em consonância com os resultados anteriores, o estudo buscou não apenas identificar momentos educacionais decisivos, mas também categorizar as variáveis relevantes em fatores estruturais e pedagógicos. Essa divisão visa tornar a análise e a implementação de estratégias de melhoria mais acessíveis a gestores educacionais, ampliando o potencial de impacto prático da pesquisa na formulação de políticas públicas e no planejamento educacional.

2 Considerações Finais

A utilização de dados para a tomada de decisões tornou-se uma necessidade essencial no modelo de sociedade contemporânea, que é cada vez mais orientada por informações quantitativas e analíticas. No entanto, essa transição para uma sociedade guiada por dados não é um processo simples e automático. Ela demanda uma série de adaptações que devem ser feitas de acordo com as especificidades de cada contexto. Ao aplicarmos essa realidade ao ambiente escolar, a complexidade aumenta ainda mais, pois estamos lidando com a diversidade de estruturas, recursos tecnológicos, infraestrutura e capital humano presentes nos mais de 5.570 municípios do Brasil. Esta diversidade cria uma enorme disparidade no modo como os dados podem ser coletados, analisados e utilizados, o que torna a tarefa de aplicar dados na gestão educacional ainda mais desafiadora. Contudo, é uma tarefa imprescindível para avançarmos em direção a uma análise precisa e à implementação de soluções eficazes para enfrentar as complexidades e os desafios de nosso sistema educacional, que é ao mesmo tempo vasto e multifacetado.

Com o objetivo de proporcionar subsídios para a tomada de decisões mais fundamentadas por parte dos gestores educacionais e outros atores envolvidos no processo de formulação de políticas públicas, esta dissertação foca na sistematização, organização e utilização dos dados educacionais. A partir de uma abordagem estruturada, busquei responder perguntas de pesquisa que pudessem orientar e apoiar esses tomadores de decisão no contexto específico da educação pública no Brasil. Além disso, explorei diferentes formas de apresentar esses dados de maneira acessível e prática, permitindo que os dados não apenas sejam analisados, mas também aplicados de maneira efetiva no planejamento e execução de políticas educacionais.

O objetivo desta dissertação é, portanto, não apenas trazer uma análise detalhada sobre o uso dos dados educacionais, mas também promover a disseminação de práticas que já se mostraram eficazes e inovadoras na compreensão e utilização dos indicadores educacionais públicos. Ao compartilhar esses métodos, esperamos contribuir para a transformação da forma como os dados são vistos e utilizados na gestão da educação, promovendo avanços significativos na qualidade e na equidade do ensino no Brasil.

2.1 Artigos aceitos

Os artigos resultantes desta pesquisa que foram aceitos em conferências e revistas da área foram os seguintes:

Barros, Abílio Nogueira, et al. "Aplicação de Learning Analytics para identificação de tomada de decisão sobre a distorção idade-série no Brasil." *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*. SBC, 2023.

Barros, Abilio Nogueira, et al. "Arquitetura de Dados Educacionais como Plataforma para Governo Inteligente-Utilizando dados abertos para apoiaa gestao educacional baseada em evidências."

Nogueira Barros, A., A. Félix de Albuquerque, A. Leite de Alencar, A. Nascimento, I. Mateus Bittencourt, and R. Ferreira Mello. "Workflow for the Acquisition, Processing, and Dissemination of Brazilian Public Data Focused on Education". *Journal of Information and Data Management*, vol. 15, no. 1, Apr. 2024, pp. 224-33, doi:10.5753/jidm.2024.3570.

Barros, Abílio Nogueira, Andrêza Leite de Alencar, and Rafael Ferreira Mello. "Criação de um panorama da Taxa de distorção idade-série com base nos dados abertos do INEP." *Anais do VIII Congresso sobre Tecnologias na Educação*. SBC, 2023.

Barros, Abílio, Gabriel Alves, and Rafael Ferreira Mello. "Entendendo os Fatores de Previsão do TDI nas Escolas Públicas Brasileiras: Uma Abordagem Usando a Técnica SHAP." *Anais do XXXV Simpósio Brasileiro de Informática na Educação, Rio de Janeiro/RJ, 2024*. SBC, 2024, pp.1529-1542.

2.2 Limitações da pesquisa

Uma das principais limitações enfrentadas nesta pesquisa foi a mudança brusca no formato dos dados divulgados pelo INEP. A mudança no formato de disponibilização dos dados em 2022 introduziu barreiras significativas para a análise direta dos alunos, o que, por sua vez, dificultou uma investigação mais detalhada sobre os estudantes e a TDI.

Outro desafio relevante foi a validação direta das análises junto aos gestores educacionais. Esse processo foi dificultado por agendas e compromissos, especialmente no contexto de eleições e mudanças de governo. No caso específico da finalização deste trabalho, as eleições municipais de 2024 representaram um obstáculo adicional, dada a sobrecarga de compromissos eleitorais e a transição administrativa.

2.3 Trabalhos Futuros

Uma das principais possibilidades de aprimoramento está na avaliação de tecnologias emergentes para armazenamento e processamento de dados. Soluções de bancos de dados colunares, como o DuckDB, oferecem vantagens significativas em consultas analíticas, especialmente no tratamento de grandes volumes de dados educacionais. Comparações entre essas tecnologias e outras arquiteturas tradicionais podem revelar novos caminhos para a otimização da eficiência e da escalabilidade na análise de dados no contexto educacional, criando, assim, oportunidades para a redução de custos operacionais e o aumento da acessibilidade tecnológica.

Outro campo promissor reside na aplicação de técnicas de interpretabilidade a modelos preditivos e análises estatísticas. Em um contexto onde os resultados precisam ser acessíveis e compreensíveis para gestores educacionais, a utilização de ferramentas como SHAP ou LIME pode facilitar a explicação de previsões individuais e a análise da relevância das variáveis. Isso assegura que os tomadores de decisão possam confiar nas informações geradas e compreendê-las plenamente, promovendo um uso mais eficaz e fundamentado dos dados.

Além disso, destaca-se a importância de alinhar as análises e materiais produzidos com as demandas dos gestores educacionais, mensurando a percepção e a aplicabilidade real dos dados na tomada de decisão. Realizar discussões técnicas e aplicar questionários de percepção com gestores são ações fundamentais para compreender e aprimorar o processo de ponta a ponta, garantindo que as ferramentas e metodologias desenvolvidas estejam alinhadas às necessidades práticas do ambiente educacional.

Referências

- BARROS, A. N.; ALENCAR, A.; NASCIMENTO, A.; ALBUQUERQUE, A. F. de; MELLO, R. F. Elaboração do conjunto de dados agregados do censo da educação básica. In: SBC. **Anais do IV Dataset Showcase Workshop**. [S.l.], 2022. p. 35–45.
- CHIAPINOTO, A. M.; LUNARDI, E. M. Análise das produções sobre distorção idade/ano e políticas públicas sobre a defasagem na aprendizagem e correção do fluxo escolar. **LUMEN ET VIRTUS**, v. 15, n. 38, p. 1067–1085, 2024.
- FIALHO, L. M. F.; NEVES, V. N. S.; OLIVA, M. F. R. Políticas públicas para o ensino superior: a produção científica brasileira em circulação internacional. **Ensaio: Avaliação e Políticas Públicas em Educação**, SciELO Brasil, v. 32, p. e0244199, 2024.
- HUSSAIN, S.; GAFTANDZHIEVA, S.; MANIRUZZAMAN, M.; DONEVA, R.; MUHSIN, Z. F. Regression analysis of student academic performance using deep learning. **Education and Information Technologies**, Springer, v. 26, n. 1, p. 783–798, 2021.
- KHOUDIER, M. M. E.; ABDELNABY, R. H. M.; ELDAMNHOURY, Z. M.; ABOUZEID, S. R. A.; EL-MONAYER, G. K.; ENAN, N. M.; EL-GHAMRY, A.; FOUAD, K.; MOAWAD, I. Prediction of student performance using machine learning techniques. In: IEEE. **2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)**. [S.l.], 2023. p. 333–338.
- MARCH, S.; SMITH, G. Design and natural science research on information technology. **Decision Support Systems**, v. 15, p. 251–266, 12 1995.
- OLIVEIRA, D. A.; DUARTE, A. W. B.; JESUS, A. M. C. et al. A nova gestão pública no contexto escolar e os dilemas dos (as) diretores (as). **Revista Brasileira de Política e Administração da Educação**, Universidade Federal de Minas Gerais, 2017.
- PORTELLA, A. L.; BUSSMANN, T. B.; OLIVEIRA, A. M. H. d. A relação de fatores individuais, familiares e escolares com a distorção idade-série no ensino público brasileiro. **Nova economia**, SciELO Brasil, v. 27, n. 3, p. 477–509, 2017.
- YAĞCI, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. **Smart Learning Environments**, Springer, v. 9, n. 1, p. 11, 2022.

3 Apêndice A - Arquitetura de Dados Educacionais como Plataforma para Governo Inteligente Utilizando dados abertos para apoio à gestão educacional baseada em evidências

Barros, A. N., de Albuquerque, A. F., Alencar, A., Mello, R. F., Alves, G., & Bittencourt, I. M. (2023, August). Arquitetura de dados educacionais como plataforma para governo inteligente-utilizando dados abertos para apoio à gestão educacional baseada em evidências. In Workshop de Computação Aplicada em Governo Eletrônico (WCGE) (pp. 130-140). SBC.

Contribuição do autor: Responsável pela definição e implementação da metodologia, avaliação dos resultados e escrita do texto.

Arquitetura de Dados Educacionais como Plataforma para Governo Inteligente - Utilizando dados abertos para apoio à gestão educacional baseada em evidências

Abílio Nogueira Barros¹, Aldéryck Félix de Albuquerque², Andreza Alencar¹,
Rafael Ferreira Mello^{1,2}, Gabriel Alves¹,
Ibsen Mateus Bittencourt³

¹Departamento de Computação - Universidade Federal Rural de Pernambuco (UFRPE)

²Centro de Estudos e Sistemas Avançados do Recife (CESAR)

³Universidade Federal de Alagoas (UFAL)

{abilionbarros,derycck}@gmail.com,{andreza.leite,rafael.mello,gabriel.alves}@ufrpe.br,
ibsen@feac.ufal.br

Abstract. *The main objective of the National Education Plan (PNE) is to guide the development of education in the country and its monitoring is characterized as a fundamental activity for educational management. To carry out this monitoring, solutions are needed that provide a reliable data set capable of assisting in evidence-based decision making. Therefore, this article presents an educational data architecture that uses open government data to support intelligent educational management. It provides various public data for monitoring the situation of municipal and state education based on the goals and indicators of the PNE version 2014-2024. This also makes it possible to cross data from different sources to find factors that influence the achievement of such goals. In order to guarantee the construction of reliable evidence, the proposed architecture relies on the use of census databases, such as the Educational Census, coordinated by the National Institute of Educational Studies and Research (INEP). The present study also presents a case study that demonstrates one of the possible applications of this architecture and highlights the potential value of its use for educational management in the country. We therefore conclude that the proposed architecture is a useful instrument to support platforms for evaluating and monitoring educational public policies, contributing to the construction of an Intelligent Government and to promoting improvements in the quality of Brazilian education.*

Resumo. *O Plano Nacional de Educação (PNE) tem como objetivo central orientar o desenvolvimento da educação no país e seu monitoramento se caracteriza como uma atividade fundamental para a gestão educacional. Para realizar esse monitoramento se fazem necessárias soluções que forneçam um conjunto de dados confiável e capaz de auxiliar na tomada de decisão baseada em evidências. Diante disso, este artigo apresenta uma arquitetura de dados educacionais que utiliza dados abertos governamentais para apoiar uma gestão educacional inteligente. Esta disponibiliza diversos dados públicos para o acompanhamento da situação da educação municipal e estadual com base nas metas*

e indicadores do PNE da versão 2014-2024. Esta também possibilita o cruzamento de dados de diversas fontes para encontrar fatores que influenciem no alcance de tais metas. A fim de garantir a construção de evidências confiáveis, a arquitetura proposta conta com o uso de bases de dados censitárias, como o Censo Educacional, coordenado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). O presente estudo também apresenta um estudo de caso que demonstra uma das aplicações possíveis dessa arquitetura e evidencia o valor potencial no uso dela para a gestão educacional no país. Concluímos assim, que a arquitetura proposta é um instrumento útil para suportar plataformas de avaliação e acompanhamento de políticas públicas educacionais, contribuindo para a construção de um Governo Inteligente e para promover melhorias na qualidade da educação brasileira.

1. Introdução

O Plano Nacional de Educação (PNE) brasileiro pretende orientar o desenvolvimento da educação no país por 10 anos [Macena et al. 2018]. O PNE estabelece metas, estratégias e ações para a melhoria da qualidade da educação em todas as etapas e modalidades de ensino, da educação infantil ao ensino superior, bem como para a formação de professores e para a gestão democrática das escolas. O monitoramento do PNE é uma atividade fundamental para manter a gestão na direção de cumprir as metas estabelecidas [Vinente and Duarte 2015] e para possibilitar ajustes ou mudanças de trajetória nas estratégias de gestão com o intuito de alcançar o plano definido.

Visando possibilitar o monitoramento do PNE e auxiliar na tomada de decisões baseadas em evidências, é necessário disponibilizar soluções capazes de fornecer um conjunto de dados confiável e capaz de gerar tais evidências. Assim, a disponibilização de uma arquitetura de dados para acompanhamento dos Indicadores do PNE em formato acessível, tanto para gestores educacionais públicos quanto para privados, pode ser uma ferramenta valiosa para nutrir esses agentes de informações para avaliar os resultados das políticas educacionais adotadas no país. Estudos anteriores já demonstraram a eficácia de técnicas de análise de dados visando otimizar o sistema educacional municipal, como o realizado por [Júnior et al. 2019] para a cidade de Maceió-AL. No entanto, ainda são necessários esforços para disponibilizar soluções que permitam esse tipo de análise para todos os municípios do Brasil.

Considerando a necessidade de lidar com diferentes fontes de dados em condições diversas de formato, estrutura e nível de qualidade, torna-se necessário discutir o processo de geração dos indicadores do PNE, que envolve a jornada para obtenção e transformação dos dados brutos oriundos dos sistemas governamentais. Para que os indicadores produzidos possam capacitar tomadas de decisões confiáveis ao nível municipal, é fundamental evitar o uso de fontes de dados geradas a partir de consultas amostrais. Em vez disso, devem ser utilizadas bases de dados censitárias, como o Censo Educacional [Diniz 1999] coordenado pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP), órgão autônomo regulamentado por instrumentos normativos que instituem a obrigatoriedade, os prazos, os responsáveis e suas responsabilidades, bem como os procedimentos para realização de todo o processo de coleta de dados.¹

¹ <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar>

Nesse contexto, o presente trabalho apresenta uma arquitetura de dados educacionais que utiliza dados abertos para apoiar a gestão educacional baseada em evidências. Visando a construção de um Governo Inteligente, essa arquitetura disponibiliza os dados dos indicadores instituídos pelo PNE às secretarias de educação e à população, possibilitando o cruzamento de informações de diversas fontes e variáveis que podem estar influenciando no alcance das metas estabelecidas. Dessa forma, a arquitetura apresenta-se como um instrumento útil no processo de tomada de decisão, possibilitando o desenvolvimento de sistemas de apoio à gestão educacional, além de aumentar a transparência e divulgação do impacto causado pelos investimentos na área.

Este documento encontra-se organizado da seguinte maneira: a Seção 2 apresenta alguns trabalhos relacionados ao uso de dados abertos, uso de dados para tomada de decisão e *pipeline* de processamento de dados educacionais e populacionais; na Seção 3 é apresentada arquitetura proposta, suas fontes de dados, ferramentas utilizadas e o processo de aquisição e tratamento de dados seguidos; na Seção 4.1 são demonstrados alguns dos resultados que já podem ser atingidos com o uso da arquitetura em um estudo de caso onde realizamos o cruzamento dos dados do PNE com outros indicadores selecionados a partir da base de perfil municipal MUNIC (Pesquisa de Informações Básicas Municipais) divulgados pelo Instituto Brasileiro de Geografia e Estatística (IBGE); e, por fim, a Seção 5 traz as considerações finais e trabalhos futuros.

2. Trabalhos Relacionados

Alguns trabalhos já foram produzidos sobre as bases educacionais oferecidas pelo INEP. Em [Conte 2019], foram utilizadas bases de perfil socioeconômico e resultados do Exame Nacional do Ensino Médio (ENEM), permitindo, assim, a realização de procedimentos de limpeza e organização dos dados para fornecer conhecimento sobre a base do exame na edição do ano de 2015. Como destacado pelo autor, é extremamente importante associar esse conhecimento adquirido com outras bases fornecidas, como a Prova Brasil, o Exame Nacional de Desempenho de Estudantes (ENADE) e o Sistema de Avaliação da Educação Básica (SAEB).

Em [Wanderley 2021], foi apresentado alguns indicadores para monitorar instituições escolares do estado do Acre. As fontes de dados utilizadas para a criação e apontamento desses indicadores foram as bases de dados educacionais fornecidas pelo INEP. Após o processamento desses dados, os resultados dos indicadores foram utilizados para a criação de painéis utilizando a ferramenta *Power BI*.

A importância dos *data lakes* lógicos como plataforma para dados governamentais em sociedades e cidades inteligentes é discutido no artigo [Ramos et al. 2022]. O autor destaca a necessidade de gerenciar grandes volumes de dados provenientes de várias fontes para promover a tomada de decisão informada e a melhoria dos serviços governamentais. O artigo também descreve as principais características de um *data lake* lógico, incluindo a capacidade de lidar com dados estruturados e não estruturados, escalabilidade e flexibilidade.

Já em [Rodrigues and Maciel 2022], destaca-se a importância dos dados abertos educacionais para a tomada de decisões informadas e a melhoria da qualidade da educação. O método proposto envolve a extração de dados brutos de várias fontes, a transformação dos dados para um formato padronizado e a carga dos dados em um re-

positório centralizado. O artigo também apresenta uma arquitetura de software para implementar o processo *ETL* (Extração, Transformação e Carga) e discute a importância da segurança dos dados e da privacidade dos usuários. Tendo como objetivo do método proposto é melhorar a eficiência da captura e compartilhamento de dados educacionais abertos, permitindo que os usuários acessem dados atualizados e relevantes para melhorar a educação.

O artigo apresentado em [Albuquerque et al. 2022] descreve um conjunto de dados que fornece estimativas populacionais desagregadas por município e idade no Brasil para o período de 2014 a 2020. Para criar esse conjunto de dados, os autores utilizaram uma metodologia que combina informações do Censo Demográfico de 2010 com as projeções populacionais do IBGE. O artigo detalha a metodologia utilizada, incluindo os ajustes realizados nas projeções populacionais para garantir a consistência temporal e espacial dos dados.

Uma abordagem para a junção histórica dos dados do censo da educação básica de 2011 a 2021 em seu novo formato da versão jan/2022, é apresentado em [Barros et al. 2022]. Seu objetivo foi o de obter uma base de dados resultante com pouco descarte de colunas que foram sendo descontinuados do Censo Educacional ao longo dos anos. O trabalho detalha o *pipeline* executado e as parametrizações realizadas para alcançar esse resultado. Além disso, destaca-se que essa abordagem pode ser replicada para determinados anos alvo de projetos educacionais.

Os estudos mencionados foram úteis para iniciar o processo de concepção e elaboração de uma arquitetura que possa lidar com grandes quantidades de dados públicos e que permita o monitoramento unificado das metas e indicadores do PNE de todos os municípios. Essa arquitetura deve fornecer uma visão geral da situação educacional ao nível federal e ser transparente desde a coleta até o fornecimento dos dados. Assim, este trabalho busca preencher a lacuna de uma arquitetura que forneça os serviços e dados necessários para esse monitoramento.

3. Arquitetura de Dados Educacionais como Plataforma para Governo Inteligente

Este estudo propõe uma arquitetura de dados para o acompanhamento dos indicadores do PNE ao longo do ciclo de 2014 a 2024. Entre as etapas para a construção dessa arquitetura e serão abordadas neste artigo, destacam-se a aquisição, armazenamento, processamento e disponibilização dos indicadores e variáveis de interesse.

A Figura 1 apresenta a arquitetura proposta, onde do lado esquerdo pode-se observar as fontes de dados públicas atualmente utilizadas, que serão detalhadas na Seção 3.1. Ao lado direito, temos o fluxo conceitual de como os dados são implantados e processados.

Após o processamento, os dados são armazenados em um banco de dados relacional no PostgreSQL². Esse banco é dividido em dois esquemas: *staging* e *serving*.

No esquema *staging*, são armazenados os dados brutos provenientes de outras fontes de dados, como arquivos Comma-separated values (CSV) originado do IBGE e do

²<https://www.postgresql.org/>

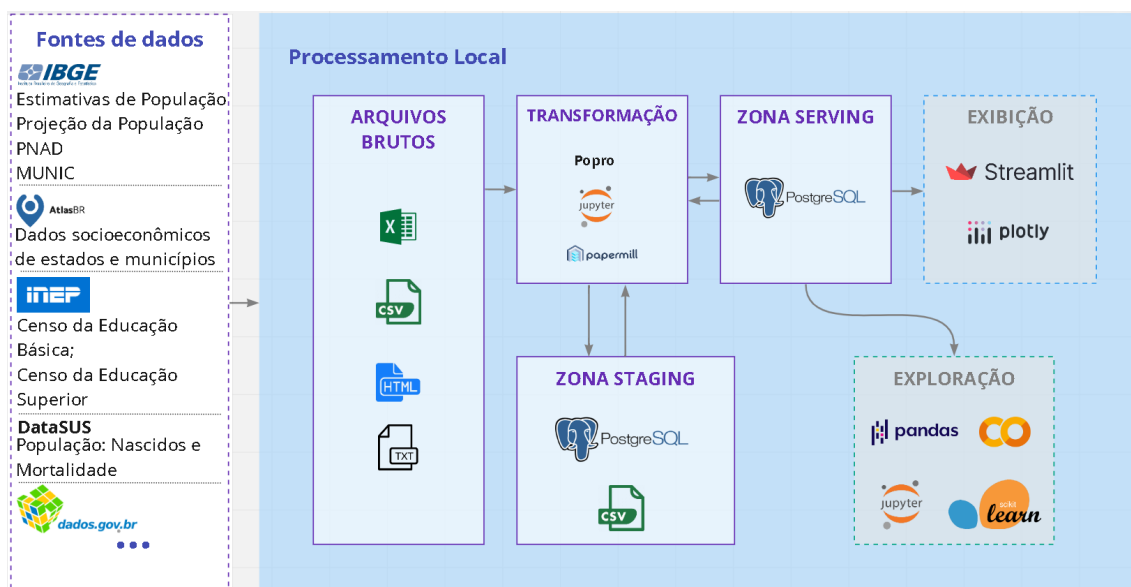


Figura 1. Estrutura da arquitetura de dados educacionais proposta

INEP, e dados de *Application Programming Interface* (API) do Ipeadata³.

No esquema *serving*, armazenam-se as tabelas com dados que foram processados e filtrados conforme os interesses de análise, estando aptas para consumo. As tabelas nessa zona são construídas com dados num formato que capacite análises e filtros por ano e níveis geográficos como estado e município. Estas tabelas também fornecem informações específicas, como os valores já calculados dos indicadores das metas do PNE, as variáveis necessárias para os cálculos e outras informações úteis, como dados socio-econômicos, que capacita uma análise multifacetada. Deste modo, esta zona fornece dados que estão prontos para serem consumidos por ferramentas de visualização de dados e Business Intelligence (BI), que estão presentes nas zonas de EXPLORAÇÃO e EXIBIÇÃO da arquitetura. Um exemplo de como esses dados podem ser consumidos será apresentado no estudo de caso da Seção 4.1.

Vale destacar que as ferramentas tecnológicas utilizadas na criação e instanciação da arquitetura são gratuitas e de código aberto, o que permite a replicação da arquitetura sem custos adicionais. Além do fator de código aberto, tais ferramentas possuiriam uma melhor curva de aprendizado e apoio a implementação de ambiente *on premise*, porém caso deseje ser replicado é possível sua aplicação em ambientes como *Amazon Web service* ou *Google Cloud platform*.

3.1. Aquisição e Processamento de dados

Os dados utilizados atualmente vêm em sua totalidade de fontes Governamentais como o INEP⁴, IBGE⁵, Instituto de Pesquisa Econômica Aplicada (Ipea)⁶, Departamento de Informática do Sistema Único de Saúde (DATASUS)⁷ e de organizações como o Atlas Bra-

³<http://www.ipeadata.gov.br/api/>

⁴<https://www.gov.br/inep/pt-br>

⁵<https://www.ibge.gov.br>

⁶<https://www.ipea.gov.br/portal/>

⁷<https://datasus.saude.gov.br>

sil⁸ mantida pelos órgãos Fundação João Pinheiro (FJP), Ipea e o Programa das Nações Unidas para o Desenvolvimento (PNUD).

Durante a etapa de aquisição de dados, foi possível notar que a disponibilização desses dados sofreu alterações desde o início do projeto em novembro de 2021. Primeiramente, algumas colunas foram atualizadas devido ao tipo de dados e à nomenclatura das colunas informadas. Em janeiro de 2022, foi divulgada a intenção de disponibilizar uma nova versão desses dados para se adequar à Lei Geral de Proteção de Dados (LGPD)⁹. Os dados que antes eram disponibilizados de forma desagregada e individual, com informações anonimizadas de alunos e professores, agora são apresentados em formato sintético, reunidos em um único arquivo e com registros a nível de instituições. Para o ensino superior, é fornecida uma tabela adicional com dados já agregados sobre os cursos. No entanto, essa agregação impede análises mais específicas.

A publicação desse novo formato de dados acarretou prejuízos, como a impossibilidade do cálculo ou atualização de alguns indicadores, pois a versão anterior foi descontinuada em 2020 e alguns indicadores dependem de filtros compostos de variáveis, como idade do aluno e etapa escolar específica. Algumas dessas variáveis necessárias não estão disponíveis no atual modelo de dados publicado. A justificativa oficial divulgada pelo INEP para a mudança estrutural da base de dados foi a necessidade de mitigar o risco de rastreamento de alunos ou professores com base nas informações relatadas no censo em seu formato anterior. Mais detalhes sobre o estudo realizado para a adequação podem ser encontrados no portal de microdados do INEP¹⁰. No entanto, dados gerais, como tipo de etapa de ensino ou qual nível de qualificação realizada pelo docente, são categorias necessárias para o cálculo de indicadores. Diante disso, para esses indicadores, a plataforma apresentará os resultados até o ano de 2020. Para as bases educacionais, foi aplicado o processo descrito em [Barros et al. 2022] para o processamento das bases vindas do INEP, sendo elas o Censo Educacional da Educação Básica e o Censo da Educação Superior.

No que diz respeito aos dados socioeconômicos e populacionais, realizou-se a coleta nos sites oficiais, como o IBGE e DATASUS, que fornecem os insumos essenciais para a criação das tabelas. Os sites do Ipea e PNUD servem como fontes para o enriquecimento dos dados.

As fontes de dados governamentais disponibilizadas pelo IBGE não fornecem estimativas populacionais segmentadas por idade para municípios, o que é fundamental para o cálculo de diversos indicadores do PNE. A divulgação de estimativas populacionais segmentadas por idade para anos posteriores ao último Censo Demográfico realizada pelo IBGE é restrita ao nível de Estado Federativo, o que representa um desafio para criação de indicadores demográficos ao nível municipal.

Para suprir essa lacuna, utilizou-se o processo descrito em [Albuquerque et al. 2022] para gerar as estimativas populacionais segmentadas tanto por idade quanto por municípios, sendo essa etapa imprescindível para o contexto do monitoramento de indicadores e metas do PNE.

Os dados socioeconômicos vindos do PNUD e Ipea apresentam granularidade

⁸<http://www.atlasbrasil.org.br/>

⁹<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados>

¹⁰<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados>

anual e municipal, possibilitando o cruzamento com os dados educacionais para viabilizar o cálculo dos indicadores que envolvem dados populacionais.

