

Integração e Análise de Dados do INEP e IBGE: um Pipeline para Estudo da Evasão no Ensino Superior

Guilherme Fortes Lordelo¹, Grinaldo Lopes de Oliveira²

¹ Análise e Desenvolvimento de Sistemas – Instituto Federal da Bahia (IFBA) – Salvador
– BA – Brasil

² Análise e Desenvolvimento de Sistemas – Instituto Federal da Bahia (IFBA) – Salvador
– BA – Brasil

guilherme211200@gmail.com, grinaldo@gmail.com

Resumo

A evasão no ensino superior constitui um fenômeno complexo, influenciado por fatores institucionais, socioeconômicos e territoriais que demanda análises munidas de dados encontrados em diferentes bases. No contexto brasileiro, os microdados do Censo da Educação Superior, produzidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), oferecem um amplo detalhamento anual capaz de acompanhar a evasão ao longo dos anos no país, porém, em sua disponibilização pública atual, apresentam limitações sob a ótica territorial. Diante desse cenário, este trabalho tem como objetivo conceber e implementar um pipeline de dados voltado a harmonização de dados INEP com informações territoriais e socioeconômicas produzidas pelo Instituto Brasileiro de Geografia e Estatística (IBGE), possibilitando a visualização da evasão sobre o território brasileiro. A metodologia adotada inclui as etapas de extração, transformação e carga dos dados, bem como a definição de métricas de evasão e agregações em nível municipal, estadual e nacional. Como resultado, é construído um conjunto de dados analítico integrado, bem como mapas e gráficos.

Palavras-chave: evasão no ensino superior; dados educacionais; análise territorial; pipeline de dados; mapas temáticos.

Abstract

Dropout in higher education is a complex phenomenon influenced by institutional, socioeconomic, and territorial factors that demands analyses supported by data from multiple sources. In the Brazilian context, microdata from the Higher Education Census, produced by the National Institute for Educational Studies and Research Anísio Teixeira (INEP), provide ample detailed annual information that enables the

monitoring of dropout on the country over time, however, in their current public releases, they present limitations from a territorial perspective. In this scenario, this work aims to design and implement a data pipeline focused on the harmonization of INEP data with territorial and socioeconomic information produced by the Brazilian Institute of Geography and Statistics (IBGE), enabling the visualization of dropout patterns across the Brazilian territory. The adopted methodology includes data extraction, transformation, and loading, as well as the definition of dropout metrics, and data aggregation at the municipal, state, and national levels. As a result, an integrated analytical dataset is produced, along with maps and charts.

Keywords: higher education dropout; educational data; territorial analysis; data pipeline; thematic maps.

1 Introdução

O desempenho educacional representa um fator importante no âmbito nacional, pois tem relação direta com diversos aspectos sociais, econômicos, de inovação e até mesmo orçamentários. Nesse contexto, a evasão no ensino superior configura um problema sério, responsável pela perda potencial de profissionais capacitados e, conseqüentemente, pela redução do potencial de crescimento econômico e da produtividade nacional.

Logo, torna-se pertinente o estudo da evasão como meio de investigar e remediar os empecilhos detectados na formação efetiva de graduantes no Brasil. Nesse campo, a principal fonte de dados de acesso público que alimenta os estudos sobre a evasão nacional é produzida pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Trata-se dos microdados publicados anualmente pela autarquia, que permitem análises que vão do perfil das instituições ao comportamento das matrículas e das taxas de evasão.

Esses dados trazem índices importantes para o cenário atual como o número de matrículas em instituições de ensino superior (IES), bem como a composição dessas IES quanto à modalidade de ensino e à natureza administrativa. Tais índices podem ajudar a entender os efeitos oriundos de mudanças estruturais que já vem sendo observadas.

Quanto à natureza administrativa, observa-se que a expansão do ensino superior brasileiro tem sido impulsionada majoritariamente pela rede privada. Em 2023, as instituições privadas concentravam cerca de 79,3% das matrículas de graduação no país [15], proporção que alcançou aproximadamente 79,8% em 2024, evidenciando a predominância e o crescimento contínuo desse segmento em relação à rede pública [16].

Esse predomínio da rede privada está associado a transformações estruturais ocorridas desde a década de 1990, quando políticas de expansão do ensino superior estimularam a participação do setor privado. Nesse contexto, intensificou-se o processo de privatização da educação superior, com o crescimento das instituições privadas em relação às públicas e a formação de grandes grupos educacionais por meio de fusões e aquisições, fato que contou com a entrada de empresas internacionais no setor. Essas mudanças contribuíram

para a consolidação de oligopólios educacionais e do capital privado no ensino superior brasileiro [5].

Já no que diz respeito à modalidade de ensino, conforme o *Acórdão 658/2023* do Tribunal de Contas da União (TCU), em 2010, aproximadamente 15% das matrículas estavam na modalidade de Educação a Distância (EaD); esse percentual aumentou para 36% em 2020. Ao mesmo tempo, as matrículas presenciais diminuíram, o que indica não apenas o crescimento do EaD, mas também a substituição parcial das modalidades presenciais. Entre 2017 e 2020, a taxa média anual de crescimento no EaD superou 20%, em contraste com uma taxa negativa de cerca de 4% no ensino presencial. Em 2020, pela primeira vez, o número de ingressantes em EaD ultrapassou o de ingressantes em cursos presenciais, o que correspondeu a 53% dos novos alunos [8].

Essa evolução evidência uma mudança de paradigma no ensino superior brasileiro. Esse passou a depender fortemente da iniciativa privada e da modalidade EaD. Os motivos por trás desse fenômeno são complexos, porém, podem ser citados os limites aos gastos governamentais como fatores importantes no concebimento do cenário atual.

O investimento governamental no ensino superior brasileiro tem enfrentado limitações nas últimas décadas. Entre elas, o *teto de gastos*, que limita a expansão das ofertas de vagas públicas [11], contribuindo para que a demanda crescente no setor de ensino superior seja capturada cada vez mais pelas instituições privadas, que favorecem a modalidade EaD. Esses fatos vão de encontro ao desempenho estudantil, pois tanto a natureza privada quanto a modalidade EaD se entrelaçam com a evasão.

Devido à relevância desse problema para as políticas públicas e organizacionais, muitos estudos se dedicam a compreender os fatores relacionados à evasão. No âmbito internacional, Tinto (1975) marcou um avanço significativo nesse campo. Esse trabalho apresentou um modelo teórico que busca explicar as causas da evasão no ensino superior. Nesse trabalho, são identificados não apenas interações sociais entre estudantes e a instituição de ensino, mas também características pré-existentes à entrada na instituição [23].

O trabalho de Barroso et al. (2022) traz a predominância de fatores familiares e socioeconômicos no processo de evasão acadêmica. Se trata de uma revisão de literatura sobre o tema, e a partir dessa, foi encontrado que 58,3% dos estudos abrangidos discutem atributos prévios à entrada no ensino superior, como a situação familiar [4]. Esses achados reforçam a importância de considerar variáveis socioeconômicas e familiares no entendimento da permanência e do abandono estudantil no Brasil.

No que diz respeito ao escopo, Lobo et al. (2007) apontam que os estudos internos realizados pelas próprias IES podem apresentar maior detalhamento, pois permitem acompanhar de perto as situações associadas à evasão. Segundo os autores, esse tipo de investigação permite acompanhar individualmente os casos e a evolução da evasão, distinguindo modalidades como cancelamento, trancamento, transferência ou desistência [20].

Apesar de sua profundidade, esses estudos institucionais apresentam limitações

quanto à replicabilidade e à comparabilidade entre diferentes instituições e, consequentemente, entre localidades. Cada IES adota metodologias próprias adaptadas ao contexto institucional, o que dificulta a comparação direta entre esses estudos. Nesse contexto, o uso dos microdados do INEP surge como uma alternativa para estudos macroscópicos, capazes de identificar padrões, tendências e relações socioeconômicas entre diferentes localidades do país.

Considerando esse cenário, o presente trabalho busca facilitar o estudo da evasão sob a ótica territorial por meio da concepção de um *pipeline* de dados (fluxo automatizado de processos).

O *pipeline* apresentado demonstra um processo de integração de informações educacionais do INEP, no nível superior, integrando séries anuais entre si e permitindo análises territorializadas sobre os conjuntos através do alinhamento de dados de abrangência nacional produzidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Para tanto, este trabalho abrange o uso de dados censitários e malhas territoriais disponibilizadas pelo IBGE.

2 Fundamentação Teórica

O desempenho no sistema educacional brasileiro é objeto de estudo e acompanhamento frequentes. O Tribunal de Contas da União (TCU) que acompanha o tema já identificou problemas como limitações nos indicadores mantidos por instituições como o INEP, falta de informações territorializadas e, recentemente, a necessidade de substituir o e-MEC [6, 7, 8].

Esses processos evidenciam a evolução da coleta de dados e de indicadores no ensino brasileiro, que ainda é jovem em comparação com muitos países desenvolvidos. Entretanto, essa evolução implica na inconsistência entre as divulgações anuais. Os microdados do INEP, por exemplo, tiveram vários campos modificados ou removidos, o que dificulta análises longitudinais [9, 19].

Esses problemas refletem não apenas a necessidade de modernização dos sistemas digitais de muitos órgãos brasileiros, mas também a necessidade de prover melhor comunicação entre órgãos pertencentes ao mesmo setor educacional. Essa situação somente se amplifica quando outros corpos governamentais que não interagem com regularidade precisam ser aliados. Como observa Schymura, cada produtor de dados atua de forma isolada e responde a estruturas ministeriais distintas. O IBGE se vincula ao Ministério do Planejamento, o INEP ao Ministério da Educação, o Datasus ao Ministério da Saúde, e o Inpe ao Ministério da Ciência e Tecnologia, entre outros [21]. As bases resultantes utilizam linguagens e padrões distintos, o que dificulta a realização de estudos que abrangem dados de entidades independentes.

2.1 Microdados públicos do Censo da Educação Superior

O INEP é responsável pela consolidação das informações estatísticas sobre o ensino superior brasileiro por meio do *Censo da Educação Superior*. Até 2008, os dados de alunos e docentes eram coletados de forma agregada por curso, o que impossibilitava acompanhar trajetórias individuais e calcular indicadores longitudinais. Porém, a partir de 2009 passou a trazer informações no nível da matrícula, permitindo análises mais detalhadas [13]. Essa mudança marcou um avanço no potencial de análise das estatísticas educacionais produzidas pela autarquia.

Contudo, a disponibilização desses dados passou a ser restrita em 2022, quando o INEP suspendeu a divulgação dos registros individualizados do Censo da Educação Superior. A decisão foi fundamentada na necessidade de adequação à Lei n.º 13.709/2018 — Lei Geral de Proteção de Dados Pessoais (LGPD) —, conforme nota oficial publicada pelo Instituto [9]. O documento esclarece que estudos conduzidos em parceria com o Laboratório Inscript da Universidade Federal de Minas Gerais (UFMG) identificaram risco de reidentificação de indivíduos a partir de variáveis combinadas, como mês e ano de nascimento, código da instituição e município de residência.

Diante disso, a autarquia removeu variáveis potencialmente identificáveis como as que informavam o município de residência e o local de nascimento dos estudantes. Desde então, o acesso aos microdados individualizados do ensino superior passou a depender de autorização, mediante o Serviço de Acesso a Dados Protegidos (SEDAP). Essa restrição dificultou a reprodutibilidade de estudos baseados nos microdados, além de limitar análises territoriais que dependem de variáveis espaciais.

2.1.1 Limitações atuais dos microdados e lacunas territoriais

Nos microdados do Censo da Educação Superior, as únicas informações de caráter territorial atualmente disponíveis são aquelas associadas aos cursos [9]. Assim, cada registro de estudante, o qual se encontra agregado, herda o município do respectivo curso, e não o de sua residência ou origem.

Essa estrutura de dados mascara a localização real dos discentes que estudam em municípios diferentes de onde residem, relevante nas regiões de pequeno porte e nas áreas metropolitanas, onde o deslocamento entre cidades é frequente. O problema se torna ainda mais acentuado nos cursos da modalidade de EaD, nos quais a localização física da instituição tem pouca relação com o contexto socioeconômico do estudante.

Como consequência, os microdados tendem a representar a distribuição das instituições, e não necessariamente a distribuição territorial dos estudantes.

2.2 Dados do IBGE e sua integração com os microdados do INEP

A ausência de dados municipais associados aos estudantes nos microdados limita a compreensão das dinâmicas regionais relacionadas à evasão no ensino superior. Além disso, em muitos municípios brasileiros de pequeno porte, não há IES registradas nos microdados, o que resulta nesses municípios não constarem nos dados.

O presente trabalho utiliza dados do Censo Demográfico publicados pelo IBGE referentes ao ano 2010 e malhas territoriais também publicadas pelo órgão como fonte complementar, visando atingir três objetivos principais: prover uma base de referência que represente todos os municípios brasileiros reconhecidos (incluindo os ausentes nos microdados do INEP); enriquecer o conjunto de dados final com indicadores demográficos; e viabilizar a representação visual da evasão por meio de mapas temáticos.

2.3 Conceito de evasão no ensino superior

A evasão no ensino superior se trata da interrupção do vínculo do estudante com o curso, instituição ou com o sistema antes da conclusão da graduação e é amplamente usada como a principal métrica para avaliar a eficiência acadêmica. Na literatura internacional, o termo *dropout* é empregado com o mesmo significado que a evasão no Brasil, e serve como base para formulação de políticas educacionais. Modelos, como o de Tinto, têm destaque no âmbito do estudo da evasão e ajudaram a alicerçar a importância desse indicador ao associá-lo aos níveis de integração acadêmica e social do estudante [23].

2.3.1 Problemática da definição

Conforme destacado por Tinto [24], a definição de evasão depende do modelo utilizado. Esses modelos não buscam explicar todas as formas de evasão, mas sim aspectos do fenômeno, os quais podem mudar a depender do objetivo e do contexto. Um exemplo de variação entre estudos são de diferentes níveis de análise da evasão: no nível do curso, quando o estudante abandona o curso; no nível institucional, quando ocorre a saída da instituição; e no nível do sistema, quando há saída do ensino superior como um todo. Na prática, a escolha do nível e do modelo dependem da disponibilidade de dados, uma vez que bases mais amplas, como aquelas abrangendo a nação inteira, tendem a dispor de dados mais agregados, enquanto pesquisas dispondo de dados internos das IES podem ter dados mais detalhados.

2.4 Indicadores socioeconômicos

A literatura internacional indica que certos fatores socioeconômicos exercem influência sobre o desempenho acadêmico e a evasão no ensino superior. A revisão feita por Aina [2] traz diversos indicadores com diversos níveis de correlação com a evasão. Entre os mencionados, fatores relacionados a idade, renda, jornadas de trabalho simultâneas aos estudos e caracterização dos pais possuem alta relevância no fenômeno da evasão.

Nesse contexto, os dados produzidos pelo IBGE podem ser úteis para contextualizar territorialmente estudos sobre evasão e desempenho no ensino superior.

2.5 Trabalhos correlatos

Os trabalhos correlatos foram identificados a partir de buscas por palavras-chave como “*pipeline de dados INEP*”, priorizando estudos que abordam o processamento ou a visualização de dados educacionais públicos.

Amador et al. [3] propõem um processo de ETL para integração de microdados do INEP com foco em dados conectados e grafos de conhecimento. Embora compartilhe a utilização de pipelines de dados educacionais, o trabalho apresenta um recorte distinto da abordagem territorial adotada neste estudo.

Adão [1] desenvolve um dashboard interativo para visualização de indicadores do Censo Escolar, utilizando dados do INEP e recursos de georreferenciamento. Em contraste, o presente trabalho concentra-se no ensino superior e na integração de dados territoriais e socioeconômicos do IBGE.

Sousa [22] apresenta uma arquitetura automatizada de ETL para microdados públicos do INEP, aplicada ao ENEM. Diferentemente dessa proposta, este trabalho busca a compatibilização territorial dos dados de ensino superior e sua exploração para mapas e gráficos.

3 Especificação do Sistema

Esta seção apresenta a formalização dos requisitos, a arquitetura do pipeline analítico e o modelo de dados que compõem o sistema desenvolvido para integração, tratamento e análise dos microdados do INEP e dos dados territoriais e socioeconômicos do IBGE.

3.1 Visão Geral do Sistema

O pipeline desenvolvido neste trabalho trata-se de um sistema de composto por módulos responsáveis pela extração, transformação, integração, análise e visualização dos dados. O escopo inclui: coleta e processamento de dados brutos; cálculo da evasão operacional; compatibilização com indicadores do IBGE; geração de arquivos analíticos finais; e produção de visualizações.

3.2 Requisitos do Sistema

3.2.1 Requisitos Funcionais (RF)

- **RF1 – Ingestão de dados:** O sistema deve extrair os dados INEP e IBGE.
- **RF3 – Limpeza e Padronização:** O sistema deve padronizar colunas, corrigir inconsistências e validar dados faltantes.
- **RF4 – Cálculo da Evasão:** O sistema deve calcular o índice de evasão conforme metodologia definida na Seção 4.1.
- **RF6 – Construção do Conjunto Analítico compatível:** O sistema deve gerar um conjunto de dados final nos níveis municipal, estadual e nacional com colunas-chave compatíveis.
- **RF7 – Construção dos Mapas:** O sistema deve produzir mapas utilizando as malhas do IBGE.

3.2.2 Requisitos Não Funcionais (RNF)

- **RNF1 – Reprodutibilidade:** O pipeline deve permitir reexecução completa a partir do código-fonte, garantindo reprodutibilidade dos resultados.
- **RNF2 – Qualidade dos Dados:** O sistema deve implementar validações para identificar valores ausentes, códigos inválidos e inconsistências.
- **RNF3 – Organização Modular:** O código deve ser organizado em módulos distintos (extração, limpeza, integração, visualizações).
- **RNF4 – Performance:** O pipeline deve ser capaz de processar os microdados e integrar as bases em tempo adequado ao volume (no máximo 1 hora em ambiente local).
- **RNF5 – Segurança e Privacidade:** A manipulação dos microdados deve garantir anonimidade e conformidade com a LGPD.
- **RNF6 – Versionamento:** Todo o código deve ser versionado em sistema de controle de versão (Git).

3.3 Arquitetura do Pipeline

A arquitetura proposta segue um modelo de processamento de dados em camadas, estruturado segundo o processo de *Extract, Transform, Load* (ETL, do inglês).

O fluxo inicia-se na **Camada de Ingestão (Raw)**, responsável pela recepção dos arquivos em seus formatos originais, preservando os dados originais.

Em seguida, a **Camada de Processamento (Interim)** funciona como um diretório para armazenamento de *Checkpoints* (do inglês, pontos de controle).

A **Camada de Consolidação (Processed)** é responsável pelo armazenamento do conjunto final de dados.

Por fim, a **Camada de Consumo** utiliza os dados finais como base para visualizações.

3.3.1 Diagrama de Arquitetura

O diagrama conceitual da arquitetura do sistema está apresentado no Diagrama [1](#).

3.4 Modelo de Dados

O modelo de dados adotado é apresentado sob a forma de um *Star Schema*. Nesse modelo, os dados educacionais possuem natureza longitudinal e representam os eventos que variam ao longo do tempo, enquanto os dados demográficos fornecem o contexto regional estático, caracterizando a separação entre fatos e dimensões.

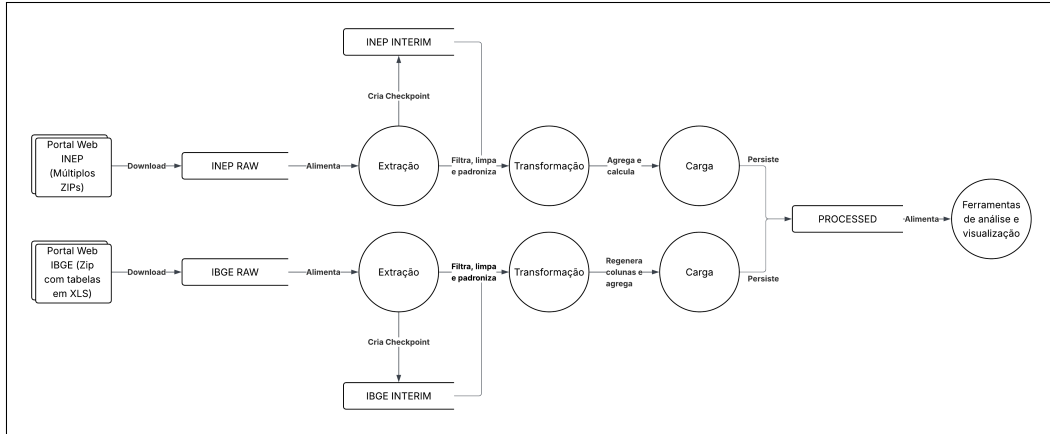


Diagrama 1: Arquitetura conceitual do pipeline analítico.

3.4.1 Tabela Fato: (INEP)

A tabela fato representa o grão da análise, chaveado por (COD_MUNICIPIO), (SG_UF) e por ano de referência (NU_ANO_CENSO). Cada registro corresponde a uma região ou municipalidade específica. Os principais atributos quantitativos são os indicadores educacionais. Esses campos são utilizados em análises estatísticas e comparações temporais. Além das métricas numéricas, a tabela fato contém atributos categóricos que qualificam os registros, como o tipo de rede de ensino, a modalidade de ensino e o grau acadêmico, permitindo segmentações dos dados operacionais.

3.4.2 Tabela Dimensão: (IBGE)

A tabela de dimensão regional fornece o contexto socioeconômico e demográfico. A chave primária da dimensão são os atributos (COD_MUNICIPIO) e (SG_UF) comuns às duas fontes.

3.5 Princípios

A modelagem foi guiada por três critérios:

1. **Expansibilidade temporal:** o pipeline deve aceitar a entrada de novos anos sem necessidade de reestruturação.
2. **Variabilidade de indicadores IBGE:** as colunas socioeconômicas podem mudar a cada extração; portanto, o modelo precisa ser agnóstico ao conjunto exato de variáveis.
3. **Capacidade de derivação analítica:** diversos indicadores como a evasão anual dependem da presença simultânea de várias colunas derivadas. O modelo deve manter coerência para cálculos posteriores.