3.2. Dicionário de dados

Para cada tabela produzida, foi criado um novo dicionário de dados para auxiliar na utilização e montagem de novas tabelas. Sua produção foi baseada nos dicionários originais fornecidos pelos órgãos oficiais e é acrescido dos novos campos calculados ou modificações aplicadas pelos autores. O dicionário completo pode ser acessado em <https://aiboxlab-pne.github.io/dados/dict/serving/> e <https://aiboxlab-pne.github.io/dados/dict/staging/>.

4. Resultados

A principal contribuição deste trabalho é a disponibilização da arquitetura de dados para consumo por sistemas e ferramentas interessados em serviços de dados para o monitoramento do PNE ou outras atividades correlatas ligadas à melhoria da qualidade da educação no Brasil.

Com o intuito de demonstrar a aplicação da arquitetura de dados desenvolvida, realizamos um experimento que consistiu no cruzamento dos dados do PNE com outras variáveis educacionais apuradas na esfera municipal.

Cada meta do PNE possui um ou mais indicadores associados a ela, que aferem quantitativamente desde o percentual de atendimento de uma faixa etária na escola até a eficiência na gestão escolar. Uma vez que os indicadores educacionais são impactados por fatores diversos [De Witte et al. 2013], convém analisar a relação e possível influência de diferentes variáveis sobre os indicadores do PNE.

Esse tipo de análise pode contribuir para aprimorar as políticas públicas, já que permite identificar fatores que podem estar influenciando no alcance das metas estabelecidas pelo PNE em âmbito federal.

4.1. Estudo de Caso

Para demonstrar a aplicação prática da arquitetura de dados desenvolvida, foi realizado um estudo de caso que será apresentado nesta Seção, visando evidenciar as possibilidades de extração de utilidade a partir dessa arquitetura, auxiliando as equipes de gestão educacional a tomarem decisões baseadas em evidências.

Nesse estudo de caso, utilizamos a base de dados de perfil municipal (MUNIC)¹¹, tendo como referência o Indicador 1A do Plano Nacional de Educação (PNE). A análise comparada dos múltiplos indicadores municipais permitiu avaliar o impacto das políticas educacionais implementadas pelos municípios. Ademais, a base MUNIC apresenta-se como uma fonte valiosa para a análise cruzada dos fatores que podem estar influenciando no alcance das metas estabelecidas em âmbito federal pelo PNE.

Para identificar alguns fatores de influência no Indicador 1A do PNE (Percentual da população de 4 a 5 anos que frequenta a escola/creche), realizou-se um experimento a partir dos indicadores educacionais da base de dados MUNIC. O histórico de dados

¹¹<https://cidades.ibge.gov.br/brasil/pesquisa/1/74454>

educacionais nessa base contemplou os anos de 2014, 2018 e 2021. Para o ano de 2021, foram disponibilizadas 102 variáveis booleanas, embora nem todas essas variáveis estivessem presentes nas bases de dados dos anos 2018 e 2014. De forma geral, essas variáveis indicam se há a incidência ou ocorrência de um dado fator.

Uma resposta verdadeira ao fator é sempre algo semanticamente positivo. Nesse sentido, quanto mais respostas marcadas como verdadeiras, melhor é a situação geral do município. Para possibilitar a avaliação de um município a partir de um conjunto dessas variáveis, foi utilizado a razão entre as respostas verdadeiras e o total de indicadores, resultando no Fator Educacional MUNIC (FEM), o que representa um desafio para criação de indicadores demográficos ao nível municipal.

Partindo dessa base, foram filtradas variáveis com equivalência histórica que possuíam alguma relação tanto com o Indicador 1A, quanto as 13 variáveis pré-selecionadas. Em seguida, ao analisar o comportamento do FEM em relação ao Indicador 1A do PNE, foi identificada uma baixa influência entre as duas métricas, o que demandou uma seleção mais criteriosa a partir das 13 variáveis para formar um subconjunto que explicasse melhor o comportamento do Indicador 1A do PNE. Para realizar essa seleção, foi aplicado um processo para busca de subconjuntos com variáveis, mapeando todas as combinações possíveis de variáveis para poder ser gerado um subconjunto mínimo de 5 elementos, esse processo gerou um total de 7.099 subconjuntos a serem analisados.

Todos os subconjuntos foram analisados considerando sua correlação com o indicador 1A do PNE e seu comportamento histórico. Para explicar o processo de escolha do subconjunto, chamamos de comportamento em “sintonia” quando um município apresenta um comportamento entre 2014 e 2021 de subida, descida ou inalteração do *FEM* acompanhado do mesmo comportamento no Indicador 1A do PNE.

De todas as 7.099 combinações possíveis, o subconjunto de variáveis que gerou maior volume de municípios com comportamento de subida e descida em sintonia entre o *FEM* e o Indicador 1A do PNE, foi composto pelas 5 variáveis listadas abaixo.

1. Plano Municipal de Educação - existência
2. Plano de Carreira para o Magistério - existência
3. Conselho de Educação - existência
4. Conselho de alimentação escolar - existência
5. Conselho de transporte escolar - existência

Considerando o subconjunto escolhido em uma análise dos 233 municípios da Paraíba, foi observado que 96 municípios apresentaram comportamento de subida em sintonia, 2 municípios tiveram comportamento de queda, e 1 município permaneceu estacionário. Assim 109 municípios identificados como em sintonia (44,3% do total) entre o *FEM* e o indicador 1A do PNE. Por outro lado, os demais 121 municípios apresentaram comportamento diferente entre o *FEM* e o Indicador 1A do PNE.

Na Figura 2, é ilustrada uma amostra da relação entre o Indicador 1A do PNE e o fator *FEM* a partir de 2014, demonstrando um movimento de sintonia na melhora ou piora deles, para os municípios de Gado Bravo e Santa Inês, respectivamente. A partir da considerável relação da composição do subconjunto de indicadores municipais e o Indicador 1A do PNE, os gestores municipais têm um ponto de partida alinhado aos

interesses do Plano Nacional de Educação, sobre quais variáveis requerem maior atenção dentre as 5 existentes no subconjunto *FEM*.

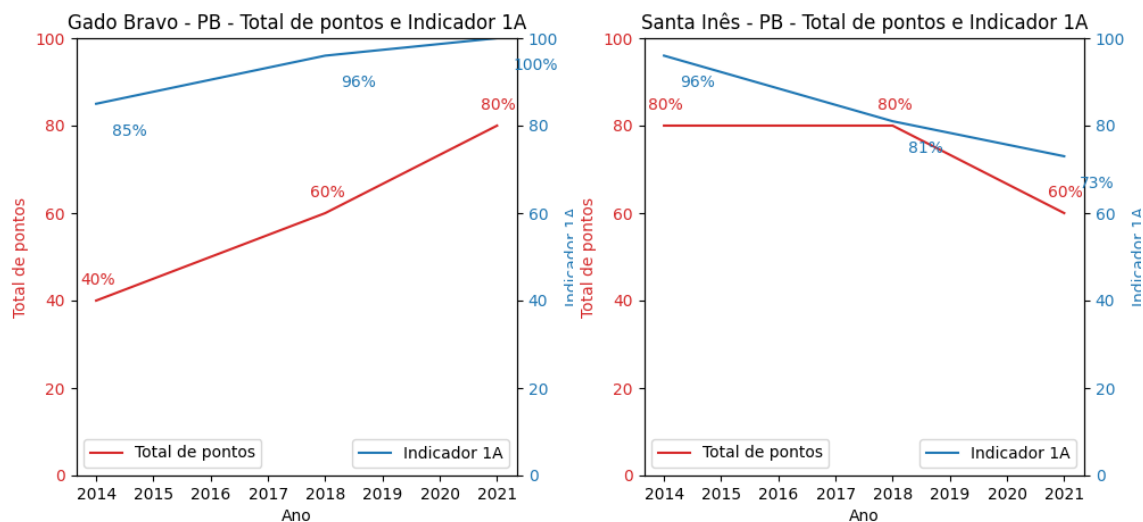


Figura 2. Amostra da relação entre o Indicador 1A do PNE e o Fator Educacional MUNIC (FEM) a partir de 2014.

Após o estudo, é importante ressaltar que uma arquitetura de dados bem estruturada e integrada pode, portanto, potencializar análises e, consequentemente, trazer mais transparência e efetividade para a definição e aplicação das políticas públicas educacionais.

Ressaltando que este é apenas um estudo de caso no qual é possível cruzar algumas das bases disponíveis na plataforma. Os dados disponíveis representam todo o território brasileiro.

5. Conclusão

O presente artigo apresentou o desenvolvimento de uma arquitetura de dados educacionais usando dados abertos governamentais, solucionando a observada necessidade de dados para sustentar serviços e plataformas de apoio a uma gestão educacional baseada em evidências. Observa-se assim a capacidade de uso dessa arquitetura para acompanhar o andamento das metas, estratégias e ações do PNE e, consequentemente, apoiar o processo de tomada de decisão.

A partir do apoio oferecido por essa arquitetura de dados, as secretarias de educação e a população em geral têm acesso aos dados dos indicadores do PNE e demais dados por metodologias censitárias (não amostrais) e todas as variáveis necessárias para a formação destes indicadores. Isso pode auxiliar na avaliação e acompanhamento das políticas educacionais adotadas, bem como contribuir na construção de um Governo Inteligente. O estudo de caso apresentado na Seção 4.1 demonstrou uma das aplicações possíveis dessa arquitetura. Diante disso, evidencia-se o valor potencial no uso dessa arquitetura por gestores educacionais para promover melhorias na qualidade da educação brasileira.

5.1. Trabalhos Futuros

A arquitetura de dados construída neste estudo pode ser expandida para incluir novas metas e serviços, além daqueles já contemplados e apresentados ao longo deste artigo, sendo um esforço a ser dedicado para trabalhos futuros. Além disso, é possível incorporar outras bases de dados que tenham relação direta ou indireta com os indicadores do PNE, como dados de empregabilidade ou de violência. Esses dados podem propiciar grupos de estudos e trabalhos conjuntos entre a secretaria de educação e outras secretarias, em sintonia com as metas definidas pelo PNE. Por fim, é possível realizar melhorias técnicas, como a automação do processo de transformação do pipeline de dados, tornando-o mais robusto e escalável.

Referências

- Albuquerque, A., Barros, A., Alencar, A., Nascimento, A., Bittencourt, I., and Mello, R. (2022). Dataset de estimativas populacionais desagregada por município e idade 2014-2020. In *Anais do IV Dataset Showcase Workshop*, pages 25–34, Porto Alegre, RS, Brasil. SBC.
- Barros, A. N., Alencar, A., Nascimento, A., de Albuquerque, A. F., and Mello, R. F. (2022). Elaboração do conjunto de dados agregados do censo da educação básica. In *Anais do IV Dataset Showcase Workshop*, pages 35–45. SBC.
- Conte, V. d. S. (2019). Mineração de dados educacionais para avaliar os fatores que influenciam no desempenho de candidatos do enem.
- De Witte, K., Cabus, S., Thyssen, G., Groot, W., and van den Brink, H. M. (2013). A critical review of the literature on school dropout. *Educational Research Review*, 10:13–28.
- Diniz, E. (1999). O censo escolar. *Revista Brasileira de Estudos Pedagógicos*, 80(194).
- Júnior, O. d. G. F., Rodrigues, W. R. M., Barbirato, J. C. C., and de Barros Costa, E. (2019). Melhoria da gestão escolar através do uso de técnicas de mineração de dados educacionais: um estudo de caso em escolas municipais de maceió. *RENOTE*, 17(1):296–305.
- Macena, J. d. O., Justino, L. R. P., and Capellini, V. L. M. F. (2018). O plano nacional de educação 2014–2024 e os desafios para a educação especial na perspectiva de uma cultura inclusiva. *Ensaio: Avaliação e Políticas Públicas em Educação*, 26:1283–1302.
- Ramos, G. S., Fernandes, D., Coelho, J. A. P. d. M., and Aquino, A. L. (2022). Data lakes lógicos como plataformas para dados governamentais em sociedades e cidades inteligentes. In *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, pages 215–226. SBC.
- Rodrigues, F. A. and Maciel, C. (2022). Um método para captura e compartilhamento de dados abertos educacionais via um processo etl. In *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, pages 133–144. SBC.
- Vinente, S. and Duarte, M. (2015). O plano nacional de educação (2014-2024) e a garantia de um sistema educacional inclusivo: possibilidade ou utopia. *Olhares: Revista do Departamento de Educação da Unifesp*, 3(2):133–151.

Wanderley, P. F. (2021). Uso de business intelligence para avaliação de indicadores de desempenho na educação básica: um estudo de caso no estado do acre.

4 Apêndice B - *Workflow* para aquisição, processamento e disseminação de dados públicos brasileiros voltados à educação

Nogueira Barros, A., A. Félix de Albuquerque, A. Leite de Alencar, A. Nascimento, I. Mateus Bittencourt, and R. Ferreira Mello. “Workflow for the Acquisition, Processing, and Dissemination of Brazilian Public Data Focused on Education”. *Journal of Information and Data Management*, vol. 15, no. 1, Apr. 2024, pp. 224-33, doi:10.5753/jidm.2024.3570.

Contribuição do autor: Responsável pela implementação da metodologia, implementação da documentação, execução e avaliação dos resultados educacionais e escrita do texto.

Workflow para aquisição, processamento e disseminação de dados públicos brasileiros voltados à educação

Abílio Nogueira Barros   [Universidade Federal Rural de Pernambuco | abilio.nogueira@ufrpe.br]

Aldéryck Félix de Albuquerque  [Cesar School | derycck@gmail.com]

Andréza Leite de Alencar  [Universidade Federal Rural de Pernambuco | andreza.leite@ufrpe.br]

André Nascimento  [Universidade Federal Rural de Pernambuco | andre.nascimento@ufrpe.br]

Ibsen Mateus Bittencourt  [Universidade Federal de Alagoas | ibsen@feac.ufal.br]

Rafael Ferreira Mello  [Universidade Federal Rural de Pernambuco | rafael.mello@ufrpe.br]

Abstract Este artigo visa demonstrar o processo de criação de bases de dados públicas voltadas a área educacional e populacional. Descrevendo o processo de obtenção, quais procedimentos de adaptação dos dados e otimização para criar sua série histórica, além quais boas práticas foram seguidas para seu desenvolvimento e metadados gerados. Destacando suas particularidades entre as temáticas de educação e população, relatando seus desafios e particularidades de cada conjunto. É relatado também quais resultados já podem ser obtidos diretamente por cada conjunto e como unidos podem realizar o acompanhamento de indicadores do Plano Nacional da Educação, uma das maiores políticas públicas brasileiras voltadas a educação.

Keywords: Conjuntos de dados, Dados abertos, Dados públicos, Dados educacionais, Governo inteligente

1 Introdução

A busca por aprimorar a qualidade e eficiência das políticas públicas educacionais é um desafio constante para governos e instituições. Nesse contexto, dados precisos e abrangentes desempenham um papel fundamental na elaboração de métricas e indicadores que embasam a construção de planos educacionais sólidos e na avaliação da efetividade das ações implementadas. Dentre os dados essenciais para esta elaboração, destacam-se as informações sobre matrículas dos alunos em diferentes municípios, aliadas ao conhecimento do volume populacional segmentado por idade, que proporcionam uma fonte de informação valiosa para orientar estratégias educacionais, elaborar métricas e indicadores que servem tanto de ponto de partida para construção de modelos que sustentam a construção de planos educacionais, como para acompanhar a execução dos planos elaborados.

Essas informações possibilitam uma maior compreensão da demanda e oferta de educação em diferentes regiões, identificando áreas com maior concentração de alunos em determinadas faixas etárias e onde as necessidades educacionais são mais prementes. Além disso, tais dados permitem a análise da distribuição da população em idade escolar, auxiliando na projeção de futuros contingentes de alunos e preparando o sistema educacional para acomodar essas demandas. A segmentação por municípios também possibilita uma abordagem mais localizada, considerando particularidades regionais e permitindo a formulação de estratégias específicas para atender às necessidades educacionais de cada localidade. Com base nesses dados analíticos sólidos, o Governo pode tomar decisões embasadas e direcionar recursos de forma mais eficiente, garantindo a implementação de políticas educacionais mais assertivas e, consequentemente, contribuindo para o avanço e a melhoria do sistema educacional do país.

O presente trabalho busca construir conjuntos de dados educacionais e populacionais segmentados por município e idade através de um detalhado processo de transformação e modelagem estruturada com o objetivo de facilitar o consumo desses dados. Observa-se nas fontes de dados públicos educacionais, uma coleção que requer um tratamento de tradução de termos com base num dicionário de dados, bem como um enriquecimento desta coleção para contextualizar estes dados e permitir futuros processo de agregação, filtro, seleção e demais operações necessárias para extrair valor deste material.

Diante do exposto, a fonte principal dos dados trabalhados neste estudo foi estabelecida com o Censo Educacional coordenado pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP), sendo este Censo Educacional regulamentado por instrumentos normativos que instituem a obrigatoriedade, os prazos, os responsáveis e suas responsabilidades, bem como os procedimentos para realização de todo o processo de coleta de dados.¹

Este artigo trata-se da versão estendida do trabalho Elaboração do conjunto de dados agregados do censo da educação básica publicado no DSW-2022. Essa versão estendida foi acrescida do trabalho Dataset de Estimativas populacionais desagregada por município e idade 2014-2020, também publicado no DSW-2022. E como uma versão estendida foi adicionada a análise sobre a utilização dos dados educacionais acrescidos desses dados populacionais do trabalho também citado, de modo a demonstrar seu impacto em políticas públicas, que no caso desse projeto se do Plano Nacional da Educação, uma das maiores políticas públicas educacionais brasileiras.

Este artigo encontra-se organizado da seguinte maneira: a Seção 2 apresenta alguns trabalhos relacionados ao uso de da-

¹<https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar>

dos educacionais e populacionais na construção de bases de dados públicas e suas utilizações, além de discutir formas de se estimar população; na Seção 3 é apresentada a forma com que os dados educacionais e populacionais foram desenvolvidos, desde sua aquisição até a disponibilização das bases de dados junto a seus metadados; na Seção 5 são demonstrados alguns dos resultados que já podem ser atingidos com o uso individual desses dados e depois aprofundando mostrando com o estudo de caso de um plano nacional; e, por fim, a Seção 6 traz as considerações finais e trabalhos futuros que possam realizar a melhoria e otimização do trabalho já realizado.

2 Trabalhos relacionados

Para criar a base de conhecimento de trabalhos anteriores, foi necessário buscar sobre o uso dados públicos gerais, sejam eles educacionais, populacionais e até da área da saúde, e principalmente como esse dados podem ter sido aplicados para o acompanhamento de políticas públicas.

No campo de construção de bases de dados de bases públicas é importante destacar o caso demonstrado em Vasconcelos *et al.* [2021], onde foram recolhidos e padronizados dados diretamente do portal do TSE. Após etapas de aquisição e processamento dos dados brutos, foram adicionados novos dados geográficos com o objetivo de fornecer mais usabilidade aos utilizadores das bases.

Transportando a temática para a área da saúde temos em Gonçalves *et al.* [2021] a coleta, tratamento e disponibilização de base de dados sobre a vacinação registrados na plataforma **OpenDataSUS**. O processo se baseou sobre características de processos de ciência de dados, realizando seu desenvolvimento de forma interativa até atingir o objetivo de disponibilização do conjunto de dados e dicionário de dados.

Para embasar a parte que trata sobre dados populacionais foi necessário buscar trabalhos que já tratassem da ótica populacional e como trabalhar com essa desagregação.

Trabalhos recentes têm investigado métodos de projeção estatística de dados demográficos essenciais para a construção de políticas públicas. Gonzaga and Schmertmann [2016] buscou desenvolver um modelo de estimativa de mortalidade baseado em idade para pequenas áreas geográficas no Brasil partindo dos dados do Censo Demográfico 2010 e dados do Sistema público de Saúde (*DataSUS*). A variável de mortalidade possui um importante papel para estimar a população de um local num dado ano, de modo que os autores propõem um sofisticado modelo estatístico para a sua projeção. Mas algumas observações sobre a cronologia para obtenção de dados bases para a construção de estimativas populacionais no cenário brasileiro precisam ser destacadas. O censo demográfico possui uma frequência de atualização decenária, enquanto os dados reais de mortalidade e nascimentos possuem frequência de atualização anual. Sendo assim, no período entre o último censo e o ano anterior à análise, não se faz necessário estimar dados de mortalidade para a composição do cálculo de estimativa populacional uma vez que esses dados podem ser obtidos como fatos.

Pode-se observar que os dados factuais atualizados que podem ser obtidos de fontes governamentais, possuem uma

forte influência no processo metodológico para gerar os dados populacionais desagregados por pequenas regiões e idade. Possibilitando inclusive, simplificar a desagregação, por não ser necessário estimar certas variáveis de entrada ou estabelecer alguma premissa. Como exemplo, podemos citar os dados nominais de "Nascidos" ou de "Mortalidade" com o nível de desagregação desejado. Caso estes dados sejam obtidos de modo atualizado, não será necessário a aplicação de um processo de estimativa de variáveis bases para compor o cálculo de estimativa populacional.

Além do processo de construção de estimativas populacionais, um outro conceito relevante que fundamenta o processo de modelagem dos dados, é o da desagregação populacional em relação à estimativa populacional. Além de destacar que modelos de projeções populacionais são instrumentos fundamentais utilizados na implementação de diversos programas e políticas públicas, especialmente os relacionados à sustentabilidade dos sistemas públicos de aposentadoria, saúde e educação, o estudo promovido pelo González *et al.* [2015] distingue esses conceitos, destacando que o processo de desagregação populacional parte de dados demográficos preexistentes fornecidos por instituições governamentais de estatística, para posteriormente implantar uma metodologia que amplie o grau de granularidade desses dados, seja por etnia, classe social, sexo, idade e entre outros.

Dado a temática populacional possuir menos abordagens disponibilizadas seu estudo de embasamento foi mais denso e focal em quais abordagens poderiam ser aplicadas para o processo desse estudo. De posse desses conceitos e ponderações sobre variáveis significantes, a metodologia de desagregação populacional que se mostrou mais aderente ao escopo temporal e ao cenário de obtenção de dados factuais, foi a metodologia de Atividades Especiais TCE-SC [2021], desenvolvida pela Diretoria de Atividades Especiais do Tribunal de Contas do Estado de Santa Catarina. Nela o cenário de ausência de dados sobre migração entre os municípios foi contornado pela variação das estimativas populacionais dos respectivos municípios num período temporal. Assim o conjunto de dados que fundamenta esta metodologia se resume em dados censitários, dados de nascimentos, mortalidade infantil e estimativas populacionais dos municípios que não possuem granularidade etária.

Sobre a temática educacional, por já termos mais fontes de dados consolidadas foi possível aplicar técnicas de mineração de dados para construir a base necessária a responder os principais indicadores alvo desse estudo.

3 Aquisição de processamento dos dados

Nessa etapa do projeto, o tipo de processamento se desvinculam para suas determinadas áreas. Para a temática educacional o principal fornecedor que o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)², já para a distribuição de dados geográficos temos o Instituto Brasileiro de Geografia e Estatística (IBGE)³. Dado as di-

²<https://www.gov.br/inep/pt-br>

³<https://www.ibge.gov.br>

vergências demonstraremos em dois fluxos diferentes o processo aplicado a cada temática de dados.

Destacaremos as principais etapas do processo que levou a criação das bases de dados, o processo detalhado e completo da base educacional pode ser encontrado em Barros *et al.* [2022], e o processamento completo da parte populacional pode ser visto em Albuquerque *et al.* [2022].

3.1 Dados educacionais

O processo de desenvolvimento desse conjunto de dados foi desenvolvido tendo como base o processo de ETL (Extract, Transform, Load) assim como definido em Ferreira *et al.* [2010] é um processo de extrair os dados de suas fontes, de preferências originais, aplicarem os processos necessários a fim de transformar aquela base da forma mais acessível e direta para a utilização, e o *Load* vem do carregamento dentro a um local de armazenamento mais definitivo, onde no escopo deste projeto será utilizado um conjunto de arquivos no formato CSV (Comma-separated values).

Todo o processamento foi elaborado e executado com o objetivo final de consolidar uma base do censo da educação básica, onde a faixa temporal escolhida para esse projeto foi a de 2010 a 2021, criando assim um conjunto que possa ser analisado e incorporado a outras bases de dados públicas. Para a correta elaboração do fluxo de processamento de dados foi levantado leituras sobre o tema e principalmente o entendimento dos dicionário de dados divulgados pela publicadora, resultando no seguinte processo ilustrado na figura 1.

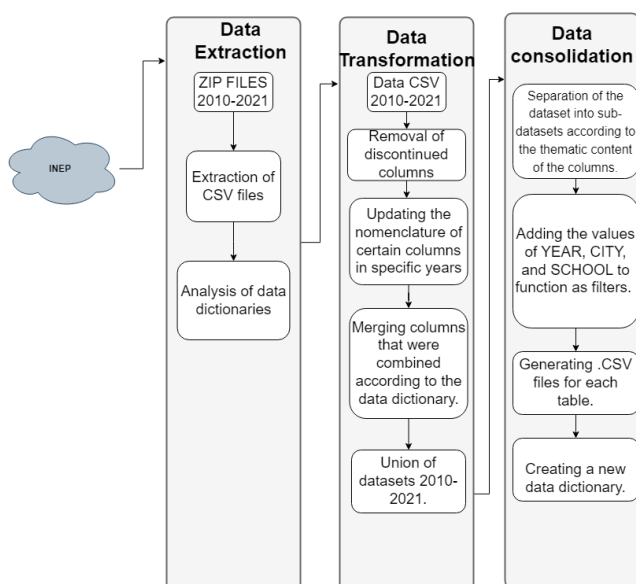


Figure 1. Processing pipeline of the used data.

3.1.1 Extração dos dados

Os dados originais deste projeto foram inteiramente retirados do site do INEP. Esses dados são fornecidos no formato compactado a fim de facilitar o descarregamento para a máquina local. Os dados originalmente são fornecidos anualmente, são fornecidos anualmente um arquivo no formato CSV contendo os dados e um dicionário de dados no formato de

planilhas do excel, além de outros *metadados* como: os questionários aplicados às instituições; um manual do usuário indicando informações para uso em determinados *softwares*

3.1.2 Transformação dos dados

Após a leitura dos dicionários de dados foi possível elencar quais alterações eram necessárias serem aplicadas visando unir todos os anos elencados como objetivo do projeto. Foi observado que grande parte das colunas informadas já foram padronizadas, visto que já existiram atualizações realizadas pelo próprio INEP nas versões através dos anos, entretanto existiam colunas que ainda impediriam essa junção e algumas outras que foram descontinuadas em relação aos anos mais recentes.

1. **Remoção de colunas descontinuadas:** Foram removidas as colunas descontinuadas a partir de 2019, pois os dados mais atuais não iriam mais refleti-las e assim não devem trazer informações tão atuais para o acompanhamento da estrutura da rede educacional. Do total das 370 variáveis disponibilizadas, 36 foram removidas. Restando assim 334 para o processamento das etapas seguintes.
2. **Atualização da nomenclatura das colunas:** Seguindo para o processamento das colunas remanescentes certas alterações precisaram ser aplicadas para que fosse possível uma adequação de nomenclatura e variáveis para a faixa temporal aplicada. Tal necessidade foi apontada pela leitura dos dicionários de dados disponibilizados junto aos arquivos.
3. **Junção das colunas:** Algumas colunas tiveram seus nomes alterados depois de 2018, com isso foi necessário padronizar essas colunas para os nomes mais atuais. Colunas essas que não tiveram seus valores alterados, apenas a nomenclatura, esse processo foi realizado seguindo o padrão indicado pelo dicionário de dados. Outras colunas tiveram que ser unidas a partir do ano de 2019, sendo assim será utilizado a estratégia do **OU**, pois essas colunas são especificações de uma mesma resposta, sendo assim basta que apenas uma das colunas tenham seu resultado atendido que será aplicado o valor de Verdadeiro para a nova coluna, caso não tenham nenhum valor positivo o valor falso será atribuído a coluna.