3.5.1 Estrutura do Conjunto Analítico

O conjunto final combina informações em nível municipal ou mais agregado, incluindo todas as métricas calculadas, variáveis de validação, bloco atributos socioeconômicos provenientes do IBGE e quaisquer variáveis que o usuário deseje preservar.

- **Tabela FatoEvasao:** contém, para cada município, a principal métrica calculada. Nesse trabalho, se trata da evasão anual.
- **DimMunicipio:** informações territoriais identificadoras.
- **DimSocioeconomicaIBGE:** contém o conjunto de variáveis provenientes das tabelas IBGE utilizadas.
- **DimTempo:** Representa a variação temporal no conjunto INEP.

3.5.2 Diagrama Conceitual

O diagrama 2 ilustra a organização conceitual.

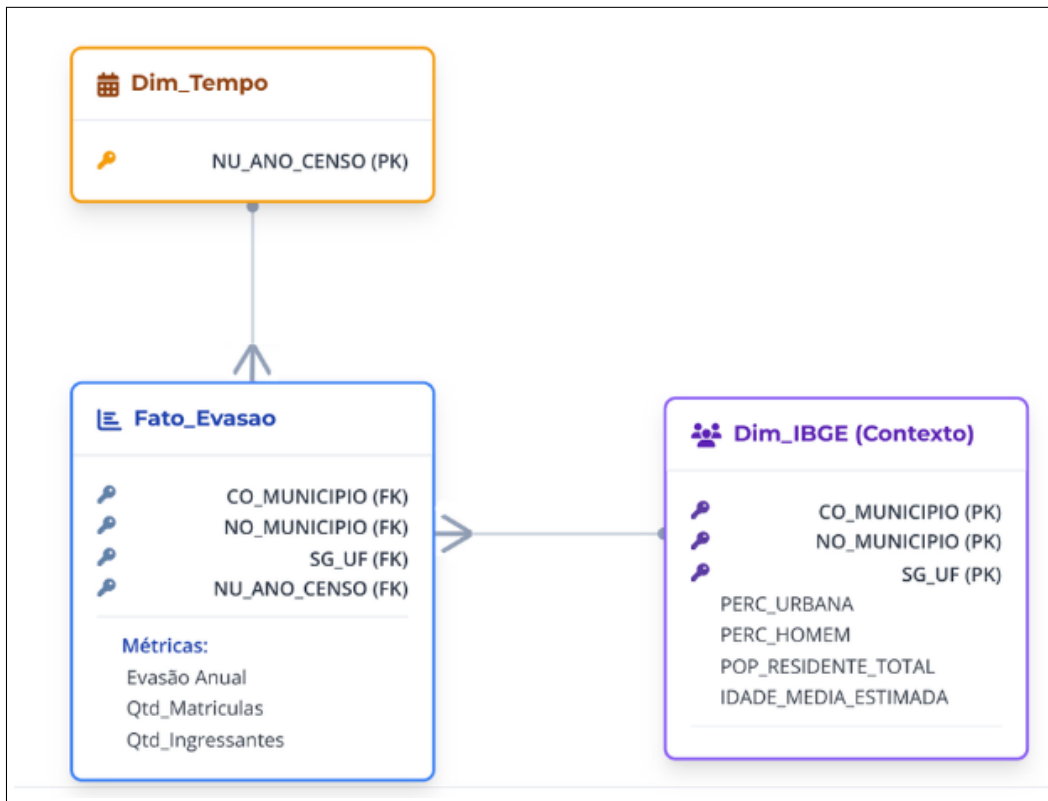


Diagrama 2: Modelo de dados para integração INEP-IBGE.

4 Metodologia

4.1 Evasão Operacional

Neste trabalho, adota-se a definição operacional de evasão proposta pelo Instituto Lobo para Desenvolvimento da Educação, da Ciência e da Tecnologia [19].

O cálculo da evasão parte da determinação da taxa de permanência (P), que representa a proporção de estudantes que permanecem matriculados entre dois anos consecutivos. Essa taxa é dada pela relação entre o número de alunos matriculados que continuam no curso e o total potencial de continuidade, conforme a seguinte expressão:

$$P = \frac{M(n) - I(n)}{M(n-1) - C(n-1)} \quad (1)$$

onde:

- $M(n)$: número de matrículas no ano de referência n ;
- $I(n)$: número de ingressantes no mesmo ano;
- $M(n-1)$: número de matrículas no ano anterior;
- $C(n-1)$: número de concluintes no ano anterior.

A partir dessa taxa, a evasão (E) é obtida como o complemento da permanência, representando o percentual de alunos que deixaram o curso entre os dois anos analisados:

$$E(n) = 1 - P = 1 - \frac{M(n) - I(n)}{M(n-1) - C(n-1)} \quad (2)$$

O Instituto Lobo traz que essa metodologia é adequada quando se trabalha com dados agregados entre múltiplos anos, pois independe de informações sobre transferências internas ou externas de estudantes, cujos registros nem sempre estão disponíveis de forma consistente nos censos [18]. Atualmente, se torna mais relevante ainda, pois não depende de dados fechados.

4.2 Configurações

O pipeline é orientado por arquivos de configuração no formato **YAML**, que permitem definir de forma declarada o comportamento do sistema sem necessidade de modificação do código-fonte.

Dentre os arquivos de configuração, **base.yml** é responsável por definir colunas-chave territoriais e o formato de saída dos dados processados. A Listagem 1 apresenta um exemplo de configuração utilizada para alinhar os pipelines.

Listagem 1: Exemplo de configuração base para alinhamento entre os pipelines

```
saída:
```

```
encoding: "utf-8"
sep: ";"
municipio:
  codigo: "CO_MUNICIPIO"
  nome: "NO_MUNICIPIO"
estado:
  sigla: "SG_UF"
nacional:
  coluna: "ID_BRASIL"
```

As chaves `codigo`, `nome` e `sigla` definem as colunas identificadoras de território utilizadas nas tabelas do INEP. Nas planilhas censitárias do IBGE empregadas neste trabalho, os campos correspondentes aparecem nas primeiras colunas, ainda que sob denominações distintas, o que permite o alinhamento entre os pipelines por meio de uma etapa de renomeação.

A configuração nacional cria uma coluna de agregação para o nível do país, sendo que, como que esse nível não consta nos dados originais, o nome dessa coluna é atribuído via arquivo de configuração. As opções especificadas em `saida` determinam parâmetros comuns aos arquivos finais gerados em formato `.csv`, assegurando uniformidade na codificação e separação dos dados produzidos.

Os dados originais e derivados não versionados seguem uma estrutura de diretórios que reflete o estágio de processamento dos arquivos ao longo do pipeline:

- `/DATA/RAW` — armazenamento exclusivo dos arquivos originais, conforme disponibilizados pelos portais do IBGE e do INEP.
- `/DATA/INTERIM` — arquivos intermediários gerados após as rotinas iniciais de extração e limpeza.
- `/DATA/PROCESSED` — conjuntos finais de dados derivados das etapas de transformação e integração.
- `/DATA/RENDERED` — arquivos de saída gráfica, como mapas e visualizações produzidas a partir dos dados consolidados.

4.3 Fontes de Dados

O pipeline utiliza dados provenientes de duas fontes de abrangência nacional: o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e o Instituto Brasileiro de Geografia e Estatística (IBGE).

Os microdados do Censo da Educação Superior foram obtidos a partir do portal oficial do INEP [12]. Para cada ano analisado, foi utilizado o arquivo de cadastro de cursos, denominado `MICRODADOS_CADASTRO_CURSOS_<ano>.CSV`, em que `<ano>` corresponde ao ano de referência do censo. Foram empregadas, nas execuções do pipeline, as bases anuais do Censo da Educação Superior correspondentes ao período de 2016 a 2024. Os arquivos censitários originais do INEP, no formato `.csv`, devem ser posicionados no diretório:

/DATA/RAW/INEP/

Os dados do Censo Demográfico de 2010 foram obtidos por meio das planilhas de tabulação disponibilizadas no portal oficial do IBGE [10]. A escolha desse censo deve-se ao fato de apresentar estrutura conhecida e de fácil identificação das colunas, o que favorece a construção do dicionário de dados empregado na demonstração do funcionamento do pipeline. Diferentemente do censo mais recente, cuja organização é orientada ao processamento automático, o Censo 2010 apresenta estrutura voltada à leitura humana, exigindo menor tratamento inicial.

Nos dados agregados por município, por exemplo, a coluna de sigla da unidade federativa não está presente, sendo obtida a partir dos dígitos iniciais do código municipal. Além disso, as colunas de indicadores requerem a consulta ao dicionário de dados disponibilizado junto às planilhas. Para a utilização de dados censitários mais recentes, seria necessário o alinhamento ao formato esperado pelo pipeline, incluindo a extração da sigla da unidade federativa e a adequação do formato de entrada. Alternativamente, a etapa de extração do pipeline IBGE pode ser estendida para suportar explicitamente os novos formatos.

Os arquivos censitários originais do IBGE, no formato `.xls`, devem ser posicionados no diretório:

/DATA/RAW/IBGE/

4.4 Processamento INEP

Os dados do INEP são processados por meio de um pipeline orientado ao processo de *Extract, Transform, Load* (ETL), que ingere dados anuais INEP.

4.4.1 Extração dos Dados INEP

A etapa de extração dos microdados do Censo da Educação Superior é responsável por ler, normalizar e persistir versões intermediárias dos arquivos originais dentro do período selecionado. A partir da configuração, permite o mapeamento de nomes divergentes de colunas entre diferentes edições do censo, remediando inconsistências entre bases anuais.

As variáveis selecionadas são organizadas em três grupos: temporais, quantitativas e categóricas. Essa classificação define o tratamento posterior na etapa de transformação, determinando o papel de cada coluna no processo de agregação e no cálculo das métricas.

As variáveis categóricas correspondem a classificações simbólicas definidas pelo INEP, como modalidade de ensino ou tipo de rede, e exigem a declaração explícita de seus valores possíveis. A configuração permite ainda a definição de filtros para exclusão de categorias específicas, restringindo o conjunto de dados final.

Considerando o tamanho dos microdados originais, filtros são aplicados ainda durante a extração, evitando a persistência de arquivos intermediários grandes. Ao final dessa etapa, os dados intermediários são armazenados no diretório:

4.4.2 Transformação dos Dados INEP

Na etapa de transformação, são realizados o tratamento de registros, as agregações territorial e temporal, e o cálculo das métricas definidas pelo usuário. Registros sem identificação territorial, comuns em cursos na modalidade EaD, podem ser preservados via atribuição de valores ou descartados. O comportamento padrão adotado neste trabalho é a atribuição dos valores "ND" para a sigla da unidade federativa, "-1" para o código municipal e "SEM_MUNICIPIO" para o nome do município.

As variáveis temporais definem o eixo cronológico da análise, enquanto que as variáveis quantitativas representam contagens absolutas e constituem a base para agregações e para o cálculo de métricas derivadas. As variáveis quantitativas são agrupadas por soma simples.

No caso das variáveis categóricas, essas são transformadas de valores simbólicos em valores percentuais representativos de cada categoria. Isso implica no desdobramento de colunas originais em colunas derivadas. Para tanto, uma das colunas contendo valores quantitativos é declarada como peso na configuração, permitindo o cálculo do percentual desses valores sobre cada categoria. Por exemplo, a variável TP_MODALIDADE_ENSINO com valores 1 para presencial e 2 para EaD, encontrada nos microdados, pode ser desdobrada em TP_MODALIDADE_ENSINO_Presencial e TP_MODALIDADE_ENSINO_EaD.

O principal indicador analisado neste trabalho é a taxa de evasão anual no ensino superior, definida na Seção 4.1. Para permitir a inclusão de anos consecutivos, necessários para o cálculo da evasão anual, as fórmulas utilizam marcadores que são substituídos pelo ano base e ano seguinte durante a execução do pipeline, sendo {p} o ano base e {n} o ano seguinte.

A Equação 3 apresenta a formula para cálculo da evasão anual usando nomes canônicos dos microdados INEP:

$$\text{Evasão}_{\text{anual}} = 1 - \frac{QT_MAT_TOTAL_n - QT_ING_TOTAL_n}{QT_MAT_TOTAL_p - QT_CONC_TOTAL_p} \quad (3)$$

A seguir, é demonstrado como essa equação é declarada na configuração.

```
formulas:
  evasao_anual:
    descricao: "Evasao_anual_considerando_todos_os_alunos"
    expressao:
      1 - ((QT_MAT_TOTAL_{n} - QT_ING_TOTAL_{n}) /
            (QT_MAT_TOTAL_{p} - QT_CONC_TOTAL_{p}))
```

Como existem, nos microdados, colunas contendo número de ingressantes, matriculados e concluintes com financiamento estudantil, é possível calcular a evasão anual para

esse grupo também. A Equação 4 apresenta a versão da métrica de evasão anual aplicada exclusivamente a estudantes com financiamento.

$$\text{Evasão}_{\text{financiados}} = 1 - \frac{QT_MAT_FINANC_n - QT_ING_FINANC_n}{QT_MAT_FINANC_p - QT_CONC_FINANC_p} \quad (4)$$

Regras adicionais podem ser associadas às métricas para invalidar registros que não as satisfaçam, como municípios com número muito baixo de matrículas. As métricas são calculadas para os três níveis territoriais e atribuídas ao registro correspondente ao ano subsequente, não sendo possível o cálculo para o primeiro ano do período.

4.4.3 Carga dos Dados INEP

A etapa de carga recebe os dados transformados nos níveis municipal, estadual e nacional e os persiste conforme as configurações de saída definidas no arquivo `base.yml`. Os arquivos finais são armazenados no diretório:

`/DATA/PROCESSED/INEP/`

4.5 Processamento IBGE

Os dados do IBGE são processados por meio de um pipeline orientado ao processo de *Extract, Transform, Load* (ETL), que ingere tabelas no formato `.xls`.

4.5.1 Extração dos Dados do IBGE

A extração dos dados do IBGE consiste na leitura de arquivos no formato `.xls` e na identificação das planilhas contidas nesses arquivos. O processo é orientado por um arquivo de configuração que reflete a estrutura interna das planilhas originais.

O número de arquivos extraídos pode ser maior que o número de arquivos fonte empregados, pois é comum nesses dados do IBGE a presença de subconjuntos dos dados. Por exemplo, subconjuntos rurais e urbanos. Os arquivos extraídos são convertidos para o formato `.csv` e armazenados no diretório:

`/DATA/INTERIM/IBGE/`

4.5.2 Transformação dos Dados IBGE

A etapa de transformação das planilhas IBGE consiste nas seguintes operações: limpeza, reconstrução semântica de colunas, definição de colunas de peso, definição de formato, remoção de colunas, agrupamento de colunas e agregação estadual e nacional. A limpeza é responsável por remover linhas resquiciais dos arquivos `.xls` como cabeçalhos múltiplos e comentários. Já a reconstrução das colunas é feita a partir de um dicionário de dados elaborado com estrutura similar às planilhas originais, porém consistente com o formato `.csv`.

A definição do nome de cada coluna é necessária pois os cabeçalhos das planilhas IBGE são compostos por mais de uma linha, e fazem uso de mesclagem de células. Esses formatos característicos de planilhas eletrônicas perdem sua forma e sentido ao serem convertidos em `.csv`. Essas colunas podem ter formato declarado dentre as opções suportadas: `CONTAGEM`, `PORCENTAGEM`, `MEDIA` e `PROPORCAO`. Se nenhum for selecionado, por padrão, o formato será `CONTAGEM`, enquanto que demais formatos requerem uso obrigatório de campo de coluna de peso como população ou domicílios para auxiliar agregações. As colunas de peso devem ser declaradas na configuração.

O pipeline também permite agrupar colunas com elevado acoplamento entre si. Por exemplo:

`colunas:`

- `POP_RESIDENTE_TOTAL`
- `PERC_0_A_5`
- `PERC_6_A_14`
- `PERC_15_A_24`
- `PERC_25_A_39`
- `PERC_40_A_59`
- `PERC_60_OU MAIS`

As colunas, desde a segunda até a última representam faixas etárias que podem ser agrupadas em uma única coluna de idade média estimada. Para tanto, é necessário a concepção dos seguintes campos no arquivo de configuração `.yaml`:

`merges_colunas:`

- `destino: IDADE_MEDIA_ESTIMADA`
- `fontes_idx: [2, 3, 4, 5, 6, 7]`
- `metodo: MEDIA_PONDERADA`
- `peso_merge: [2.5, 10, 19.5, 32, 49.5, 70]`
- `formato: MEDIA`
- `coluna_peso: populacao`

Esses campos dentro de `merges_colunas` definem, no arquivo de configuração `.yaml`, o procedimento para criação de uma coluna contendo a estimativa de idade média a partir de faixas etárias identificadas por índice e associadas a pesos.

4.5.3 Carga dos Dados IBGE

A etapa de carga recebe os dados transformados nos níveis municipal, estadual e nacional e os persiste conforme as configurações de saída definidas no pipeline. Os arquivos finais são armazenados no diretório:

`/DATA/PROCESSED/IBGE/`

4.6 Visualizações

São gerados dois tipos de gráficos: séries temporais (gráficos de linhas) e gráficos de dispersão (*scatter plots*).

Os gráficos de linhas são utilizados para representar a evolução temporal de indicadores sobre o tempo, como a evasão anual. O sistema oferece suporte à conversão de formato, permitindo que métricas, as quais são expressas em proporção, sejam convertidas para porcentagem nesses gráficos e não permite formatos que não possam ser harmonizados como contagens absolutas juntamente com proporções ou porcentagens. Já os gráficos de dispersão utilizam duas variáveis como eixo e não representam o tempo visualmente.

O sistema também gera mapas temáticos construídos a partir das malhas territoriais oficiais do IBGE referentes ao ano de 2024 [14]. Os arquivos extraídos dessa fonte devem estar armazenados nos seguintes diretórios, segundo o nível territorial:

```
/DATA/SHAPEFILES/BR_Municipios_2024
```

```
/DATA/SHAPEFILES/BR_UF_2024
```

4.6.1 Tratamento de Ausência de Dados nos Mapas

Os mapas reconhecem três tipos de ausências de dados:

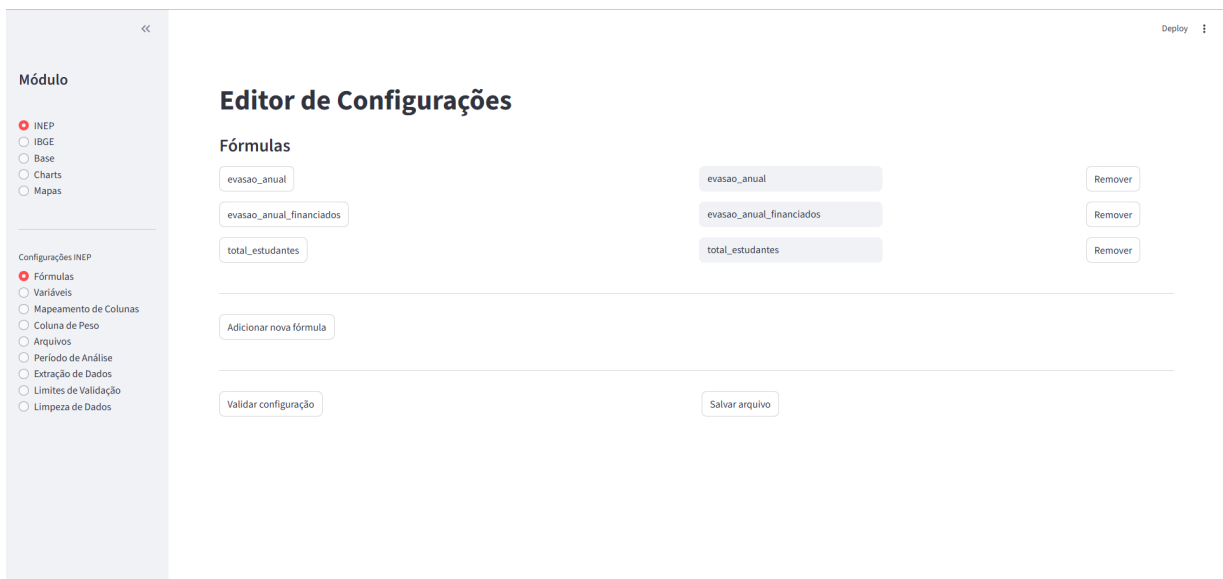
- **Fora do INEP:** territórios que não aparecem em nenhum ano do período processado, indica ausência completa de registros nos microdados.
- **Sem dados no ano:** territórios que existem em algum ano dentro do período, mas não estão presentes no ano representado.
- **Sem métrica no ano:** territórios que possuem registros no ano analisado, porém onde as métricas não puderam ser calculadas por motivo de regra aplicada ou dados inválidos.

4.7 Interface Gráfica do Sistema

A interface gráfica possibilita a configuração do pipeline, evitando a edição direta de arquivos de configuração. Conta com formato padrão carregado dos arquivos `.yaml` já disponibilizados no repositório. As telas são organizadas de modo que cada arquivo de configuração represente um módulo e cada subtela dentro de um módulo representa um campo de nível mais alto no respectivo arquivo `.yaml`.