4. **Consolidação dos dados:** A última etapa desse processo se propôs a unir todo o dado que foi processado anualmente em uma única tabela de dados. Diferentemente da abordada elencada pelo INEP, foi proposta uma separação das colunas agrupando-as por temática para que o dado, caso deseje-se ser carregado apenas algum grupo de colunas, que nesse caso foi separado por temáticas como matrículas, docentes e turmas (separação já aplicada anteriormente pelo INEP em versões anteriores de divulgação do censo), resultado. As colunas removidas, combinadas e remapeadas estão disponíveis nos metadados que podem ser encontrado junto a base de dados em <https://zenodo.org/record/6666613#.YrioP3bMLnY>.

Table 1. Tabelas geradas por tema.

Tabela	Descrição
TB_GEOGRAFICA	Dados geográficos onde a instituição de ensino está localizada.
TB_TURMA	Informações sobre as turmas de uma instituição de ensino específica.
TB_DOCENTE	Dados quantitativos sobre os docentes pertencentes à instituição de ensino.
TB_MATRICULA	Matrículas realizadas na instituição de ensino.
TB_ESCOLA_ADMINISTRATIVO	Dados sobre a instituição de ensino.
TB_ESCOLA_ESTRUTURA	Dados sobre a infraestrutura da instituição de ensino.
TB_ESCOLA_FUNCIONARIO	Dados sobre os funcionários não docentes da instituição de ensino.
TB_ESCOLA_PEDAGOGICO	Dados sobre a estrutura pedagógica da instituição de ensino.

3.2 Dados populacionais

Os dados para a implantação da metodologia de desagregação populacional escolhida foram obtidos do IBGE e do Ministério da Saúde. Como destacado na nota técnica que formaliza a metodologia, os dados obtidos foram:

1. Obtidos pelo IBGE:

- Dados censitários segmentados por ano e idade;⁴
- Estimativas populacionais anuais dos municípios. (Não possui segmentação etária.)⁵

2. Obtidos pelo Ministério da saúde:

- Dados do Sistema de Informações de Nascidos Vivos;⁶
- Dados do Sistema de informações sobre Mortalidade.⁷

3.2.1 Aquisição e processamento dos dados

O processo ilustrado na figura 2 segue um fluxo onde dados de entrada apresentados na seção anterior são limpos e normalizados uma vez que alguns deles são disponibilizados em planilhas eletrônicas com observações textuais em células de valores e fazem referência a municípios de forma textual ao invés de seu código único normalizado pelo IBGE.

De modo geral, o primeiro desafio em comum no tratamento dos dados de entrada para todos os relatórios apurados, está na presença de textos explicativos no cabeçalho e rodapé dos relatórios CSV extraídos tanto do *DataSUS* quanto do *IBGE*, não havendo um padrão na quantidade de linhas de comentários existentes nestes dois locais.

Entrando em detalhes, os relatórios de "Nascimentos" e "Mortalidade com idade de zero ano" extraídos do *DataSUS* em formato CSV possuem apenas duas colunas, onde a primeira delas é rotulada como "Município" e contém um texto que concatena o código do município normalizado pelo *IBGE* e o nome do município. Além desta característica de concatenação que exige tratativa, existem algumas poucas linhas em cada relatório em que o código do município não é válido quando consultado na tabela de código de município do *IBGE*. Mas o somatório da quantidade populacional

de linhas com código de município inválido representa uma parcela insignificante, sendo por volta de 1 a cada 30.000 pessoas contabilizadas (0.0033%).

No relatório do Censo Demográfico fornecido pelo *IBGE*, as idades são dispostos em diferentes colunas e os municípios em diferentes linhas, formando uma matriz de dados com milhares de linhas e dezenas de colunas. É observado que para algumas raras combinações de municípios e idade, a quantidade populacional não é apresentada, sendo informado um valor nulo. Considerando o volume populacional de cada idade nos municípios em que apresentam casos do tipo, parece pouco provável que a causa destes valores nulos seja a ausência de pessoas daquela idade no município em questão, mas sim que a causa reside no processo de apuração dos dados de pesquisa ou de processamento na construção do relatório, passiva de resolução pelo responsável por seu desenvolvimento.

Por fim, o relatório de Estimativas Populacionais também fornecido pelo *IBGE*, tal como os anteriores, possui informações textuais no cabeçalho e rodapé, mas também possui na coluna de quantidade certas marcações indicando notas de rodapé que esclarecem algum detalhe metodológico que levou àquele número reportado. Estas marcações são compostas por números entre parênteses, o que modifica a estrutura da coluna deixando de ser puramente numérica, o que mais uma vez demanda transformação dos dados do relatório em formato tabulado com tipo de dados consistentes em cada coluna.

Com o término do processo de limpeza, são construídas três tabelas auxiliares normalizadas que juntamente com o metadado do ano em que os dados censitários estão relacionados, serão utilizadas como variáveis de entrada na ferramenta de desagregação populacional apresentada na seção 3.2.2.

As tabelas auxiliares são de Nascidos sobreviventes (Ano, Local, Nascidos), População real do último censo segmentada por idade e local (Idade, Local, População) e Estimativa Populacional por Município (Ano, Local, Estimativa).

3.2.2 Desagregação Populacional

O processo de desagregação populacional possui duas abordagens principais dependendo do ano objetivo para a análise e o ano de nascimento do grupo etário em análise. Em resumo, a metodologia parte de uma base populacional e a envelhece com base da flutuação da estimativa populacional para o município em análise ao longo do período necessário.

Esta base populacional pode ser de duas fontes depen-

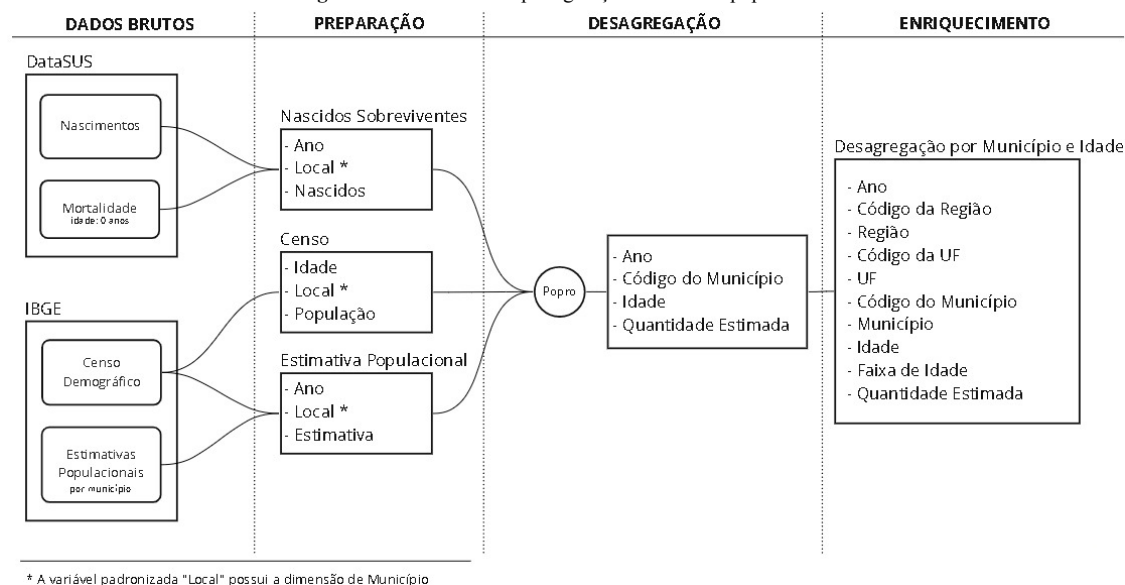
⁴<https://sidra.ibge.gov.br/tabela/200>

⁵https://ftp.ibge.gov.br/Estimativas_de_Populacao

⁶<https://datasus.saude.gov.br/nascidos-vivos-desde-1994>

⁷<https://datasus.saude.gov.br/mortalidade-desde-1996-pela-cid-10>

Figure 2. Fluxo de dados para geração do Dataset populacional



dendo do ano de nascimento do grupo etário em análise. A base populacional será censitária quando o conjunto populacional em um determinado ano objetivo, possuir um ano de nascimento igual ou anterior a execução do último Censo. A base populacional será baseada em volume de nascidos sobreviventes para os demais casos, ou seja quando o conjunto populacional em um determinado ano objetivo possuir um ano de nascimento posterior à execução do último Censo. O modelo matemático e uma discussão mais detalhada sobre ele podem ser encontrados na nota técnica de Atividades Especiais TCE-SC [2021] mencionada que formaliza esta metodologia.

Esta desagregação se deu através da utilização da ferramenta Popro Albuquerque [2022], uma biblioteca de desagregação populacional desenvolvida pelos autores na linguagem de programação Python e expansível para diferentes metodologias, com o modelo estatístico da metodologia citada sendo implementado como *plugin* para a geração do presente *dataset*.

3.2.3 Dataset resultante

Ao final das etapas de extração, tratamento, desagregação e enriquecimento dos dados, o *dataset* resultante é capaz de sustentar planos de ações e balizar o estabelecimento de metas em políticas públicas. Devido a metodologia de desagregação implantada possuir um grau de confiança que decai na medida que a idade estimada aumenta, é relevante ressaltar que a base de dados disponibilizada é mais apropriada para suportar planos envolvendo faixas etárias de crianças e adolescentes, como políticas educacionais de nível fundamental e médio.

O *Dataset* resultante possui por granularidade as variáveis ANO, MUNICÍPIO e IDADE. Com adição de dados importantes para futuras análises e agrupamentos, como o código do município, nome da unidade federativa e o nome da região geográfica, ambas com seus respectivos códigos normatizados pelo IBGE.

4 Boas práticas de divulgação de conjunto de dados

Essa etapa por ser comum a ambos os conjuntos de dados seguiu um mesmo procedimento, mesmo que produzindo artefatos ligeiramente diferentes buscaram atender as seguintes boas práticas de publicação de dados listadas em ⁸

Ao alcançar as metas apontadas na 2 conseguimos fornecer uma base de dados que seja possível ser rastreada tanto quanto a seus dados de origem (BP5), quanto qual versão está sendo utilizada e qual época de publicação (BP7), além da disponibilização dos metadados atualizados para uma coerente utilização dos dados (BP1, BP2, BP3).

A disponibilização dos dados bem como os metadados responsáveis pela aplicação das boas práticas podem ser encontrados na plataforma ZENODO⁹:

- Educacional: <https://zenodo.org/record/6666613#.YrioP3bMLnY>
- Populacional: <https://zenodo.org/record/6689160#.YrdBeHbMK5c>

Os arquivos estão dispostos sobre a licença CC-BY-4.0 onde, esta licença permite que outros baixem, ajustem e desenvolvam esse trabalho, mesmo que seja utilizado para fins comerciais, contando que esse projeto seja referenciado como base.

5 Aplicações das bases de dados

Quando abordamos a temática de resultado sobre a criação de conjunto de dados públicos devemos pensar primeira sobre a progressão ou regressão de certas faixas etárias, acesso

⁸<https://ceweb.br/media/docs/publicacoes/1/fundamentos-publicacao-dados-web.pdf>

⁹<https://zenodo.org/>

Table 2. Boas práticas atingidas pelas bases

Código	Definição
BP1	FORNECER METADADOS
BP2	FORNECER METADADOS DESCRITIVOS
BP3	FORNECER METADADOS ESTRUTURAIS
BP5	FORNECER INFORMAÇÕES DE PROVENIÊNCIA DOS DADOS
BP7	FORNECER INDICADOR DE VERSÃO

à educação, quantitativo de professores em determinada localidade, todos esses dados de forma isolada já produzem um certo teor informativo e pode ajudar numa formação de sociedade colaborativa. Uma segunda ótica de observação é no cruzamento desses dados para realizar o acompanhamento de planos e políticas públicas que esses conjunto de dados podem fornecer uma maneira de enxerga-los até a nível municipal e compreendendo todos os municípios brasileiros, criando assim uma forma plural e não excludente de acompanhamento.

A seguir temos demonstrações de utilizações nesses dois tópicos, tanto os conjuntos de dados de forma atômica quanto combinados para realizar o acompanhamento das metas do Plano Nacional da Educação (PNE), umas das principais políticas públicas na área educacional do Brasil.

5.1 Utilização das bases para acompanhamento educacional

Isoladamente, os dados já podem nos fornecer um panorama quanto a presença de certos fatores durante a série histórica abordada nos estudos.

Na temática educacional são inúmeras possibilidades de aplicação desses dados, como em Balbinot and Haubert [2017] a aplicação desses microdados censitários foi utilizada para o acompanhamento de indicadores sobre educação especial no Rio de Janeiro, já em Balbinot and Haubert [2015] foi realizada uma análise temporal das matrículas do estado do Paraná. São alguns exemplos sobre trabalhos já realizados com base nesses dados educacionais.

Isoladamente podemos utilizar ferramentas de visualização de dados como a Plotly(<https://plotly.com/>) onde junto a ferramentas da linguagem Python¹⁰ nos possibilita algumas visualizações sobre a situação escolar de um determinado município, buscando entender qual a situação refletida no tempo levantado pela base.

Na figura 4 podemos ver o decréscimo da quantidade de professores do município de Moreno-PE, os professores das três etapas de ensino (Azul representando ensino básico, vermelho o fundamental e verde o ensino médio), é de fácil constatação que naquela localidade a maior quantidade de professores disponíveis é do ensino básico e podemos buscar observar a quantidade de matrículas se estabelece esse mesmo tipo de padrão.

Seguindo na utilização da base, podemos acompanhar na figura 5 que, no mesmo exemplo de cores e legendas Azul

representando matrículas do ensino básico, vermelho matrículas do ensino fundamental e verde matrículas do ensino médio), a quantidade majoritária de matrículas é pertencente a faixa de educação básica e que, igualmente a quantidade de professores da rede de ensino, a quantidade desse tipo de faixa educacional vem caindo, porém ainda se mantendo em maior número em relação aos demais.

5.1.1 Acompanhamento de faixas populacionais entre os anos

O utilização de um *Dataset* populacional segmentado por idade em pequenas áreas geográficas tem por objetivo principal apoiar a elaboração de planejamentos de políticas públicas, bem como aferir métricas em vários tipos de instituições, sejam públicas ou privadas.

Como exemplo de utilidade na esfera privada no segmento educacional, uma instituição de ensino básico com base neste *Dataset* e seu próprio banco de dados de alunos matriculados poderá mensurar a representatividade de seus alunos em relação aos potenciais estudantes do município onde a instituição está localizada, calculando assim o indicador de quota do mercado (*Market Share*) como discutido no estudo Ozkan et al. [2022], sendo importante para guiar as decisões estratégicas que transitam entre investir com maior foco na captação de novos alunos, na retenção dos alunos já matriculados ou na qualificação do serviço ofertado.

Já na esfera pública, como exemplo de coleta de dados para capacitar a análise da segurança pública por uma abordagem metodológica, o *Dataset* apresentado possibilita a formulação de indicadores como taxa de crimes cometidos contra determinada faixa etária da população, aferindo seu comportamento ao longo do tempo e capacitando a identificação de padrões e tendências.

O desenvolvimento desse tipo de base dados traz a possibilidade do acompanhamento da progressão das faixas de idade, afinal o censo populacional do IBGE é realizado a cada década, criando assim a necessidade projeções dessas outras faixas de modo a realizar um maior entendimento de como está o comportamento de certas faixas de idade através dos anos, fazendo com que seja observado o avanço ou recuo da quantidade de população estimada para aquela determinada faixa estipulada. Podemos observar para o município do Recife, na Figura 5.

As tabelas foram organizadas de tal modo a possibilitar a realização de agregações por meio do código do estado e região. Com isso basta que sejam aplicados agrupamentos com o auxílio de alguma ferramenta ou linguagem para processamento de dados. Na Figura 6 podemos observar o

¹⁰<https://www.python.org/>

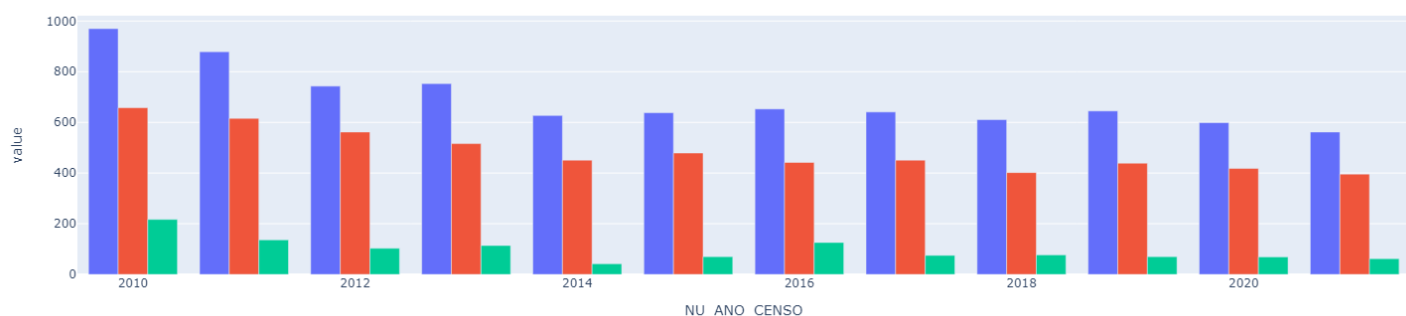


Figure 3. Quantidade de Professores por ano e grau do Município de Moreno-PE

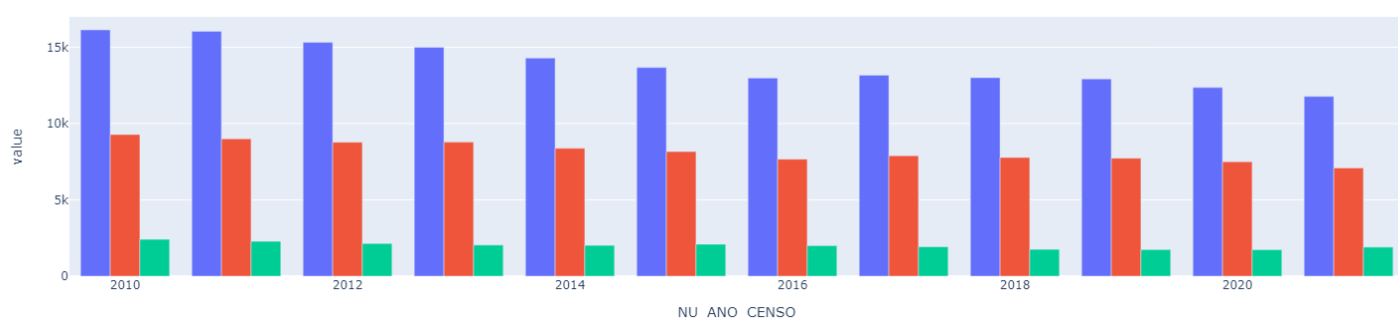
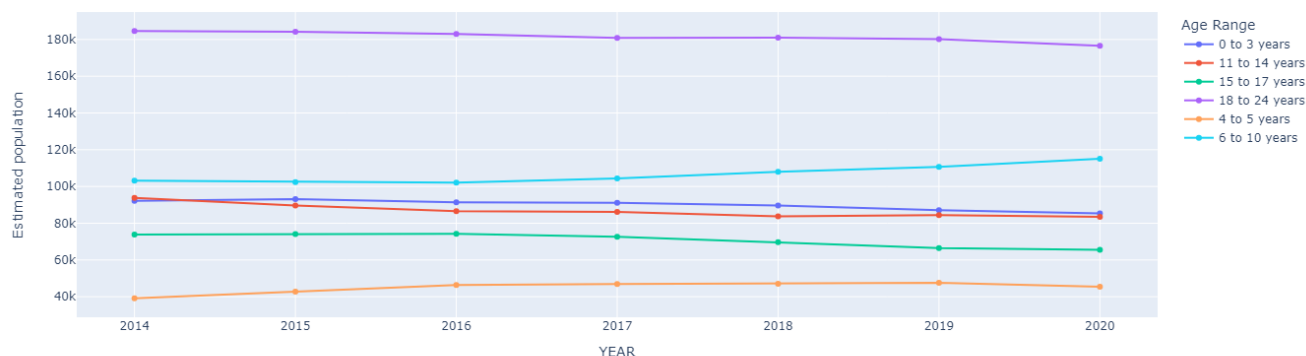


Figure 4. Quantidade de Matrículas por ano e grau do Município de Moreno-PE

Figure 5. Progressão de estimativa populacional por faixa do município do Recife.



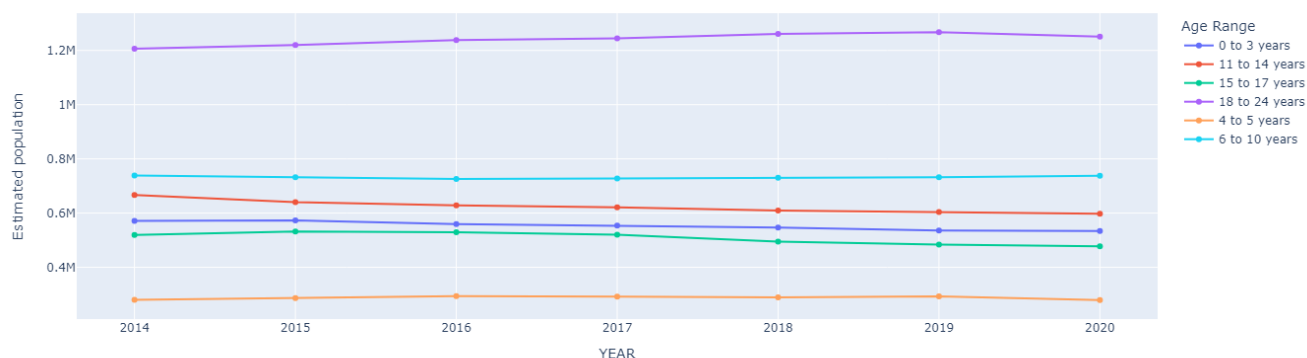
mesmo exemplo aplicado anteriormente, desta vez para o estado de Pernambuco.

5.2 Acompanhamento ao PNE

Todo o desenvolvimento desse trabalho foi baseado em realizar o acompanhamento de uma das maiores políticas públicas voltadas a educação, o PNE, é uma política pública de longo prazo que estabelece diretrizes e metas para o desenvolvimento da educação no Brasil ao longo de um período de dez anos. Ele foi criado visando promover uma educação de qualidade, inclusiva e equitativa em todos os níveis, desde a educação infantil até o ensino superior. Sendo assim uma fer-

ramenta fundamental para orientar as ações do governo e da sociedade civil na busca por uma educação mais igualitária e capaz de impulsionar o desenvolvimento social e econômico do país.

A importância do acompanhamento do Plano Nacional de Educação reside no fato de que esse monitoramento é essencial para garantir que as metas estabelecidas sejam alcançadas efetivamente. Ao avaliar periodicamente o progresso das ações, é possível identificar avanços e também eventuais desafios e obstáculos a serem superados. O acompanhamento permite ajustar estratégias, redistribuir recursos e direcionar esforços para áreas prioritárias, garantindo que a educação alcance as populações mais vulneráveis e carentes

Figure 6. Progressão de estimativa populacional por faixa do estado de Pernambuco.

de investimentos.

Por ser um plano de vigência de uma década, sua meta a serem atingidas e planos de ação para atingimento de seus indicadores precisam ser compreendidos por todos os governos que atuem nesse espaço de tempo, independente de qual partido seja o atual responsável pela governabilidade dos entes federais, estaduais e municipais. Necessitando assim de um compromisso firme de cooperação entre todos os entes governamentais através dos anos.

Ao avaliar tudo isso, o acompanhamento do PNE é de vital importância tanto para que ele seja efetivado como um instrumento de governo colaborativo, quanto para sua funcionalidade como um avaliador geral da educação pública e a efetividade das ações que buscam realizar sua melhoria. Ao falarmos do Brasil e seus 5570 municípios, já se é possível mensurar a dificuldade do acompanhamento de todas as 20 metas com em média 2 indicadores presentes no PNE vigente (2014-2024)¹¹.

O INEP fornece relatórios bianuais¹² destacando os resultados ao nível federal, estadual e de algumas capitais brasileiras, porém o intuito da criação desses conjuntos de dados é poder mapear a totalidade da quantidade de municípios onde os dados estejam disponíveis desde sua fundação (dado que alguns municípios foram emancipados posterior ao censo de 2010 gerando a falta de dados em determinados anos).

Para esses relatórios desenvolvidos pelo INEP são utilizadas algumas fontes amostrais como a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD), que não reflete a totalidade dos municípios brasileiros, com isso uma das principais criações e união dessas bases é reproduzir o cálculo dos indicadores 1A, 1B, 2A, 3A e 3B para representar todos os municípios brasileiros, inicialmente de 2014-2020.

Com o conjunto de dados gerado, é possível calcular os seguintes indicadores para todos os municípios brasileiros:

$$1A = \frac{\text{Quantidade de matrículas 4 a 5 anos}}{\text{População estimada de 4 a 5 anos}} \times 100$$

$$1B = \frac{\text{Quantidade de matrículas 0 a 3 anos}}{\text{População estimada de 0 a 3 anos na escola}} \times 100$$

$$2A = \frac{\text{Quantidade de matrículas 6 a 14 anos}}{\text{População estimada de 6 a 14 anos}} \times 100$$

$$3A = \frac{\text{Quantidade de matrículas 15 a 17 anos}}{\text{População estimada de 15 a 17 anos}} \times 100$$

$$3B = \frac{\text{Quantidade de matrículas 15 a 17 anos no ensino médio}}{\text{População estimada de 15 a 17 anos}} \times 100$$

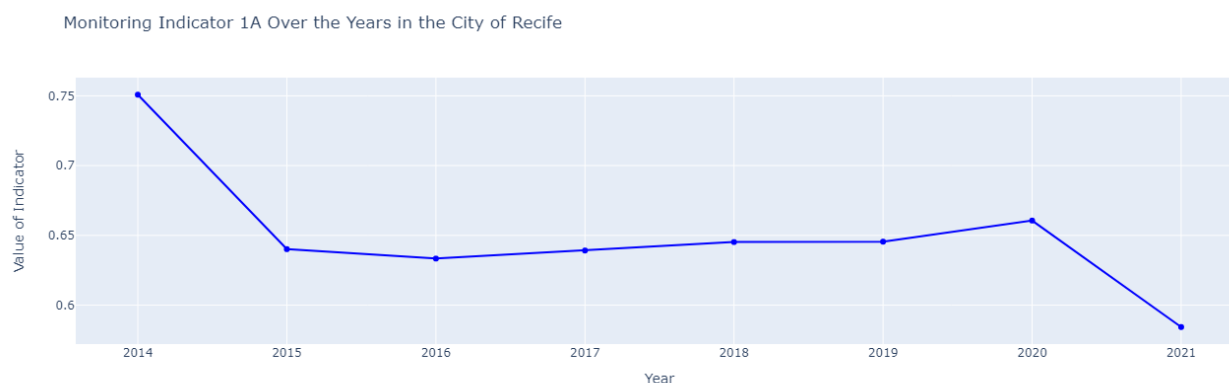
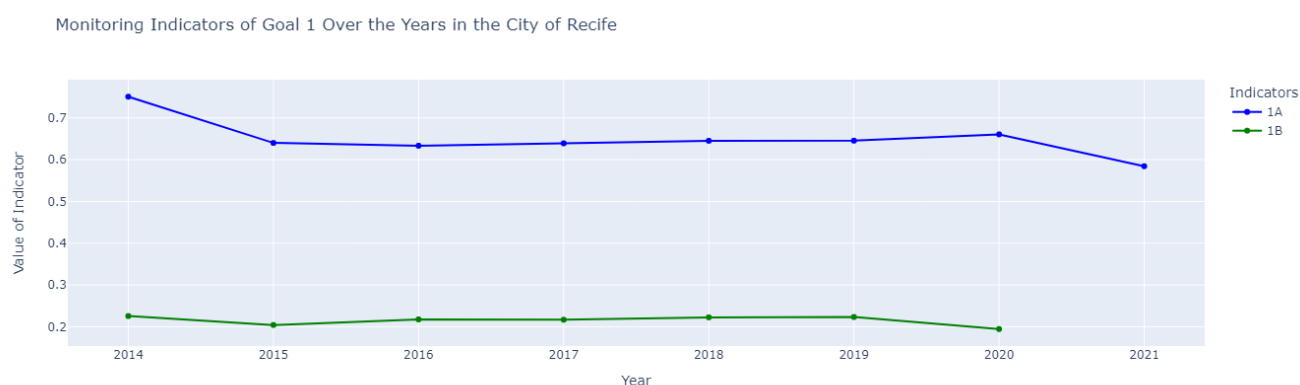
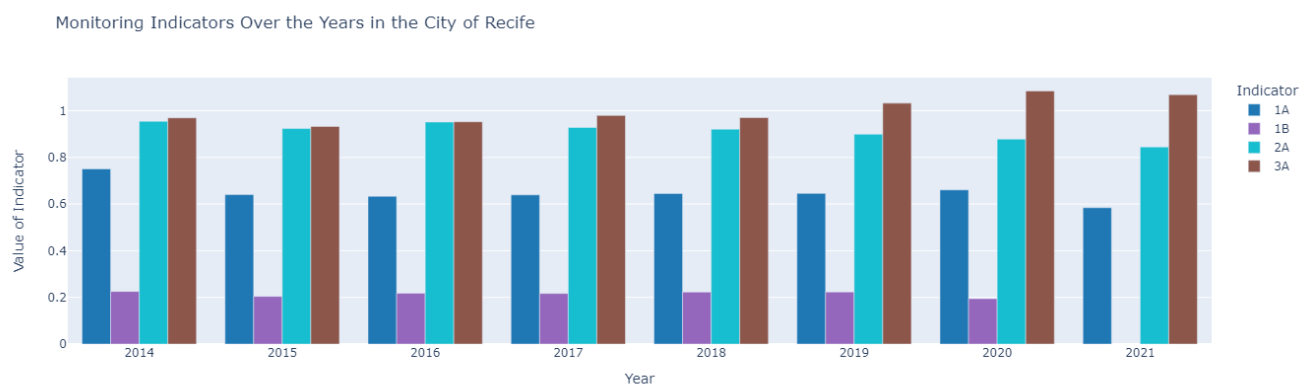
A utilização desses dados para o cálculo dessas metas permite que esses dados sejam analisados, conferidos, acompanhados, sejam de um mesmo município ou até municípios vizinhos, trazendo ainda mais possibilidades para a existência de governos colaborativos. Os dados permitem a replicação por funcionários diretos ligados à administração pública, grupos de pesquisa em temas educacionais ou governamentais, ou até mesmo o cidadão que deseje acompanhar a situação da sua cidade ao passar dos anos.

O uso desses dados para calcular essas metas permite sua análise, verificação e monitoramento, seja do mesmo município ou de municípios vizinhos, ampliando ainda mais as possibilidades de governança colaborativa. Os dados permitem a replicação por autoridades governamentais, grupos de pesquisa em temas educacionais ou governamentais, ou mesmo por cidadãos que desejam acompanhar a situação de sua cidade ao longo dos anos.

Destacamos três leituras baseadas nessa união de conjunto de dados, onde, buscando realizar a leitura de um município, que para esse caso foi Recife, capital de Pernambuco, onde na figura 7 onde podemos analisar as variações de resultados apenas do indicador 1A. Já na figura 8 temos a comparação entre indicadores de uma mesma meta, podendo assim avaliar a sua situação, onde numa mesma localidade, para a composição de uma meta, são necessárias duas faixas de idade distintas de modo a mensurar a inserção escolar na pré-escola e creche. Por último, temos em uma comparação geral com todos os indicadores, que foram citados nesse texto, de modo a ilustrar o conjunto completo que pode ser gerado com esses dados.

¹¹<https://pne.mec.gov.br/>

¹²<https://pne.mec.gov.br/publicacoes/itemlist/category/4-monitoramento-e-avaliacao>

Figure 7. Monitoring Indicator 1A Over the Years in the City of Recife**Figure 8.** Monitoring Indicators of Goal 1 Over the Years in the City of Recife**Figure 9.** Monitoring Indicators Over the Years in the City of Recife

6 Conclusão

Os conjuntos de dados apresentados ao decorrer desse trabalho foram desenvolvidos com o intuito de sanar essa lacuna de dados educacional e populacional de forma desagregada na faixa temporal de 2014-2021, faixa específicas para alguns projetos, como o PNE. O acompanhamento desse e de outros projetos governamentais pode acarretar cada vez mais na participação popular na criação, aplicação e revisão

de políticas públicas.

Em trabalhos futuros buscamos a melhoria contínua dos conjuntos de dados. No tocando as dados educacionais, temos como objetivo o processamento dos anos seguintes, tal como a expansão para outros conjuntos de dados como o censo da educação superior¹³ e o Sistema de avaliação da

¹³<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>

educação básica (SAEB) <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/saeb> dados esses que apresentam estruturas semelhantes ao conjunto de dados educacional gerado por esse estudo. Além disso, temos como objetivos futuros desenvolver uma *API (Application Programming Interface)* para o fornecimento desses dados utilizando um banco de dados relacional e facilitando consultas direto a determinados municípios sem que existe a necessidade de carregamento e seleção da base de dados, buscando facilitar ainda mais o acesso das secretarias municipais.

Já na ótica populacional buscaremos o aperfeiçoamento da metodologia de desagregação populacional podem ser realizados com o objetivo de aumentar o grau de precisão das estimativas. Também é possível expandir o período de anos do processo de desagregação através de premissas para estimação da taxa de fertilidade, mas que neste estudo o período de anos desagregados está limitado ao relatórios de dados factuais obtidos do Sistema de Saúde Brasileiro (*DataSus*), especificamente de Nascimentos e Mortalidade. Uma etapa também em desenvolvimento é a utilização do Censo demográfico realizado em 2022 para a validação das projeções e aperfeiçoamento da base.

Uma das oportunidades de melhoria da metodologia de desagregação se dá através do consumo dos dados de mortalidade de todas as faixas etárias obtidos pelo *DataSus*, não se limitando apenas à mortalidade infantil. Desta forma o processo de envelhecimento populacional refletirá as naturais diferenças no crescimento vegetativo de populações com idades distintas, uma vez que cada idade tem uma particular taxa de mortalidade. Esta e outras linhas de pesquisa deverão ser submetidas a um rigoroso processo de análise estatística e conceitual à luz da literatura sobre projeção e desagregação demográfica para garantir que a metodologia ajustada forneça mais precisão e seja racionalmente consistente, explorando ao máximo os dados relevantes que os órgãos governamentais dispõem.

Por fim, temos como objetivo a disponibilização desses conjuntos de dados no formato de dados abertos conectados para que possibilite a utilização de consultas SPARQL e a disponibilização contínua e aberta dos datasets atualizados junto a seus metadados sempre que passível de atualização.

Authors' Contributions

A parte educacional foi escrita principalmente por Abílio Barros, enquanto a parte de dados populacionais foi desenvolvida pelo autor Aldéryck Albuquerque. Os demais autores trabalharam no texto, e todos concordamos com a versão publicada.

Competing interests

Os autores declaram não ter conflitos de interesse.

Availability of data and materials

Os conjuntos de dados gerados e/ou analisados durante o estudo atual estão disponíveis em

- Educacional: <https://zenodo.org/record/6666613#.YrioP3bMLnY>

- Populacional: <https://zenodo.org/record/6689160#.YrdBeHbMK5c>.

References

- Albuquerque, A. (2022). A population projection engine. url: <https://pypi.org/project/poppro/>.
- Albuquerque, A., Barros, A., Alencar, A., Nascimento, A., Bittencourt, I., and Mello, R. (2022). Dataset de estimativas populacionais desagregada por município e idade 2014-2020. In *Anais do IV Dataset Showcase Workshop*, pages 25–34, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/dsw.2022.225525.
- Balbinot, A. D. and Haubert, A. (2015). Análise temporal das matrículas em educação especial entre 2005 e 2013 no estado do paran . *Revista Pr ksis*, 2:121–132.
- Balbinot, A. D. and Haubert, A. (2017). An lise de matr culas como indicadores da evolu  o da educa  o especial no estado do rio de janeiro. *REVISTA ELETR NICA PESQUISEDUCA*, 9(19):663–673.
- Barros, A. N., Alencar, A., Nascimento, A., de Albuquerque, A. F., and Mello, R. F. (2022). Elabora  o do conjunto de dados agregados do censo da educa  o b sica. In *Anais do IV Dataset Showcase Workshop*, pages 35–45. SBC.
- de Atividades Especiais TCE-SC, D. (2021). Metodologia estim  o populacional. url: <https://www.tcsc.tc.br/sites/default/files/2021-06/Metodologia>
- Ferreira, J., Miranda, M., Abelha, A., and Machado, J. (2010). O processo etl em sistemas data warehouse. In *INForum*, pages 757–765.
- Gon alves, M. V. F., dos Santos, J. S., Ferreira, C. Z., Zavaleta, J., da Cruz, S. M. S., and Sampaio, J. O. (2021). Datasets curados e enriquecidos com proveni ncia da campanha nacional de vacina  o contra covid-19. In *Anais do III Dataset Showcase Workshop*, pages 148–159. SBC.
- Gonzaga, M. R. and Schmertmann, C. P. (2016). Estimativa de taxas de mortalidade por idade e sexo para pequenas  reas com regress o de topals: uma aplica  o para o brasil em 2010. *Revista Brasileira de Estudos de Popula  o*, 33(3):629–652.
- Gonz lez, M., Fern ndez V zquez, E., and Moroll n, F. (2015). A methodological note for local demographic projections: A shift-share analysis to disaggregate official aggregated estimations. 16:43–50.
- Ozkan, K. S., Khan, H., Deligonul, S., Yenyurt, S., Gu, Q. C., Cavusgil, E., and Xu, S. (2022). Race for market share gains: How emerging market and advanced economy mnes perform in each other’s turf. *Journal of Business Research*, 150:208–222. DOI: <https://doi.org/10.1016/j.jbusres.2022.04.040>.
- Vasconcelos, F. F., Tavares, J. V., Ribeiro, M. U., Coutinho, F. J., and Clarindo, J. P. (2021). Candidata: um dataset para an lise das elei  es no brasil. In *Anais do III Dataset Showcase Workshop*, pages 160–168. SBC.

Workflow for the acquisition, processing, and dissemination of Brazilian public data focused on education

Abílio Nogueira Barros   [Universidade Federal Rural de Pernambuco | abilio.nogueira@ufrpe.br]

Aldéryck Félix de Albuquerque  [Cesar School | deryck@gmail.com]

Andréza Leite de Alencar  [Universidade Federal Rural de Pernambuco | andreza.leite@ufrpe.br]

André Nascimento  [Universidade Federal Rural de Pernambuco | andre.nascimento@ufrpe.br]

Ibsen Mateus Bittencourt  [Universidade Federal de Alagoas | ibsen@feac.ufal.br]

Rafael Ferreira Mello  [Universidade Federal Rural de Pernambuco | rafael.mello@ufrpe.br]

Received: DD Month YYYY • Accepted: DD Month YYYY • Published: DD Month YYYY

Abstract This article aims to demonstrate the process of creating public databases focused on the educational and population areas. It describes the process of obtaining data from official government sources such as INEP (National Institute for Educational Studies and Research) and IBGE (Brazilian Institute of Geography and Statistics), the procedures for data adaptation and optimization to create their historical series, as well as the best practices followed for their development and the generated metadata. Highlighting the specificities between the themes of education and population, reporting their challenges and peculiarities of each dataset. It also reports the results that can already be directly obtained from each dataset and how, when combined, they can track indicators of the National Education Plan, one of the largest Brazilian public policies focused on education.

Keywords: Datasets, Open data, Public data, Education data, Smart government

1 Introduction

The pursuit of improving the quality and efficiency of public education policies is a constant challenge for governments and institutions. In this context, accurate and comprehensive data play a fundamental role in developing metrics and indicators that underpin the construction of robust educational plans and in evaluating the effectiveness of implemented actions. Among the essential data for this purpose, information about student enrollments in different municipalities, coupled with knowledge of population volumes segmented by age, provides a valuable source of information to guide educational strategies. It also helps in devising metrics and indicators that serve as a starting point for building models supporting the development of educational plans and monitoring their implementation.

These informations enable a better understanding of the demand and supply of education in different regions, identifying areas with a higher concentration of students in specific age groups and where educational needs are more pressing. Furthermore, such data enables the analysis of the distribution of the school-age population, assisting in projecting future student contingents and preparing the educational system to accommodate these demands. The segmentation by municipalities also allows for a more localized approach, considering regional specificities and formulating specific strategies to meet the educational needs of each locality. Based on these robust analytical data, the government can make informed decisions and allocate resources more efficiently, ensuring the implementation of more effective educational policies and thereby contributing to the advancement and improvement of the country's educational system.

The present study aims to construct educational and pop-

ulation datasets for Brazil, segmented by municipality and age through a detailed process of transformation and structured modeling to facilitate data consumption. In public educational data sources, a collection is observed that requires translation treatment based on a data dictionary, as well as an enrichment of this collection to contextualize the data and enable future processes of aggregation, filtering, selection, and other operations needed to extract value from this material.

In light of the above, the main source of data used in this study was established with the Educational Census coordinated by the Brazilian National Institute for Educational Studies and Research (INEP), which is regulated by normative instruments that establish the obligation, deadlines, responsibilities, and procedures for the entire data collection process.¹

This article is an extended version of the paper Elaboration of the aggregated dataset of the basic education census Barros *et al.* [2022], published in DSW-2022. As an extended version, new analysis of the use of educational data added to population data was made, in order to demonstrate its impact on public policies, which in the case of this project is the National Education Plan, one of the major Brazilian educational public policies. Additionally, this extended version was also added the process of a work Dataset of population estimates disaggregated by municipality and age 2014-2020 Albuquerque *et al.* [2022], fundamental to the new analysis performed.

The article is organized as follows: Section 2 presents some related works on the use of educational and population data in the construction of public databases and their applications, as well as discusses methods for estimating the

¹<https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar>

population; Section 3 presents the process through which the educational and population data were developed, from their acquisition to the provision of the databases along with their metadata; Section 4 demonstrates some of the results that can already be achieved with the individual use of these data and then delves into a case study of a national plan; and finally, Section 5 provides the final considerations and future work that may improve and optimize the work already done.

2 Related Works

To create the knowledge base of previous works, it was necessary to search for general public data usage, including educational, population, and health data, and particularly how these data have been applied for monitoring public policies.

In the context of constructing databases from public sources, an important case demonstrated in Vasconcelos *et al.* [2021] involved collecting and standardizing data directly from the TSE (Superior Electoral Court) portal. After acquiring and processing the raw data, new geographical data were added to enhance the usability for database users.

Transitioning to the field of health, Gonçalves *et al.* [2021] collected, processed, and made available a database on vaccination records registered in the *OpenDataSUS* platform. The process followed data science principles and was developed interactively until achieving the objective of providing the dataset and data dictionary.

To support the section on population data, it was necessary to search for works that focused on population perspectives and how to work with this disaggregation.

Recent works have explored statistical projection methods for demographic data, essential for constructing public policies. Gonzaga and Schmertmann [2016] aimed to develop an age-based mortality estimation model for small geographic areas in Brazil, using data from the 2010 Demographic Census and the Public Health System (*DataSUS*). The mortality variable plays a crucial role in estimating a location's population in a given year, so the authors proposed a sophisticated statistical model for its projection. However, some observations regarding the timeline for obtaining base data for constructing population estimates in the Brazilian scenario need to be highlighted. The demographic census is updated once every ten years, while actual mortality and birth data are updated annually. Therefore, during the period between the last census and the year prior to the analysis, it is not necessary to estimate mortality data for the population estimate calculation, as this data can be obtained as facts.

It can be observed that the updated factual data obtained from government sources strongly influence the methodological process to generate population data disaggregated by small regions and age. This can even simplify the disaggregation process, as there is no need to estimate certain input variables or establish assumptions. For instance, nominal data on "Births" or "Mortality" at the desired level of disaggregation. If this data is obtained in a timely manner, there is no need for the application of an estimation process for base variables to compose the population estimate calculation.

In addition to the process of constructing population estimates, another relevant concept that underpins the data mod-

eling process is that of population disaggregation in relation to population estimates. The study conducted by González *et al.* [2015] highlights that population projection models are essential tools used in various programs and public policies, especially those related to the sustainability of public pension, health, and education systems. The study distinguishes these concepts, emphasizing that the population disaggregation process starts from pre-existing demographic data provided by government statistical institutions, and subsequently implements a methodology to enhance the granularity of this data, considering factors such as ethnicity, social class, gender, age, among others.

Due to the limited availability of approaches on the population topic, the study's foundational research was more extensive, focusing on which approaches could be applied to this study's process. Based on these concepts and considerations of significant variables, the population disaggregation methodology that proved most suitable for the temporal scope and data acquisition scenario was the methodology proposed by de Atividades Especiais TCE-SC [2021], developed by the Special Activities Directorate of the Court of Auditors of the State of Santa Catarina. In this methodology, the absence of data on migration between municipalities was addressed by considering the variation of population estimates for the respective municipalities over a period of time. Thus, the dataset underlying this methodology consists of census data, birth and infant mortality data, and population estimates for municipalities without age granularity.

Regarding the educational topic, since we already have more consolidated data sources, data mining techniques were applied to construct the necessary foundation to address the main indicators targeted in this study.

3 Data Acquisition and Processing

In this stage of the project, the processing types are separated for their respective areas. For the educational data, the main provider is the National Institute for Educational Studies and Research Anísio Teixeira (INEP)², while for geographic data distribution, we have the Brazilian Institute of Geography and Statistics (IBGE)³. Given these differences, we will demonstrate the process applied to each data theme in two distinct flows.

The detailed and complete process of the complete processing of the population part can be seen in Albuquerque *et al.* [2022].

3.1 Educational Data

The development process of this dataset was based on the Extract, Transform, Load (ETL) process, as defined in Ferreira *et al.* [2010]. This process involves extracting data from its sources, preferably in its original form, applying the necessary transformations to make the data more accessible and direct for use, and loading it into a more definitive storage location. In the scope of this project, a set of files in CSV

²<https://www.gov.br/inep/pt-br>

³<https://www.ibge.gov.br>

format (Comma-separated values) was used as the final storage format.

All processing was designed and executed with the ultimate goal of consolidating a basic education census dataset, with the chosen time frame for this project being from 2010 to 2021. This created a dataset that can be analyzed and integrated with other public databases. To correctly design the data processing flow, readings on the subject were conducted, and, most importantly, an understanding of the data dictionaries published by the publisher. This resulted in the following process, as illustrated in Figure 1

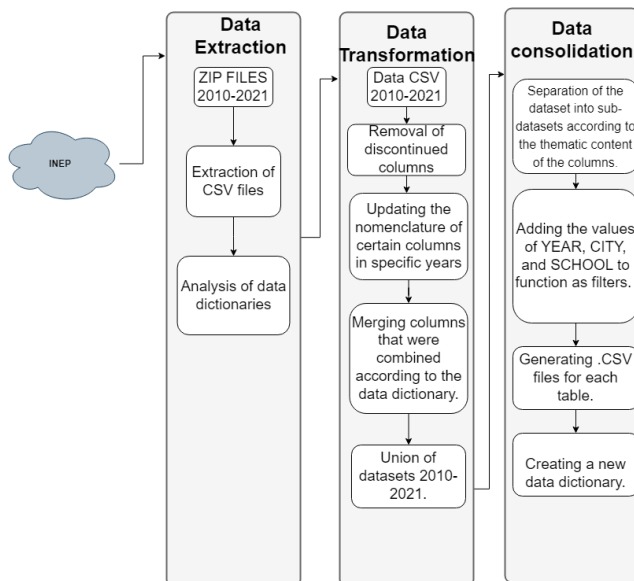


Figure 1. Processing pipeline of the used data.

3.1.1 Data Extraction

The original data for this project was entirely retrieved from the INEP website. These data are provided in a compressed format to facilitate downloading to the local machine. The data is originally provided annually, with a CSV file containing the data and an excel spreadsheet data dictionary, along with other metadata, such as questionnaires applied to institutions and a user manual indicating information for use in specific software.

3.1.2 Data Transformation

After reading the data dictionaries, the necessary changes were identified to unite all the years targeted in the project. It was observed that a large part of the columns had already been standardized through updates made by INEP over the years. However, there were still some columns that would hinder this unification, and some others were discontinued in more recent years.

1. **Removal of discontinued columns:** Discontinued columns from 2019 onwards were removed as the most current data no longer reflects them and, therefore, may not provide up-to-date information for monitoring the educational network structure. Out of the total of 370 available variables, 36 were removed, leaving 334 for further processing.

2. **Update of column names:** During the processing of the remaining columns, certain changes were applied to adapt the nomenclature and variables to the applied time frame. This need was identified through the examination of the data dictionaries provided with the files.
3. **Combining columns:** Some columns had their names changed after 2018, and it was necessary to standardize these columns to the most current names. These columns had their names altered, but not their values. This process followed the standard indicated in the data dictionary. Other columns had to be combined starting from the year 2019, and the strategy of using the *OR* was applied. Since these columns represent specifications of the same response, if any of these columns have a positive value, the new column will be assigned a value of True. If none of the columns have a positive value, the value False will be assigned to the new column.
4. **Data consolidation:** The final step of this process aimed to merge all the data that was processed annually into a single data table. In contrast to the approach adopted by INEP, a separation of columns was proposed, grouping them by theme, as can be observed in the table 1, to allow for loading only specific groups of columns. The data was separated by themes such as enrollment, teachers, classes (previously separated by INEP in earlier versions of the census), and results. The removed, combined, and remapped columns are available in the metadata, which can be found alongside the database at <https://zenodo.org/record/6666613#.YrioP3bMLnY>.

3.2 Population Data

The data for the chosen population disaggregation methodology was obtained from IBGE and the Ministry of Health. As highlighted in the technical note that formalizes the methodology, the obtained data includes:

1. Obtained from IBGE:
 - Census data segmented by year and age;⁴
 - Annual population estimates of municipalities. (No age segmentation available.)⁵
2. Obtained from the Ministry of Health:
 - Data from the Live Birth Information System;⁶
 - Data from the Mortality Information System.⁷

3.2.1 Data Acquisition and Processing

The process illustrated in Figure 2 follows a flow where input data from the previous section are cleaned and normalized, as some of them are provided in spreadsheets with textual observations in cell values and refer to municipalities in a

⁴<https://sidra.ibge.gov.br/tabela/200>

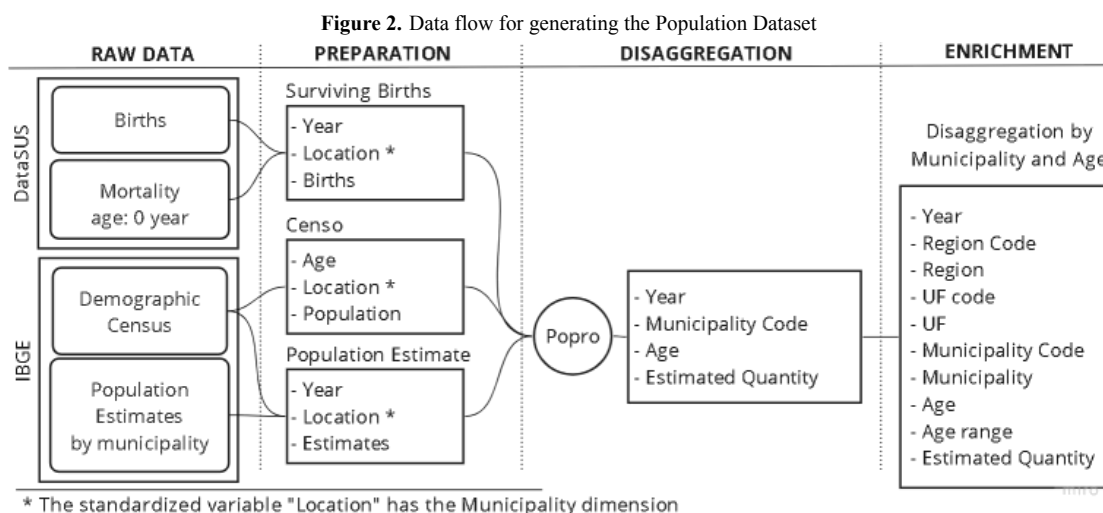
⁵https://ftp.ibge.gov.br/Estimativas_de_Populacao

⁶<https://datasus.saude.gov.br/nascidos-vivos-desde-1994>

⁷<https://datasus.saude.gov.br/mortalidade-desde-1996-pela-cid-10>

Table 1. Tables generated by theme.

Table	Description
TB_GEOGRAFICA	Geographical data where the educational institution is located.
TB_TURMA	Information about the classes of a specific educational institution.
TB_DOCENTE	Quantitative data about the teachers belonging to the educational institution.
TB_MATRICULA	Enrollments made at the educational institution.
TB_ESCOLA_ADMINISTRATIVO	Data about the educational institution.
TB_ESCOLA_ESTRUTURA	Data about the infrastructure of the educational institution.
TB_ESCOLA_FUNCIONARIO	Data about non-teaching staff at the educational institution.
TB_ESCOLA_PEDAGOGICO	Data about the pedagogical structure of the educational institution.



textual format instead of their unique code standardized by IBGE.

In general, the first common challenge in processing the input data for all the analyzed reports lies in the presence of explanatory texts in the header and footer of the CSV reports extracted from both *DataSUS* and *IBGE*, with no standardized number of comment lines in these two locations.

Going into detail, the "Births" and "Mortality with age of zero year" reports extracted from *DataSUS* in CSV format have only two columns. The first column is labeled "Município" and contains text that concatenates the municipality code standardized by *IBGE* with the municipality name. Besides this concatenation characteristic requiring treatment, there are a few lines in each report where the municipality code is not valid when queried in the *IBGE* municipality code table. However, the total population quantity of rows with invalid municipality codes represents an insignificant fraction, around 1 in every 30,000 individuals counted (0.0033%).

In the Census Demographic report provided by *IBGE*, the ages are arranged in different columns, and municipalities are presented in different rows, forming a data matrix with thousands of rows and dozens of columns. It is observed that for some rare combinations of municipalities and age groups, the population quantity is not provided, and a null value is given. Considering the population volume for each age group in the municipalities where these cases occur, it seems unlikely that the cause of these null values is the absence of people of that age in the respective municipality. Instead, the cause is likely related to the data collection or

processing process during the report construction, which can be resolved by the responsible party for its development.

Lastly, the Population Estimates report, also provided by *IBGE*, like the previous ones, contains textual information in the header and footer. Additionally, in the quantity column, there are specific markings indicating footnotes that clarify some methodological details leading to the reported numbers. These markings consist of numbers within parentheses, altering the structure of the column, making it non-purely numeric. This, once again, requires transforming the report data into a tabular format with consistent data types in each column.

With the completion of the cleaning process, three normalized auxiliary tables are constructed. Along with the meta-data indicating the year to which the census data relates, these tables will be used as input variables in the population disaggregation tool presented in Section 3.2.2.

The auxiliary tables consist of: "Surviving Births" (Year, Location, Births), "Real Population from the last census segmented by age and location" (Age, Location, Population), and "Population Estimate by Municipality" (Year, Location, Estimate).

3.2.2 Population Disaggregation

The population disaggregation process follows two main approaches, depending on the target year for analysis and the birth year of the age group under consideration. In summary, the methodology starts with a population base and ages it

based on the fluctuation of the population estimate for the municipality under analysis over the required period.

This population base can be from two sources, depending on the birth year of the age group under analysis. The population base is census-based when the population set for a specific target year has a birth year equal to or earlier than the last Census execution. The population base is derived from the volume of surviving births for other cases, i.e., when the population set for a specific target year has a birth year subsequent to the last Census execution. The mathematical model and a more detailed discussion can be found in the technical note de Atividades Especiais TCE-SC [2021], which formalizes this methodology.

The disaggregation process was carried out using the Popro tool Albuquerque [2022], a population disaggregation library developed by the authors in the Python programming language. The tool is expandable to different methodologies, and the statistical model of the mentioned methodology was implemented as a plugin for generating the current dataset.