4.7.1 Tela INEP

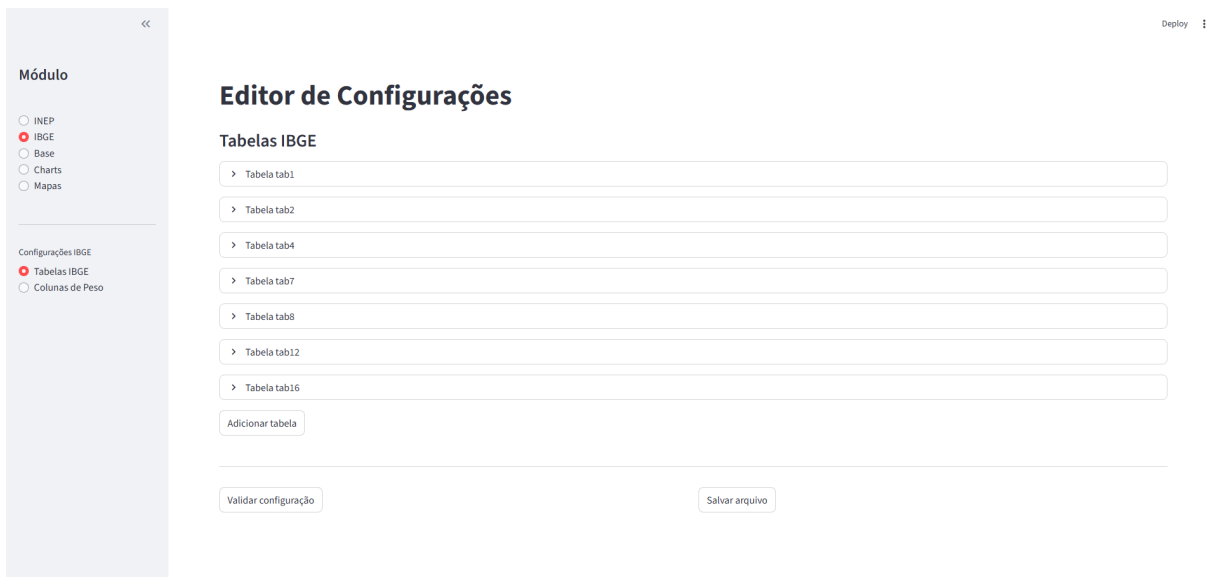
A tela INEP permite a seleção, mapeamento e filtro das variáveis provenientes dos microdados do Censo da Educação Superior, além da definição de métricas derivadas e outras configurações como formato dos dados de entrada e procedimento de limpeza de campos ausentes.



Tela 1: Configuração dos dados do INEP.

4.7.2 Tela IBGE

A tela IBGE é responsável por guardar a estrutura exata das tabelas IBGE e as transformações realizadas, incluindo colunas de peso requeridas.



Tela 2: Configuração dos dados do IBGE.

4.7.3 Tela Base

A tela Base permite definir as colunas-chave e o formato de saída.

Tela 3: Definição da base analítica e formatos de saída.

4.7.4 Tela Gráficos

A tela Gráficos define a criação de linhas temporais e diagramas de dispersão selecionando o tipo de gráfico, variáveis utilizadas e território.

Tela 4: Visualização de gráficos temporais e diagramas de dispersão.

4.7.5 Tela Mapas

A tela Mapas define a configuração dos mapas gerados.



Tela 5: Visualização de mapas temáticos coropléticos.

4.8 Stack Tecnológico

O desenvolvimento do pipeline e das visualizações foi realizado utilizando a linguagem Python, escolhida pela ampla disponibilidade de bibliotecas voltadas ao processamento, análise e visualização de dados. A Tabela 1 apresenta as principais tecnologias e bibliotecas empregadas no projeto.

Tabela 1: Stack tecnológico utilizado no projeto

Tecnologia	Versão	Finalidade
Python	3.x	Linguagem principal para implementação do pipeline de dados.
PyYAML	—	Leitura e interpretação de arquivos de configuração no formato YAML.
ruamel.yaml	—	Manipulação de arquivos YAML, preservando estrutura para uso como modelo.
Pandas	—	Processamento, limpeza, transformação e agregação de dados tabulares.
NumPy	—	Suporte a operações numéricas e vetorializadas utilizadas no processamento dos dados.
Matplotlib	—	Geração de gráficos estáticos, incluindo séries temporais e diagramas de dispersão.
GeoPandas	—	Manipulação de dados geoespaciais e integração com malhas territoriais do IBGE.
xlrd	—	Leitura de planilhas em formato .xls utilizadas em dados públicos.
Streamlit	≥ 1.30	Desenvolvimento da interface gráfica interativa para configuração e visualização dos resultados.

4.9 Exemplos de Códigos Principais

4.9.1 Códigos do Pipeline INEP

Listagem 2: Resolução automática de schema INEP

```
def resolver_schema_entrada(header):
    colunas_fisicas = []
    mapeamento = {}
    faltantes = set()
    for var in variaveis_esperadas:
        if var in header:
            colunas_fisicas.append(var)
            continue
        encontrado = False
        for alternativo, destino in map_dict.items():
            if destino == var and alternativo in header:
                colunas_fisicas.append(alternativo)
                mapeamento[alternativo] = var
                encontrado = True
                break
    if not encontrado:
```

```

        faltantes.add(var)
    return colunas_fisicas,mapeamento,faltantes

```

Listagem 3: Padronização de variáveis categóricas

```

def padronizar_categoricas(df):
    df_out=df.copy()
    for var in categoricas:
        if var not in df_out.columns:
            continue
        serie=df_out[var].apply(
            lambda x:
                x if x in valores_validos
                else "OUTROS"
        )
        dummies=pd.get_dummies(
            serie,
            prefix=var,
            dtype=float
        )
        df_out=pd.concat(
            [df_out,dummies],
            axis=1
        )
        df_out.drop(
            columns=[var],
            inplace=True
        )
    return df_out

```

Listagem 4: Agregação quantitativa INEP

```

def agrega_quantitativas(
    df,
    nivel="municipal"
):
    if nivel=="municipal":
        colunas_groupby=
            campos_padrao+[ano]
    elif nivel=="estadual":
        colunas_groupby=[UF,ano]
    elif nivel=="nacional":
        colunas_groupby=[ano]
    agg=
        df.groupby(
            colunas_groupby
        )[colunas_quant]
        .sum()

    return agg.reset_index()

```

Listagem 5: Cálculo automático de indicadores

```

def calcular_formulas(

```

```

df,
anos,
formulas,
col_ano,
col_chave
):
    for nome_formula, config in formulas.items():
        nome_coluna =
            nome_formula.upper()
        if nome_coluna not in df.columns:
            df[nome_coluna] = np.nan
        for ano_base, ano_seguinte in zip(
            anos[:-1],
            anos[1:]
        ):
            valores =
                avaliar_expressao(
                    config.expressao,
                    df,
                    ano_base,
                    ano_seguinte
                )
            mask =
                df[col_ano] == ano_seguinte
            df.loc[mask, nome_coluna] =
                valores

    return df

```

4.9.2 Códigos do Pipeline IBGE

Listagem 6: Limpeza de registros

```

def limpar(df):
    mask = df.iloc[:, 0].apply(
        is_codigo_municipio
    )
    df = df[mask].copy()
    df = df.map(
        lambda x:
            x.strip()
            if isinstance(x, str)
            else x
    )
    return df

```

Listagem 7: Integração de planilhas

```

def integrar_sheets_tabela(
    dfs_sheets,
    colunas_base,
):
    if not dfs_sheets:
        return None
    colunas_base = list(colunas_base)

```

```

dfs = iter(dfs_sheets.values())
df_final = next(dfs).set_index(
    colunas_base,
    drop=False
)
for df in dfs:
    df_idx = df.set_index(
        colunas_base,
        drop=False
    )
    df_final = df_final.join(
        df_idx.drop(
            columns=colunas_base
        ),
        how="outer",
    )
return df_final.reset_index(
    drop=True
)

```

Listagem 8: Cálculo de média ponderada na agregação

```

def media_ponderada(
    df: pd.DataFrame,
    coluna: str,
    peso: str,
) -> float:
    valores = pd.to_numeric(df[coluna], errors="coerce")
    pesos = pd.to_numeric(df[peso], errors="coerce")
    mask = valores.notna() & pesos.notna() & (pesos > 0)
    valores_f = valores[mask]
    pesos_f = pesos[mask]
    if valores_f.empty:
        return float("nan")
    total_peso = pesos_f.sum()
    if total_peso == 0:
        return float("nan")
    return (valores_f * pesos_f).sum() / total_peso

```

Listagem 9: Agregação estadual a partir dos dados municipais

```

def agregar_estadual(
    df: pd.DataFrame,
    agregacoes: list[AgregacaoColuna],
) -> pd.DataFrame:
    registros = []
    UF = get_colunas_municipio()[1]
    for uf, g in df.groupby(UF):
        row = {UF: uf}
        for agg in agregacoes:
            if agg.nome not in g.columns:
                continue
            if agg.tipo == TipoAgregacao.SOMA:
                row[agg.nome] = soma_numerica(g, agg.

```

```

        nome)
    elif agg.tipo == TipoAgregacao.MEDIA_PONDERADA:
        if not agg.coluna_peso:
            continue
        row[agg.nome] = media_ponderada(
            g,
            coluna=agg.nome,
            peso=agg.coluna_peso,
        )
    registros.append(row)
return pd.DataFrame(registros)

```

Listagem 10: Pipeline de agregação

```

def executar_agregacoes(
    df_municipal: pd.DataFrame,
    config_ibge: dict,
) -> None:
    agregacoes_raw = iterar_sheets_ibge(
        tabelas=config_ibge,
        fn=coletar_agregacoes,
    )
    agregacoes = []
    if agregacoes_raw:
        for sub in agregacoes_raw:
            agregacoes.extend(sub)

    # Persistencia municipal
    persistir_tabela_final(
        df_municipal,
        nivel="municipal"
    )

    # Agregacao estadual

    df_estadual = agregar_estadual(
        df_municipal,
        agregacoes
    )
    persistir_tabela_final(
        df_estadual,
        nivel="estadual"
    )

    # Agregacao nacional
    df_nacional = agregar_nacional(
        df_municipal,
        agregacoes
    )
    persistir_tabela_final(
        df_nacional,
        nivel="nacional"
    )

```

4.10 Scripts Utilizados

Todo o código-fonte desenvolvido neste trabalho, incluindo os pipelines de extração, transformação e carga dos dados, a interface gráfica de configuração e os scripts de geração de visualizações, encontra-se versionado em repositório público na plataforma GitHub¹.

Os scripts permitem a reprodução das etapas de processamento de dados, desde a extração até a geração das visualizações. Os comandos utilizados para execução dos pipelines e da interface gráfica encontram-se descritos no Apêndice 7.

5 Resultados

O produto principal do pipeline, sobre o qual as visualizações são construídas é o conjunto de dados nos níveis municipal, estadual e nacional. Com o objetivo de enriquecer os conjuntos finais, os dados processados do IBGE são harmonizados com as mesmas colunas-chaves presentes nos dados do INEP, como descrito na metodologia. Para demonstrar a compatibilidade dos dados, foi realizada uma análise de correlação entre métricas de evasão e variáveis socioeconômicas municipais. A Tabela 2 apresenta alguns dos resultados.

Tabela 2: Principais correlações entre evasão anual e indicadores municipais

Métrica	Indicador municipal	Correlação
Evasão anual	Percentual de habitantes do sexo masculino	0.128
Evasão anual	Idade média	-0.052
Evasão anual	Percentual com renda até 1/2 salário mínimo	-0.057
Evasão anual	Percentual de habitantes analfabetos com saneamento inadequado	-0.053
Evasão anual	Renda per capita	0.001

A correlação positiva implica que a evasão cresce em conjunto com o indicador socioeconômico, e negativa o contrário. É importante ressaltar que os indicadores socioeconômicos mostrados pertencem aos dados censitários IBGE de 2010, caracterizando uma defasagem temporal, e que englobam a população como um todo, não apenas estudantes. Logo, a tabela produzida não indica, por exemplo, que estudantes com idade mais avançada são menos propensos a evadir, mas sim que municípios com idade elevada têm menor índice de evasão.