3.2.3 Resulting Dataset

Upon completion of the data extraction, processing, disaggregation, and enrichment steps, the resulting dataset is capable of supporting action plans and guiding the establishment of goals in public policies. Due to the implemented disaggregation methodology having a decreasing level of confidence as the estimated age increases, it is essential to emphasize that the provided database is more suitable for supporting plans involving age groups of children and adolescents, such as primary and secondary education policies.

The resulting *Dataset* has granularity in the variables YEAR, MUNICIPALITY, and AGE. Important data additions were made for future analysis and grouping, including the municipality code, name of the federative unit, and the name of the geographic region, each with their respective codes standardized by IBGE.

3.2.4 Best Practices for Dataset Dissemination

This step, being common to both datasets, followed the same procedure, albeit producing slightly different artifacts, aiming to comply with a set of data publication best practices⁸.

By achieving the goals outlined in Table 2, we were able to provide a dataset that is traceable both in terms of its source data (BP5) and the version being used and publication date (BP7), along with updated metadata for coherent data utilization (BP1, BP2, BP3).

The data and metadata responsible for applying these best practices are available on the ZENODO platform⁹:

- Educational: <https://zenodo.org/record/6666613#.YrioP3bMLnY>
- Populational: <https://zenodo.org/record/6689160#.YrdBeHbMK5c>.

The files are made available under the CC-BY-4.0 license, which allows others to download, adapt, and build upon this work, even for commercial purposes, as long as proper attribution is given to this project as the source.

4 Applications of the datasets

When addressing the theme of outcomes regarding the creation of public datasets, we should first consider the progression or regression of certain age groups, access to education, and the quantity of teachers in specific locations. All these data, even in isolation, already provide informative insights and can contribute to the formation of a collaborative society. Another perspective is to cross-reference these data to monitor plans and public policies. These datasets can offer a way to observe them at the municipal level, encompassing all Brazilian municipalities, thus creating a plural and inclusive approach to monitoring.

Below are demonstrations of uses in these two areas, both using the datasets individually and in combination, to monitor the goals of the National Education Plan (PNE), one of the main public policies in the educational field in Brazil.

4.1 Utilization of the datasets for educational monitoring

Individually, the data can already provide us with an overview of the presence of certain factors during the historical series covered in the studies.

In the educational field, there are countless possibilities for applying these data. For example, in Balbinot and Haubert [2017], the application of these census microdata was used to monitor indicators related to special education in Rio de Janeiro. Similarly, in Balbinot and Haubert [2015], a temporal analysis of enrollments in the state of Paraná was conducted. These are some examples of works already carried out based on these educational data.

Individually, we can use data visualization tools like Plotly¹⁰, which, combined with Python¹¹, allows us to create visualizations about the educational situation of a specific municipality, helping us understand the situation reflected over the time covered by the dataset.

In Figure 3, we can observe the decrease in the number of teachers in the municipality of Moreno, PE. The teachers from the three education levels (Blue representing basic education, red representing elementary education, and green representing high school) show that in that locality, the largest number of available teachers is in basic education. We can also observe if the number of enrollments follows the same pattern.

Continuing to use the dataset, in Figure 4, with the same color and legend scheme (Blue representing enrollments in basic education, red representing enrollments in elementary education, and green representing enrollments in high school), we can see that the majority of enrollments belong to the basic education level. Similarly to the number of teachers

⁸<https://ceweb.br/media/docs/publicacoes/1/fundamentos-publicacao-dados-web.pdf>

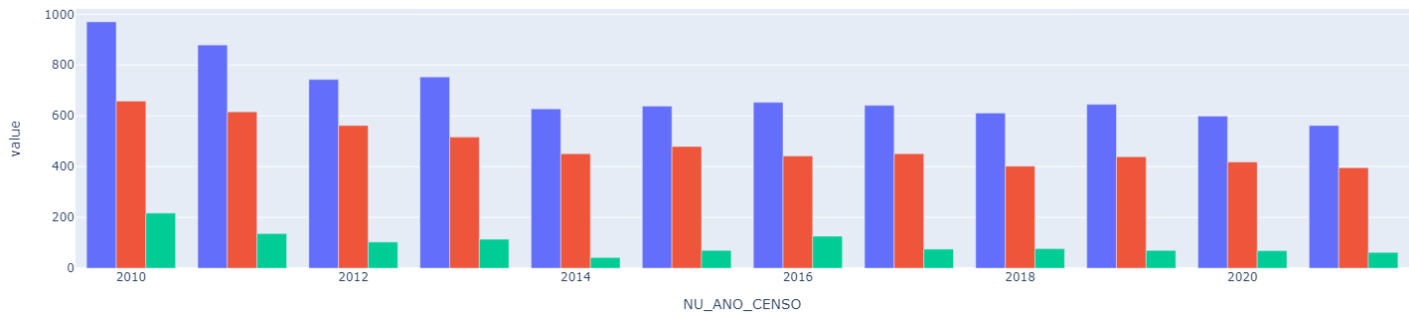
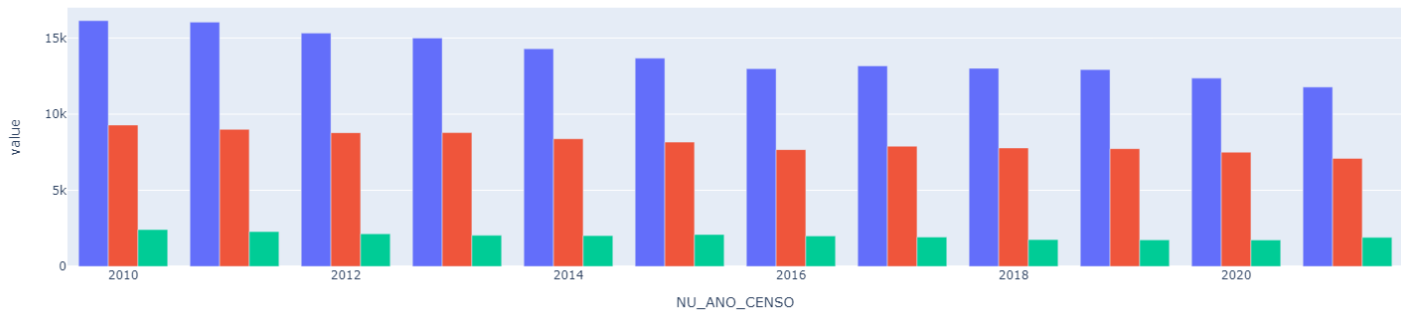
⁹<https://zenodo.org/>

¹⁰<https://plotly.com/>

¹¹<https://www.python.org/>

Table 2. Achieved best practices for the datasets

Code	Definition
BP1	PROVIDE METADATA
BP2	PROVIDE DESCRIPTIVE METADATA
BP3	PROVIDE STRUCTURAL METADATA
BP5	PROVIDE DATA PROVENANCE INFORMATION
BP7	PROVIDE VERSION INDICATOR

**Figure 3.** Number of Teachers per year and grade in the Municipality of Moreno-PE**Figure 4.** Number of Enrollments per Year and Grade in the Municipality of Moreno-PE

in the education system, the number of enrollments in this educational level has been decreasing, but it still remains higher than the others.

4.1.1 Monitoring age groups over the years

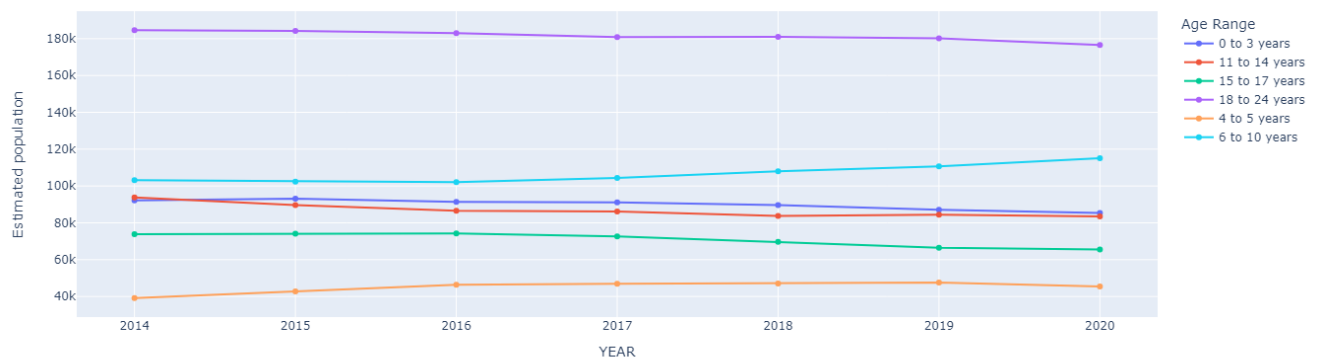
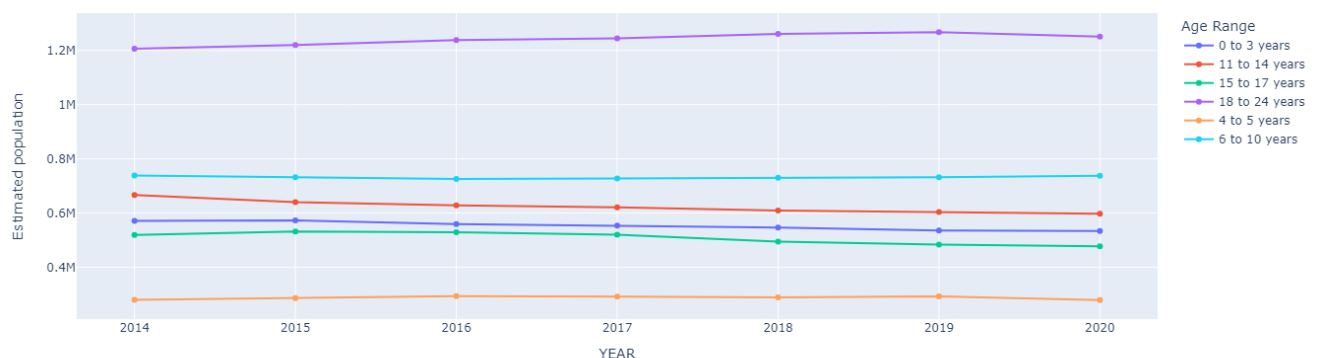
The use of a population dataset segmented by age in small geographical areas aims to support the development of public policy planning and the measurement of metrics in various types of institutions, both public and private.

For instance, in the private educational sector, a basic education institution, using this dataset and its own database of enrolled students, can measure the representation of its students compared to potential students in the municipality where the institution is located. This allows for the calculation of the market share indicator, as discussed in the study Ozkan *et al.* [2022], which is essential to guide strategic decisions, such as focusing on attracting new students, retaining enrolled students, or improving the quality of services offered.

On the other hand, in the public sphere, this dataset can be used, for example, to collect data for analyzing public safety through a methodological approach. It allows the formulation of indicators, such as the crime rate against specific age groups of the population, to assess their behavior over time and identify patterns and trends.

The development of this type of dataset enables the tracking of age group progressions since the IBGE population census is conducted once every decade, creating the need for projections of these other age groups. This understanding is crucial to observe the advancement or decline in the estimated population for each specific age group over the years. We can observe this for the municipality of Recife in Figure 5.

The tables were organized to enable aggregations based on state and region codes. This allows for applying grouping operations with the help of data processing tools or languages. In Figure 6, we can observe the same example applied earlier, this time for the state of Pernambuco.

Figure 5. Population estimate progression by age group in the municipality of Recife-PE.**Figure 6.** Population estimate progression by age group in the state of Pernambuco.

4.2 Monitoring the PNE

The entire development of this work was based on monitoring one of the most significant public policies focused on education, the PNE. It is a long-term public policy that establishes guidelines and goals for the development of education in Brazil over a ten-year period. It was created to promote quality, inclusive, and equitable education at all levels, from early childhood education to higher education. As such, it is a fundamental tool to guide government and civil society actions in the pursuit of more equitable education capable of driving the country's social and economic development.

The importance of monitoring the National Education Plan lies in the fact that this monitoring is essential to ensure that the established goals are effectively achieved. By periodically assessing the progress of actions, it is possible to identify advances and also potential challenges and obstacles to be overcome. Monitoring allows for adjusting strategies, reallocating resources, and directing efforts to priority areas, ensuring that education reaches the most vulnerable populations in need of investment.

Due to its ten-year duration, its goals and action plans for achieving its indicators need to be understood and continued by all governments operating during this period, regardless of the political party in power at the federal, state, and municipal levels. This requires a firm commitment to cooperation

among all government entities throughout the years.

In evaluating all of this, monitoring the PNE is of vital importance, both for its effectiveness as a collaborative government instrument and for its functionality as an overall evaluator of public education and the effectiveness of actions aimed at improving it. When we consider Brazil and its 5,570 municipalities, we can already gauge the difficulty of monitoring all 20 goals with an average of 2 indicators present in the current PNE (2014-2024)¹².

INEP provides biennial reports¹³ highlighting results at the federal, state, and some Brazilian capital levels. However, the purpose of creating these datasets is to fully map the quantity of municipalities where data is available from their foundation (given that some municipalities were emancipated after the 2010 census, leading to a lack of data in certain years).

For these reports developed by INEP, some sampling sources such as the Continuous National Household Sample Survey (PNAD) are used, which does not cover all Brazilian municipalities. Therefore, one of the main objectives of creating and merging these datasets is to reproduce the calculation of indicators 1A, 1B, 2A, 3A, and 3B to represent all

¹²<https://pne.mec.gov.br/>

¹³<https://pne.mec.gov.br/publicacoes/itemlist/category/4-monitoramento-e-avaliacao>

Brazilian municipalities, initially from 2014 to 2020.

With the generated dataset, it is possible to calculate the following indicators for all Brazilian municipalities:

$$1A = \frac{\text{Enrollments for 4 to 5 yo.}}{\text{Estimated population for 4 to 5 yo.}} \times 100$$

$$1B = \frac{\text{Enrollments for 0 to 3 yo.}}{\text{Estimated population for 0 to 3 yo. in school}} \times 100$$

$$2A = \frac{\text{Enrollments for 6 to 14 yo.}}{\text{Estimated population for 6 to 14 yo.}} \times 100$$

$$3A = \frac{\text{Enrollments for 15 to 17 yo.}}{\text{Estimated population for 15 to 17 yo.}} \times 100$$

$$3B = \frac{\text{Enrollments for 15 to 17 yo. in high school}}{\text{Estimated population for 15 to 17 yo.}} \times 100$$

The use of these data to compute these targets allows for their analysis, verification, and monitoring, whether from the same municipality or neighboring municipalities, bringing even more possibilities for collaborative governance. The data enables replication by government officials, research groups in educational or governmental topics, or even citizens who wish to track the situation of their city over the years.

We highlight three readings based on this dataset combination, aiming to perform the analysis of a municipality, which, in this case, is Recife, the capital of Pernambuco. In Figure 7, we can observe variations in the results of only indicator 1A. In 8, we have a comparison of indicators within the same goal, allowing us to assess their situation. In the same location, to compose a goal, two different age groups are required to measure school enrollment in preschool and daycare. Finally, we have a general comparison of all the indicators mentioned in this text 9, illustrating the complete set that can be generated with this data.

5 Conclusion

The datasets presented throughout this work were developed with the aim of addressing the gap in disaggregated educational and population data for the time period of 2014–2021, a specific range for certain projects like the PNE. Monitoring this and other government projects can increasingly lead to public participation in the creation, implementation, and review of public policies.

In future work, we seek continuous improvement of the datasets. Regarding educational data, our objectives include processing the following years, as well as expanding to other datasets such as the Higher Education Census¹⁴ and the Basic Education Assessment System (SAEB)¹⁵. These datasets

have similar structures to the educational dataset generated in this study. Additionally, our future goals involve developing an Application Programming Interface (API) for data delivery using a relational database, facilitating direct queries to specific municipalities without the need to load and select the database, thereby further enhancing access for municipal secretaries.

Regarding the population perspective, we will seek to refine the population disaggregation methodology to increase the precision of estimates. Expanding the disaggregation period through assumptions for estimating the fertility rate is also possible, but in this study, the disaggregated years are limited to factual data reports obtained from the (*DataSus*), specifically Births and Mortality. Another ongoing stage involves using the 2022 demographic census for validating projections and updating the dataset.

One improvement opportunity for the disaggregation methodology lies in utilizing mortality data for all age groups obtained from *DataSus*, not just infant mortality. Thus, the population aging process will reflect the natural differences in the natural increase of populations with different ages, as each age has a particular mortality rate. This and other research avenues should undergo rigorous statistical and conceptual analysis in light of the literature on demographic projection and disaggregation to ensure that the adjusted methodology provides greater precision and rational consistency, making the most of the relevant data available to government agencies.

Finally, our aim is to make these datasets available in the connected open data format so that SPARQL queries can be used, and the updated datasets can be made continuously and openly available, along with their metadata whenever it is possible to update them.

Authors' Contributions

For the educational part, the author was mainly Abilio Barros, while for the population data, the part was developed by the author Aldéryck Albuquerque. The other authors worked on the text, and we all agree with the version published.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in

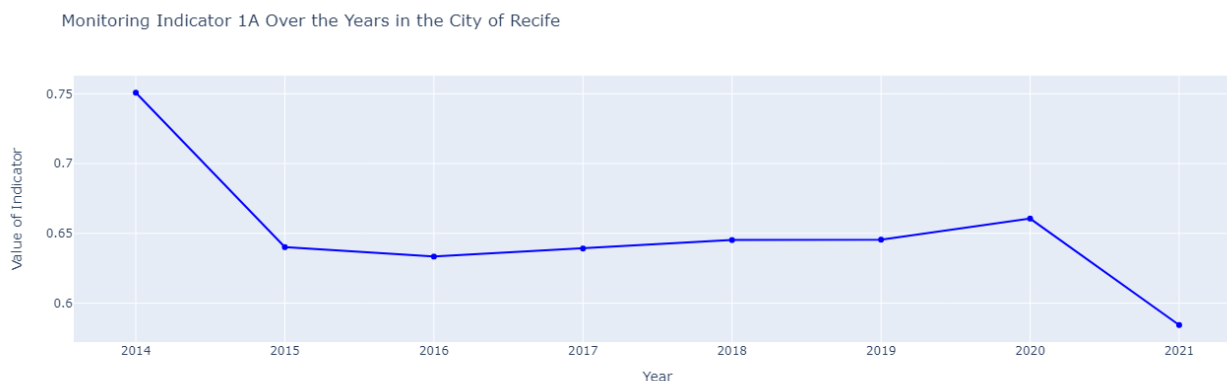
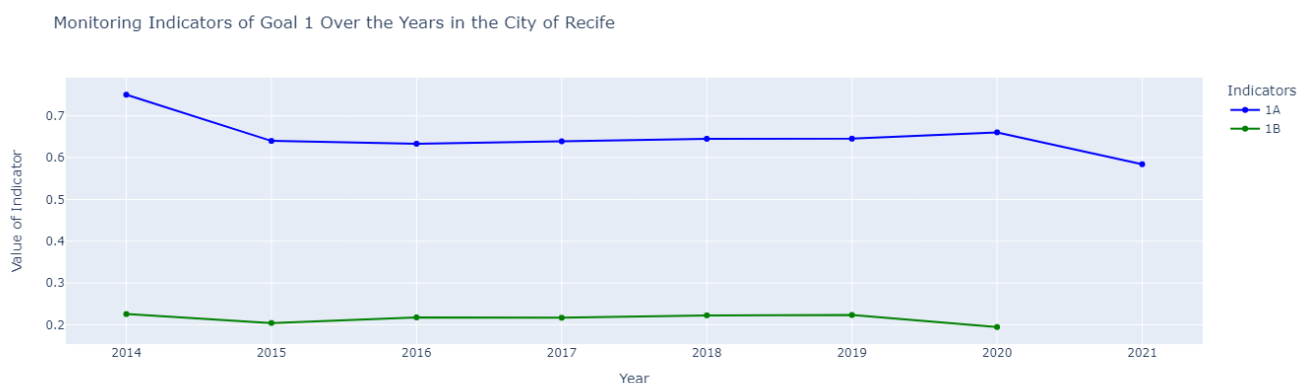
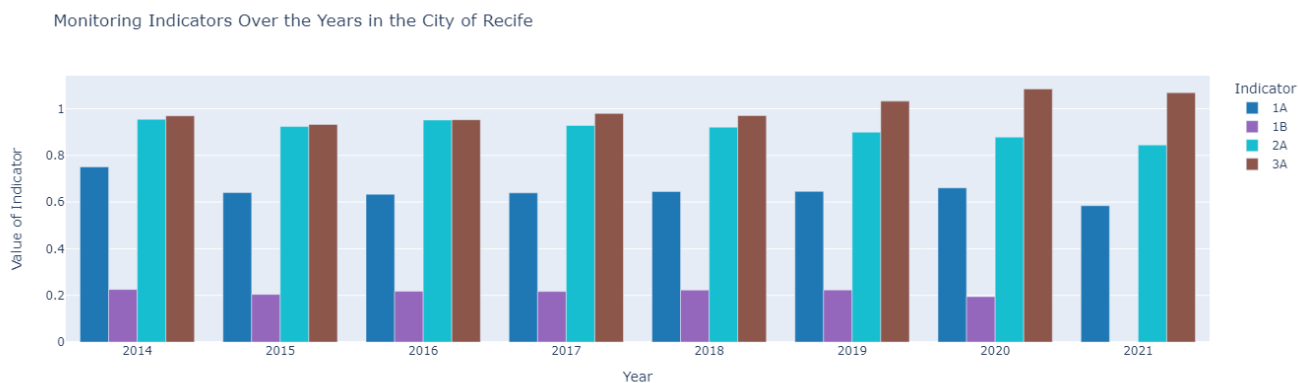
- Educational: <https://zenodo.org/record/6666613#.YrioP3bMLnY>
- Population: <https://zenodo.org/record/6689160#.YrdBeHbMK5c>.

References

- Albuquerque, A. (2022). A population projection engine. url: <https://pypi.org/project/poppro/>.

¹⁴<https://www.gov.br/inep/pt-br/acao-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>

¹⁵<https://www.gov.br/inep/pt-br/acao-a-informacao/dados-abertos/microdados/saeb>

Figure 7. Monitoring Indicator 1A Over the Years in the City of Recife**Figure 8.** Monitoring Indicators of Goal 1 Over the Years in the City of Recife**Figure 9.** Monitoring Indicators Over the Years in the City of Recife

Albuquerque, A., Barros, A., Alencar, A., Nascimento, A., Bittencourt, I., and Mello, R. (2022). Dataset de estimativas populacionais desagregada por município e idade 2014-2020. In *Anais do IV Dataset Showcase Workshop*, pages 25–34, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/dsw.2022.225525.

Balbinot, A. D. and Haubert, A. (2015). Análise temporal das matrículas em educação especial entre 2005 e 2013 no

estado do paran . *Revista Pr xis*, 2:121–132.

Balbinot, A. D. and Haubert, A. (2017). An lise de matr culas como indicadores da evolu  o da educa  o especial no estado do rio de janeiro. *REVISTA ELETR NICA PESQUISEDUC*, 9(19):663–673.

Barros, A. N., Alencar, A., Nascimento, A., de Albuquerque, A. F., and Mello, R. F. (2022). Elabora  o do conjunto de dados agregados do censo da educa  o b sica. In *An is*

- do IV Dataset Showcase Workshop, pages 35–45. SBC.
- de Atividades Especiais TCE-SC, D. (2021). Metodologia estimação populacional. [urlhttps://www.tcesc.tc.br/sites/default/files/2021-06/Metodologia](https://www.tcesc.tc.br/sites/default/files/2021-06/Metodologia)
- Ferreira, J., Miranda, M., Abelha, A., and Machado, J. (2010). O processo etl em sistemas data warehouse. In *INForum*, pages 757–765.
- Gonçalves, M. V. F., dos Santos, J. S., Ferreira, C. Z., Zavaleta, J., da Cruz, S. M. S., and Sampaio, J. O. (2021). Datasets curados e enriquecidos com proveniência da campanha nacional de vacinação contra covid-19. In *Anais do III Dataset Showcase Workshop*, pages 148–159. SBC.
- Gonzaga, M. R. and Schmertmann, C. P. (2016). Estimativa de taxas de mortalidade por idade e sexo para pequenas áreas com regressão de topals: uma aplicação para o brasil em 2010. *Revista Brasileira de Estudos de População*, 33(3):629–652.
- González, M., Fernández Vázquez, E., and Morollón, F. (2015). A methodological note for local demographic projections: A shift-share analysis to disaggregate official aggregated estimations. 16:43–50.
- Ozkan, K. S., Khan, H., Deligonul, S., Yenyurt, S., Gu, Q. C., Cavusgil, E., and Xu, S. (2022). Race for market share gains: How emerging market and advanced economy mnes perform in each other’s turf. *Journal of Business Research*, 150:208–222. DOI: <https://doi.org/10.1016/j.jbusres.2022.04.040>.
- Vasconcelos, F. F., Tavares, J. V., Ribeiro, M. U., Coutinho, F. J., and Clarindo, J. P. (2021). Candidata: um dataset para análise das eleições no brasil. In *Anais do III Dataset Showcase Workshop*, pages 160–168. SBC.

5 Apêndice C - Criação de um panorama da Taxa de distorção idade-série com base nos dados abertos do INEP

Barros, Abílio, Andrêza Leite de Alencar, and Rafael Ferreira Mello. "Criação de um panorama da Taxa de distorção idade-série com base nos dados abertos do INEP." Anais do VIII Congresso sobre Tecnologias na Educação, Santarém/PA, 2023. SBC, 2023, pp.457-462.

Contribuição do autor: Responsável pela definição e implementação da metodologia, desenvolvimento e avaliação dos resultados e escrita do texto.

Criação de um panorama da Taxa de distorção idade-série com base nos dados abertos do INEP

Abílio Nogueira Barros¹, Andréza Leite de Alencar¹,
Rafael Ferreira Mello^{1,2}

¹Departamento de Computação - Universidade Federal Rural de Pernambuco (UFRPE)

²Centro de Estudos e Sistemas Avançados do Recife (CESAR)

abilionbarros@gmail.com, andreza.leite, rafael.mello@ufrpe.br

Abstract. *This article aims to demonstrate how the Taxa de Distorção Idade-série (TDI) can be used to monitor the Brazilian education scenario using open data. It also presents how these data can be acquired, processed, and classified in order to provide an overview of this rate for Brazilian municipalities. With only the TDI indicator, it is already possible to observe fluctuations in the development of Brazilian education during the years 2018 to 2022, which was the chosen time frame for the development of this work.*

Resumo. *Este artigo busca demonstrar como a Taxa de Distorção Idade-série (TDI) pode ser utilizada para acompanhar o cenário da educação brasileira utilizando dados abertos. É apresentado também como esses dados podem ser adquiridos, processados e classificados de modo a fornecer um panorama dessa taxa para municípios brasileiros. Com apenas o indicador TDI é possível já observar flutuações do desenvolvimento da educação brasileira durante os anos 2018 a 2022 que foi a faixa temporal escolhida para o desenvolvimento deste trabalho.*

1. Introdução

A educação básica no Brasil enfrenta diversos desafios que impactam diretamente a qualidade do ensino e o desenvolvimento dos estudantes. Entre os principais desafios, destacam-se a desigualdade educacional, a infraestrutura precária das escolas, a formação e valorização dos professores, a deficiência na gestão escolar e a falta de recursos financeiros [Schwartzman and Brock 2005]. Sobre o contexto educacional a análise de dados já é amplamente utilizada como em [Fonseca and Namen 2016] o artigo destaca a utilidade da mineração de dados para explorar e analisar os dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), a fim de identificar padrões e relações entre as variáveis. O que foi apresentado no trabalho pode fornecer uma maneira de gerar melhorias no sistema educacional brasileiro e para aprimorar a qualidade dos dados disponíveis para análise.

Dentro da vasta quantidade de dados e indicadores fornecidos pelo INEP, este trabalho busca fornecer um panorama sobre um de seus indicadores, a Taxa de Distorção Idade-série (TDI). A TDI é calculada pelo Ministério da Educação (MEC) como uma medida que indica a proporção de alunos que estão em séries diferentes daquelas adequadas à sua idade cronológica. O cálculo da TDI envolve a comparação entre a idade dos alunos

e a série em que estão matriculados. O MEC utiliza os dados do Censo Escolar, que é realizado anualmente, para obter informações sobre a idade dos alunos e a série em que estão matriculados. Com base nessas informações, é possível calcular a TDI utilizando a seguinte fórmula:

$$TDI = \frac{\text{Número de alunos com idade superior à recomendada para a série}}{\text{Total de alunos matriculados na série}} \times 100$$

A fórmula permite determinar a proporção de alunos que estão em séries superiores às recomendadas para sua idade. Quanto maior o valor da TDI, maior é a taxa de distorção idade-série. É importante destacar que a TDI é apenas um indicador e não reflete necessariamente a qualidade do ensino.

Outros trabalhos já abordaram o TDI a nível municipal e estadual, destacando o trabalho de [Bezerra 2019], que analisa dados específicos da etapa final do Ensino Fundamental e investiga os principais fatores que contribuem para a ocorrência desse fenômeno, como a idade dos alunos, a série em que estão matriculados, a evasão escolar e a repetência. O objetivo é compreender a magnitude da distorção idade-série na rede municipal de Natal e identificar possíveis estratégias para enfrentar esse desafio educacional.

Diferentemente do trabalho citado, o objetivo deste é fornecer esse panorama a nível de todo o território brasileiro, podendo assim fornecer esse panorama e realizar o acompanhamento em anos importantes para a educação básica.

2. Aquisição e Processamento dos dados

O responsável pela publicação desses dados educacionais é o INEP, a entidade fornece os dados a nível de microdados, como o censo da educação básica e o censo da educação superior, como outros indicadores educacionais como taxas de aprovação, reprovação e abandono. Os dados de contexto educacional foram retirados em sua totalidade do site do INEP. Este trabalho busca criar um mapeamento do TDI para assim criar um panorama geral desses resultados levando em consideração a faixa temporal de 2018 a 2022, anos estratégicos que representam períodos pré e pós eventos pandêmicos vivenciados em 2020 e 2021, anos esses que trouxeram novos desafios [Dias et al. 2020] além dos já mencionados no começo desse trabalho, tanto para o momento de vivência da educação remota, quanto para sua retomada a sala de aulas.

Os dados foram adquiridos direto do portal do INEP, no portal é possível selecionar a nível federal, estadual, municipal e escolar. Dado que o primeiro objetivo deste trabalho é analisar o panorama das cidades a nível estadual.

2.1. Processamento de dados

Destacamos inicialmente que não é disposto nenhum dicionário de dados para esse tipo de recurso, porém suas colunas são de fácil entendimento e com isso é possível realizar sua utilização. Para essa primeira etapa do projeto foi utilizado a linguagem de programação Python¹ e sua biblioteca para análise de dados Pandas² determinando assim que todo o processo pode ser feito com ferramentas totalmente *open source*, trazendo a possibilidade de replicação dessa processo de forma gratuita.

¹<https://www.python.org/>

²<https://pandas.pydata.org/>

As etapas seguidas no processamento dos dados estão ilustradas na figura 1, demonstrando cada etapa do processo que foi seguido para criação da base de dados.

Extração dos dados

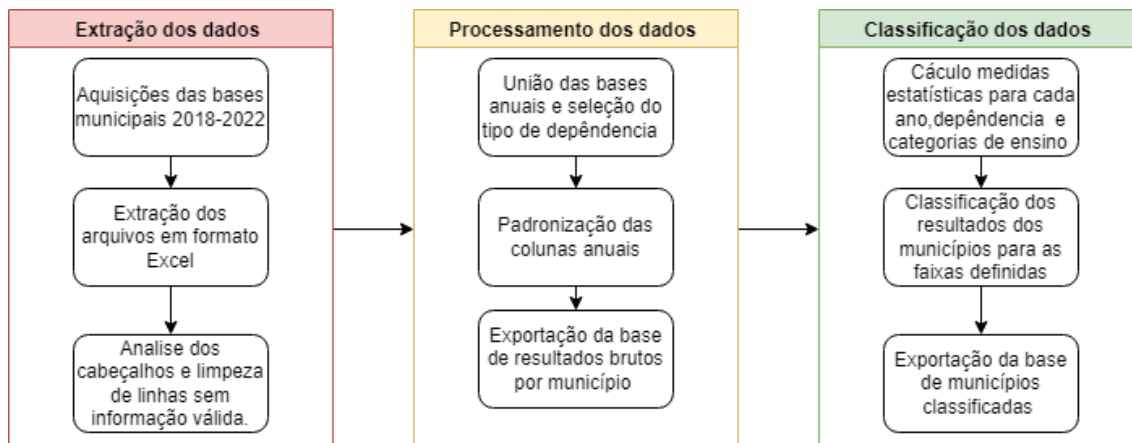


Figura 1. Fluxo de processamento dos dados utilizado.

Os dados foram extraídos diretamente do INEP³, onde foram selecionados os anos de interesse, e foi avaliado as colunas que seriam o alvo do estudo bem como a retirada de linhas desnecessárias e outros dados que não tratariam do escopo desse projeto.

Processamento dos dados

Nessa etapa, foi preciso analisar os dados buscando compreender quais etapas educacionais eram informadas e quais seriam escolhidas para esse trabalho, a dependência administrativa e localidade. Com base na Lei de Diretrizes e Bases da Educação Nacional⁴, é sabido que ao município só compete a estrutura escolar até os anos finais do ensino fundamental, pois a gerência de ensino do ensino médio é de responsabilidade do governo estadual. Além disso, ao criar a base de dados será utilizado o filtro de dependência administrativa para pública, pois o foco do estudo está na rede pública de ensino.

Classificação dos dados

Os municípios foram classificados individualmente por ano, por etapa escolar (Fundamental anos iniciais, Fundamental anos finais e Ensino médio) e tipo da localidade analisada. Municípios que por algum motivo não tiveram seus resultados apresentados não foram contabilizados.

A classificação é baseada na divisão dos dados em quartis, que é uma técnica estatística para dividir um conjunto de valores em quatro partes iguais, cada uma representando uma porcentagem da distribuição dos dados. Fazendo com que assim os dados sejam sempre calculados baseado naquele ano e com a realidade expressa naquele espaço de tempo, buscando assim levar em conta as flutuações do TDI através dos anos estudados. A classificação foi definida da seguinte forma:

- Valor TDI no intervalo de 0 a 25% (primeiro quartil), será classificado como 1.
- Valor TDI no intervalo de 25% a 50% (segundo quartil), será classificado como 2.
- Valor TDI no intervalo de 50% a 75% (terceiro quartil), será classificado como 3.

³<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/indicadores-educacionais/taxas-de-distorcaao-idade-serie>

⁴https://www.planalto.gov.br/ccivil_03/leis/l9394.htm

- Valor TDI acima de 75% (quarto quartil), será classificado como 4.

Formalizando assim a classificação dos municípios em sua determinada faixa educacional em determinado ano. Registrando que quanto menor a faixa em que o município for classificado, menor foi seu TDI, sendo assim tendo um melhor desempenho educacional neste indicador.

Exemplificando a necessidade de gerar as faixas dinamicamente para cada ano, para o ano de 2018 o TDI foi de 2.9 para o ensino fundamental anos iniciais, o primeiro quartil. Então os melhores serão os resultados menores que esse, já para 2020 o primeiro quartil foi de 3.5, onde mais uma vez os melhores que receberam a faixa 1 foram apontados com valores abaixo dessa faixa, garantindo assim que o sistema de classificação irá respeitar as flutuações educacionais que existiram ao decorrer da faixa de tempo analisada.

3. Resultados

Após a conclusão da etapa de aquisição e processamento é possível realizar novas agregações buscando obter algumas informações sobre os dados dispostos. Por se tratar da primeira etapa desse estudo, as informações apresentadas a seguir vão ser sobre a mesma localidade *Total*, que corresponde ao resultado do urbano e rural juntos. O objetivo da classificação nos intervalos de quantil é poder dividir em grupos com em média 1380 a 1400 municípios em cada faixa classificatória, e assim podendo observar quais os valores que definiram o limiar de classificação e assim observar sua mudança durante os anos.

3.1. Ensino fundamental anos iniciais e finais

Para os anos iniciais podemos observar na tabela 1 que indica os valores que serviram como base para essa classificação que a variação entre o primeiro quantil, onde estão alocado os municípios com menores índices, teve seu valor reduzido pela metade ao compararmos o ano de 2018 com o de 2022, efeito que pode ser observado para as demais faixas, demonstrando assim que mesmo com as adversidades pode-se reduzir o valor do TDI. Já para os anos finais podemos visualizar, na tabela 2, que houve uma queda crescente do TDI para o limiar da classificação das faixas. Mesmo que não da mesma magnitude do anterior, vemos uma queda crescente.

Ano	Q1	Q2	Q3
2018	4.3	9.2	17.4
2019	4.0	8.5	16.4
2020	3.6	7.9	15.0
2021	2.7	5.6	10.8
2022	2.3	4.8	9.0

Tabela 1. Valores dos quartis por ano para o Ensino Fundamental anos iniciais

Ano	Q1	Q2	Q3
2018	16.7	25.9	37.2
2019	15.5	24.3	35.575
2020	15.4	23.9	34.275
2021	13.2	21.3	32
2022	11.0	18.2	28.2

Tabela 2. Valores dos quartis por ano para o Ensino Fundamental anos finais

3.2. Estudo de caso do estado de Pernambuco

Outra possibilidade que temos é analisar a nível estadual a taxa TDI, visando mensurar a quantidade de municípios em cada faixa. Selecionando o estado de Pernambuco, todos os seus 185 municípios foram relacionados. As colunas representam a quantidade de municípios em cada faixa. Na tabela 3 podemos observar que o estado, de forma coletiva,

ANO	F1	F2	F3	F4
2018	3	30	99	53
2019	2	35	102	46
2020	3	38	107	37
2021	4	35	108	38
2022	3	36	93	53

Tabela 3. Quantidade de municípios em cada faixa para Ensino Básico Anos Iniciais

ANO	F1	F2	F3	F4
2018	0	10	73	102
2019	0	14	79	92
2020	1	20	76	88
2021	3	32	69	81
2022	5	29	60	91

Tabela 4. Quantidade de municípios em cada faixa para Ensino Básico Anos Finais

conseguiu manter mais da metade de seus municípios entre as faixas 2 e 3 não conseguindo alavancar seus resultados durante o período de tempo analisado. Já na observação da tabela 4, que reporta os dados relativo aos anos finais, mostra que existiu uma progressão para outras faixas onde após 2020, mesmo sendo um ano pandêmico, se conseguiu avançar na quantidade de municípios em faixas melhores como a 2 e 3, e pela primeira vez nos anos observados no estudo alcançar a faixa 1.

4. Conclusão

Em conclusão, este estudo permitiu-nos apresentar uma visão abrangente sobre o TDI e ressaltar a importância do seu acompanhamento para o monitoramento efetivo da educação básica brasileira. Ao compreendermos as nuances e desafios associados ao TDI, podemos desenvolver estratégias mais eficazes para aprimorar o sistema educacional. Ressalta-se que esse não é um indicador direto de qualidade de ensino e sim um indicador para acompanhamento sobre a idade que os estudantes estão sendo matriculados na rede pública de ensino. Existem outros indicadores como Sistema Nacional de Avaliação da Educação Básica (SAEB) que podem atestar a qualidade do ensino das redes de ensino. O acompanhamento do TDI tende a servir de guia para que outros fenômenos escolares, como evasão e reprovações, possam ser investigadas mais de perto. Podendo assim buscar entender qual é a real faixa escolar que aquela localidade está servindo e não apenas a faixa ideal que é relatada na literatura.

O desenvolvimento da base de dados que foi desenvolvida para o desenvolvimento desse panorama pode ser encontrada nesse endereço <https://zenodo.org/record/8105876> podendo ser aplicados em algoritmos inteligentes de classificação ou agrupamento.

Na continuidade desse projeto temos o objetivo de incorporar outros indicadores definidos pelo INEP, tal como as taxas de aproveitamento, como também dados de quantitativos e qualitativos sobre as matrículas disponíveis no Censo educacional da educação básica com o intuito de realizar o cruzamento de todas as informações que compõem o TDI a fim de rastrear quais estão trazendo mais impacto para sua redução de forma positiva e contínua através dos anos.

Referências

- Bezerra, A. R. d. F. (2019). Distorção idade-série nos anos finais do ensino fundamental: o caso da rede municipal de ensino do natal/rn. Master's thesis, Brasil.
- Dias, G. N., Vogado, G. E. R., Barreto, W. D. L., da Silva Junior, W. L., da Silva Barbosa, E., Rodrigues, A. E., Junior, A. F. S., and Costa, C. A. C. (2020). Retorno às aulas presenciais no sistema educacional do estado do pará-brasil: Obstáculos e desafios durante a epidemia de covid-19 (sars-cov-2). *Brazilian Journal of Development*, 6(6):37906–37924.
- Fonseca, S. O. d. and Namen, A. A. (2016). Mineração em bases de dados do inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. *Educação em Revista*, 32:133–157.
- Schwartzman, S. and Brock, C. (2005). Os desafios da educação no brasil. *Rio de Janeiro: Nova Fronteira*, 1320.

6 Apêndice D - Aplicação de Learning Analytics para Identificação de Tomada de Decisão sobre a Distorção Idade-Série no Brasil

Barros, Abílio, Elyda Laisa Soares Xavier, Gabriel Alves, and Rafael Ferreira Mello. "Aplicação de Learning Analytics para Identificação de Tomada de Decisão sobre a Distorção Idade-Série no Brasil." Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil, Passo Fundo/RS, 2023. SBC, 2023, pp.21-31.

Contribuição do autor: Responsável pela definição e implementação da metodologia, desenvolvimento e avaliação dos resultados e escrita do texto.

Aplicação de Learning Analytics para Identificação de Tomada de Decisão sobre a Distorção Idade-Série no Brasil

Abílio Nogueira Barros¹, Elyda Laisa Soares Xavier², Gabriel Alves¹, Rafael Ferreira Mello¹

¹Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)

²Universidade Estadual de Pernambuco (UPE)

{abilio.nogueira,rafael.mello,gabriel.alves}@ufrpe.br, elyda.freitas@upe.br

Abstract. *The Age-Grade Distortion Rate (from the Portuguese Taxa de Distorção Idade-Série - TDI) is a key measure for assessing the number of students enrolled in academic years that diverge from their age-appropriate levels. This study leverages Learning Analytics and multi-source data to extract insights to support informed decision-making in this context. Utilizing data from the Basic Education Census, we conducted predictive analytics and feature importance investigation to perform an in-depth analysis. This analysis aims to elucidate the primary factors influencing TDI, enabling a discussion on potential areas for improvement in schools to reduce TDI.*

Resumo. *A Taxa de Distorção Idade-Série (TDI) é um indicador que mede a quantidade de alunos que estejam em um ano curricular diferente de sua idade esperada. Este artigo propõe utilizar Learning Analytics e dados de diferentes fontes para extrair informações relevantes para a tomada de decisão no contexto do TDI. Utilizando fundamentalmente dados do censo da educação básica, foram realizadas previsões e análise de importância de características para realizar uma análise detalhada do que mais influencia esse indicador e assim prover a discussão sobre o que pode ser mantido ou melhorado em escolas para reduzir o TDI.*

1. Introdução

O sistema educacional brasileiro é permeado por uma série de desafios que impactam diretamente em sua qualidade e paridade entre os entes que compõe o sistema educacional [Schwartzman and Brock 2005, Palomino et al. 2022]. Por exemplo, o Brasil tem um território com proporções continentais e seus mais de 47 milhões de alunos matriculados apenas na educação básica 2022, distribuídos em mais de 178 mil escolas em todo o Brasil¹.

Para auxiliar no acompanhamento de um sistema educacional tão complexo, vários indicadores foram definidos, como taxas de rendimento escolar [Justino 2022], complexidade de gestão escolar [de Andrade et al. 2020], taxa de alfabetização [Bernardi and Luchese 2020], taxa de distorção idade-série (TDI) [NOGUEIRA and Silva 2022], índice de desenvolvimento da educação básica (IDEB) [Rodrigues et al. 2016]. Tais indicadores são calculados e divulgados pelo INEP.

¹<https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar/resultados/>

A disponibilização dos dados utilizados para o cálculo desses indicadores permite que sejam realizadas diversas análises, como a desigualdade na educação brasileira [Vitelli et al. 2019], a relação da educação com programas sociais como o Bolsa Família [Santos et al. 2019] e o desempenho educacional dos estudantes através de avaliações como a Prova Brasil [Lacruz et al. 2019].

Neste contexto, a TDI é calculada pelo INEP como uma medida que indica a proporção de alunos em séries diferentes em mais de dois anos da idade correta para a série em questão. O cálculo da TDI envolve a comparação entre a idade dos alunos e a série em que estão matriculados.

Os dados fornecidos pelo INEP podem ser analisados em um contexto histórico através de técnicas de Learning Analytics para extração e processamento desses dados, processo esse utilizado em trabalhos como o de [do Nascimento et al. 2018], que aplicou técnicas de mineração de dados voltados aos dados buscando aplicar modelos de regressão visando encontrar padrões preditivos para evasão escolar. Outro exemplo é [Fonseca and Namen 2016], que de forma mais exploratória visa buscar fornecer apontamentos para a melhoria do sistema educacional baseado no censo da educação básica.

Este estudo propõe utilizar técnicas de Learning Analytics para identificar informações relevantes para a tomada de decisão no contexto de TDI. Para isso, foram utilizados dados do censo da educação básica para identificar as características que podem exercer influência sobre variações no valor desse indicador. Como resultado, é possível identificar pontos de melhoria, que podem ser discutidos gestores educacionais e governamentais, a fim de serem aperfeiçoados.

2. Trabalhos Relacionados

O TDI é um indicador bem consolidado na temática de mapeamento educacional por meio de dados abertos. Para melhor formalizar o cálculo deste indicador, a fórmula a seguir ilustra como a TDI é calculada [Soares and Sátyro 2008]:

$$TDI = \frac{\text{Número de alunos com idade superior à recomendada para a série}}{\text{Total de alunos matriculados na série}} \times 100$$

Já existem trabalhos que utilizam o monitoramento do TDI para o acompanhamento, não só educacional como o de outras áreas como segurança pública. Como visto em [Ferreira and Teixeira 2018], que retratou como foi possível cruzar os indicadores de violência em localidades que possuíam um alto TDI buscando indicar relação de causa-efeito.

Esse indicador também pode ser usado para avaliar a defasagem entre a educação urbana e rural. Em [NOGUEIRA and Silva 2022] o indicador foi utilizado como ferramenta de acompanhamento de escolas situadas em zona rural e expor assim fatores intra e extraescolares a seus gestores.

Também existem trabalhos que aplicam técnicas de Learning Analytics em microrredes educacionais que tratam, desde o agrupamento de estudantes de ensino superior [da Silva Vieira et al. 2022], onde aplicou o algoritmo *k-means* buscando avaliar a qualidade do ensino superior brasileiro. Como em [Viana et al. 2022], que propôs uma forma de classificação de evasão e graduação aplicados a cursos de computação da Universidade Federal do Piauí (UFPI).

Esses trabalhos serviram para nortear a forma com que esse indicador já foi utilizado e para fortalecer o entendimento do processo já utilizado por outros grupos de pesquisa. Não encontramos trabalhos específicos sobre a utilização de Learning Analytics para tomada de decisão sobre o TDI. Com isso, nosso trabalho visa apresentar um panorama de todas as escolas brasileiras entre 2018 a 2022, pois anos esse que foram anos e transições tanto governamentais quanto sociais, com a educação remota de emergência durante a pandemia [de Souza et al. 2021].

3. Metodologia

O presente estudo emprega técnicas de Learning Analytics (LA) na base de dados do censo educacional para alcançar o objetivo. Inicialmente, foi construída a base de dados, compreendendo todas as variáveis necessárias para a investigação. Através da aplicação de LA visando avaliar diversas combinações de características visando aprimorar a precisão preditiva do modelo. Ao final dos experimentos foram realizadas análise dos resultados, destacando a identificação das características mais relevantes para aquele resultado. Cada etapa está descrita em mais detalhes nas próximas seções.

3.1. Aquisição e Processamento dos Dados

Os dados utilizados para esse projeto foram todos adquiridos do INEP respeitando a faixa temporal definida por esse estudo. As bases necessárias foram o TDI ao nível de entidade escolar ² e o censo escolar da educação básica ³.

Para o processamento do censo escolar foi realizada a extração e transformação de dados, com base no que foi descrito em [Barros et al. 2022, de Albuquerque et al. 2022], a fim de unificar os anos e assim preparar uma base para os algoritmos de LA. O processo necessário para a criação da base envolveu a remoção de colunas que tivessem uma quantidade de valores nulos que ultrapassem 80% dos dados [Barros et al. 2022, de Albuquerque et al. 2022]. Além disso, foram removidas as colunas de identificação direta da instituição escolar, como colunas que falassem sobre localidade geográfica [Barros et al. 2023]. Ficando assim a base final com cerca de 335 atributos, a lista de colunas removidas pode ser encontrada nesse endereço ⁴. Foi necessário criar apenas uma coluna não original do censo para que pudesse se adequar a próxima etapa do estudo, sendo ela DIAS LETIVOS, pois se trata da diferença entre as datas informadas no censo para o cômputo em dias, criando assim uma coluna quantitativa onde anteriormente haviam duas do tipo data.

Para a padronização dos arquivos originais do indicador TDI, coluna alvo para esse estudo, o processo a extração dos arquivos focou no ensino fundamental anos iniciais, por questões de limitação de espaço no artigo. Com isso, apenas escolas que apresentaram o resultado desse indicador em seu respectivo ano. A partir disto, uniu-se os resultados do censo em que existia correspondência direta entre instituição e ano, formando assim a base entre o resultado que visamos entender com as informações escolares dispostas no censo. Formando assim uma base de pouco mais de 533 mil registros.

²<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/indicadores-educacionais/taxas-de-distorcão-idade-serie>

³<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-escolar>

⁴<https://www.zenodo.org/record/6666613>

Assim como outros indicadores já citados nesse trabalho, o TDI é expresso como o percentual de alunos com matrículas escolares que estão atrasados em relação à idade escolar adequada. Neste caso, quanto menor sua porcentagem, melhor é a situação da instituição. Em outras palavras, escolas com TDI próximo de 0% tem menos alunos com a distorção de idade e série.

3.2. Análise dos Dados

Após a preparação da base de dados, foi definido que seria utilizado algoritmos de regressão, visto que o valor alvo é numérico, para prever o TDI. Para essa etapa do projeto foi utilizado um algoritmo caixa branca a fim de aumentar a explicabilidade de seus resultados, mais especificamente utilizamos o *Decision Tree Regressor* [Mahbooba et al. 2021]. Para buscar a melhoria no algoritmo escolhido, foi utilizada a otimização de hiperparâmetros, visando melhorar o algoritmo antes de sua execução com base nos dados que vão ser executados, podendo assim gerar resultados mais satisfatórios [Luo 2016]. A fim de gerar o melhor modelo possível, utilizamos a biblioteca *GridSearchCV*, que implementa o algoritmo Grid Search com o objetivo de otimizar os parâmetros do modelo [Huang et al. 2012].

Com a escolha dos melhores parâmetros para aplicação do modelo, foram realizados três configurações de experimentos nos dados coletados:

Experimento 1: Nessa etapa, todas as colunas da base de dados foram mantidas.

Experimento 2: Foram selecionadas apenas as características que informavam sobre matrícula, sendo assim todas de teor quantitativo.

Experimento 3: Utilização das características qualitativas do censo, excluindo as matrículas.

Para cada experimento, os classificadores foram executados com a configuração determinada pelo otimizador de hiperparâmetros. Após a execução foram escolhidas as métricas tradicionais para avaliação de algoritmos de regressão para medir o desempenho dos modelos desenvolvidos:

- Correlação de Pearson: Mede a força e direção da correlação linear entre duas variáveis contínuas, variando de -1 a 1 [Stoean et al. 2013];
- Correlação de Spearman: Avalia a relação monotônica entre duas variáveis, atribuindo classificações ordinais aos dados [Budach et al. 2022];
- RMSE (Root Mean Square Error): Métrica de erro que calcula a diferença média entre os valores previstos e os valores reais, onde quando menor o valor, melhor o desempenho do modelo [Arghandabi and Shams 2020].

Além disto, no contexto deste estudo, foram selecionadas as dez melhores características apontadas por cada um dos algoritmos em seus respectivos experimentos. Para o levantamento dessas melhores características foi utilizado o método *feature_importances_*⁵, as características são calculadas com base na média e desvio padrão das do acúmulo da diminuição das impurezas geradas nas classes das árvores.

O objetivo dessa etapa não é de fato selecionar apenas as dez colunas para novas rodadas de processamento em busca de prever o TDI de escolas em novos anos, e sim

⁵https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

poder investigar as características apontadas pelo modelo com maior potencial de caracterizar o TDI.

Junto à descrição do que cada característica representa, teremos uma coluna informando os valores das escolas que apontaram uma melhor situação, dado que o TDI menor indica uma melhor situação educacional, e as que estão no grupo em pior situação, instituições que estejam superior à média de TDI. Para a análise a seguir, o TDI médio da educação básica foi de 11.71%, dividindo assim a base de dados de mais de 500 mil registros entre TDI inferior à média (indicando um bom índice) e TDI superior à média (indicando um índice ruim).

Em resumo, para a avaliação dos resultados, foi aplicada a seguinte abordagem:

1. Calcular a média do TDI das entidades escolas;
2. Após definido o valor médio, a base de dados é adicionada a coluna se a entidade escolar está abaixo (valor de TDI menor que a média, situação melhor) ou acima (valor de TDI maior que a média, situação pior);
3. São selecionadas as dez características mais importantes de cada experimento, as colunas podem ser de dois tipos:
 - Quantitativos: Onde será exibida a quantidade média e o desvio padrão da característica.
 - Categórica: Onde será exibida a quantidade de escolas em cada categoria daquela característica.

4. Resultados

Como cada experimento possuía diferente conjunto de dados, foram gerados os resultados com base em cada rodada de execução do algoritmo e seus respectivos dados. As tabelas 1, 2 e 3 mostram os resultados para a melhor execução do modelo após a utilização do grid search nos parâmetros⁶.

Tabela 1. Resultados das Métricas — Experimentos 1, 2 e 3

Experimento	Descrição	Métrica		
		Pearson	Spearman	RMSE
1	Base completa	0,786225	0,740435	8,571647
2	Base matrículas	0,797232	0,735176	8,358507
3	Base atributos	0,582022	0,510149	11,289482

Após a execução dos algoritmos é calculada a importância das características. Essa abordagem permite identificar quais características têm maior contribuição para a previsão e, potencialmente, descartar características menos relevantes, simplificando o modelo e melhorando sua capacidade de generalização. Em cada experimento as colunas categóricas são destacadas a fim de exemplificar como aquele resultado foi aplicado na tabela relacionado a cada experimento, buscando fornecer uma maior interpretabilidade da característica que foi destacada.

⁶Não foi possível mostrar o resultado de todos os parâmetros devido à limitação de espaço do artigo

Ao observarmos os resultados na Tabela 2, conseguimos observar que do grupo de entidades escolares que possuem um TDI inferior à média, temos mais matrículas, com quantidades maiores de turmas, em todas as etapas do ensino fundamental. A tabela também aponta que as entidades escolares que possuem internet disponível na entidade escolar são 80% dentre os posicionados abaixo da média. Já as entidades de dependência pública representam 90% das escolas que possuem esse indicador mais elevado que a média. Observa-se também que, embora esteja consideravelmente distante do público-alvo deste estágio educacional, há uma significativa quantidade de matrículas de estudantes com 18 anos ou mais. Essas matrículas indicam que esses indivíduos encontram-se além da faixa etária prevista para a conclusão do ciclo escolar completo. Já nesse primeiro experimento podemos destacar também que a quantidade de escolas públicas é mais de 90% das instituições com valores de TDI acima da média nessa faixa, necessitando assim um maior olhar ao sistema público de ensino.