5.1 Linhas Temporais

As linhas temporais permitem observar a evolução dos dados ao longo dos anos, tanto em nível nacional quanto estadual ou municipal. O Gráfico 1 apresenta a evolução da evasão anual em nível nacional, incluindo a evasão entre estudantes com financiamento, enquanto o Gráfico 2 traz o comportamento dessas métricas apenas no estado da Bahia.

¹Disponível em: <https://github.com/guilherme-lordelo/analise-evacao-inep-ibge>. Acesso em: jan. 2026.

Os gráficos trazem uma tendência de crescimento da evasão ao longo do período, apesar de estabilizar nos três últimos anos representados.

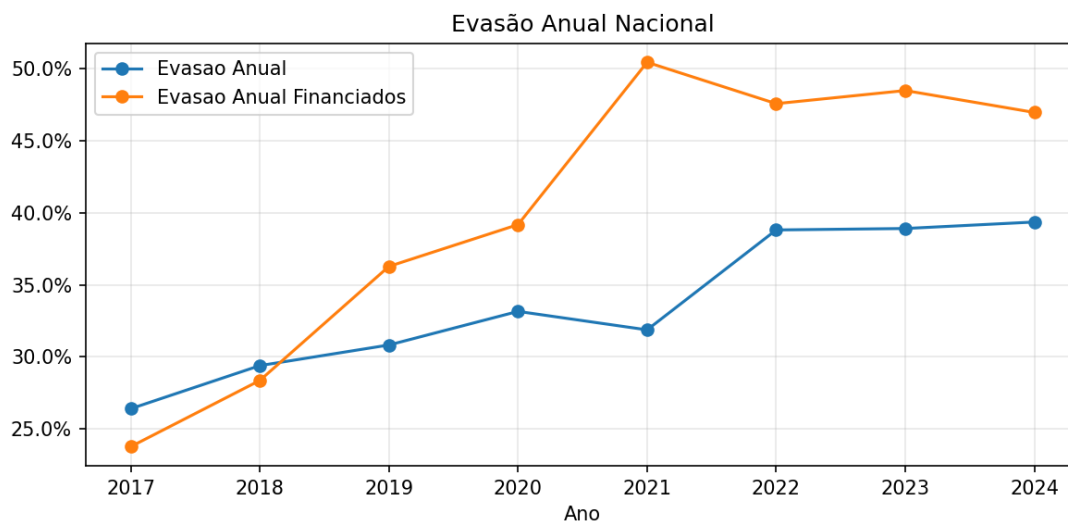


Gráfico 1: Evolução da evasão anual em nível nacional.

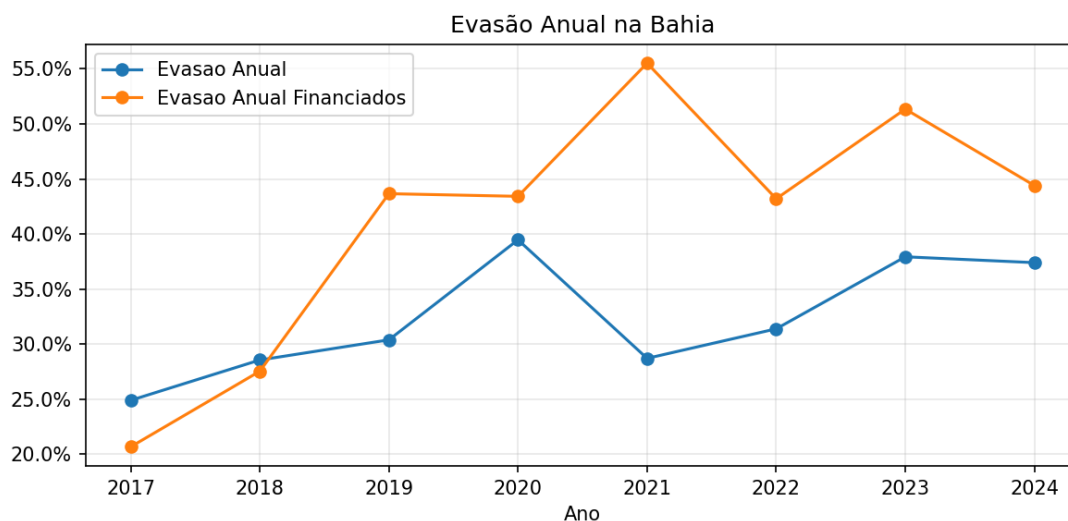


Gráfico 2: Evolução da evasão anual no estado da Bahia.

5.1.1 Percentual de Matrículas por Modalidade

Usando número de matrículas como coluna de peso, foi possível acompanhar percentuais de matrículas por categoria. O Gráfico 3 apresenta a evolução dos percentuais de matriculados por tipo de modalidade de ensino. Observa-se um crescimento da modalidade EaD, passando de aproximadamente 20% das matrículas em 2016 até ultrapassar, pela primeira vez, a modalidade presencial no ano final da série temporal, em 2024. Esse

comportamento é constatado pelo resultado do Censo da Educação Superior 2024 divulgado pelo INEP [17].

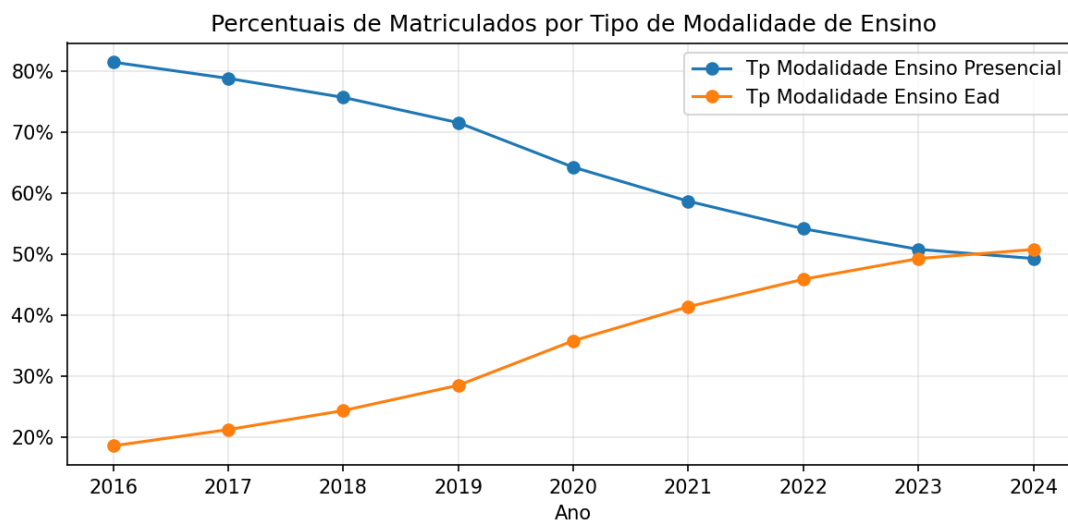


Gráfico 3: Evolução dos percentuais de matriculados por modalidade de ensino.

5.1.2 Evasão Anual por Modalidade de Ensino

Com o objetivo de observar o comportamento da evasão por modalidade de ensino, foram realizadas novas execuções do pipeline após ajustes na configuração. A Figura 1 e a Figura 2 mostram as modificações realizadas na configuração de arquivos do pipeline INEP através da interface gráfica para gerar um conjunto de dados contendo agregações de cursos presenciais, excluindo a modalidade a distância. As mudanças realizadas na primeira figura modificam o nome dos arquivos intermediários e finais enquanto que na segunda, foi aplicado um filtro sobre os cursos à distância.

Extração

Prefixo input

MICRODADOS_CADASTRO_CURSOS_

Extensão input

.CSV

Prefixo output

inep_reduzido_

Extensão output

.CSV

Transformação

Prefixo output

inep_

Extensão output

.CSV

Nível 1

municipios_

Nível 2

estados_

Nível 3

nacional_

Extração

Prefixo input

MICRODADOS_CADASTRO_CURSOS_

Extensão input

.CSV

Prefixo output

inep_reduzido_presencial_

Extensão output

.CSV

Transformação

Prefixo output

inep_

Extensão output

.CSV

Nível 1

municipios_presencial_

Nível 2

estados_presencial_

Nível 3

nacional_presencial_

Figura 1: Alteração de nomes de arquivos por interface gráfica.

TP_MODALIDADE_ENSINO

Descrição

Modalidade de ensino

Código	Descrição
1	Presencial
2	EaD

Adicionar valor

Excluir valores

Choose options

TP_MODALIDADE_ENSINO

Descrição

Modalidade de ensino

Código	Descrição
1	Presencial
2	EaD

Adicionar valor

Excluir valores

2 - EaD x

Figura 2: Introdução de filtro de cursos à distância por interface gráfica.

A partir desse conjunto, foram geradas novas linhas temporais representando a evasão anual e a evasão anual entre estudantes financiados, considerando exclusivamente cursos presenciais em nível nacional. A Figura 3 captura os dados, na interface gráfica, para a criação do gráfico de linha temporal intitulada **Evasão Anual Nacional Presencial** e o Gráfico 4 apresenta os resultados obtidos.

▼ Evasão Anual Nacional Presencial

Nome

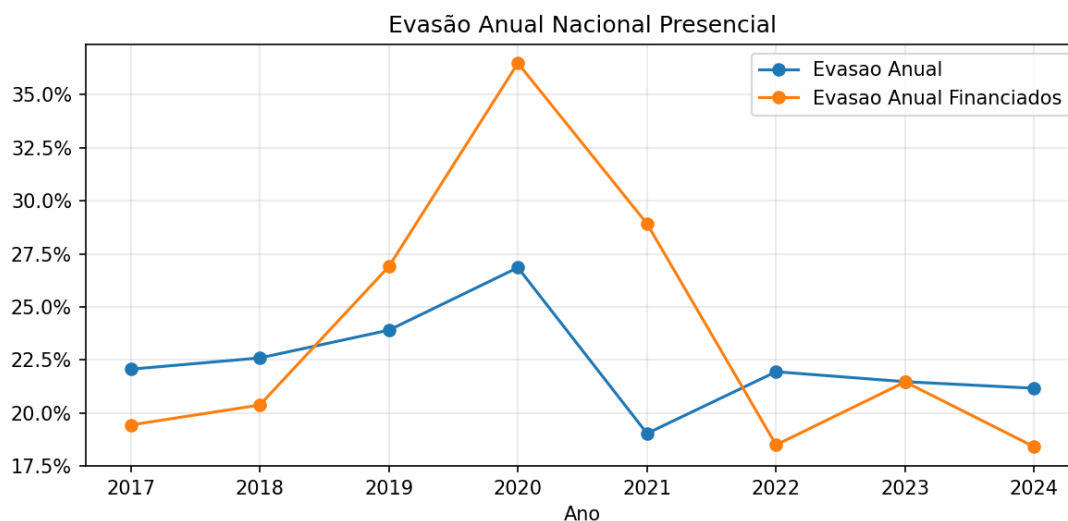
Evasão Anual Nacional Presencial

Nível

nacional

Variáveis (separadas por vírgula)

evasao_anual,evasao_anual_financiados



Em seguida, foi realizado um procedimento similar, com um filtro inverso para construir o conjunto de dados mantendo apenas cursos na modalidade EaD. Novas linhas temporais de evasão anual e evasão anual entre estudantes financiados geradas a partir desse conjunto são representadas no Gráfico 5.