Tabela 2. Importância das Características — Experimento 1

Característica	Descrição	Importância	TDI	
			Inferior	Superior
QT_MAT_BAS_6_10	Qtd de matrículas 6 a 10 anos	0.24	148.25 (160.94)	93.53 (120.36)
QT_MAT_BAS_11_14	Qtd de matrículas 11 a 14 anos	0.19	65.97 (112.12)	57.99 (101.51)
IN_INTERNET	Possuem acesso a internet	0.016	82.26 (17.74)	56.62 (43.38)
QT_MAT_BAS_15_17	Qtd de matrículas 15 a 17 anos	0.09	22.34 (65.55)	18.50 (47.75)
QT_MAT_FUND_AI	Qtd de matrículas no fundamental anos iniciais	0.05	153.74 (168.38)	110.46 (139.38)
QT_MAT_FUND_AF	Qtd de matrículas no fundamental anos finais	0.04	65.44 (118.97)	53.23 (113.78)
QT_MAT_BAS_BRANCA	Qtd de matrículas de alunos declarados brancos	0.04	108.03 (163.59)	36.78 (81.35)
TP_DEPENDENCIA	Tipo de dependência da instituição	0.03	71.41 (28.58)	92.87 (7.13)
IN_ENERGIA	Utiliza fornecimento de energia pública	0.02	97.80 (2.20)	88.08 (11.92)
QT_MAT_BAS_18 MAIS	Qtd de matrículas 18 anos ou mais.	0.02	9.85 (40.14)	16.39 (45.99)

Avaliando agora os resultados do experimento 2, apresentados na tabela 3. Podemos observar que em números, que apenas 3 das 10 características mais importantes apontadas pelo modelo indicam que mais matrículas trazem um TDI menor que a média geral dos casos analisados. Trazendo assim a leitura que o problema não está diretamente ligado a quantidade de matrículas da instituição nem a quantidade de diferentes faixas etárias de matrículas na mesma instituição. Ainda na mesma tabela vemos que a questão da declaração de etnia se destacando, onde instituições com uma maior quantidade de alunos declarados brancos estariam abaixo da média do indicador, obtendo assim um melhor resultado, indicativo desse que pode remeter a questões mais uma vez sobre o sistema escolar público/privado.

Tabela 3. Importância de Característica - Experimento 2

Característica	Descrição	Importância	TDI	
			Inferior	Superior
QT_MAT_BAS_6_10	Qtd de matrículas 6 a 10 anos	0.28	148.25 (160.94)	93.53 (120.36)
QT_MAT_BAS_11_14	Qtd de matrículas 11 a 14 anos	0.22	65.97 (112.12)	57.99 (101.51)
QT_MAT_BAS_BRANCA	Qtd de matrículas de alunos declarados brancos	0.15	108.03 (163.59)	36.78 (81.35)
QT_MAT_FUND_AI	Qtd de matrículas no fundamental anos iniciais	0.09	153.74 (168.38)	110.46 (139.38)
QT_MAT_BAS_15_17	Qtd de matrículas 15 a 17 anos	0.07	22.34 (65.55)	18.50 (47.75)
QT_MAT_FUND_AF	Qtd de matrículas no fundamental anos finais	0.04	65.44 (118.97)	53.23 (113.78)
QT_MAT_EJA_FUND	Qtd de matrículas na modalidade EJA no ensino fundamental.	0.03	6.61 (30.31)	14.98 (46.0)
QT_MAT_BAS_18 MAIS	Qtd de matrículas 18 anos ou mais.	0.03	9.85 (40.14)	16.39 (45.99)
QT_MAT_MED	Qtd de matrículas no ensino médio.	0.01	18.06 (69.54)	5.45 (41.91)
QT_MAT_BAS_PRETA	Qtd de matrículas de alunos declarados pretos.	0.01	7.43 (14.98)	8.36 (19.27)

Por fim, temos os resultados expressos na Tabela 4 e nesse resultado podemos destacar que a maioria de características diretamente ligadas a estrutura escolar sejam mais relevantes. Foi destacado o uso de internet tanto no geral como diretamente ligada ao processo de ensino e aprendizagem, demonstrando que a utilização dessa tecnologia, que também requer outras atreladas a seu uso, dando assim um destaque a temática de tecnologias na educação que deve ser considerado por gestores e professores. A quantidade de dias letivos também é mostrada, onde uma maior quantidade de dias traria um TDI abaixo da média. A quantidade turmas e de docentes em uma maior quantidade pode acarretar valores de indicadores mais baixos, levantando a discussão sobre quantidade de alunos por turma e um maior quadro de docentes.

5. Implicações Práticas

As implicações práticas destes resultados podem refletir em diferentes contextos. A primeira, e principal implicação, está relacionada à tomada de decisão para políticas públicas. Por exemplo, a identificação que a quantidade de matrículas por escola influencia a TDI pode levar a decisões de redistribuição de alunos. Ainda neste contexto, o experimento 3 lista uma grande quantidade de características que estão impactando negativamente o TDI que podem ser direcionadoras de políticas públicas eficientes.

O segundo ponto relevante das análises estão relacionadas ao contexto dos resultados. As tabelas apontam para algumas características determinantes para maiores (ou

Tabela 4. Feature importance experimento 3

Característica	Descrição	Importância	TDI	
			Inferior	Superior
IN_INTERNET	Possuem acesso a internet	0.22	82.26 (17.73)	56.62 (43.38)
IN_ENERGIA_REDE_PUBLICA	Utiliza fornecimento de energia pública	0.10	97.80 (2.20)	88.08 (11.92)
TP_DEPENDENCIA	Tipo de depência da instituição	0.06	71.41 (28.59)	92.87(7.13)
IN_LOCAL_FUNC_SOCIOEDUCATIVO	A escola disponibiliza atendimento socioeducativo	0.04	0.03 (99.97)	0.23 (99.77)
QT_TUR_FUND_AI	Número de Turmas de Ensino Fundamental - Anos Iniciais	0.03	7.08 (6.39)	4.82 (5.62)
TP_INDIGENA_LINGUA	Língua indígena que é ministrada.	0.03	99.04 (0.96)	95.56 (4.42)
DIAS_LETIVOS	Quantidade de dias letivos.	0.05	312.08 (19.66)	305.49 (25.27)
QT_DOC_FUND_AI	Número de Docentes do Ensino Fundamental - Anos Iniciais	0.03	8.98 (8.37)	5.45 (6.52)
IN_INTERNET_APRENDIZAGEM	Possuem acesso à internet para uso nos processos de ensino e aprendizagem	0.02	45.50 (54.50)	19.73 (80.27)
CO_LINGUA_INDIGENA_1	Escolha da língua indígena	0.03	99.04 (0.96)	95.56 (4.44)

seja, piores) TDIs para escolas como a restrição no acesso à internet e energia (por exemplo). Além disto, grupos desfavorecidos também influenciam bastante no resultado. Por fim, os resultados apresentados também podem ser utilizados por gestores escolares podem validar aspectos administrativos da escola, como quantidade de alunos por turma ou utilização de laboratórios.

6. Conclusão

Os resultados apresentados neste trabalho mostram que o *Decision Tree Regressor* consegue identificar as variáveis que mais impactaram o TDI. A divisão das bases de dados, para assim gerar diferentes experimentos, pôde trazer uma melhor visão dos resultados, visto que a quantidade de matrículas foi um dos elementos mais apontados desde a primeira rodada de execução do algoritmo. Levantando assim que escolas com uma maior quantidade de turmas e matrículas tendem a possuir um TDI menos elevado em relação as com menos matrículas e turmas, não só apenas da faixa escolar levantado no estudo.

Com vistas à continuidade desta pesquisa, almejamos ampliar a análise para abranger os anos finais do ensino fundamental e o ensino médio. Adicionalmente, planejamos efetuar uma comparação das características mais significativas em cada etapa educacional, acompanhando seus resultados por entidade escolas. Paralelamente, aspiramos validar tais resultados junto a gestores educacionais, que, devido ao seu conhecimento especializado na área, poderão enriquecer a avaliação dos resultados apontados por este algoritmo ou futuras otimizações do mesmo [Spanol et al. 2022].

Referências

- Arghandabi, H. and Shams, P. (2020). A comparative study of machine learning algorithms for the prediction of heart disease. *Int J Res Appl Sci Eng Technol*. <https://doi.org/10.22214/ijraset>.
- Barros, A. N., Alencar, A., Nascimento, A., de Albuquerque, A. F., and Mello, R. F. (2022). Elaboração do conjunto de dados agregados do censo da educação básica. In *Anais do IV Dataset Showcase Workshop*, pages 35–45. SBC.
- Barros, A. N., de Albuquerque, A. F., Alencar, A., Mello, R. F., Alves, G., and Bittencourt, I. M. (2023). Arquitetura de dados educacionais como plataforma para governo inteligente-utilizando dados abertos para apoio à gestão educacional baseada em evidências. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 130–140. SBC.
- Bernardi, M. C. and Luchese, T. Â. (2020). A taxa de alfabetização de antônio prado, rio grande do sul (1895-1920). *Revista Educação em Questão*, 58(56).
- Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., and Harmouch, H. (2022). The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529*.
- da Silva Vieira, A., Bertolini, D., and Schwerz, A. L. (2022). Análise do desempenho no enade dos concluintes de computação usando técnica de agrupamento. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 834–845. SBC.
- de Albuquerque, A. F., Barros, A. N., Alencar, A., Nascimento, A., Bittencourt, I. M., and Mello, R. F. (2022). Dataset de estimativas populacionais desagregada por município e idade 2014-2020. In *Anais do IV Dataset Showcase Workshop*, pages 25–34. SBC.
- de Andrade, M. C. B., Silva, L. F., Fecury, A. A., de Oliveira, E., Dendasck, C. V., de Araújo, M. H. M., da Souza, K. O., da Silva, I. R., de Medeiros Moreira, E. C., Pascoal, R. M., et al. (2020). Indicadores de complexidade de gestão em escolas públicas e privadas de duas cidades do estado do amapá entre 2014 e 2018. *Research, Society and Development*, 9(9):e856998112–e856998112.
- de Souza, G. H. S., Jardim, W. S., Marques, Y. B., Junior, G. L., dos Santos, A. P. S., and de Paula Liberato, L. (2021). Educação remota emergencial (ere): um estudo empírico sobre capacidades educacionais e expectativas docentes durante a pandemia da covid-19. *Research, Society and Development*, 10(1):e37510111904–e37510111904.
- do Nascimento, R. L. S., da Cruz Junior, G. G., and de Araújo Fagundes, R. A. (2018). Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. *RENOTE*, 16(1).
- Ferreira, V. B. and Teixeira, E. C. (2018). O impacto da distorção idade-série sobre a criminalidade nos municípios de minas gerais. *Revista Brasileira de Segurança Pública*, 12(2):269–291.
- Fonseca, S. O. d. and Namen, A. A. (2016). Mineração em bases de dados do inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. *Educação em Revista*, 32:133–157.

- Huang, Q., Mao, J., and Liu, Y. (2012). An improved grid search algorithm of svr parameters optimization. In *2012 IEEE 14th International Conference on Communication Technology*, pages 1022–1026. IEEE.
- Justino, M. R. (2022). A relação do esforço docente e da infraestrutura escolar nas taxas de rendimento escolar: uma análise para a cidade do natal no ano de 2019. B.S. thesis, Universidade Federal do Rio Grande do Norte.
- Lacruz, A. J., Américo, B. L., and Carniel, F. (2019). Indicadores de qualidade na educação: análise discriminante dos desempenhos na prova brasil. *Revista brasileira de educação*, 24:e240002.
- Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5:1–16.
- Mahbooba, B., Timilsina, M., Sahal, R., and Serrano, M. (2021). Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021:1–11.
- NOGUEIRA, M. D. O. E. and Silva, L. C. (2022). Escolarização em áreas rurais: a distorção idade-série na ótica dos gestores. *Estudos em Avaliação Educacional*, 33.
- Palomino, P., Falcao, T. P., Medeiros, R., Uehara, M., Bittencourt, I., and Mello, R. F. (2022). Plataformas de dados educacionais: Análise com foco no plano nacional de educação. In *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 60–68. SBC.
- Rodrigues, E. C. et al. (2016). Indicadores educacionais e contexto escolar: uma análise das metas do ideb. *Estudos em Avaliação Educacional*, 27(66):662–688.
- Santos, M. C. S., Delatorre, L. R., Ceccato, M. d. G. B., and Bonolo, P. d. F. (2019). Programa bolsa família e indicadores educacionais em crianças, adolescentes e escolas no brasil: revisão sistemática. *Ciência & Saúde Coletiva*, 24:2233–2247.
- Schwartzman, S. and Brock, C. (2005). Os desafios da educação no brasil. *Rio de Janeiro: Nova Fronteira*, 1320.
- Soares, S. and Sátyro, N. (2008). O impacto de infra-estrutura escolar na taxa de distorção idade-série das escolas brasileiras de ensino fundamental: 1998 a 2005. Technical report, Texto para Discussão.
- Spanol, M., Oliveira, E., Alves, G., Bittencourt, I. M., Falcao, T. P., and Mello, R. F. (2022). Uso de agrupamento para avaliação de desempenho educacional e apoio à gestão em áreas de investimento. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 944–955. SBC.
- Stoean, C., Preuss, M., and Stoean, R. (2013). Ea-based parameter tuning of multimodal optimization performance by means of different surrogate models. In *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*, pages 1063–1070.
- Viana, F. S., Santana, A. M., and Rabêlo, R. d. A. L. (2022). Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 908–919. SBC.

Vitelli, R. F., Fritsch, R., and da Silva, R. D. (2019). A desigualdade brasileira revelada pelo resultado de indicadores educacionais.

7 Entendendo os Fatores de Previsão do TDI nas Escolas Públicas Brasileiras: Uma Abordagem Usando a Técnica SHAP

Barros, Abílio, Gabriel Alves, and Rafael Ferreira Mello. "Entendendo os Fatores de Previsão do TDI nas Escolas Públicas Brasileiras: Uma Abordagem Usando a Técnica SHAP." Anais do XXXV Simpósio Brasileiro de Informática na Educação, Rio de Janeiro/RJ, 2024. SBC, 2024, pp.1529-1542.

Contribuição do autor: Responsável pela definição e implementação da metodologia, desenvolvimento e avaliação dos resultados e escrita do texto.

Entendendo os Fatores de Previsão do TDI nas Escolas Públicas Brasileiras: Uma Abordagem Usando a Técnica SHAP

Abílio Nogueira Barros¹,
Gabriel Alves¹, Rafael Ferreira Mello^{1,2}

¹Departamento de Computação - Universidade Federal Rural de Pernambuco (UFRPE)

²Centro de Estudos e Sistemas Avançados do Recife (CESAR)

abilionbarros@gmail.com, gabriel.alves, rafael.mello@ufrpe.br

Abstract. *The Age-Grade Distortion Rate (TDI) measures the number of students not in the expected grade for their age. This article employs machine learning techniques to predict TDI values from 2018 to 2023, covering pre-pandemic, pandemic, and post-pandemic periods. Using data from the Basic Education Census, we demonstrate the algorithm selection process and the application of SHAP for interpreting its metrics. Our objective is to identify the most important features highlighted by the predictive model and to stimulate discussion on qualitative and quantitative improvements in educational institutions, addressing structural, planning, and pedagogical aspects.*

Resumo. *A Taxa de Distorção Idade-Série (TDI) mede a quantidade de alunos fora do ano curricular esperado para sua idade. Este artigo utiliza técnicas de aprendizagem de máquina para prever os valores do TDI entre 2018 e 2023, abrangendo períodos pré, durante e pós-pandemia. Utilizando dados do Censo da Educação Básica, detalhamos a seleção do algoritmo e o uso do SHAP para interpretar suas métricas. Nosso objetivo é identificar as características mais importantes apontadas pelo modelo preditivo e fomentar a discussão sobre melhorias qualitativas e quantitativas nas instituições de ensino, abordando aspectos estruturais, de planejamento e pedagógicos.*

1. Introdução

Segundo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) apenas no ano de 2023 o sistema básico educacional brasileiro registrou 47,3 milhões de matrículas dividido em suas 178,5 mil escolas¹. Tendo esses números, o sistema educacional brasileiro abarca a mesma quantidade de alunos quanto os habitantes da Espanha, e com esses números vários desafios já são vivenciados há décadas e que ainda permeiam o dia a dia escolar como citado em [Schwartzman and Brock 2005, Palomino et al. 2022].

Quando um sistema numeroso é atingido por um dos maiores desafios enfrentados pela humanidade nesse século, a pandemia de COVID-19 [Brito et al. 2020], novos desafios foram criados [Peres 2020], como também a necessidade de novas

¹<https://download.inep.gov.br/publicacoes>

políticas educacionais [Galzerano 2021] e até metodologias foram desenvolvidas fosse possível manter o ensino durante tempos de educação remota e híbrida [Silva et al. 2022, Corrêa and Brandemberg 2021].

Após superar esse momento crítico é essencial avaliar fazer estudos para analisar os impactos da pandemia considerando indicadores registrados antes, durante e depois desse período. Isso nos permitirá acompanhar as mudanças e gerar aprendizados em um contexto histórico por meio de técnicas de Learning Analytics para extração e processamento desses dados, processo esse utilizado em trabalhos como o de [do Nascimento et al. 2018]. Diversos indicadores são disponibilizados pelo INEP, como taxas de rendimento escolar [Justino 2022], complexidade de gestão escolar [de Andrade et al. 2020], taxa de alfabetização [Bernardi and Luchese 2020], taxa de distorção idade-série (TDI) [NOGUEIRA and Silva 2022] e índice de desenvolvimento da educação básica (IDEB) [Rodrigues et al. 2016].

Entre os indicadores fornecidos pelo INEP, este trabalho focará na Taxa de Distorção Idade-Série (TDI). A TDI é calculada pelo Ministério da Educação (MEC) e mede a proporção de alunos que estão matriculados em séries com pelo menos dois anos de diferença em relação à idade adequada. O cálculo da TDI envolve a comparação entre a idade dos alunos e a série em que estão matriculados. O MEC utiliza dados do Censo Escolar, realizado anualmente, para obter essas informações. Com base nesses dados, a TDI é calculada usando a seguinte fórmula:

$$TDI = \frac{\text{Número de alunos com dois anos ou mais de atraso escolar}}{\text{Total de alunos matriculados na série}} \times 100$$

O TDI é um indicador bem consolidado na temática de mapeamento educacional por meio de dados abertos. Esse indicador também pode ser usado para avaliar a defasagem entre a educação urbana e rural. Em [NOGUEIRA and Silva 2022], o TDI foi utilizado como ferramenta de acompanhamento de escolas situadas em zonas rurais, expondo fatores intra e extraescolares para seus gestores. Foi utilizado também em outras áreas que não diretamente para resultados educacionais, em [Ferreira and Teixeira 2018] por exemplo, foi possível cruzar os indicadores de violência em localidades com alto TDI, buscando indicar uma relação de causa e efeito. Em [Barros et al. 2023] o TDI foi utilizado junto ao censo da educação básica e por meio de técnicas de Learning Analytics de modo a avaliar como esse indicador pode sugerir melhorias com base em valores médios de TDI.

Em busca de utilizar as técnicas de Learning Analytics para gerar essas informações com base no contexto histórico do TDI entre os anos 2018, 2019 como anos antes da pandemia, 2020 e 2021 como anos pandêmicos e 2022 e 2023 os posteriores a pandemia. E buscando guiar esse estudo foram desenvolvidas duas perguntas de pesquisa para serem respondidas com esse trabalho:

Pergunta de Pesquisa 1 (PP1): *As características mais importantes na previsão do TDI permanecem as mesmas ao longo dos anos?*

Pergunta de Pesquisa 2 (PP2): *Quais as características que estão mais relacionadas a um menor TDI?*

Diante disso, este trabalho visa demonstrar todo o processo de tomada de decisão incluindo à escolha do algoritmo, sua execução e, principalmente, a elucidação de seus

resultados visando embasar a tomada de decisão com base nos registros dos contextos temporais apresentados.

2. Método

Para o desenvolvimento deste experimento, seguimos várias etapas essenciais. Primeiramente, selecionamos as bases de dados e as preparamos para a aplicação dos algoritmos de aprendizado de máquina. Em seguida, escolhemos o algoritmo de regressão mais adequado para o problema. Aplicamos esses algoritmos aos dados de cada ano, permitindo uma análise detalhada ao longo do tempo. Finalmente, avaliamos os resultados utilizando a técnica *SHAP*, que nos ajudou a entender a contribuição de cada variável no modelo de regressão. O objetivo dessa etapa é gerar os dados necessários para responder às perguntas levantadas na 1. Nessa seção apresentamos como os dados foram adquiridos e processados, o processo de escolha do algoritmo e a extração do *feature importance*.

2.1. Aquisição e processamento dos dados

Os dados utilizados para este projeto foram todos adquiridos do INEP, respeitando a faixa temporal definida por este estudo. As bases necessárias foram o TDI ao nível de entidade escolar² e o censo escolar da educação básica³.

Em [Barros et al. 2022], é demonstrado como realizar o processamento do Censo da Educação Básica utilizando o dicionário de dados para avaliar a ausência de colunas ao longo dos anos. No entanto, em nosso experimento, a junção dos dados não foi necessária, uma vez que os resultados foram avaliados individualmente para cada ano. Isso permitiu simplificar o método de processamento dos dados do censo em comparação com o proposto no trabalho de base.

Assim, foram removidas as colunas do tipo *string* ou *char* que identificassem a escola, como nome da entidade e código definido pelo INEP, para evitar o enviesamento dos dados. Em seguida, também foram excluídas as colunas que apresentavam 80% de registros nulos. Colunas que apresentassem um valor não registrado como:8888 para variáveis quantitativas e 9 para qualitativas (que apresentava resultados entre 0, 1, 2) foram zeradas. Essa operação foi guiada pela informação disponível no dicionário de dados de modo a garantir que não fossem prejudicadas outras colunas. Por fim foram removidas as colunas que tratavam sobre a quantidade de matrículas nas instituições buscando não induzir uma separação via quantidade desses elementos.

Para a extração dos resultados do TDI, selecionamos os dados a nível escolar e, após esse filtro, aplicamos um filtro adicional para instituições públicas, pois nosso objetivo é avaliar a rede pública de forma geral, englobando os níveis municipal, estadual e federal. Por fim, selecionamos os valores do TDI do ensino fundamental anos finais (*TDI_FUN_AF*).

Os dados foram armazenados utilizando o banco de dados colunar DuckDB [Mühleisen and Raasveldt 2024], com o objetivo de realizar os testes de treinamento dos algoritmos de aprendizado de máquina de forma mais eficaz e confiável, simplificando o

²<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/indicadores-educacionais/taxas-de-distorcão-idade-serie>

³<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-escolar>

processo em comparação com outros bancos de dados que requerem uma infraestrutura mais complexa de instalação ou a utilização de contêineres, como MySQL, ClickHouse ou MariaDB.

2.2. Algoritmo de aprendizagem de máquina

Nesta etapa, foram aplicados algoritmos de aprendizagem de máquina para prever o valor que uma instituição de ensino pode alcançar no TDI. É importante destacar que, ao contrário de outros indicadores como taxa de aprovação e taxa de alfabetização, onde valores mais altos indicam melhor desempenho, no caso do TDI, valores mais baixos são melhores, afinal indicam que naquela instituição de ensino existe uma menor quantidade de matrículas com idade em desacordo ao nível escolar esperado. Assim, quando uma instituição apresenta um TDI menor que a do ano anterior, é considerado uma melhoria na situação escolar daquela instituição. Por exemplo, se uma escola apresentou um *TDI_FUN_AF* de 50% em 2020 e, após três anos, esse indicador caiu para 35% em 2023, isso indica uma melhora, mostrando que existiu uma redução de 15% nas matrículas fora da faixa escolar esperada para sua idade.

Por esse trabalho buscamos a explicabilidade dos resultados dos algoritmos e não apenas o melhor desempenho, optamos por selecionar algoritmos de árvore de decisão, pois em sua natureza são algoritmos de melhor interpretação e explicabilidade [Mahbooba et al. 2021]. Em nosso caso foi utilizado um conjunto de algoritmos da família das árvores de regressão:

- **CatBoost:** É um algoritmo de gradient boosting desenvolvido pela Yandex que se destaca no manuseio de variáveis categóricas sem a necessidade de pré-processamento extenso [Prokhorenkova et al. 2018].
- **LightGBM:** É uma variante do gradient boosting, desenvolvido pela Microsoft, que foca em alta eficiência e escalabilidade [Ke et al. 2017].
- **XGBoost:** É uma implementação otimizada de gradient boosting que tem sido amplamente utilizada em competições de machine learning devido ao seu desempenho superior [Chen and Guestrin 2016].
- **Random Forest:** É um *ensemble* (uma técnica de aprendizado de máquina que combina múltiplos modelos-base para melhorar a precisão e a robustez das previsões em comparação com um único modelo individual) de árvores de decisão que se destaca pela sua simplicidade e robustez. Ele constrói múltiplas árvores de decisão durante o treinamento e produz a média das previsões de todas as árvores para regressão [Breiman 2001].
- **Extra Trees:** Extra Trees (Extremely Randomized Trees) é uma variante do Random Forest que randomiza ainda mais a construção das árvores. Em vez de procurar o ponto de corte ótimo, ele seleciona pontos de corte aleatoriamente, o que pode reduzir a variância do modelo e aumentar a diversificação entre as árvores [Geurts et al. 2006].
- **Decision Tree:** O Decision Tree Regressor é um modelo de árvore de decisão simples que divide o espaço dos dados em regiões homogêneas, com base em um conjunto de regras de decisão [Quinlan 1986].

Após a definição dos algoritmos foram definidas quais as métricas seriam utilizadas pelo *framework pycaret*⁴, responsável por realizar os testes de cada algoritmo

⁴<https://pycaret.gitbook.io/docs>

ponderando sobre as seguintes métricas:

- **MAE (Mean Absolute Error):** O MAE calcula a média das diferenças absolutas entre valores previstos e observados.
- **MSE (Mean Squared Error):** O MSE calcula a média dos quadrados dos erros entre valores previstos e observados.
- **RMSE (Root Mean Squared Error):** O RMSE é a raiz quadrada do MSE e representa o desvio padrão dos resíduos.
- **R² (Coefficient of Determination):** O R² indica a proporção da variância dos dados dependentes explicada pelo modelo.
- **RMSLE (Root Mean Squared Logarithmic Error):** O RMSLE calcula o erro entre os logaritmos dos valores previstos e observados.
- **MAPE (Mean Absolute Percentage Error):** O MAPE calcula a média das diferenças percentuais absolutas entre valores previstos e observados.

A busca pelo melhor algoritmo foi feita com base nos dados de 2022, que foi a base anual que apresentou uma menor quantidade de mudança de colunas entre toda a faixa temporal avaliada, vide inclusão e exclusão, com base em nossa faixa temporal escolhida.

2.3. Análise de características

Como última etapa do processo, foi utilizado o *SHAP values*, mais especificamente sua versão para algoritmos de regressão. Os SHAP values, ou valores SHAP (SHapley Additive exPlanations), são uma técnica para a interpretação de modelos de aprendizado de máquina, baseada no conceito de valores de Shapley da teoria dos jogos cooperativos [Leite 2022]. Técnica essa já consolidada para a avaliação de modelos de aprendizagem de máquina como vemos em [Wang et al. 2022], [Lubo-Robles et al. 2020] e [Hamilton and Papadopoulos 2023].

A utilização dessa técnica foi escolhida para esse trabalho justamente trazer uma interpretação dos valores SHAP é direta e intuitiva. Cada valor SHAP atribuído a uma variável representa a mudança esperada na predição quando essa variável é incluída no modelo. Onde:

- **Valores Positivos e Negativos:** Um valor SHAP positivo indica que a variável aumenta a predição do modelo, enquanto um valor negativo sugere que a variável reduz a predição.
- **Magnitude:** A magnitude do valor SHAP reflete a importância da variável para a predição específica. Valores absolutos maiores indicam maior influência.

Fornecendo para a nossa interpretação, diferentemente da técnica de *feature_importance* dos modelos, é sua explicabilidade em detrimento de uma quantidade maior ou menor de cada variável, para variáveis quantitativas como também para valores 0 e 1 na definição de se a entidade escolar possui ou não possui aquela característica informada, fazendo com que seja possível essa visualização já direta nos gráficos.

3. Resultados

Uma vez que todos os dados tenham sido calculados e os gráficos gerados, podemos responder às perguntas elencadas na seção 1. O dicionário de dados das *características* apontadas podem ser encontrados em dicionário de características.

Modelo	MAE	MSE	RMSE	R2	RMSLE	MAPE	Tempo de Execução (Sec)
CatBoost	5.83	76.98	8.77	0.832	0.441	0.342	13.05
LGBM	6.001	81.33	9.01	0.822	0.456	0.355	258.20
XGBoost	6.260	89.25	9.44	0.805	0.470	0.384	1.740
Random Forest	6.600	93.01	9.63	0.797	0.475	0.416	3.650
Extra Trees	6.989	111.07	10.53	0.757	0.485	0.418	3.630
Decision Tree	8.695	168.50	12.97	0.632	0.609	0.496	2.780

Tabela 1. Resultados das Métricas de Desempenho dos Modelos Testados

3.1. Seleção do regressor

Como pode ser visto na Tabela 1, o algoritmo *CatBoost* apresentou o melhor resultado. A escolha desse algoritmo é corroborada por artigos que destacam sua eficácia em comparação a outros regressores para dados categóricos, como mostrado em [Hancock and Khoshgoftaar 2020] e [Prokhorenkova et al. 2018]. Tendo um MAE de 5,83 sugere que, em média, as previsões do modelo estão a 5,83 unidades de distância dos valores reais. O MSE de 76,98 é um bom indicativo de que o modelo está conseguindo minimizar erros grandes. Um RMSLE de 0,441 é bastante competitivo e indica que o modelo lida bem com grandes variações. Um valor de R^2 de 0,832 significa que o modelo explica 83,2% da variação nos dados, indicando um bom ajuste sobre os dados.

Uma vez definido o *CatBoost* como algoritmo a ser utilizado, foi aplicado o protocolo de separação dos dados em 80%, da base de dados criada, para treino e o restante para teste. Com isso para cada ano um modelo foi gerado e treinado com os dados do ano correspondente, gerando assim seis modelos já treinados.

Feature	Importância por Ano					
	2018	2019	2020	2021	2022	2023
IN_DESKTOP_ALUNO		8				
IN_ENERGIA_REDE_PUBLICA	3	6	6	4	3	2
IN_EQUIP_MULTIMIDIA						9
IN_FUNDAMENTAL_CICLOS	1	1	1	3		AUSENTE
IN_INTERNET	7					
IN_INTERNET_APRENDIZAGEM		3	3	2	2	5
IN_LABORATORIO_INFORMATICA	8					
IN_NOTURNO	6					
IN_ORGAO_ASS_PAIS_MESTRES					7	6
IN_ORGAO_GREMIO_ESTUDANTIL		2	2	1	1	1
IN_QUADRA_ESPORTES_COBERTA	2	4	4	9	9	7
QT_DESKTOP_ALUNO		7	5	5	6	4
QT_DOC_FUND_AF			10		10	
QT_DOC_FUND_AI	5					
QT_EQUIP_MULTIMIDIA	4	10	9	6	5	3
QT_EQUIP_TV	9					
QT_TRANSP_PUBLICO	AUSENTE	AUSENTE	AUSENTE	AUSENTE	AUSENTE	8
QT_TUR_FUND_AF	10			10		
QT_TUR_FUND_INT	AUSENTE	AUSENTE	AUSENTE	AUSENTE	AUSENTE	10
TP_ATIVIDADE_COMPLEMENTAR		5	7	7	8	
TP_PROPOSTA_PEDAGOGICA		9	8	8	4	

Tabela 2. Importância das Características por Ano

3.2. Pergunta de pesquisa 1

Para respondermos essa primeira pergunta, foram levantadas as dez principais características apontadas pelo método SHAP. E que podem ser conferidas na Tabela 2.

Ao analisarmos o contexto dos anos pré-pandemia e pós-pandemia, observamos que a característica `IN_FUNDAMENTAL_CICLOS`, que define o funcionamento da escola em ciclos do ensino fundamental, foi a mais importante para a previsão do indicador TDI antes da pandemia. No entanto, essa variável não aparece mais nos anos pós-pandêmicos. Em 2021, por exemplo, a característica mais relevante passou a ser `IN_ORGAO_GREMIO_ESTUDANTIL`, que indica a existência ou não de um grêmio estudantil na instituição educativa. Essa variável já mostrava um alto grau de impacto desde 2019.

Ao analisar as três primeiras características, vemos a presença constante de variáveis estruturais, como `IN_QUADRA_ESPORTES_COBERTA` e `IN_ENERGIA_REDE_PUBLICA`, além de uma principal característica pedagógica: `IN_INTERNET_APRENDIZAGEM`, que indica se a escola utiliza a internet como ferramenta educacional.

Expandindo a análise para as demais características, observamos a importância da quantidade de equipamentos eletrônicos, número de turmas, apresentação da proposta pedagógica, e a existência de organizações de alunos e de pais e mestres.

O ano de 2023 trouxe mais mudanças, principalmente devido à introdução de novas variáveis, como transporte público e a quantidade de equipamentos por aluno, incluindo desktops, portáteis e multimídia em geral. Em contrapartida, onde em 2018 ainda existiam características que eram relacionados ao funcionamento da instituição escolar, `IN_NOTURNO`, onde esse tipo de característica não foi vista mais ao decorrer dos anos.

Por fim, para formulamos uma resposta a nossa pergunta, a evolução das características entre os anos sofreu poucas alterações quando falamos do objetivo geral, onde a organização escolar e participação dos alunos (`IN_ORGAO_GREMIO_ESTUDANTIL`, `IN_ORGAO_ASS_PAIS_MESTRES`), técnicas pedagógicas (`IN_INTERNET_APRENDIZAGEM`, `TP_PROPOSTA_PEDAGOGICA`, `TP_ATIVIDADE_COMPLEMENTAR`), estrutural (`IN_QUADRA_ESPORTES_COBERTA`, `IN_ENERGIA_REDE_PUBLICA`, `IN_FUNDAMENTAL_CICLOS`) e por fim, na quantidade de equipamentos da escola (`QT_DESKTOP_ALUNO`, `QT_EQUIP_MULTIMIDIA`, `QT_EQUIP_TV`).

3.3. Pergunta de Pesquisa 2

Para responder a essa pergunta, analisaremos um dos tipos dos gráficos SHAP, o *Beeswarm*. É uma importante ferramenta do conjunto SHAP que auxilia na interpretação de modelos menos transparentes como Redes Neurais, ou mais transparentes como árvore de decisão, que foi o escolhido para esse trabalho. O gráfico permite identificar quais características têm o maior impacto nas previsões do modelo e se essas características estão aumentando ou diminuindo a previsão. Esse tipo de rastreamento é o que precisamos para avaliar a interação das características mais importantes com a previsão do TDI de cada unidade escolar entre cada dupla de anos, caracterizados como: pré-pandêmico, pandêmico e pós-pandêmico.

A estrutura do gráfico SHAP *Beeswarm* é composta por dois eixos principais e um conjunto de pontos que fornecem uma visualização detalhada das contribuições das características para as previsões do modelo. O eixo X representa os valores SHAP, as quais são as contribuições individuais das características para a previsão do modelo; valores

positivos indicam um aumento na previsão, que para nosso caso é um pior desempenho, enquanto valores negativos indicam uma diminuição, que em nosso caso indicam um melhor desempenho da instituição. O eixo Y lista as dez principais características do modelo, permitindo uma análise comparativa de sua influência. Cada ponto no gráfico representa uma instância de dados, que em nosso caso é uma instituição de ensino, e a cor dos pontos indica o valor da característica correspondente, com tons vermelhos representando valores altos e tons azuis representando valores baixos.

Para respondermos a essa pergunta analisaremos a presença das características de duas formas diferentes, pois em nosso conjunto de análise, para variáveis com prefixo **IN**, que indicam a existência utilizando o valor 1, e a ausência com o valor 0. Já para variáveis com prefixo **QT** demonstraram variáveis quantitativas. Como a observação dessas variáveis são divergentes, a seguir iremos abordá-las como qualitativas (IN) e quantitativas (QT).

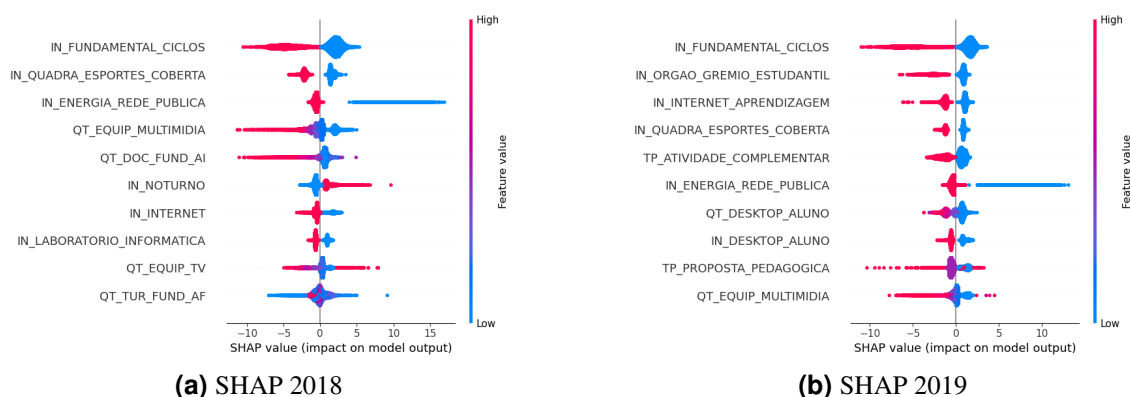


Figura 1. Análise SHAP 2018 e 2019: Pré pandêmico

Iniciando pelas Figuras 1 a e b, podemos analisar as variáveis qualitativas, destacando a manutenção do **IN_FUNDAMENTAL_CICLOS** como 1, indicando que instituições que mantêm o ensino fundamental em ciclos preveem um menor TDI. A implementação de um sistema escolar de ensino fundamental em ciclos representa uma abordagem progressista que visa promover um aprendizado contínuo e inclusivo, ao evitar a repetência direta, o sistema de ciclos tem em vista reduzir a evasão escolar e os impactos negativos associados à reprovação. Esse tipo de estratégia tem um alto teor de impacto junto ao TDI visto que reduzirá a quantidade de alunos, que são reprovados e caso já sejam o segundo ano de atraso escolar.

Outras variáveis qualitativas demonstram que: possuir quadras cobertas, laboratórios de informática, acesso à internet, uso da internet como ferramenta pedagógica, atividades complementares e proposta pedagógica atualizada anualmente são fatores que indicam uma previsão de menor TDI, impactando positivamente para o resultado do indicador na instituição.

Em contraste, o **IN_NOTURNO**, que indica funcionamento noturno, mostrou um impacto inverso. A existência desse turno se correlaciona com previsões de maior TDI, levantando discussões sobre os desafios do ensino noturno, conforme abordado em [dos Santos and Pouchain 2011]. Isso sugere a necessidade de revisar estratégias pe-

dagógicas diferenciadas para os turnos diurnos e noturnos.

Na análise das variáveis quantitativas, especialmente de equipamentos físicos, observamos que uma maior quantidade de equipamentos multimídia (QT_EQUIP_MULTIMIDIA) e computadores (QT_DESKTOP_ALUNO) tende a reduzir o TDI. Em 2018, a quantidade de TVs (QT_EQUIP_TV) não teve grande importância, pois esses equipamentos se tornaram obsoletos com a adoção de projetores, os quais são mais versáteis e adequados a diferentes espaços escolares, como auditórios e laboratórios.

Por fim, ao analisar o fator humano, em 2018, a quantidade de docentes no ensino fundamental anos iniciais (QT_DOC_FUND_AI) indicou que uma maior quantidade de professores correlaciona-se com um menor TDI. No mesmo ano, a quantidade de turmas do fundamental anos finais (QT_TUR_FUND_AF) também foi importante, embora com um impacto variável. Menos turmas podem afetar o TDI tanto positiva quanto negativamente, mas, em geral, uma menor oferta de vagas não é necessariamente positiva, pois o TDI é analisado com base na totalidade das matrículas e sua conformidade com o fluxo regular esperado.

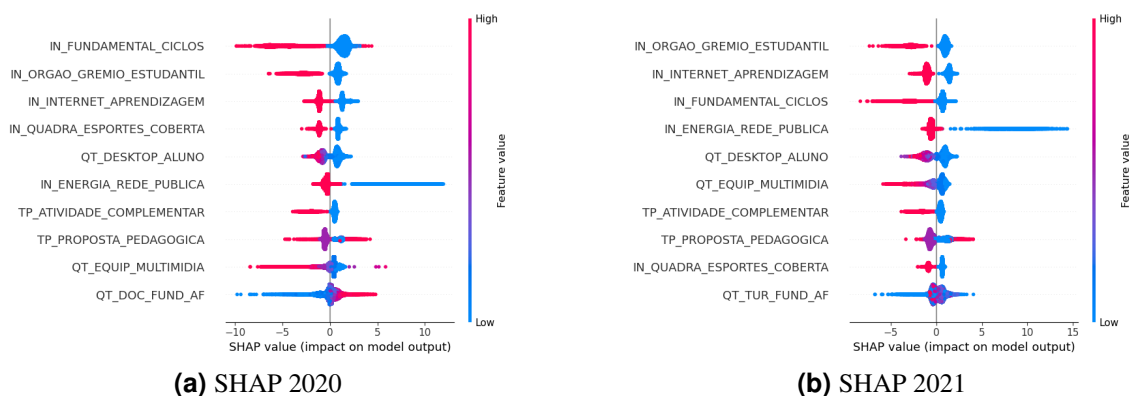


Figura 2. Análise SHAP 2020 e 2021: Durante a pandemia

Analisando agora o que foi disposto na figura 2 a e b, temos a avaliação das características mais importantes para anos mais críticos no que tange a educação brasileira [Ludovico et al. 2020].

Semelhante ao modo apresentado na seção anterior, ao analisarmos primeiramente as características qualitativas apresentadas. Temos a manutenção das variáveis visto anteriormente, onde a presença dos fatores levantados favorecem uma previsão de menores valores do TDI. Destacando o aparecimento da variável que mensura a existência de atividades complementares oferecidas pela instituição, sejam de forma exclusiva ou parcial, onde avaliando ambos os anos mostra que esse fator impacta em sua totalidade para a previsão de menores valores de TDI.

Já quando avaliamos as quantitativas temos a manutenção da quantidade de desktops por aluno, onde prevalece que uma maior quantidade favorece uma previsão menor do indicador. O mesmo é observado para equipamento de multimídia, característica essa que em 2021 se torna mais importante, podendo assim estar relacionada com a vota parcial dos alunos para a sala de aula.

Indo para a análise do fator humano dentre os quantitativos, temos o aparecimento

da quantidade de docentes do fundamental anos finais(QT_DOC_FUND_AF) em 2020, onde ao interpretarmos vemos que uma quantidade maior de professores impactaria num maior TDI, trazendo uma análise contrária a lógica, de que mais professores trariam um melhor acompanhamento dos alunos e consequentemente um melhor resultado, mas nesse ano em específico esse não foi o resultado apresentado. Onde essa, possa ser uma peculiaridade vista nesse ano, pois ao visitarmos novamente a figura 2 b, vemos que essa variável em questão já não mais figura nosso gráfico. Com isso podemos interpretar também que a maior quantidade de professores já se deem em virtude de uma maior quantidade de alunos e um esforço para a melhoria situacional da escola e buscando assim reduzir o índice, visto que não é mantido para os anos sucessores.

Por fim, vemos em 2021, que a quantidade de turmas do fundamental anos finais novamente contrasta entre a quantidade menor de turmas para um TDI maior ou menor, onde no gráfico define uma pequena diferença ao favor de uma quantidade menor de turma trazer um resultado ligeiramente menor.

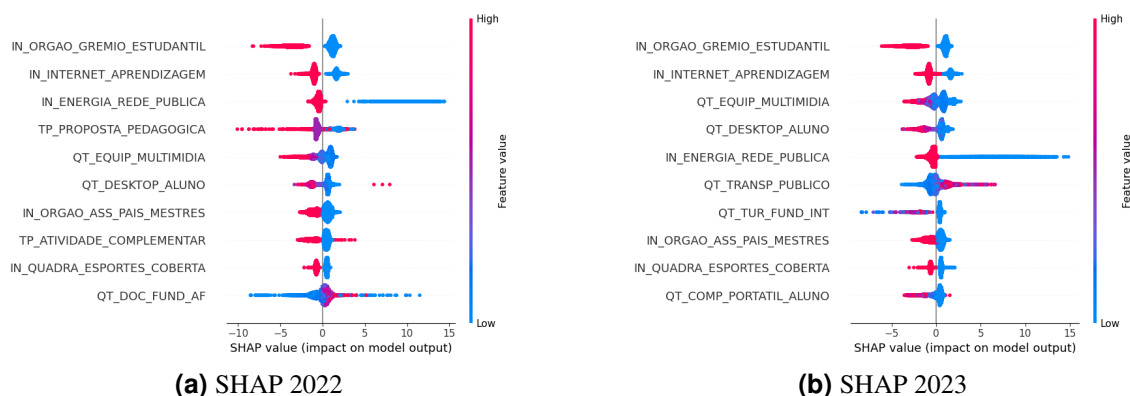


Figura 3. Análise SHAP 2022 e 2023: Pós pandemia

Seguindo a forma da análise anteriores, analisemos a última faixa temporal definida, e a mais recente durante o desenvolvimento desse trabalho. Os gráficos estão presentes na figura 3 a e b.

Ao começarmos pelas características qualitativas, a existência do grêmio estudantil se perpetuou nesses anos, sempre para impactar em um TDI mais baixo. Como também a utilização da internet como ferramenta educacional. O impactado dessas variáveis quantitativas seguem positivo, com a adição de uma que trate sobre organização de pais e mestres (IN_ORGAO_ASS_PAIS_MESTRES) onde também atua positivamente na previsão de um menor valor para o indicador alvo.

O desaparecimento da variável IN_FUNDAMENTAL_CICLOS da lista de variáveis deve-se que para o ano de 2022 ela já não teve 80% dos dados preenchidos, e em 2023 a variável foi definitivamente removida dos campos disponibilizados no censo da educação básica.

Passando para as variáveis quantitativas relacionado a objetos pertencentes a escola, temos a manutenção do comportamento, que onde tivermos uma maior quantidade de equipamentos multimídia, desktops, e para 2023, computadores portáteis trazem uma previsão menor do modelo sobre o TDI.

E ao analisarmos agora quanto a fatores quantitativos e não de patrimonio da instituição, temos a quantidade de docentes do ensino fundamental anos finais ainda trazendo um impacto, mesmo que pequeno, para valores maiores do TDI, comportamento visto anteriormente, favorecendo a interpretação de que a simples inserção de profissionais de ensino seja a resposta para a regressão do indicador. Para o ano de 2023 tivemos a ausência dessa feature entre as dez mais importantes, dando lugar a quantidade de turmas fundamentais de ensino integral (QT_TUR_FUND_INT), onde temos uma certa sobreposição dos valores, mas indicando que valores mais alto desse aspecto tendem a ter um valor de diminuição durante a previsão do TDI. Nesse mesmo ano temos também a adição da variável que trata da quantidade de transporte público utilizado pelos alunos da instituição (QT_TRANSP_PUBLICO), ele oferece uma relação oposta ao que pode ser indicado em [Evangelista et al. 2017] que ressalta os desafios enfrentados pelos alunos entre a ida e volta do âmbito escolar. Essa feature, criada nesse ano de 2023, carece de um melhor acompanhamento em seus próximos resultados, pois hoje traz seu resultado que quanto matriculas nesse tipo de serviço indicaria um resultado maior na previsão de seu TDI.

4. Considerações Finais e Trabalhos Futuros

Este trabalho se propôs a análise do indicador TDI das escolas públicas brasileiras durante os dois anos que precederam a pandemia de COVID-19, os dois anos de ápice da pandemia e os dois anos subsequentes. Após testar algoritmos e adequar a base de dados, utilizamos a técnica de interpretação de modelos SHAP para identificar as dez principais características de cada ano, evidenciando suas mudanças e impactos, positivos ou negativos, na previsão do indicador.

A discussão dessas características ao longo dos anos visa fornecer um ponto de vista baseado em dados para gestores escolares, gestores de redes educacionais e a comunidade em geral. O objetivo é promover melhorias na educação, considerando características como a presença de quadras cobertas. Destacamos a importância da participação da comunidade escolar, por meio de associações de pais e mestres e grêmios estudantis. Vimos também a estrutura disponível na escola, e também seus equipamentos se mostraram de fundamental importância da redução do TDI, por fim a utilização da internet como ferramenta educacional e manter sua proposta pedagógica atualizada complementa um perfil pedagógico mais proveniente a menores resultados do indicador.

Aspiramos continuar este trabalho, analisando e mapeando como essas características se comportam em outros níveis educacionais, comparando escolas em zonas rurais e urbanas. Também aspiramos aplicar o método para categorizar variáveis do censo da educação básica em pilares administrativos, estruturais e pedagógicos. Por fim, visamos validar os resultados com gestores para captar suas opiniões e aprimorar a aplicação prática das descobertas.

Referências

- [Barros et al. 2022] Barros, A. N., Alencar, A., Nascimento, A., de Albuquerque, A. F., and Mello, R. F. (2022). Elaboração do conjunto de dados agregados do censo da educação básica. In *Anais do IV Dataset Showcase Workshop*, pages 35–45. SBC.
- [Barros et al. 2023] Barros, A. N., Xavier, E. L. S., Alves, G., and Mello, R. F. (2023). Aplicação de learning analytics para identificação de tomada de decisão sobre a

- distorção idade-série no brasil. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 21–31. SBC.
- [Bernardi and Luchese 2020] Bernardi, M. C. and Luchese, T. Â. (2020). A taxa de alfabetização de antônio prado, rio grande do sul (1895-1920). *Revista Educação em Questão*, 58(56).
- [Breiman 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- [Brito et al. 2020] Brito, S. B. P., Braga, I. O., Cunha, C. C., Palácio, M. A. V., and Take-nami, I. (2020). Pandemia da covid-19: o maior desafio do século xxi. *Vigilância Sanitária em Debate: Sociedade, Ciência & Tecnologia*, 8(2):54–63.
- [Chen and Guestrin 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.
- [Corrêa and Brandemberg 2021] Corrêa, J. N. P. and Brandemberg, J. C. (2021). Tecnologias digitais da informação e comunicação no ensino de matemática em tempos de pandemia: desafios e possibilidades. *Boletim Cearense de Educação e História da Matemática*, 8(22):34–54.
- [de Andrade et al. 2020] de Andrade, M. C. B., Silva, L. F., Fecury, A. A., de Oliveira, E., Dendasck, C. V., de Araújo, M. H. M., da Souza, K. O., da Silva, I. R., de Medeiros Moreira, E. C., Pascoal, R. M., et al. (2020). Indicadores de complexidade de gestão em escolas públicas e privadas de duas cidades do estado do amapá entre 2014 e 2018. *Research, Society and Development*, 9(9):e856998112–e856998112.
- [do Nascimento et al. 2018] do Nascimento, R. L. S., da Cruz Junior, G. G., and de Araújo Fagundes, R. A. (2018). Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. *RENOTE*, 16(1).
- [dos Santos and Pouchain 2011] dos Santos, M. J. C. and Pouchain, J. F. (2011). Evasão escolar no ensino médio noturno: Um estudo de caso na escola de ensino fundamental e médio prof. jáder moreira de carvalho. *Conhecer: debate entre o público e o privado*, 1(01):295–329.
- [Evangelista et al. 2017] Evangelista, J. C. S., Santos, C. R., Silva, L. R., and Santos, A. R. d. (2017). A política do transporte escolar na educação do campo: impactos e desafios na realidade escolar. *Seminário Nacional e Seminário Internacional Políticas Públicas, Gestão e Práxis Educacional*, 6(6).
- [Ferreira and Teixeira 2018] Ferreira, V. B. and Teixeira, E. C. (2018). O impacto da distorção idade-série sobre a criminalidade nos municípios de minas gerais. *Revista Brasileira de Segurança Pública*, 12(2):269–291.
- [Galzerano 2021] Galzerano, L. S. (2021). Políticas educacionais em tempos de pandemia. *Argumentum*, 13(1):123–138.
- [Geurts et al. 2006] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.
- [Hamilton and Papadopoulos 2023] Hamilton, R. I. and Papadopoulos, P. N. (2023). Using shap values and machine learning to understand trends in the transient stability limit. *IEEE Transactions on Power Systems*.

- [Hancock and Khoshgoftaar 2020] Hancock, J. T. and Khoshgoftaar, T. M. (2020). Catboost for big data: an interdisciplinary review. *Journal of big data*, 7(1):94.
- [Justino 2022] Justino, M. R. (2022). A relação do esforço docente e da infraestrutura escolar nas taxas de rendimento escolar: uma análise para a cidade do natal no ano de 2019. B.S. thesis, Universidade Federal do Rio Grande do Norte.
- [Ke et al. 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154.
- [Leite 2022] Leite, G. B. (2022). Jogos cooperativos: uma introdução ao valor de shapley. *Informe Econômico (UFPI)*, 44(1).
- [Lubo-Robles et al. 2020] Lubo-Robles, D., Devegowda, D., Jayaram, V., Bedle, H., Marfurt, K. J., and Pranter, M. J. (2020). Machine learning model interpretability using shap values: Application to a seismic facies classification task. In *SEG international exposition and annual meeting*, page D021S008R006. SEG.
- [Ludovico et al. 2020] Ludovico, F. M., Molon, J., Barcellos, P. D. S. C. C., Franco, S. R. K., et al. (2020). Covid-19: desafios dos docentes na linha de frente da educação. *Interfaces Científicas-Educação*, 10(1):58–74.
- [Mahbooba et al. 2021] Mahbooba, B., Timilsina, M., Sahal, R., and Serrano, M. (2021). Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021:1–11.
- [Mühleisen and Raasveldt 2024] Mühleisen, H. and Raasveldt, M. (2024). *duckdb: DBI Package for the DuckDB Database Management System*. R package version 1.0.0.9000, <https://github.com/duckdb/duckdb-r>.
- [NOGUEIRA and Silva 2022] NOGUEIRA, M. D. O. E. and Silva, L. C. (2022). Escolarização em áreas rurais: a distorção idade-série na ótica dos gestores. *Estudos em Avaliação Educacional*, 33.
- [Palomino et al. 2022] Palomino, P., Falcao, T. P., Medeiros, R., Uehara, M., Bittencourt, I., and Mello, R. F. (2022). Plataformas de dados educacionais: Análise com foco no plano nacional de educação. In *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 60–68. SBC.
- [Peres 2020] Peres, M. R. (2020). Novos desafios da gestão escolar e de sala de aula em tempos de pandemia. *Revista de Administração Educacional*, 11(1):20–31.
- [Prokhorenkova et al. 2018] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.
- [Quinlan 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.
- [Rodrigues et al. 2016] Rodrigues, E. C. et al. (2016). Indicadores educacionais e contexto escolar: uma análise das metas do ideb. *Estudos em Avaliação Educacional*, 27(66):662–688.
- [Schwartzman and Brock 2005] Schwartzman, S. and Brock, C. (2005). Os desafios da educação no brasil. *Rio de Janeiro: Nova Fronteira*, 1320.

- [Silva et al. 2022] Silva, D. S. M. d., Sé, E. V. G., Lima, V. V., Borim, F. S. A., Oliveira, M. S. d., and Padilha, R. d. Q. (2022). Metodologias ativas e tecnologias digitais na educação médica: novos desafios em tempos de pandemia. *Revista Brasileira de Educação Médica*, 46:e058.
- [Wang et al. 2022] Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., and Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on shap values for tree-based machine learning methods. *Journal of Environmental Management*, 301:113941.