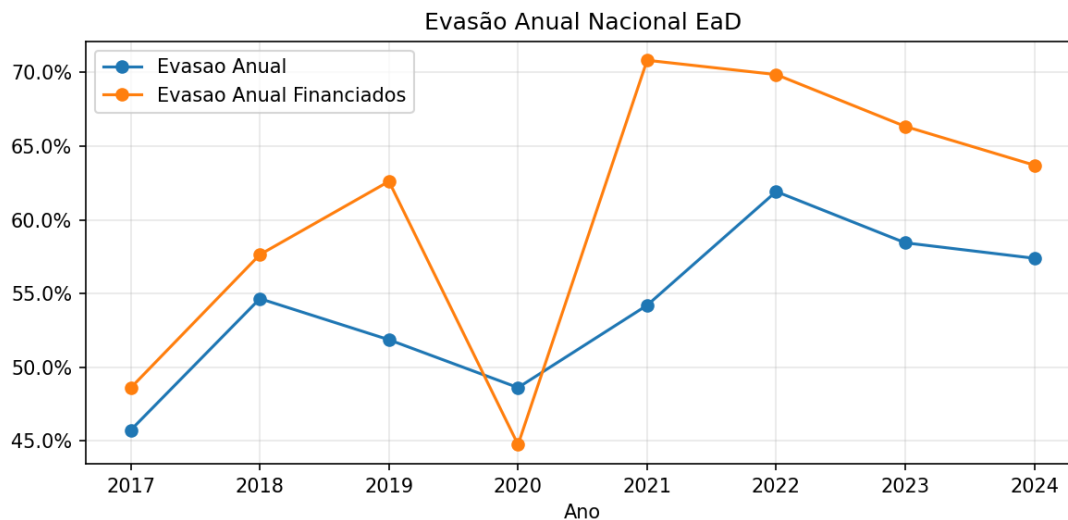


Gráfico 5: Evolução da evasão anual em nível nacional para cursos a distância (EaD).

A comparação entre os Gráficos 4 e 5 indicam que a evasão em cursos a distância permanece substancialmente superior àquela verificada em cursos presenciais ao longo de toda a série temporal. Pode ser observado também o comportamento atípico no ano de 2020, no qual os cursos presenciais apresentam aumento acentuado da evasão, enquanto os cursos a distância registram redução dessa métrica. Esse fato pode ser explicado pelas mudanças ocorridas durante o pico da pandemia de COVID-19. Também foi verificado que a evasão entre estudantes financiados apresenta variações maiores em ambas as modalidades, sugerindo que esse grupo é sensível a mudanças estruturais.

5.2 Diagramas de Dispersão

O diagrama de dispersão apresentado no Gráfico 6 demonstra uma relação entre evasão e percentual de matrículas na rede privada no nível nacional. Assim como as linhas temporais, esses gráficos permitem representações nacionais ou por território.

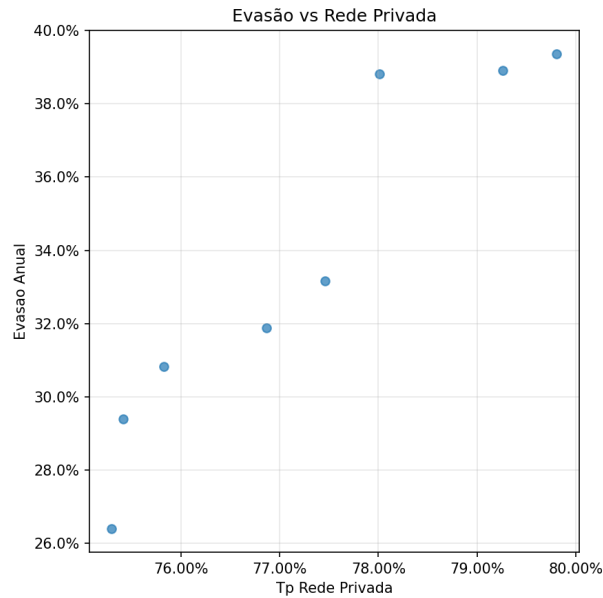


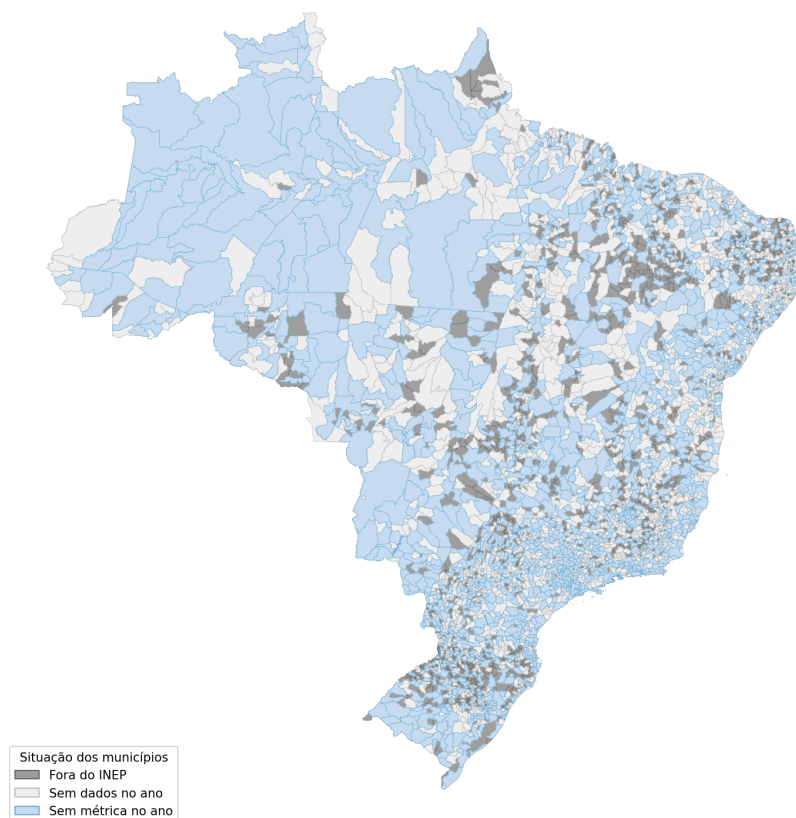
Gráfico 6: Relação entre evasão e rede privada em nível nacional.

5.3 Mapas Temáticos

A criação de mapas temáticos coropléticos para as métricas declaradas foi realizada pelo sistema nos níveis municipais e estaduais.

5.3.1 Mapas Representativos do Ano Inicial

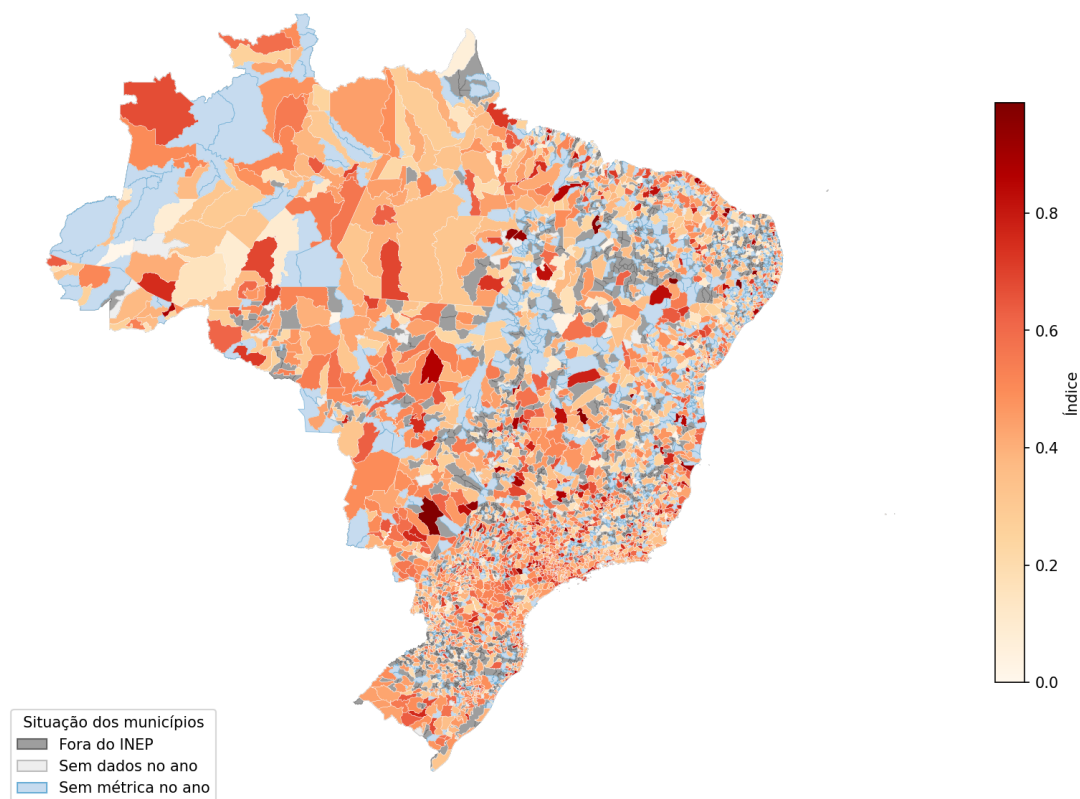
No ano inicial do período, não existem métricas calculadas, uma vez que não há valores de referência de anos anteriores. Nesse caso, os municípios sem métricas calculadas aparecem como parte do grupo **Sem métrica no ano**. Isso pode ser observado no Mapa [1](#).



Mapa 1: Distribuição da evasão anual por município em 2016, destacando diferentes situações de ausência de dados.

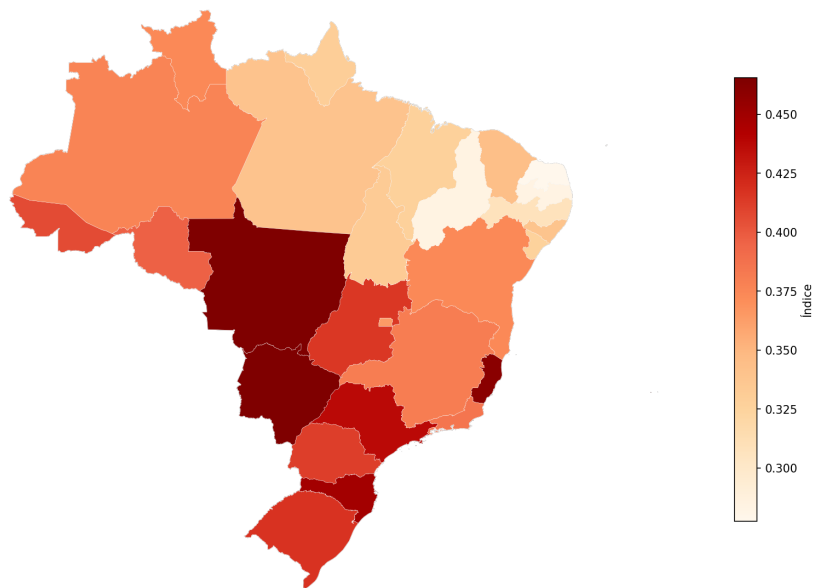
5.3.2 Mapas Representativos do Ano Final

Em contraste com o mapa anterior, o referente ao ano 2024 é apresentado no Mapa 2 e demonstra a aplicação do gradiente da evasão anual sobre os municípios.



Mapa 2: Distribuição da evasão anual por município em 2024.

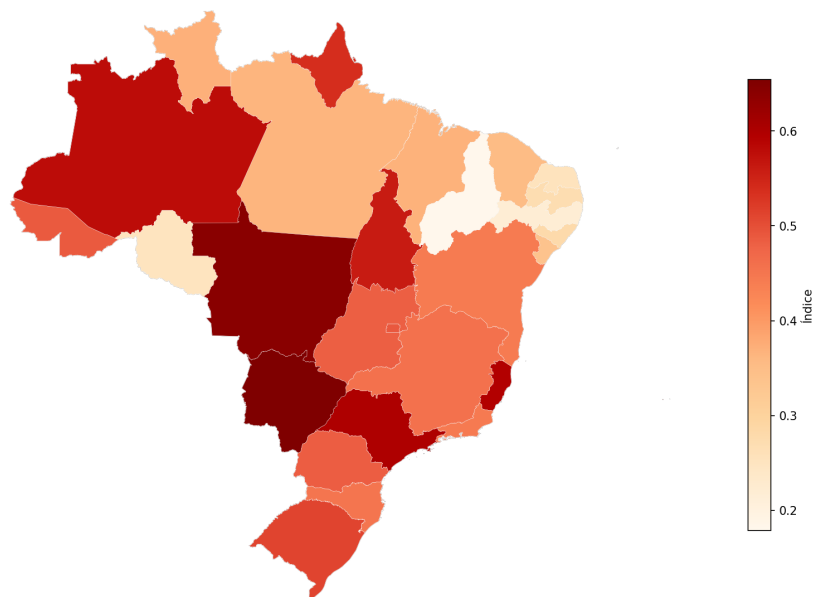
Complementarmente, o Mapa 3 apresenta a distribuição da evasão anual por unidade da federação no mesmo ano, oferecendo uma visão agregada que facilita a comparação entre estados.



Mapa 3: Distribuição da evasão anual por unidade da federação em 2024.

5.3.3 Evasão entre Estudantes Financiados

Adicionalmente, foram gerados mapas específicos para a análise da evasão entre estudantes com financiamento estudantil, conforme exemplificado no Mapa 4.



Mapa 4: Distribuição da evasão entre estudantes financiados por unidade da federação em 2024.

6 Considerações Finais

Este trabalho apresentou a concepção e a implementação de um pipeline de dados voltado à integração e análise territorializada da evasão no ensino superior brasileiro a partir da combinação de microdados do Censo da Educação Superior, produzidos pelo INEP, com dados territoriais disponibilizados pelo IBGE. A proposta buscou reduzir barreiras técnicas associadas ao uso dessas bases públicas.

De forma geral, os resultados demonstram que o pipeline proposto é capaz de integrar múltiplas bases anuais dos microdados de ensino superior nos três níveis territoriais definidos e gerar gráficos e mapas sobre a métrica definida.

Como limitações, destaca-se a defasagem temporal dos dados censitários do IBGE selecionados na demonstração e o grau de acoplamento do pipeline a essas planilhas na versão atual.

Como trabalhos futuros, sugere-se a flexibilização do pipeline para contemplar novas fontes de dados públicos, principalmente os dados censitários de 2022 do IBGE. O sistema demonstra habilidade de tipagem de dados usada para converter internamente as métricas definidas no formato de proporção para porcentagens durante a geração de gráficos, outras variáveis também possuem tipagem e podem ser convertidas para outros formatos de dado. O uso dessa operação pode ser expandido e as métricas podem adotar novos formatos de dado de entrada além de proporção. Espera-se que o pipeline desenvolvido possa contribuir como uma base extensível para estudos sobre evasão e outros fenômenos educacionais no contexto brasileiro.

Referências

- [1] Vitória Raupp Adão et al. Visualização de dados do censo escolar: Uma análise dos indicadores de tecnologia. 2024.
- [2] Carmen Aina, Eliana Baici, Giorgia Casalone, and Francesco Pastore. The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, 79:101102, 2022.
- [3] Bruna Oenning Amador and Rita Cristina Galarraga Berardi. Processo de etl para inserção de dados estatísticos do ensino superior brasileiro com foco em mulheres nas áreas stem no grafo de conhecimento da plataforma ellas. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 15–21. SBC, 2024.
- [4] Paula Cristina Freitas Barroso, Íris Martins Oliveira, Dulce Noronha-Sousa, Ana Noronha, Cristina Cruz Mateus, Enrique Vázquez-Justo, and Cristina Costa-Lobo. Fatores de evasão no ensino superior: uma revisão de literatura. *Psicologia Escolar e Educacional*, 26:e228736, 2022.
- [5] Vera Lúcia Jacob Chaves. Expansão da privatização/mercantilização do ensino superior brasileiro: a formação dos oligopólios. *Educação & Sociedade*, 31:481–500, 2010.

- [6] Tribunal de Contas da União. Decisão nº 408/2002 – plenário. Relator: Iram Saraiva. Processo: 016.229/1999-1. Tipo de processo: Relatório de Acompanhamento (RACOM). Data da sessão: 24 abr. 2002. Ata nº 13/2002 – Plenário., 2002. Disponível em: https://pesquisa.apps.tcu.gov.br/documento/acordao-completo/*/NUMACORDAO%253A408%2520ANOACORDAO%253A2002%2520COLEGIADO%253A%2522Plen%25C3%25A1rio%2522/DTRELEVANCIA%2520desc%252C%2520NUMACORDAOINT%2520desc/1/%2520?uuid=22f70ab0-d33e-11ea-8f5b-7d0876e6c039. Acesso em: 30 out. 2025.
- [7] Tribunal de Contas da União. Acórdão nº 461/2022 – plenário. Relator: Walton Alencar Rodrigues. Processo: 026.147/2020-3. Tipo de processo: Relatório de Auditoria (RA). Data da sessão: 09 mar. 2022. Ata nº 8/2022 – Plenário. Efeitos afetados por Acórdão de Relação nº 1712/2023 – Plenário., 2022. Disponível em: https://pesquisa.apps.tcu.gov.br/documento/acordao-completo/*/NUMACORDAO%253A461%2520ANOACORDAO%253A2022%2520COLEGIADO%253A%2522Plen%25C3%25A1rio%2522/DTRELEVANCIA%2520desc%252C%2520NUMACORDAOINT%2520desc/0. Acesso em: 30 out. 2025.
- [8] Tribunal de Contas da União. Acórdão nº 658/2023 – plenário. Relator: Walton Alencar Rodrigues. Processo: 033.402/2021-3. Tipo de processo: Relatório de Auditoria (RA). Data da sessão: 05 abr. 2023. Ata nº 13/2023 – Plenário., 2023. Disponível em: https://pesquisa.apps.tcu.gov.br/documento/acordao-completo/*/NUMACORDAO%253A658%2520ANOACORDAO%253A2023%2520/DTRELEVANCIA%2520desc,%2520NUMACORDAOINT%2520desc/0?utm_source=. Acesso em: 30 out. 2025.
- [9] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Nota de esclarecimento | divulgação dos microdados. Portal Gov.br – Centrais de Conteúdo / Notícias / Institucional, 2022. Publicado em: 22 fev. 2022. Atualizado em: 30 jun. 2025. Disponível em: <https://www.gov.br/inep/pt-br/centrais-de-conteudo/noticias/institucional/nota-de-esclarecimento-divulgacao-dos-microdados>. Acesso em: 30 out. 2025.
- [10] IBGE. Censo demográfico 2010 – indicadores municipais. <https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html>, 2010. Acesso em: 22 out. 2025.
- [11] Letícia Inácio, Luciana da Silva Ferreira, and Roberto de Souza Rodrigues. Educação superior no brasil: o impacto do “teto dos gastos” no orçamento das universidades federais. *Revista de Políticas Públicas*, 28(1):339–359, 2024.
- [12] INEP. Microdados do censo da educação superior (2020–2024). <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>, 2025. Atualizados em 22/09/2025. Acesso em: 22 out. 2025.
- [13] Diretoria de Estatísticas Educacionais INEP. Metodologia de cálculo dos indicadores de fluxo da educação superior, 2017.
- [14] Instituto Brasileiro de Geografia e Estatística. Malhas territoriais, 2024. Acesso em: 2026.

- [15] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Mec e inep divulgam resultado do censo da educação superior 2023. <https://www.gov.br/inep/pt-br/centrais-de-conteudo/noticias/censo-da-educacao-superior/mec-e-inep-divulgam-resultado-do-censo-superior-2023>, 2024. Acesso em: 04 fev. 2026.
- [16] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Notas estatísticas do censo da educação superior 2024. https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/notas_estatisticas_censo_escolar_da_educacao_superior_2024.pdf, 2024. Acesso em: 04 fev. 2026.
- [17] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Inep divulga resultado do censo da educação superior 2024, 2025.
- [18] MBCM Lobo. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, 25(14):32, 2012.
- [19] RL LOBO and MBCM Lobo. Esclarecimentos metodológicos sobre os cálculos de evasão. *Instituto Lobo*, 2012.
- [20] Roberto Leal Lobo, Paulo Roberto Motejunas, Oscar Hipólito, Maria Beatriz Carvalho Melo Lobo, et al. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, 37(132):641–659, 2007.
- [21] Luiz Guilherme Schymura. É hora de criar um sistema integrado de estatísticas e geoinformação no brasil. *Revista Conjuntura Econômica*, 77(9):6–9, 2023.
- [22] Ialy Cordeiro de Sousa. Arquitetura conceitual para automação de etl de microdados públicos: um estudo de caso com o enem. B.S. thesis, 2025.
- [23] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1):89–125, 1975.
- [24] Vincent Tinto. Limits of theory and practice in student attrition. *The journal of higher education*, 53(6):687–700, 1982.

7 Apêndice A – Comandos de Execução do Sistema

Este apêndice apresenta os comandos utilizados para execução dos pipelines de processamento de dados, geração das visualizações e inicialização da interface gráfica, permitindo a reprodução dos resultados deste trabalho.

7.1 Inicialização da Interface Gráfica

A interface gráfica de configuração do pipeline pode ser iniciada com o seguinte comando:

```
streamlit run src/ui/app.py
```

7.2 Execução dos Pipelines

Os pipelines de extração e transformação dos dados do INEP podem ser executados com os seguintes comandos:

```
python -m brpipe.scripts.inep.executar_extracao  
python -m brpipe.scripts.inep.executar_transformacao
```

Já os pipelines de extração e transformação dos dados do IBGE podem ser executados com os seguintes comandos:

```
python -m brpipe.scripts.ibge.executar_extracao  
python -m brpipe.scripts.ibge.executar_transformacao
```

7.3 Geração de Visualizações

As visualizações utilizadas neste trabalho podem ser geradas com os seguintes comandos:

```
python -m brpipe.scripts.viz.gerar_linha_temporal  
python -m brpipe.scripts.viz.gerar_scatter  
python -m brpipe.scripts.viz.render_mapas
```

7.4 Teste de Harmonização dos Dados

O teste de harmonização entre os dados do INEP e do IBGE pode ser executado com o seguinte comando:

```
python -m brpipe.scripts.testar_correlacao
```